Statistics **NST IB Psychology 2004–5**

Rudolf N. Cardinal

Overview

Welcome to the statistics part of the NST IB Experimental Psychology practical course. NST psychology students generally learn statistics in each of the three years, and the courses fit together roughly like this:

NST IA Elementary Maths for Biologists (EMB) (or better)	*Material including algebra, powers, logarithms, trigonometry, calculus, descriptive statistics, basic hypothesis testing. (More advanced NST students will have taken Quantitative Biology, or NST Maths proper.)*
NST IB Psychology	Some students joining Part IB Psychology have not done NST IA, and some doing Part II have not done Part IB. Therefore, we state the background knowledge required for Part IB and Part II explicitly. For Part IB, the mathematical level assumed is that taught in the NST IA EMB course; the actual background knowledge required is set out in section 1. • **Please read through the Background Knowledge section** (section 1) before the first statistics practical. It revises GCSE, A-Level, and NST IA material (including statistical terminology and principles of experimental design, plotting data, descriptive statistics with measures of central tendency and variation, the normal distribution, probability, and the logic of null hypothesis testing). • **Please bring this booklet and a calculator to every statistics practical.** • **Practical 1** (Thu 11 & Fri 12 Nov 2004) will cover correlation and regression (section 2). • **Practical 2** (Tue 30 Nov & Wed 1 Dec 2004) will cover parametric difference tests (section 3). • **Practical 3** (Thu 10 & Fri 11 Feb 2005) will cover nonparametric difference tests (section 4). • **Practical 4** (Tue 1 & Wed 2 Mar 2005) will cover χ^2 tests (section 5). • The **revision practical** (Tue 3 & Wed 4 May 2005) will revisit the important points covered in the course, discuss exam technique, and look at past paper questions.
NST II Psychology	• *Revision of IB material (background; difference tests; χ^2 tests; regression).* • *Data handling; transformations and dealing with outliers.* • *Analysis of variance (ANOVA) techniques.*

In each practical, we will

- cover the basic theory behind a statistical procedure or test (supported by this comprehensive handout);
- use the procedure or test to analyse **data that you have collected in a practical class (so bring them along).** We try to make sure that we pair the statistics practicals with the experimental classes so that we cover statistical tests as you start to need them for the experimental write-ups. This will become apparent as we go along.
- have a go at some examples relevant to the topic.

Sample questions are provided at the end of each section, with further examples and past exam questions in sections 6 / 7. **Worked answers** to the questions are also provided (section 8), including worked answers to past numerical exam questions. Section 9 contains tips on **experimental design,** and a glossary of experimental design terminology. All the material here, and all the slides I'll use during the practicals, will be available from the Psychology web site and at

www.pobox.com / ~rudolf / psychology

For each practical, you will need (1) your data; (2) a calculator; (3) this booklet.

There is a **compulsory statistics question in the exam.** You will be allowed to use your calculator **so long as it is an approved model** (see p. 5). You will also be supplied with a clean copy of the **statistical tables and formulae** in the exam (section 10, p. 120→), so you don't have to memorize formulae or procedural details, but you do have to appreciate what a test might tell you, and which test is appropriate for a particular situation. For more details of the format of the IB exams, see p. 85 and www.psychol.cam.ac.uk → Undergraduate Information → Examination Information.

RNC, 30 July 2004.

Contents

1. Background knowledge

Objectives

In this section I'll cover the background mathematical knowledge required for the IB psychology course, and the background knowledge that will underpin the statistics course. I'll also cover some basics of experimental design.

The problems we face are these. (1) People come to IB psychology with a huge range of maths backgrounds — from GCSE Maths followed by NST IA Elementary Maths for Biologists all the way up to A-Level Further Maths followed by NST IA Maths level 'B'. The advanced mathematicians will find the statistics in IB psychology a walk in the park or will have covered them already. (2) Nobody normal thinks stats is tremendously exciting; it's merely a tool for doing research. (3) Many people think that statistics is hard and / or obscure. So let's divide the essential from the rest:

Stuff with wavy borders, like this, is advanced or for interest only and may be ignored. You will NOT be examined on it. Please DON'T get upset if it looks difficult; in places, it is. You do NOT have to understand it. Although the wavy-line stuff may improve your understanding if you are a mathematician, you can understand everything that you need to do good statistics and pass the exams with flying colours even if you ignore the wavy-line stuff ENTIRELY.

Double-wavy stuff is harder than single-wavy.

The 'Basic Mathematics' section (p. 6) covers material that is assumed for IB Psychology in general (not just the statistics course). We won't revise it in the practicals.

Statistics books

You shouldn't *need* a maths or statistics book for this course. Should you *want* one, undoubtedly the best statistics book I've come across is Howell (1997) [see References on p. 131 for full reference]. It'll cover pretty much all the statistics you need for Part IB and Part II and is fairly easy to read — as stats books go. Another good book that doesn't tell you *how*, but tells you *why*, is Abelson (1995).

Calculators and computers

For the exams: an excerpt from the University Reporter, 9 June 2004:

> *'… in 2004–05 the only models of electronic calculators that candidates will be permitted to take into the examination room will be as follows:*
> *(A)… Natural Sciences Tripos, Parts IA, IB, II, II (General), and III;*
> *For the above examinations candidates will be permitted to use only the standard University calculator* ***CASIO fx 100D, CASIO fx 115 (any version) or CASIO fx 570 (any version except the fx 570MS). Each such calculator must be marked*** *in the approved fashion. Medical and veterinary students who have previously had a calculator of similar or inferior specification marked as approved will be permitted to use this calculator in biological examinations in Part II of the Medical and Veterinary Sciences Tripos and of the Natural Sciences Tripos.*
> …
> *Standard University calculators CASIO fx 115MS marked in the approved fashion will be on sale at the beginning of Full Michaelmas Term 2004 at £12 each as follows:*
> …
> *Department of Chemistry, Part IA laboratory preparation room (for the Natural Sciences Tripos); …*
> *Department of Physiology (for medical and veterinary students);*
> *Board of Examinations Office (for any subject), 10 Peas Hill, Tuesday, 5 October and Wednesday, 6 October from 9.30 a.m. to 12.30 p.m. and from 2.30 p.m. to 4.30 p.m.*
> *Candidates are strongly advised to purchase calculators at the beginning of Full Michaelmas Term at the centres named above. At other times calculators may be purchased from the institutions named above, and also from the Department of Physics. Candidates already possessing a CASIO fx 100D, CASIO fx 115 (any version) or CASIO fx 570 (any version except the fx 570MS) will be able to have it marked appropriately at no cost at one of the above centres.'*

1.1. Basic mathematics

If any of this (apart from the stuff in wavy lines) causes you problems, because for some reason you haven't done NST IA Elementary Maths, you should speak to your Director of Studies about catching up to this level. Some of it isn't used in the stats course but is common in psychology (e.g. logarithms are used in psychophysics).

Fractions, percentages

$$\frac{5}{100} \equiv 5\% \equiv 0.05$$

Notation to be familiar with

Δx	A small change in x (pronounced 'delta-x').
$\sum x$	The sum of x (i.e. add up all the xs that you have).
$\sum_{i=1}^{n} x_i$	A more precise way of specifying summation: this means 'for every value of i from 1 to n take the sum of x_i', or '$x_1 + x_2 + x_3 + \ldots + x_n$'.
$<<, <, \leq, =, \geq, >, >>$	Much less than, less than, less than or equal to, equal to, greater than or equal to, greater than, much greater than.
$\neq, \approx, \cong, \equiv$	Does not equal, approximately equals, approximately equals, is equivalent / identical to
$\Rightarrow, \Leftarrow, \Leftrightarrow$	Implies, is implied by, implies and is implied by
$\propto$	Is proportional to
∞	Infinity

Powers (a summary) — though nothing beyond x^2 and $\sqrt{x}$ used in IB statistics

$$x^1 \equiv x \qquad x^0 \equiv 1 \qquad x^{\frac{1}{2}} \equiv \sqrt{x} \qquad x^a \cdot x^b \equiv x^{a+b} \qquad (xy)^n = x^n y^n$$

$$x^2 \equiv x \cdot x \qquad x^{-1} \equiv \frac{1}{x} \qquad x^{\frac{1}{3}} \equiv \sqrt[3]{x} \qquad \frac{x^a}{x^b} \equiv x^{a-b} \qquad \left(\frac{x}{y}\right)^n = \frac{x^n}{y^n}$$

$$x^3 \equiv x \cdot x \cdot x \qquad x^{-2} \equiv \frac{1}{x^2} \qquad x^{\frac{1}{n}} \equiv \sqrt[n]{x} \qquad (x^a)^b \equiv x^{ab} \qquad \left(\frac{x}{y}\right)^{-n} = \frac{y^n}{x^n}$$

$$x^n \equiv x \cdot x \cdots x_n \qquad x^{-n} \equiv \frac{1}{x^n} \qquad x^{\frac{a}{b}} \equiv \sqrt[b]{x^a}$$

$$x^{\frac{-a}{b}} \equiv \frac{1}{\sqrt[b]{x^a}}$$

Logarithms (a summary) — though not needed for IB statistics

$$\log_a b = c \Leftrightarrow b = a^c \qquad \log_a xy \equiv \log_a x + \log_a y \qquad \log_a b \equiv \frac{1}{\log_b a}$$

$$\log_a x^n \equiv n \qquad \log_a\left(\frac{x}{y}\right) \equiv \log_a x - \log_a y \qquad \log_a x \equiv \frac{\log_b x}{\log_b a}$$

$$\log_{10}(x) \equiv \lg(x) \qquad \log_a x^y \equiv y \log_a x \qquad \log_a x \equiv \log_b x \cdot \log_a b$$

$$\log_e(x) \equiv \ln(x)$$

$$e = 2.718281828$$

Calculus

If $f(x)$ is some function of x, then the function giving the *gradient* of $f(x)$ is the *first derivative of f(x) with respect to x*, written variously $f'(x) = \dot{f} = \frac{d}{dx} f(x)$. If $f(x)$ is some function of x, then the *area under the curve* of $f(x)$ is given by the *integral* of $f(x)$ with respect to x, written $\int f(x)dx$. This is called the *indefinite integral*, because it doesn't specify which parts of the curve we want the area under. The area under the curve $f(x)$ from $x = a$ to $x = b$ is given by the *definite integral* $\int_a^b f(x)dx$.

1.2. Basic terminology

Variables and measurement

When we measure something that can vary, it is termed a **variable.** We can distinguish between **discrete variables,** which can only take certain values (e.g. in mammals, sex is a discrete variable which can take one of the two values male and female), and **continuous variables,** which can take any value (such as height).

We can also distinguish between **quantitative** data and **frequency** data (also called **categorical** or **qualitative** data). Height is measured (quantified), and is therefore quantitative. If we count the number of males and females in the room, each person falls into one category or the other, and the data we end up with are frequencies (e.g. there are 26 males and 29 females).

While we're at it, we can also distinguish several types of measurement scale. **Nominal** scales aren't really 'scales' at all, they're categories (e.g. male / female, Labour / Conservative / Lib Dem). The categories are different, but the nature of their difference isn't relevant. **Ordinal** scales rank things, but do not specify how 'far apart' they are on a scale. For example, in the Army a lieutenant ranks lower than a captain, who ranks lower than a major; however, it doesn't make sense to ask whether a major is more or less above a captain than a captain is above a lieutenant. **Interval** scales have meaningful differences; 10°C is as far above –10°C as 40°C is above 20°C. However, interval scales do not have a meaningful zero point (0°C is not the 'absence' of temperature), so we can't say that 40°C is 'twice as hot' as 20°C. **Ratio** scales have a true zero point. 40 K *is* twice as hot as 20 K (because 0 K *is* the absence of heat); 3 m is twice as far as 1.5 m.

Frequently we come across a variable that can take many values. For example, suppose we have a group of 30 people and we want to know something about their heights. We might call X the variable that represents their height. We'll be able to make 30 different measurements of X; we might call them X_1, X_2… X_{30}. Each measurement is a single **observation** *drawn from* our variable. (Variables are often referred to by upper-case letters, such as X. Individual values of a variable are referred to by corresponding lower-case letters, such as x, or by the upper-case letter with a subscript, such as X_1, X_2, X_i, or by the lower-case letter with a subscript, such as x_1, x_2, x_i.)

Populations and samples

Taking this a step further, we can distinguish **populations** from **samples.** If all we want to know is the height of our 30 people, we can measure it and that's the end of the matter. Our measured sample is the same as our total population. But very often, we want to **estimate** something about a population by measuring a sample of that population that is very far from being the whole population. For example, if we want to know the height of 20-year-old human males in general, then we'd be unable in practice to measure the whole population, but we could measure 30 male 20-year-old Cambridge psychology undergraduates. This would be convenient, and we would get a number that would be a definitive measurement of our particular set of subjects, but would also be an **estimator** of the height of all 20-year-old male Cambridge undergraduates, and an estimator of the height of all 20-year-old male humans. However, it wouldn't necessarily be a very good estimator of the latter — the sample may not be very *representative* of the whole population (average height in the UK is shorter than in Germany but taller than for Japan) and, more importantly, may be systematically different from the population mean (university students might be taller than similarly-aged UK males in general). The latter is called **bias.** If we want to obtain a sample that is likely to be a good estimator of the whole population, we should draw a **random sample** — one where every member of the population has an equal chance of being picked to be in our sample. Studies based on nonrandom samples may lack **generality** (or **external validity**) — so studying the effects of a potential memory-enhancing drug on Cambridge students might tell you a lot about what it'll do to other university students, but not the adult population as a whole.

Descriptive and inferential statistics

'Statistics' itself can mean a couple of things. **Descriptive statistics** is the business of describing things, you'll be shocked to learn; newspapers are full of it ('Henman's average serving speed was X…'). In research, it also includes the business of looking at the **distribution** of your data ('is there an even spread of ability in my subjects or do I have a high-performing subgroup and a low-performing subgroup?'). The job of having a look at the distribution of a data set before analysing it in detail is called **exploratory data analysis (EDA),** a set of techniques developed by a statistician called Tukey. **Inferential statistics** is the business of inferring conclusions about a population from studies conducted with a sample. When we measure an attribute (such as height) from a whole population, we've measured a **parameter** of the population. If we measure the same thing with a sample, we've measured a **statistic** of the sample. So inferential statistics is also the business of inferring parameters from statistics (in this specialized sense). We tend to use Greek letters for parameters, such as μ and σ, but Roman letters for statistics (such as $\bar{x}$ and s).

Exerting control: independent and dependent variables, between- and within-subject designs

If we manipulate or control a variable, it is termed an **independent variable.** We might test the reaction times of a group of people having given them one of three different doses of a drug; drug dose would then be a (discrete) independent variable. We might want to know how the drug's effect depends on their body weight; body weight would then be a (continuous) independent variable. The thing that we measure is the **dependent variable,** in this case reaction time.

When we come to manipulate independent variables, we must consider randomness, just as we do when we choose samples from populations. If we are going to give our drug to some of our subjects and no drug to other subjects, we must consider several factors. First, we probably do not want the subjects to know whether they are receiving the drug or not, because this knowledge might in some way affect their performance; we would therefore give the 'non-drug' group a placebo (Latin for 'I shall please' — a sugar pill given by doctors to placate patients they think don't need drug treatment). The groups should be unaware or 'blind' to whether they receive drug or placebo; ideally, the person running the experiment should also be unaware, so he / she can't bias performance in any way. This would make the study a double-blind, placebo-controlled study. However, we must also make sure that our drug group does not differ from the placebo group in some important way. If the drug group were male and the placebo group were all female, any potential effects of our drug would be **confounded** with the effects of the subjects' sex; our study would be uninterpretable; it would not have **internal validity.** Similarly, if the subjects who are going to receive the drug have better reaction times to begin with than the subjects who are going to receive placebo, our results might not mean what we think they mean. Ideally, we would like our two groups to be **matched** for all characteristics other than the variable we want to manipulate (drug v. placebo). We can try to craft matched groups by measuring things that we think are relevant (e.g. reaction time on the task we're going to use or a similar task, age, IQ, sex…). But we probably can't explicitly match groups on every variable that might potentially be a confound; eventually we need a mechanism to decide which group a subject goes in, and that method should be **random assignment.** So in our example, if we have plenty of subjects, we could just randomly assign them to the drug group or the placebo group. Or we could match them a bit better by ranking them in order of reaction time performance and, working along from the best to the worst, take pairs of subjects (from the best pair to the worst pair), and from each pair assign one to the drug group and one to the placebo group at random. Random assignment takes care of all the factors you haven't thought of — for example, if your subjects are all going to do an IQ test in your suite of testing rooms, you should seat them randomly, in case one room's hotter than the others, or nearer the builders' radio outside, or whatever. Common confounding factors it is always worth thinking about are **time** and **who collects the data.**

If you're not in full control of the independent variable, your conclusions may be limited. For example, suppose you find your drug improves reaction-time performance in people whose (pre-drug or 'baseline') performance was bad, but not in people whose baseline performance was good. You might conclude that your drug improves performance up to some sort of ceiling. However, suppose that all your 'good performers' were women and all the 'bad performers' were men. In that case, you can't distinguish a performance-dependent effect from a sex-dependent effect.

So far, we've been talking about **between-subjects designs,** in which you do one thing to some subjects (e.g. giving them drug) and another to others (e.g. giving them placebo). A very powerful method that you might consider is to use a **within-subjects design,** in which every person gets tested on drug *and* on placebo, at separate times. The two types of design require different statistical analysis, which we'll discuss later — basically, in a within-subjects design, two measurements from the same person are related / similar in a way that two measurements from two different people aren't, and you have to take account of that. Within-subjects designs are very powerful, but they do have some problems to do with *time:* **order** and **practice effects.** If everybody does your task on placebo first and then on drug, and they get better, the effect might be due to practice rather than the drug. There are other kinds of effects that can arise if everyone experiences treatments in a particular order. You must design your experiment to avoid such potential confounds.

1.3. Plotting data: histograms

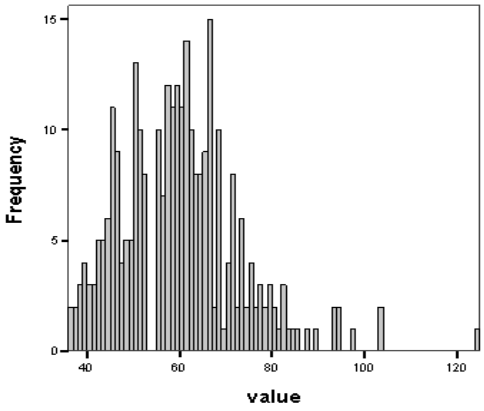
The first thing we should do before analysing any set of data is to look at it. For this, it's helpful to have some kind of graphical way of representing it. Here's one.

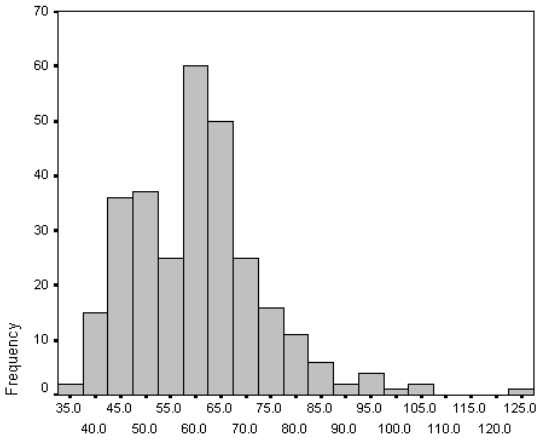
Histograms and grouped histograms

Data set 1

36	37	38	38	39	39	39	40	40	40	40	41	41	41	42	42	42	43	43	43
43	43	44	44	44	44	44	45	45	45	45	45	45	46	46	46	46	46	46	46
46	46	46	46	47	47	47	47	47	47	47	47	47	48	48	48	48	49	49	49
49	49	50	50	50	50	50	51	51	51	51	51	51	51	51	51	51	51	51	51
52	52	52	52	52	52	52	52	52	52	53	53	53	53	53	53	53	53	56	56
56	56	56	56	56	56	56	56	57	57	57	57	57	57	57	58	58	58	58	58
58	58	58	58	58	58	58	59	59	59	59	59	59	59	59	59	59	59	60	60
60	60	60	60	60	60	60	60	60	60	61	61	61	61	61	61	61	61	61	61
61	62	62	62	62	62	62	62	62	62	62	62	62	62	62	63	63	63	63	63
63	63	63	63	63	64	64	64	64	64	64	64	64	65	65	65	65	65	65	65
65	66	66	66	66	66	66	66	66	66	67	67	67	67	67	67	67	67	67	67
67	67	67	67	67	68	68	69	69	69	69	69	69	69	69	69	69	70	71	71
71	71	72	72	72	72	72	72	72	72	73	73	74	74	74	74	74	74	75	75
76	76	76	76	77	77	78	78	78	79	79	80	80	80	81	81	82	83	83	83
84	85	86	88	90	94	94	95	95	98	104	104	125							

Here we have a large list of measurements of something (it doesn't matter what), but we don't get much sense of the distribution. A histogram plots the *frequency* with which observations fall into a particular category. If there's a category for each possible value of the observation, we get a histogram like that on the left of the figure (above); this is rather silly. If the categories are made a bit bigger, we get a histogram like that on the right (below). These allow us to visualize the data readily and we get a sense of its **central tendency** (most observations are around the 45–70 range), the **distribution** (observations are clustered around the left-hand side with a 'tail' to the right), and any **extreme values** or **outliers** (there are a couple of observations that are much higher than the others).





Left: *Frequency histogram. The x axis (abscissa) shows values or categories; the y axis (ordinate) shows the frequency with which an observation fell into the appropriate category. This histogram looks rather 'noisy' because there are too many categories.* ***Right:*** *Histogram with data grouped in more sensible categories. The same data as on the left. Each category (on the x axis) represents an* ***interval.*** *In this example, the value printed on the x axis is the midpoint of the interval; thus, '45' denotes those values falling into the range 42.5–47.5 (this is just done to save a bit of space). Choose your own interval size to make the histogram look sensible —* $\sqrt{n}$ ***categories*** *is often a good choice when then are n observations. If you ever choose to make the intervals not all equal in width (you might call this asking for trouble), you should make the* ***area*** *of each bar proportional to the number of observations, rather than the height.*

1.4. Measures of 'central tendency' — taking the average

Data set 2

12	18	19	15	18	14	17	20	18	15	17	11	23	19	10

Let's take a set of 15 numbers (above). Where's the 'middle' or the 'average'? There are several ways we might answer this question. The **mode** is the value that occurs most commonly — in this case, 18. If we wanted to be formal, we could say that these data are from a variable we measured called *X*. We could therefore say that Mo(*X*) = 18. If there are two modes and they're in some sense 'adjacent', we might use the mean of the two, $\frac{Mo_1 + Mo_2}{2}$. If they're far apart, then the distribution is **bimodal** and we'd report both modes. Why use the mode? It can be applied to nominal (categorical) data. It isn't affected by extreme scores. It may be the most meaningful; if you want to buy a job-lot of shoes that are all the same size, you should buy shoes that are the modal size of the population you're going to sell them to. By definition, for an observation x_i taken at random from a variable *X*, $P(x_i$ = mode) > $P(x_i$ = any other score). Why might you not use it? If your categories are not particularly meaningful, nor will be your mode. It is also less amenable to mathematical analysis than the mean.

The **median** is the value that's in the middle if we lined all the values up in order. (More precisely, it's the value at or below which 50% of the scores fall when the data are arranged in numerical order, as below.) Here, it's 17. This is written Med(*X*) = 17, or sometimes $\tilde{x}$ = 17.

Data set 2, reordered

10	11	12	14	15	15	17	**17**	18	18	18	19	19	20	23

This was easy to find, because we had an odd number of observations. If we had an even number of observations then we'd add up the two closest to the middle and divide by two:

Data set 3

10	11	12	14	15	15	17	**17**	**18**	18	18	19	19	20	21	23

the two middle values

The median is (17+18) ÷ 2 = 17.5

Why use the median? Like the mode, it isn't affected by extreme scores ('outliers'). For example, the median number of legs on people in Britain is 2, but the mean is not. However, it is also less amenable to mathematical analysis than the mean.

The **mean** is most people's idea of the 'average'. For a sample with *n* observations $x_1, x_2, \ldots x_n$, the **sample mean** of *X* is written $\bar{x}$ and calculated as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}$$

(The two notations are simply different ways of saying 'sum all of the observations and divide by the number of observations.) The mean of data set 2 above is 16.4. The **population mean** is written μ (but we don't normally measure this directly, as discussed earlier). The mean of a given sample may not match the population mean (measure ten tuna fish — is the mean of your sample identical to the mean of all the tuna in the world, or have you caught tuna that are slightly bigger / smaller than average?) — but on average, if you took a *lot* of samples, the average of all the sample means would be the same as the population mean. We say the sample mean is a **good estimator** of the population mean (in fact, it's the best estimator).

The mean has certain disadvantages. It is influenced strongly by extreme values (try changing just one datum to 10,000 in the data set above and recalculating the mean). There may well be no individual datum whose value is the same as the mean. Interpreting it requires some justification that the underlying data is being measured on an interval scale. However, it is eminently amenable to mathematical analysis and has certain other properties which make it the most widely-used measure of central tendency; for example, it includes information from every observation.

1.5. Measures of dispersion (variability)

Knowing a measure of central tendency doesn't tell us all we need to know about a set of data. Two data sets can have the same mean but very different variability — for example, {9,10,11} and {5,10,15} both have a mean of 10. It's often very important to have a measure of variability; there are several.

Range

This is simply the distance from the lowest to the highest point. The range of {9,10,11} is 2; the range of {5,10,15} is 10. The range is simple, but is easily distorted by extreme values.

Interquartile range

We talked about this when considering boxplots. It is the range of the middle 50% of observations; it is the distance between the first and third quartiles (the 25th and 75th percentiles). This is not distorted by extreme values; in fact, it may not pay enough attention to values at the edge of a distribution!

The average deviation… is approximately zero and therefore useless.

We could measure how much each observation, x_i, deviates from the mean, $\bar{x}$, and take the average of each deviation. However, since some deviations will be positive and an equal number will be negative, the average deviation is about zero.

The mean absolute deviation… nobody uses.

One stage further: we take the deviation from the mean for each observation, and take its absolute value (dropping any minus sign), i.e. $|x_i - \bar{x}|$. We then take the mean of these values:

$$m.a.d. = \frac{\sum |x_i - \bar{x}|}{n}$$

Though this one makes some sense, nobody uses it. Instead, they use the **variance,** the **standard deviation,** and the **standard error of the mean.** We'll cover the last of these when we look at difference tests, but we'll consider the other two here.

The variance — IMPORTANT

The **population variance, σ^2** is worked out as follows. Take each deviation from the mean; square it (this eliminates negative values); sum all these together; divide by *n*, the number of observations (this gives the average squared deviation per observation).

$$\sigma_X^2 = \frac{\sum (x_i - \mu)^2}{n}$$

However, since we rarely measure whole populations, we rarely use the population variance. Instead, we usually measure samples of the population (and therefore estimate the population variance from a sample variance). The **sample variance, s^2** is just the same except we divide by *n*–1, not *n*. The formula on the far right is one that's mathematically identical but a bit easier to use for calculations by hand.

$$s_X^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

The standard deviation (SD) — IMPORTANT

The standard deviation (SD) is the square root of the variance (so it's sort of an average deviation from the mean). So the **population standard deviation, σ** is

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

and the **sample standard deviation, *s*** is

$$s_X = \sqrt{s_X^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i{}^2 - \frac{(\sum x_i)^2}{n}}{n-1}}$$

If the data are *normally distributed* (see below), 68% of observations fall within one SD of the mean, and 95% of cases fall within 2 SD. For example, if the age of a group of subjects is normally distributed, and the mean age is 45 with a standard deviation of 10, then 95% of the cases would be between 25 and 65.

Some calculators refer to the population SD as σ_n and the sample SD as σ_{n-1}.

The coefficient of variation (CV) — not often used

The coefficient of variation is the standard deviation divided by the mean:

$$CV = \frac{s_X}{\bar{x}}$$

The standard deviation often increases with the mean. For example, if you rate something on a scale with a range of 0–10 (perhaps with a mean of 5) then the (population) SD can't be bigger than 5. If your scale was 0–100, with a mean of 50, your SD could be as high as 50. By dividing the SD by the mean, the CV becomes independent of this sort of thing. But the CV is rarely used.

Discrete random variables, treated formally

(A-Level Further Maths.) A **random variable (RV)** is a measurable or countable quantity that can take any of a range of values and which has a **probability distribution** associated with it, i.e. there is a way of giving the probability of the variable taking a particular value. If the values an RV can take are real numbers (i.e. an infinite number of possibilities) then the RV is said to be **continuous;** otherwise it is **discrete.** The probability that a discrete RV X has the value x is denoted $P(x)$. We can then define the mean or **expected value:**

$$E[X] = \sum xP(x)$$

and the **variance:**

$$Var[X] = E\left[(x - E[X])^2\right] = \sum (x - E[X])^2 P(x)$$
$$= \sum x^2 P(x) - (E[X])^2 = E\left[X^2\right] - (E[X])^2$$

and the **standard deviation:**

$$\sigma^2 = Var[X]$$

Why is the sample variance calculated differently from the population variance?

What's all this 'divide by n–1' business? Suppose we have a large population and we know its mean (μ) and variance (σ^2) precisely (they are **parameters;** see above). If we were to take an infinite number of samples, each containing n observations, we can calculate **statistics** of each sample. For example, we can calculate the mean, $\bar{x}$, as usual, for each sample. We would like the sample mean $\bar{x}$ to be an **unbiased estimator** of the population mean μ (i.e. we'd like $\bar{x}$ to be the same as μ, on average), and it is. However, this isn't so simple for the variance. If we used the **(wrong)** formula for the sample variance

$$\frac{\sum (x - \bar{x})^2}{n}$$

we'd find that, on average, we'd *underestimate* σ^2 — our estimator is **biased.**

(a) Demonstration

When we calculate the variance, we calculate a whole load of values of $(x - \bar{x})^2$. These are called *summed squared deviations from the mean*, or **summed squared errors (SSE).** Suppose we have a population whose mean we know to be zero. Sup-

pose that we take three samples and find that they're {1, –4, 3}. The SSE is $(1 – 0)^2 + (–4 – 0)^2 + (3 – 0)^2 = 26$, whether we use the population mean or the sample mean to calculate it, because for this particular sample the sample mean (0) happened to be the same as the population mean (0). But suppose it wasn't; suppose our sample was {1, –1, 2}, which has a sample mean of $^2/_3$. Then if we calculated the SSE around the population mean, it'd be $(1 – 0)^2 + (–1 – 0)^2 + (2 – 0)^2 = 6$. But if we calculated the SSE around the sample mean, it'd be $(1 – {}^2/_3)^2 + (–1 – {}^2/_3)^2 + (2 – {}^2/_3)^2 = 4.67$. For a given sample, the SSE calculated using the sample mean will always be smaller than (or equal to, but never greater than) the SSE calculated using the population mean. Since we divide the population SSE by *n* to get the population variance, if we divide the sample SSE by *n* we shall get something that on average is smaller than the population variance. Some complicated maths is needed to tell us *how much* smaller, but it turns out that on average we'll be wrong by a factor of $(n–1)/n$. So if we divide our SSE by *n*–1 instead of *n*, we'll get the right answer.

(b) Explanation: degrees of freedom

The difference between calculating the sample variance and the population variance is that when we calculate the sample variance, *we already know the mean*, but when we calculate the sample variance, *we have to estimate the mean from the data.* This leads us to consider something called **degrees of freedom (*df*).** Let's use an example. Suppose you have three numbers: 6, 8, and 10. Their mean is 8. You are now told that you may change any of the numbers, so long as the mean is kept constant at 8. How many numbers are you free to vary? You can't vary all three *freely* — the mean won't be guaranteed to be 8. You can only vary two freely; you need the third to adjust the mean to 8 again. Once you've adjusted two, you have no control over the third. If you had *n* numbers and had to keep the mean constant, you could only vary *n*–1 numbers.

Let's restate that in several ways, because people generally find it hard. Estimates of parameters can be based upon different amounts of information.

- The number of *independent* pieces of information that go into the estimate of a parameter is called the degrees of freedom (*df*).
- Alternatively, the *df* is the number of observations free to vary (as in our three-numbers-and-a-mean example, above).
- Alternatively, the *df* is the number of measurements exceeding the amount absolutely necessary to measure the 'object' (or parameter) in question. To measure the length of a rod requires 1 measurement. If 10 measurements are taken, then the set of 10 measurements has 9 *df*.
- In general, the *df* of an estimate is the number of independent scores that go into the estimate minus the number of parameters estimated from those scores as intermediate steps.

When we calculate σ^2, we already know μ; we don't use up any *df* calculating it, so the denominator remains *n*. (In our example above, we *knew* the population mean was 0, regardless of the numbers in our sample, so when we calculated the population SSE we didn't need to 'use any of the sample data up' in estimating the mean.) But when we calculate s^2, we must use up one *df* calculating the sample mean $\bar{x}$, so we only have *n*–1 *df* left (*n*–1 scores free to vary). Since the denominator is the number of scores on which our estimate is based, it should reflect this restriction, and be decreased by 1 — so in all cases we're dividing the total variability by the number of places (independent observations) it could have come from.

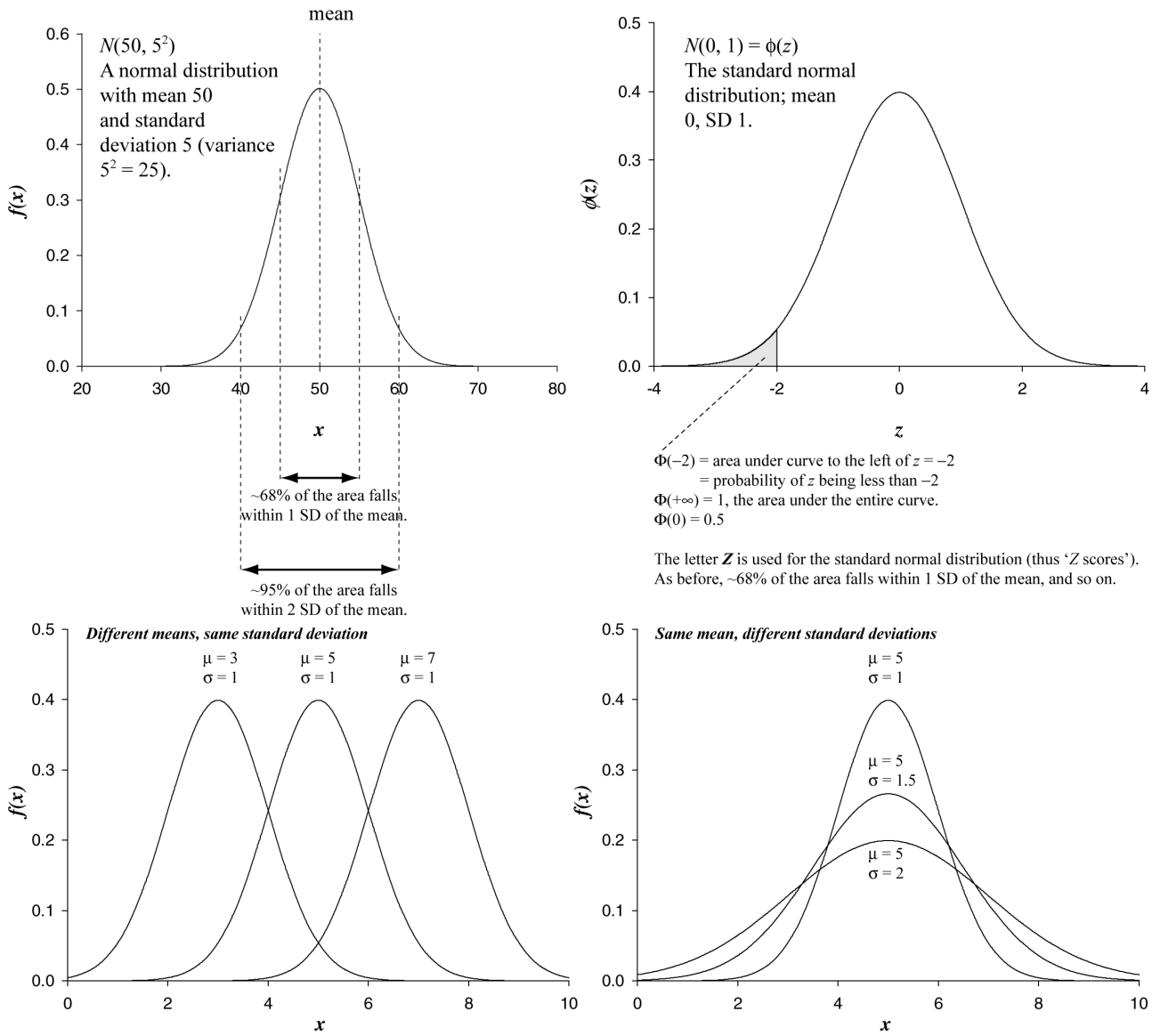
$$\sigma^2 = \frac{\sum (x-\mu)^2}{n} \quad s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

(c) Proof

Neither of the explanations above is complete. The full proof that we'll be out by a factor of $(n–1)/n$ unless we divide by *n*–1 rather than *n* is more complicated (see Frank & Althoen, 1994, pp. 301-305; or Myers & Well, 1995, p. 592, or graduate handouts at www.pobox.com/~rudolf/psychology).

1.6. The normal distribution

Many things in nature are *normally distributed*. If we plot a histogram or a probability distribution of them, the shape is something like that shown in the figure below: a 'bell curve'. It might be people's reaction times to respond to a race's starting gun, the number of barnacles found on a given area of rock, or the heights of French soldiers. Things that are normally distributed can have different means, and different standard deviations (see examples below), but once we know the mean and the standard deviation, we know all there is to know about the way that they're distributed.



Top left: *a normal distribution, which we describe as $N(\mu,\sigma^2)$ where μ is the mean and σ^2 is the variance (σ is the standard deviation).* ***Top right:*** *the 'standard' normal distribution, which always has a mean of 0 and a standard deviation of 1, and which is referred to by the letter Z. Both curves are perfectly symmetrical about the mean.* ***Below:*** *examples of normal distributions with different means and SDs.*

The normal distribution is sometimes called the Gaussian distribution (despite being invented by de Moivre in 1733). Why is it important?

(1) Z scores

First, we can calculate how likely a particular measurement is to have come from a particular population. **The area under bits of a probability distribution curve (such as the normal distribution) represents the *probability* or *proportion* of observations falling into a particular range.** Suppose healthy people have a mean plasma potassium concentration of 4.25 mM, with a standard deviation of 0.383

mM, and that this is normally distributed. Since I've told you that about 95% of the population fall within 2 SD of the mean, we can work out that 95% of healthy people have a potassium concentration in the range 3.5–5.0 mM. Furthermore, if a patient has a potassium concentration of 5.5 mM, we can work out the probability of this concentration or higher being found in the healthy population. The way we do that is as follows. It would be very tedious to work out the mathematical properties of the plasma-potassium normal distribution, which we'd call $N(4.25,\ 0.383^2)$, whenever we wanted to answer a question like this. It would certainly not be quick with pen and paper. So we convert **('transform')** our potassium score from a number from $N(4.25,\ 0.383^2)$, which we know nothing about, to a special distribution called the **standard normal distribution,** which we write $N(0,1)$ or ***Z*,** that we know everything about. This is important and very easy: if x is our potassium measurement, μ is our potassium mean, and σ is our potassium standard deviation, then

$$z = \frac{x - \mu}{\sigma}$$

In our example, $z = (5.5 - 4.25) / 0.383 = 3.26$. We have converted our potassium level of 5.5 mM to a ***Z* score** of 3.26. We can then use our **tables of the standard normal distribution** (you've got a copy — see p. 123) to find out how likely a Z score of 3.26 (or higher) is to have come from the standard normal distribution. This is answering the *same question* as 'how likely is a potassium level of 5.5 mM to have come from the distribution of plasma potassium in healthy people?' Our tables tell us that we want the probability that $Z \geq 3.26$, and that's 1 minus the probability that $Z \leq 3.26$, which is 0.9994; so the answer to our question is $1 - 0.9994 = 0.0006$. In other words, it's highly *unlikely* that a plasma potassium of 5.5 mM would be found in a healthy population. Our patient's probably not healthy — better watch it, because if the potassium level goes too high, he'll have a cardiac arrest.

***Z* scores carry information on their own,** because you automatically know what the mean and standard deviation are (they're 0 and 1, respectively). ***Z* scores tell you how far a score is from the mean, in terms of the number of standard deviations:** a Z score of +2.4 means '2.4 standard deviations above the mean'; a Z score of –1.5 means '1.5 standard deviations below the mean'.

Extreme *Z* scores (big positive numbers or big negative numbers) are *unlikely* to have come from the distribution in question.

Sometimes, information is presented in a normalized form. For example, IQ scores are transformed to a distribution with a mean of 100 and an SD of 15; knowing this, you can work out what proportion of the population have an IQ over 120.

(2) Assumptions of statistical tests

Second, many statistical tests assume that the data being tested are normally distributed. We will return to this point later.

(3) Confidence intervals

Third, we can work out **confidence intervals** for any measurement we make. We saw an example above: we said that 95% of healthy people have a plasma potassium concentration in the range 3.5–5.0 mM. That is the same as saying the **95% confidence interval (CI)** for healthy people's potassium is 3.5–5.0 mM.

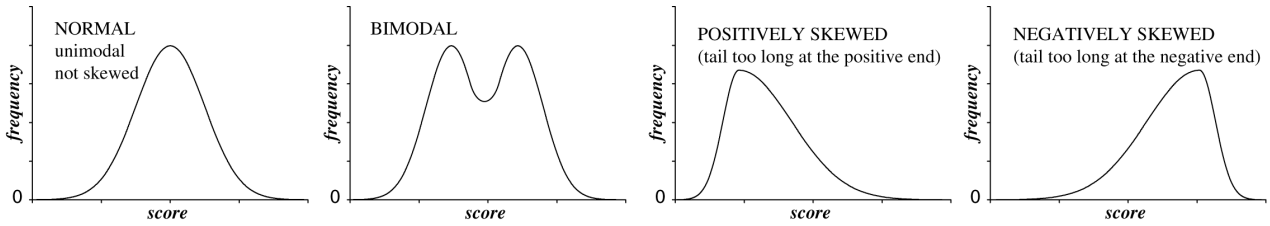
For any given set of data X, we can work out 95% confidence intervals as follows:

1. **Calculate the mean, μ, and standard deviation, σ.**
2. The Z scores that enclose 95% of the population are –1.96 and +1.96. Why? Well, our tables (see p. 123) tell us that the area (probability) under the Z curve to the left of $z = -1.96$, written, $\Phi(-1.96)$, is 0.025. Similarly, they tell us that $\Phi(+1.96) = 0.975$. Therefore the area under the normal curve between $z = -1.96$ and $z = +1.96$ is $\Phi(+1.96) - \Phi(-1.96) = 0.95$.
3. $Z = (X - \mu) / \sigma$, therefore $X = \mu + Z\sigma$. Therefore the X scores corresponding to Z scores of ±1.96 are **$\mu \pm 1.96\ \sigma$, the 95% confidence intervals.**

For our potassium example, we had a mean of 4.25 and an SD of 0.383; therefore, our 95% confidence intervals are 4.25 – (1.96 × 0. 383) and 4.25 + (1.96 × 0. 383), or 3.5 and 5.0. Try working out the 95% confidence intervals for IQ scores.

Deviations from normality

Not everything you measure will be normally distributed. Here's a normal distribution and some non-normal distributions:



Figures illustrating bimodality and ***skew.***

Continuous random variables; probability density functions

(A-Level Further Maths.) For a continuous random variable X, the probability of an exact value x occurring is zero, so we must work with the probability density function (PDF), $f(x)$. This is defined as

$$P(a \le x \le b) = \int_a^b f(x)dx \text{ where } \int_{-\infty}^{\infty} f(x)dx = 1 \text{ and } \forall x : f(x) \ge 0$$

($\forall x$ means 'for all values of x'). The mean or expected value E[X] is defined as

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

The variance, Var[X] or V[X], is given by

$$V[X] = \int_{-\infty}^{\infty} x^2 f(x)dx - (E[X])^2$$

The cumulative distribution function (CDF, also known as the 'distribution function' or 'cumulative density function'), $F(a)$, is given by

$$F(a) = \int_{-\infty}^{a} f(x)dx$$

i.e.

$$F(a) = P(x \le a)$$

$$P(a \le x \le b) = F(b) - F(a)$$

Definition of a normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

This distribution is often abbreviated to $N(\mu, \sigma^2)$.

The standard normal distribution

The 'standard' normal distribution is $N(0,1)$, i.e. a normal distribution in which μ = 0 and $\sigma = \sigma^2 = 1$. A standard normal random variable is frequently referred to as Z. The PDF is frequently referred to as $\phi(z)$, and the CDF as $\Phi(z)$. So

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \qquad \Phi(z) = \int_{-\infty}^{z} \phi(t)dt$$

Transforming any normal distribution to the standard normal distribution

As we've seen, if X is a normally-distributed random variable with mean μ and standard deviation σ, and Z is a standard normal random variable, then

$$z = \frac{x - \mu}{\sigma}$$

1.7. Probability

How much probability do you have to know? Not very much. You need to know what a probability is, what P(A) and $P(\neg$A) mean, and preferably what P(B|A) means. If you're not keen on probability, you can skip the rest of this section and move on to the logic of null hypothesis testing. If you're a bit more capable mathematically, you may like to read this section — probability is at the heart of statistical testing and you'll be streaks ahead of many researchers if you have a solid grasp of probabilistic reasoning.

Basic notation in probability

$P(A)$	probability of an event A
$P(\neg A)$	probability of the event 'not-A', the opposite of A. This is variously written as ¬A, ~A or $\overline{A}$.
$P(A \vee B)$	probability of A *or* B (or both) happening (the notation is like set union: $\cup$). Sometimes written P(A or B).
$P(A \wedge B)$	probability of A *and* B both happening (the notation is like set intersection: $\cap$). Sometimes written P(A, B) or P(A and B).
$P(B \mid A)$	probability of B, given that A has already happened, known as the **conditional probability** of B given that A has already happened

Basic laws of probability

If P(A) = 0, then A will never happen (is impossible); if P(A) = 1, then A is certain to happen. Probabilities are always in this range:

$$0 \le P(\text{A}) \le 1 \qquad [1]$$

Pick a card; there are 52 equally-likely outcomes; 13 are clubs, so $P(\clubsuit) = {}^{13}/_{52}$:

$$P(\text{A}) = \frac{\text{number of ways in which A occurs}}{\text{number of ways in which all equally likely events, including A, occur}} \qquad [2]$$

Either A happens or ¬A happens (I flip a coin, it either comes up heads or tails):

$$P(\text{A}) + P(\neg\text{A}) = 1 \qquad [3]$$
$$P(\neg\text{A}) = 1 - P(\text{A})$$

Odds

Odds are another way of expressing probability: they're the ratio of $P(A)$ to $P(\neg A)$. For example, Tiger Woods might be the favourite to win a tournament at odds of 9:5, often stated '9 to 5 on' ($= {}^{9}/_{5} = 1.8$). This means that for every 14 times he plays the tournament, he'd be expected to win 9 times and lose 5. If the event that Tiger Woods wins is A and his odds are x, we can write

$$\frac{P(A)}{P(\neg A)} = x$$

Therefore

$$\frac{P(A)}{1-P(A)} = x \ldots \quad \frac{1-P(A)}{P(A)} = \frac{1}{x} \ldots \quad x - xP(A) = P(A) \ldots \quad x = (1+x)P(A) \ldots$$

$$P(A) = \frac{x}{1+x}$$

So in the case of Tiger Woods, since $x = 1.8$, $P(A) = 0.64$. In general

$$probability = \frac{odds}{1+odds}$$

If the odds on a player were quoted as '3 to 1 *against*', the odds on them losing are 3:1 so the odds on them winning are 1:3 (i.e. probability of them winning is ¼ = 0.25).

The rest of the basic laws of probability

If A and B are **mutually exclusive** events ($\Rightarrow P(A \wedge B) = 0$) then

$$P(A \vee B) = P(A) + P(B) \qquad [4]$$

In the more general case,

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) \qquad [5]$$

If A and B are **independent** events — that is, the fact that A has happened doesn't affect the likelihood that B will happen, and vice versa: $P(B) = P(B \mid A)$ and $P(A) = P(A \mid B)$ — then

$$P(A \wedge B) = P(A) \times P(B) \qquad [6]$$

If I toss a fair coin and roll a fair die, the probability of getting a six and a head is $^1/_6 \times {}^1/_2 = {}^1/_{12}$. The probability of getting a six *or* a head *or* both is $^1/_6 + {}^1/_2 - {}^1/_{12} = {}^7/_{12}$.

In the more general case:

$$P(A \wedge B) = P(A) \times P(B \mid A) \qquad [7]$$

If I have a bag that initially contains 4 red marbles and 6 blue marbles, and I withdraw marbles one by one, the probability of picking a red marble first (event A) and a blue marble second (event B) is $^4/_{10} \times {}^6/_9 = {}^4/_{15}$.

A bit more advanced: Bayes' theorem

From [7],

$$P(B \mid A) = \frac{P(A \wedge B)}{P(A)} \qquad [8]$$

We also know, from [7],

$$P(A \wedge B) = P(B \wedge A) = P(B) \times P(A \mid B)$$

Therefore, from [8],

$$P(B \mid A) = \frac{P(B) \times P(A \mid B)}{P(A)} \qquad [9]$$

This is the simplest statement of **Bayes' theorem.** Suppose event A is discovering an improperly-sealed can at a canning factory. We know there are *k* assembly lines at which cans are sealed, and we'd like to know which one produced the faulty can. Let's call B_1 the event in which assembly line 1 produced the faulty can, B_2 that in which line 2 produced the faulty can, and so on. What's the probability that the can came from line *i*?

We know that a faulty can must have came from one of the assembly lines:

$$P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + \ldots + P(B_k)P(A \mid B_k)$$

or to write that in a shorter form:

$$P(A) = \sum_{j=1}^{k} P(B_j)P(A \mid B_j)$$

Therefore, from [9],

$$P(B_i \mid A) = \frac{P(B_i) \times P(A \mid B_i)}{\sum_{j=1}^{k} P(B_j) P(A \mid B_j)} \qquad [10]$$

So suppose there are three assembly lines; lines X, Y and Z account for 50%, 30% and 20% of the total output. Quality control records show that line X produces 0.4% faulty cans, Y produces 0.6% faulty cans, and Z produces 1.2% faulty cans. Using Bayes' theorem in the form of [10] will tell us that the chance our faulty can comes from assembly line X is 0.32 (similarly, 0.29 for line Y and 0.39 for line Z).

Let's take a simple, fictional example in which only two things may happen. **Q.** The prevalence of a disease in the general population is 0.005 (0.5%). You have a blood test that detects the disease in 99% of cases: *P*(positive | disease) = 0.99. However, it also has a false-positive rate of 5%: *P*(positive | no disease) = 0.05. A patient of yours tests positive. What is the probability he has the disease? **A.** We'd like to find *P*(disease | positive). By [9],

$$\begin{aligned} P(dis \mid pos) &= \frac{P(dis) \times P(pos \mid dis)}{P(pos)} \\ &= \frac{P(dis) \times P(pos \mid dis)}{P(dis) P(pos \mid dis) + P(\neg dis) P(pos \mid \neg dis)} \\ &= \frac{0.005 \times 0.99}{0.005 \times 0.99 + 0.995 \times 0.05} \\ &= 0.09 \end{aligned}$$

So even though our test is pretty good and has a 99% true positive rate or 'sensitivity' (a 1% false negative rate) and a 5% false positive rate (a 95% true negative rate or 'specificity'), our positive-testing patient still only has a 9% chance of having the disease — because it's rare in the first place.

Bayesian inference

Suppose we have a hypothesis H. Initially, we believe it to be true with probability *P*(H); we therefore believe it to be false with probability *P*(¬H). We conduct an experiment that produces data D. We knew how likely D was to arise if H were true — *P*(D|H) — and we knew how likely D was to arise if H were false — *P*(D|¬H). We can therefore use Bayes' theorem [9] to *update* our view of the probability of H:

$$P(H \mid D) = \frac{P(H) P(D \mid H)}{P(D)}$$

$$P(H \mid D) = \frac{P(H) P(D \mid H)}{P(H) P(D \mid H) + P(\neg H) P(D \mid \neg H)}$$

This can be expressed another way (Abelson, 1995, p. 42):

$$\frac{P(H \mid D)}{P(\neg H \mid D)} = \frac{P(H)}{P(\neg H)} \times \frac{P(D \mid H)}{P(D \mid \neg H)} \qquad [11]$$

or

posterior odds = prior odds × relative likelihood

1.8. The logic of null hypothesis testing; interpreting *p* values

We will come across a range of statistical tests. Most produce a *test statistic* and an associated *p* value; you will see these quoted in scientific journals time and time again (like this: $F_{2,47} = 10.7$, $p < .001$… $F_{3,18} = 4.52$, $p = .016$… $t_{60} = 1.96$, $p = .055$). They all work on the same principle: that of **null hypothesis testing.**

Null hypothesis testing approaches the questions we want to ask *backwards*. We typically obtain some data. Let's say we measure the weight of a hundred 18-year-old women who are either joggers (50) or non-joggers (50). We would like to know whether the mean weights of these two group differ. Obviously, it's highly unlikely that the means will be *exactly* the same. Suppose the joggers are slightly lighter on average. How big a difference counts as 'significantly' different? The conventional logic is as follows. Either the difference arises through chance, or there is some systematic difference (such as that jogging makes you thin, or that being thin encourages you to take up jogging). Our **research hypothesis** (sometimes written H_1) is that the joggers are different from the non-joggers (that our two samples come from different underlying populations). We'll invent a corresponding **null hypothesis** (sometimes written H_0) that the observed differences arise purely through chance. We'll then test the likelihood that our data could have been obtained if this null hypothesis were true. If this probability (the so-called ***p* value**) is very low, we will **reject** the null hypothesis — chance processes don't appear to be a sufficient explanation for our data, so something systematic must be going on; we'll say that there is a significant difference between our two groups. If the *p* value isn't low enough, we will **retain** the null hypothesis (applying Occam's razor — because the null hypothesis is the simplest on offer) and say that the groups do not differ significantly.

The exact meaning of a p value

Let's say we run a statistical test to examine whether these two groups differ. It produces a test statistic (such as a *t* value; we'll consider how this works later) and a *p* value — let's say 0.01. What does this mean? For shorthand, let's call D the event of obtaining a set of data, H be the research hypothesis, and ¬H the null hypothesis.

- **Correct:** "If the null hypothesis were true [if it were true that there were no systematic difference between the means in the populations from which the samples came], the probability that the observed means would have been as different as they were, or more different, is 0.01. This being strong grounds for doubting the viability of the null hypothesis, the null hypothesis is rejected."
- **Correct:** "The probability of these data (or something more extreme) being observed if the null hypothesis were true is 0.01."
- **Correct:** $P(D \mid \neg H) = 0.01$.
- **Wrong:** "The probability that the null hypothesis is true is 0.01."
- **Wrong:** "The probability that the results are a 'fluke' is 0.01."
- **Wrong:** "The probability that the research hypothesis is false is 0.01."
- **Wrong:** $P(\neg H \mid D) = 0.01$.
- **Wrong:** $P(\neg H) = 0.01$.
- **Wrong:** "The null hypothesis is disproved."
- **Wrong:** "We have established the probability of the null hypothesis being true."
- **Wrong:** "We know, if we decide to reject the null hypothesis, the probability that we are making an error."
- **Wrong:** "The probability that the null hypothesis is false is 0.99."
- **Wrong:** "The probability that the research hypothesis is true is 0.99."
- **Wrong:** $P(H \mid D) = 0.99$.
- **Wrong:** $P(H) = 0.99$.
- **Wrong:** "The research hypothesis has been proved."
- **Wrong:** "We have established the probability of the research hypothesis being true."
- **Wrong:** "If the experiment were to be replicated, there is a 99% chance that a significant result would again be found." [To see why this is wrong, consider the situation where you run an experiment and obtain $p = 0.05$ — just on the threshold of 'significance', if we use the conventional $\alpha = 0.05$. What is the

chance of getting a 'significant' result again if you repeated or replicated the experiment exactly? Well, half the time you'd expect a bigger effect and half the time you'd expect a smaller effect. A bigger effect would give $p < 0.05$, while a smaller effect would give $p > 0.05$. So the chance that the replicated experiment would produce a 'significant' result would be 50%, not 95%. (See Oakes, 1986, p. 18; Abelson, 1995, p. 75.)]

You will find countless web sites, articles, people, and even occasionally statistics textbooks that make one or more of these mistakes (for details, see Abelson, 1995, p. 40). In one study, only 3% of academic psychologists answered correctly a series of six true / false questions like the statements above (Oakes, 1986, p. 79).

It's easy to think that all these statements are saying the same thing, but they're not. The main problem is understanding the difference between P(¬H | D), which you'd really like to know, and P(D | ¬H), which is what statistical tests tell you. Compare (1) the probability of testing positive for a very rare disease if you have it, *P*(positive | diseased), with (2) the probability of having it if you test positive for it, *P*(diseased | positive). If you think the two should be the same, you're neglecting the 'base rates' of the disease: typically, the second probability is less than the first, as it's very unlikely for anybody to have a very rare disease, even those who test positive. Doctors intuitively get this wrong all the time. Substitute in *P*(rich | won the lottery) and *P*(won the lottery | rich)… the first probability is much higher, because winning the lottery is so rare.

Bayes' theorem and Bayesian statistics

The formal way to relate what we get from significance tests, *P*(data | ¬hypothesis), to what we really want, *P*(hypothesis | data), is by using Bayes' theorem (see p. 19). This is perhaps the simplest expression to use in this case:

$$\frac{P(H \mid D)}{P(\neg H \mid D)} = \frac{P(H)}{P(\neg H)} \times \frac{P(D \mid H)}{P(D \mid \neg H)}$$

posterior odds = prior odds × relative likelihood

For example, suppose that a climatologist calculates that a 1°C rise in temperature one summer had a probability of 0.01 of occurring by chance ($p = 0.01$). What does that tell us? It does *not* tell us that there's a 99% probability that it was due to the greenhouse effect. It does not even tell us that there's a 99% probability that it was not due to chance. The Bayesian approach would be this: suppose that reasonable people believed the odds were 2:1 in favour of the greenhouse hypothesis (H) before this new evidence was collected — these are the *prior odds*. Now, we've been told that $P(D|\neg H) = 0.01$. We need to know the probability that a 1°C temperature rise would occur if the greenhouse hypothesis were true; that is, $P(D|H)$. Suppose this is 0.03. Then the *relative likelihood* is 0.03 / 0.01 = 3. So the *posterior odds* are 2 × 3 = 6 in favour of the greenhouse hypothesis; odds of 6:1 equate to $P(H|D) = {}^{6}/_{7} = 0.86$.

Type I and Type II error; power

Although *p* values speak for themselves in one sense, it's very common for researchers to use them as a yes / no decision-making device. I won't debate the wisdom of this now, but this is how it works. A threshold probability, usually called ***α* (alpha),** is chosen; typically, $\alpha = 0.05$. If a given *p* value is less than α, the null hypothesis is rejected; if $p \geq \alpha$, the null hypothesis is retained. You might see this logic described in papers like this: 'the two groups were significantly different ($p < 0.05$),' or 'a significance level of $\alpha = 0.05$ was adopted throughout our study… the two groups were significantly different.'

Obviously, if $\alpha = 0.05$, then *there is a 0.05 (one in twenty) chance that an effect we label as 'significant' could have arisen by chance if the null hypothesis were true*. If this happens, and we accidentally decide that a effect was not attributable to chance when actually it did arise by chance, we're said to have made a **Type I error.** The

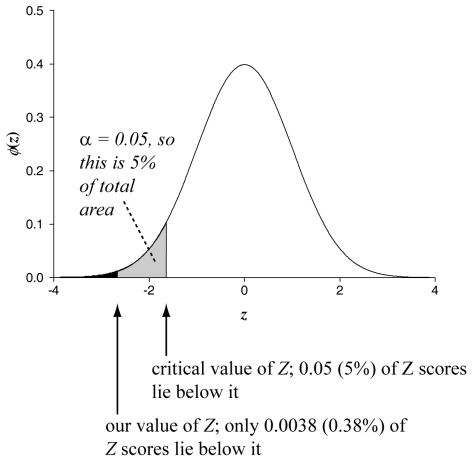
probability of making a Type I error is α. Conversely, the probability of correctly not rejecting the null hypothesis when it is true is $1 - \alpha$.

The opposite mistake is failing to reject the null hypothesis when it is false — that is, ascribing your data to chance when they actually arose from a systematic effect. This is called a **Type II error;** its probability is labelled ***β* (beta).** Conversely, the probability of correctly rejecting the null hypothesis when it is in fact false is $1 - \beta$; this is called the **power** of the test. If your power is 0.8, it means that you will detect 'genuine' effects with $p = 0.8$.

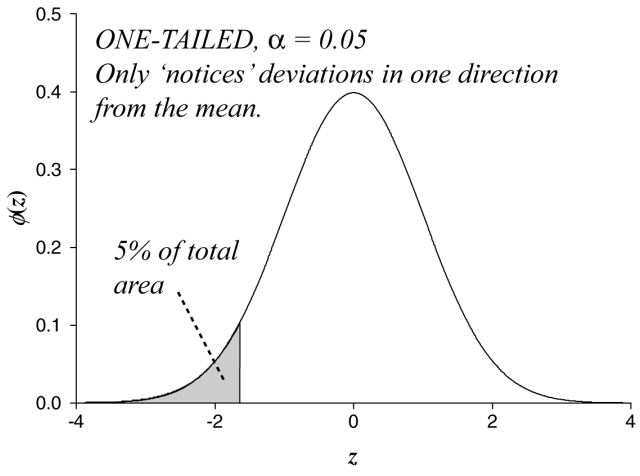
	True state of the world	
Decision	**H_0 true**	**H_0 false**
Reject H_0	Type I error $p = \alpha$	Correct decision $p = 1 - \beta$ = power
Do not reject H_0	Correct decision $p = 1 - \alpha$	Type II error $p = \beta$

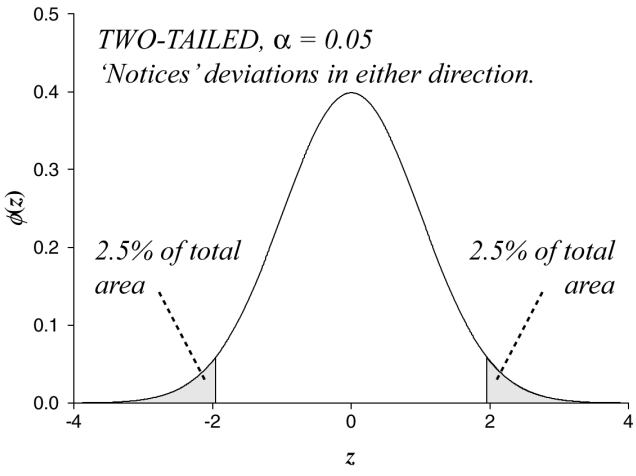
One-tailed and two-tailed tests

There's one other thing we should consider when we talk about α and Type I error. Let's go back to the example of our joggers. Presumably our leading hypothesis is that joggers will be thinner than non-joggers, so we want to be able to detect if the mean weight of joggers is less than that of non-joggers, and we might choose $\alpha = 0.05$. But what will we do if the joggers actually weigh *more*? Well, this depends on what kind of test we decided on. If we were only interested in the difference between the groups if the joggers weighed less, we would use a **one-tailed (directional) test,** so that if there was less than a 5% probability that chance alone could have produced a difference *in the direction we expect* then we would reject the null hypothesis. But if we want to be able to detect a difference in either direction, we must use a **two-tailed (nondirectional) test.** In that case, we must 'allocate' our 5% α to the two ways in which we could find a difference (joggers weigh more; joggers weigh less) — so we'd allocate 2.5% to each **tail** of the distribution. This is shown



One-tailed and two-tailed tests. ***Left:*** *what it means to declare a Z score of –2.667 to be 'significant', using a* one-*tailed Z test (*$\alpha = 0.05$*). For this example, we are only interested in Z scores that are* less *than zero. The score being tested is in the extreme lower 5% of the distribution, so the probability of obtaining a score that extreme (or more) if the null hypothesis were actually correct (if the distribution shown was the true distribution) is less than 5%.* ***Below:*** *the difference between one- and two-tailed tests is that the criterion (or critical) values must be altered. If the deviation can be* either *way (less than or greater than the value predicted by the null hypothesis), then, in order to reject the null hypothesis no more than 5% of the time when it is actually true, we must allocate 2.5% to each 'tail' of the distribution (and thereby alter the critical values).*





in the figure (plotted on a normal distribution; you might like to think of it in terms of the joggers and the potassium examples). In general, unless you would genuinely not be interested in both possible outcomes (quite a rare situation), you should use a two-tailed test. What you must *not* do is to run a one-tailed test ($\alpha = 0.05$), find a non-significant result, then look at the data, realize the difference is in the other direction to the one you predicted, and decide then to do a two-tailed test ($\alpha = 0.05$) — because what you have actually done is to allocate 5% to one tail, *then* allocate another 2.5% to the other tail, meaning that you have actually run a sort of asymmetric two-tailed test with a total α of 0.075 (7.5%) (see Abelson, 1995, p. 58). Decide what test you want **in advance** of analysing the data.

The danger of running multiple significance tests

Every time you run a test, if the null hypothesis is true, you run the risk of making a Type I error with probability α. So if you run n tests, you have n chances to make a Type I error. What's the probability that you don't make any Type I errors when you run n tests? Well, the probability that you don't make a Type I error on each test is $1 - \alpha$, so the probability you make no Type I errors when you run n tests is $(1 - \alpha)^n$. So the probability that you make at least one Type I error when you run n tests when the null hypothesis is true is $1 - (1 - \alpha)^n$.

If you set $\alpha = 0.05$, you must expect on average one in every 20 tests to come up 'significant' when it isn't (Type I error) **if the null hypothesis is in fact true.** If you run 20 tests **and the null hypothesis is true,** the probability of making at least one Type I error is $1 - (1 - 0.05)^{20} = 0.64$. This is why running lots of tests willy-nilly is a Bad Idea — eventually, something will 'turn up significant', but that doesn't mean it really is.

This doesn't mean that 5% of all your significant results are 'wrong'. You can only make Type I errors when the null hypothesis is true! In practice, on some occasions the null hypothesis will be false, so we can't make a Type I error. Therefore, something less than 5% of our 'significant' results will be Type I errors; α is the **maximum Type I error rate.**

Is there a difference between p = 0.04 and p = 0.0001?

Yes. Whether you look on p values as expressing the degree of confidence with which you reject the null hypothesis, or as information you can use to update your opinions of the world in Bayesian fashion, p values have real meaning. Some people will argue that so long as $p < \alpha$ you needn't report the actual p value, but this approach takes information away from the reader.

p = 0.06

What happens if you run a well-designed experiment in which you give a treatment to one group of people and not another, measure some aspect of their performance, test for a difference between your groups and get $p = 0.06$? You could do one of several things. (1) Re-run your experiment with more subjects; perhaps you did not have enough statistical *power* to detect the size of effect that your treatment produced. You might have been spared this embarrassment if you had tried to calculate your statistical power in advance; you might then have realised your experiment was under-powered in the first place. (2) Report your experiment as showing a 'trend' towards an effect; it's not like $p = 0.04$ is somehow magically better than $p = 0.06$, after all. (3) Use $\alpha = 0.1$ rather than $\alpha = 0.05$. However, not only will journal editors definitely be upset with this (for no real reason — there's nothing magical about $\alpha = 0.05$), but it is highly dubious to change your α only *after* you've run your experiment — after all, you're only doing it to shore up a not-quite-significant result, and you're therefore distorting the results. You should have chosen α in advance. Similarly, it is very dubious to add subjects to your original experiment 'until it reaches significance' — you're only doing this because your original data was 'near' significance and you want it to be significant. If you had a compelling reason to want your treatment to have no effect, you wouldn't be doing this — so you're biasing the experiment by this kind of *post-hoc* fiddling. (4) Retain the null hypothesis; see below.

What does 'not significant' mean?

What happens when you want to prove that a hypothesis is *not* true? Suppose your contention is that jogging doesn't affect body weight; you take two identical groups of people, set half of them jogging for a couple of months while the rest eat pies, and measure their weights. You find no difference between the groups ($p = 0.12$). What does this mean? It means that you have *failed to reject the null hypothesis* — there is a fair chance (0.12) that your observed difference could have arisen by chance alone. *It does not mean that you have proven the null hypothesis.* Take an extreme example: your null hypothesis is that all people have two arms. Just because the next 5,000 people you meet all have two arms (failure to reject the null hypothesis) does not mean that you have proved the null hypothesis.

You can do two things when you fail to reject the null hypothesis: (1) view it as an *inconclusive* result, or (2) act *as if* the null hypothesis were true until further evidence comes along.

Really, you should consider your level of α and β to meet the needs of your study. If you want to avoid Type I errors (e.g. telling someone they have an ulcer when they don't), set α low. If you want to avoid Type II errors (e.g. telling them to go home and rest when they're about to die from a gastric haemorrhage), set α higher. The other thing you can do when you're designing an experiment is to make sure the *power* is high enough to detect effects with a reasonable probability — such as by using enough subjects. If you take two people and make one jog, you'll never find a 'significant' difference between the jogging and non-jogging groups, but that doesn't mean people should believe you when you say that jogging doesn't reduce weight. If you used half a million people and still found no effect, your study might command more attention.

A statistical fallacy to avoid: A differs from C, B doesn't differ from C…

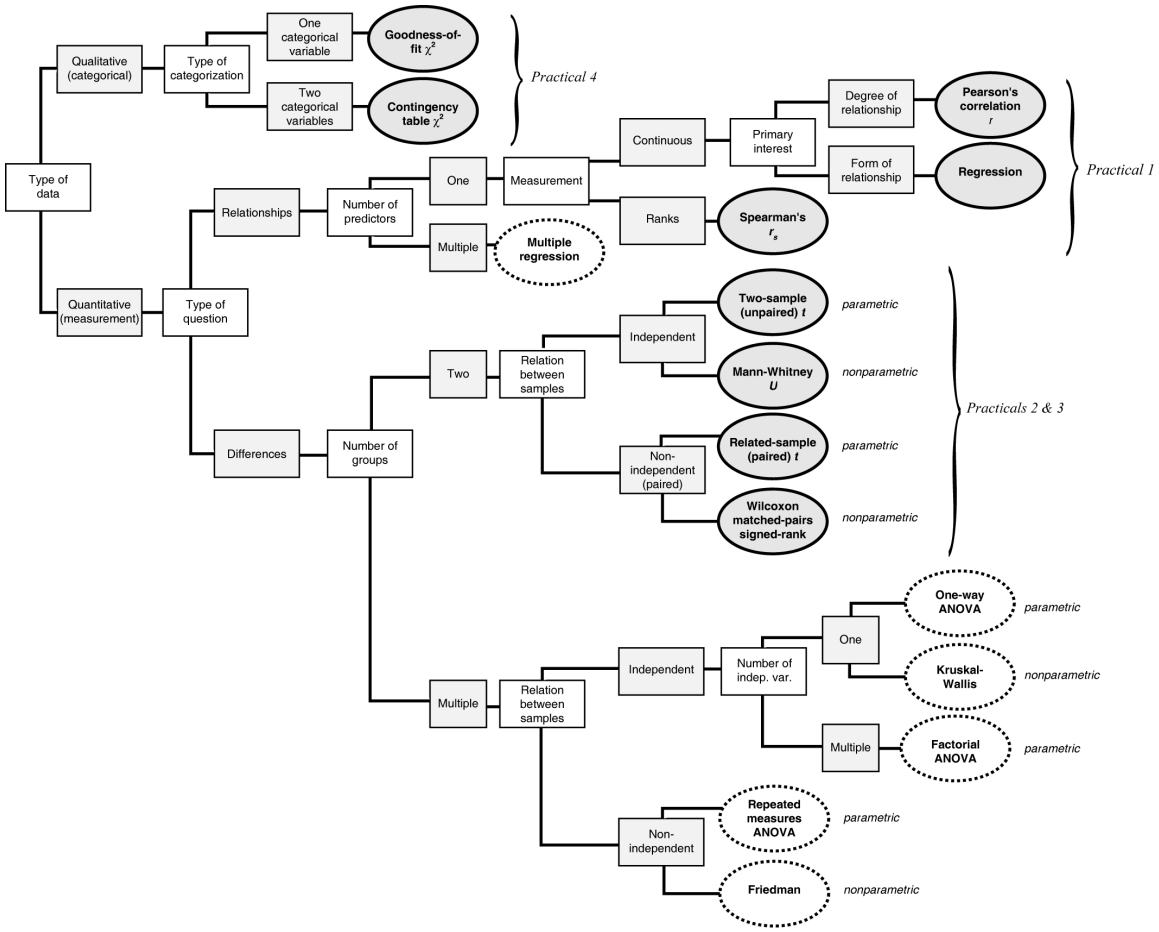
If you test three groups and find that A is significantly different from C, but B is not significantly different from C, *do not conclude that A is significantly different from B.* To see why, imagine that A is smaller than B, and B is smaller than C. Then we might find a difference between A and C ($p = 0.04$) and no difference between B and C ($p = 0.06$) — but the p values are just on either side of our threshold of 0.05 and A and B might be nearly the same! Making this conceptual mistake is quite common.

Similarly, just because A isn't significantly different from B, and B isn't significantly different from C, doesn't mean that A isn't significantly different from C.

1.9. For future reference… how the different statistical tests fit together

Overview of statistical tests

This flowchart (based on Howell, 1997, p.11) should help you fit the various statistical tests we'll cover into a coherent framework. **It's NOT intended to be a prescriptive 'use this test in this circumstance' chart —** once you understand what a test does, you can apply it whenever *you* feel it's appropriate. And **DON'T TRY TO LEARN IT! Tests with dotted lines around them aren't covered in the IB course.**



Descriptive statistics in Excel — relevant functions (see Excel help for full details)

Excel does basic analysis (especially if you switch on the Analysis ToolPak, available in Excel 97 from the *Tools* → *AddIns* menu, and thereafter from *Tools* → *Data Analysis*) and can generate quite good graphs, with a little playing. But in the exams you'll be required to do basic statistical tests with a calculator, so don't become reliant on a computer yet.

AVERAGE(…)	Mean $\bar{x}$ of a group of cells; e.g. AVERAGE(A1:A6) gives the mean of cells A1, A2… A6.
MEDIAN(…)	$\tilde{x}$ or Med(X)
MODE(…)	Mo(X)
COUNT(…)	n
VARP(…)	population variance σ^2
VAR(…)	sample variance s^2
STDEVP(…)	population standard deviation σ
STDEV(…)	sample standard deviation s
STANDARDIZE()	converts a value X into a standardized normal value Z (you have to supply X, μ and σ).
NORMSDIST()	the standard normal cumulative distribution function, $\Phi(z)$. Give it a z score and it returns a cumulative probability, i.e. $p(Z \le z) = \Phi(z) = \int_{-\infty}^{z} \phi(t)dt$.
NORMSINV()	the inverse standard normal cumulative distribution function, $\Phi^{-1}(z)$. Give it a cumulative probability $p(Z \le z)$ and it'll tell you the z score associated with that probability.

1.10. Examples 1: background, normal distribution

Q1. Given that a variable *X* is distributed normally with mean value 23.5 and standard deviation 3.0, find the probability that:

(a) $X < 28$
(b) $X < 17$
(c) $X > 30$
(d) $26 < X < 28$
(e) X differs from its mean by more than 6.0

Q2. IQ scores are derived in such a way that the mean for the population is 100 and the standard deviation is 15. In a population of 60 million, how many have IQs

(a) >145
(b) <80
(c) within one standard deviation of the mean?

Q3. African meerkats (*Suricata suricatta*) have a mean height of 30 cm with a variance of 4 cm. Their heights are normally distributed.

(a) What is the standard deviation of the heights of African meerkats?
(b) What proportion of meerkats are between 30 and 31 cm tall?
(c) If you took a thousand randomly-selected meerkats, how many would you expect to be shorter than 28.5 cm?
(d) What are the 95% confidence intervals for meerkat heights (the heights, centred around the mean, within which 95% of meerkat heights fall)?
(e) Pilchard the meerkat is 33 cm tall. What is the probability that a meerkat of Pilchard's height (or greater) could come from the population of African meerkats?
(f) What is the probability that a meerkat whose height is less than Pilchard's could come from this population?
(g) What is the approximate probability that a meerkat whose height is *exactly* that of Pilchard's could come from this population?

Q4. A researcher reporting the results of a functional imaging study states that blood flow in the left cerebellum decreased while subjects thought about music. The researcher calculated this change to be equivalent to a *Z* score of –2.4.

(a) What is the probability that this *Z* score (or one still more extreme in the same, negative, direction) could have arisen by chance?
(b) What is the probability that a *Z* score of $\geq +2.4$ or ≤ -2.4 could have arisen by chance?
(c) If mean left cerebellar blood flow is 50 ml per minute with a standard deviation of 5 ml per minute while the subjects were resting, what was the left cerebellar blood flow while the researcher's subjects were thinking about music?
(d) If the researcher had simultaneously scanned 100 areas of the brain and calculated *Z* scores for each of them (by comparing 'music' blood flow to 'resting' blood flow in each case), what is the probability of obtaining at least one *Z* score at least as extreme as ±2.4 if listening to music did not in fact affect brain blood flow at all?

Q5. A traveller one day was making his way
through a woods both wild and deep.
The road split in two and he muttered, '*Mon Dieu*!
Where in the world shall I sleep?'
To the south lay an inn most noted for sin,
but a constable shrewd and upright.
To the north (said the sign), a village benign,
a safer abode for the night.
The village, it seemed, was a traveller's dream,
for thieves chose the inn, as a rule,
But the sheriff in town had earned wide renown
as a lax and incompetent fool.

Suppose that the probability of being robbed is .60 at the inn and .20 in the village. On the other hand, the probability that the constable at the inn will recover the traveller's money is .70, but the probability that the sheriff in the village will recover the traveller's money is only .10. Which is the better choice?

2. Correlation and regression

Objectives

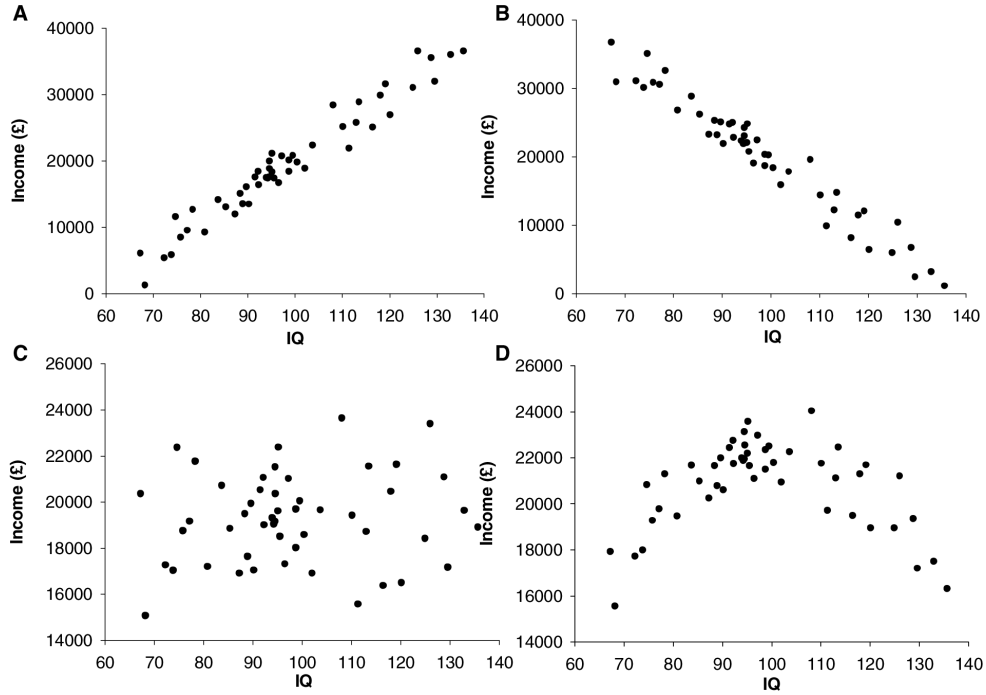
We'll examine two ways to examine the relationship between two variables — correlation and regression. They're conceptually very similar.

Stuff with a solid edge, like this, is important.

But remember — you can totally ignore stuff with single / double wavy borders.

2.1. Scatter plots

Suppose you measure two things about a group of subjects — IQ and income, say. How can we establish if there's any relationship between the two? The first thing to do is to draw a **scatter plot** of the two variables. To do this, we take one of our variables (e.g. IQ) as the *x* axis, and the other as the *y* axis. Each subject is then plotted as one point, representing an {IQ, income} pair. This might show us any of several things:



*Fictional scatterplots. **A: positive correlation** between IQ and income. As IQ goes up, income goes up. **B: negative correlation** between IQ and income. As IQ goes up, income goes down. **C: no relationship** between the two. **D:** there's a relationship, but it's not a straight line (it's not a **linear** relationship). People with high IQs and people with low IQs both earn less than those with middling IQs.*

It's always worth plotting the data like this first. However, for our next trick we'd like a statistical way to work out **if** there's a relationship, **how big** it is, and **in what direction** it goes. Please note that we'll only talk about ways to establish things about a **linear relationship** between two variables; if it's non-linear (e.g. the bottom right figure), it's beyond the scope of this course.

2.2. Correlation

We will call the degree to which they are related the **correlation** between the two variables. If *Y* gets bigger when *X* gets bigger, there's a **positive correlation;** if *Y* gets smaller when *X* gets bigger, there's a **negative correlation;** if there's no linear relationship, there's a **zero correlation.** Here's how we work it out.

The covariance

First, we need some sort of number that tells us how much our two variables vary together. Let's suppose we have *n* observations. Let's call our two variables *X* and *Y*. We first find the two means, $\bar{x}$ and $\bar{y}$. Then we can calculate something called the **sample covariance:**

$$\text{cov}_{XY} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

Look at the first part of the equation first — it's very like the sample variance (if we changed all the *y*s to *x*s in this equation, we'd have s_X; if we changed all the *x*s to *y*s, we'd have s_Y). (And yes, if you're wondering, if we wanted the population covariance, we'd divide by *n* rather than *n*–1, but we don't.)

Perhaps you can see from the equation how it works. For a given {*x*, *y*} point, if *x* is very far above the *x*-mean ($\bar{x}$), and *y* is very far above the *y*-mean ($\bar{y}$), then a big number gets added to our covariance. Similarly, if *x* is very far below the *x*-mean ($\bar{x}$), and *y* is very far below the *y*-mean ($\bar{y}$), then a big number gets added to our covariance. Both these occurrences suggest a positive linear relationship (like the top-left part of our figure). On the other hand, if *x* is very far above $\bar{x}$, and *y* is very far below $\bar{y}$, then a large *negative* number gets added to our covariance; the same's true if *x* is very far below $\bar{x}$ and *y* is very far above $\bar{y}$. Points near the mean don't tell us so much about the relationship between *x* and *y*, and they don't contribute much to the covariance score. If there's no relationship between *X* and *Y*, then when *x* is above $\bar{x}$, about half the time *y* will be above $\bar{y}$ and the covariance will get bigger, but about half the time *y* will be below $\bar{y}$ and the covariance will get smaller, so the covariance ends up being about zero.

The Pearson product–moment correlation coefficient, r

The covariance tells us how much the two variables are related, but it has a problem — the actual value of the covariance depends on the standard deviations of our two variables as well as the correlation between them. A covariance of 140 might be an high correlation if the standard deviations are small, but a poor correlation if the standard deviations are large. We can get round this problem by calculating *r*:

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y}$$

It turns out that *r* varies from –1 (perfect negative correlation), through 0 (no correlation), to +1 (perfect positive correlation).

Incidentally, the correlations in our picture were +0.79 (figure A), –0.81 (figure B), 0.10 (figure C), –0.04 (figure D).

'Zero correlation' doesn't imply 'no relationship'

That should be immediately apparent: I've just told you that the correlation between IQ and income in figure D was –0.04, nearly zero, and yet there's clearly a very strong relationship — it just isn't a linear one. **Always plot your data** to avoid drawing mistaken conclusions from *r* values.

Correlation does not imply causation

Finding that X and Y are related does not mean that X and Y are **causally** related. It's easy to jump to this assumption if the relationship is plausible — we might intuitively think that clever people get better jobs, for example, and thus accept a positive correlation between IQ as income as indicating causation. It doesn't. Maybe the causal relationship is backwards: having more money might improve your IQ. Maybe the two are connected through a third variable: Z causes X and Z causes Y (e.g. maybe having rich parents means you're more likely to have a high IQ, and

also makes you more likely to get a well-paid job as an adult). The point is, we just can't tell from the plain correlation.

Adjusted r

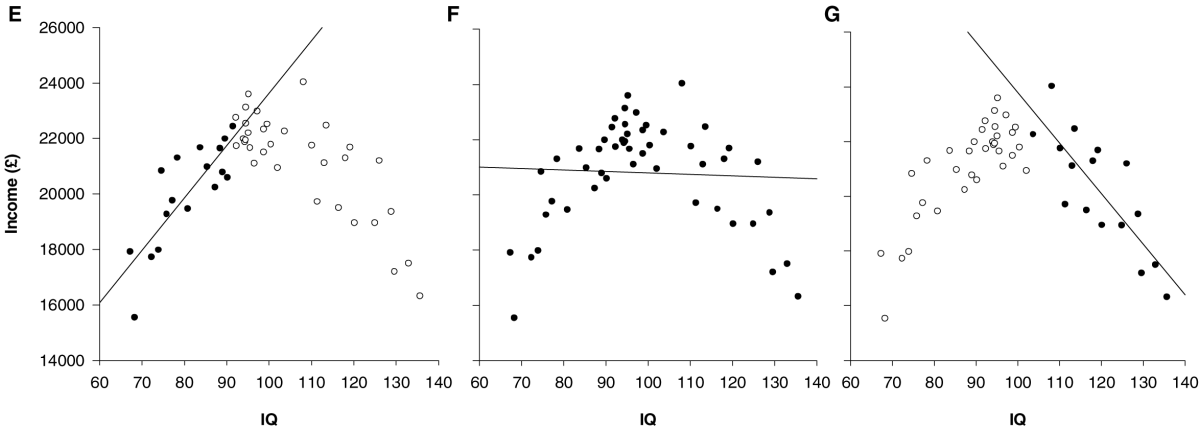
If we measured IQ and income for a *sample* of just two people — let's say {IQ 110, £20,000} and {IQ 120, £25,000} — and calculate *r*, we'll find that there's a perfect correlation, +1. If you plot only two points on a scatterplot, you can always join them perfectly with a straight line. This doesn't mean that the correlation is +1 in the *population!* So there's something slightly wrong with our sample correlation statistic, *r* — it's a *biased estimator* of the **population correlation,** which we write as ρ (Greek letter rho). We can to do something to make it a better **(unbiased)** estimator. We can calculate the **adjusted *r*, r_{adj}:**

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}}$$

If the sample size is large, *r* and r_{adj} will be about the same. Please note that doing this will give you a *positive value* for r_{adj}, since square roots can't be negative… so you need to look at the original data or *r* value to work out *which way* (+ or –) the correlation should be.

Beware if your correlation is based on a restricted range of data

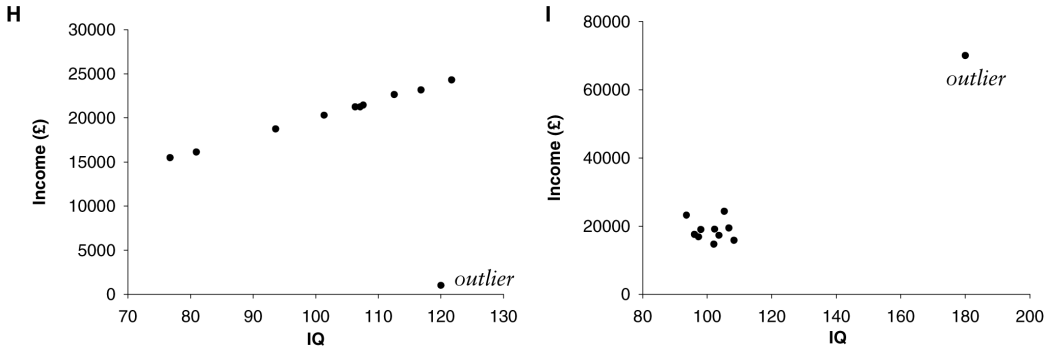
It's obvious that if we sample too few data points, we won't get a very good estimate of *r* for our population (that's what calculating r_{adj} is meant to sort out). But it should also be clear that if we sample from a **restricted range,** we can also get the wrong answer, even if we sample many observations within that restricted range. Here's an extreme example (figures E–G below): depending on the range of data we sample, we can contrive to find a negative, zero, or positive correlation between our two variables.



*Sampling a restricted range of data can overestimate r **(E)** or underestimate r **(G)** compared to sampling the whole range **(F).** Black dots are part of the sample; white dots are part of the population that wasn't sampled. The straight line represents the correlation.*

Beware outliers

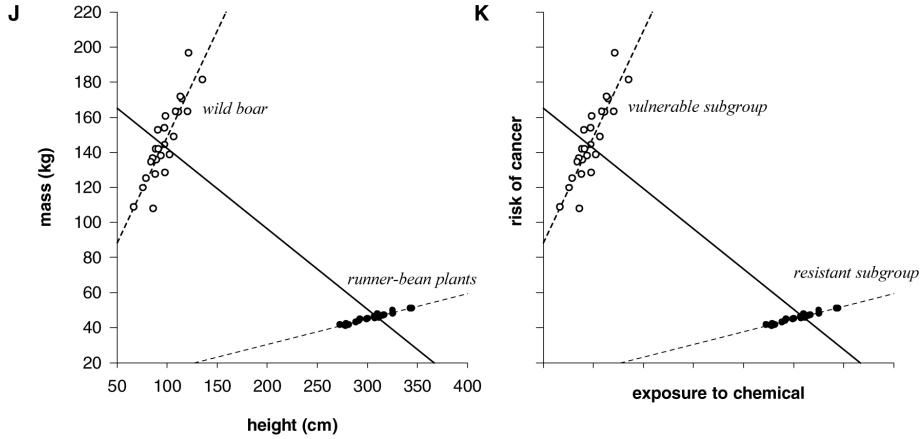
Extreme values, or *outliers*, can have large effects on the correlation coefficient. (We won't talk about what to do with them in the IB course, but you should be aware of the problems they can cause.) Two examples are shown in figures H–I.



*Outliers can have large effects on r. In **(H)** the outlier makes r nearly 0; without it, r would be nearly 1. In **(I),** the outlier makes r nearly 1; without it, r would be nearly 0.*

Beware if your population has distinct subgroups

We can also encounter problems if our measurements aren't from one homogeneous population. A couple of examples are shown in figures J–K (but subgroup effects can be a good deal more subtle than this!).



*Fictional data illustrating problems with subgroups. **(J)** Correlation between height and weight for various things we found in a magic forest. If we measure an overall correlation, we may find that tall things weigh less (negative correlation between height and mass), but this is only because we have two very different subgroups. But we have **heterogeneous subsamples** — within each subgroup (wild boar and runner beans) there is a positive correlation. **(K)** A less stupid example. If we are investigating whether something is carcinogenic, we might find a negative correlation, suggesting (if we have designed our experiment so that we know that the chemical **caused** any observed change in cancer rate) that the chemical protects from cancer. But we must check, because this could be due to a subgroup effect: a more detailed analysis may reveal a vulnerable subgroup (who get high rates of cancer) and a resistant subgroup (who aren't as likely to get cancer); in this example, the rates of cancer are actually increased by the chemical in both subgroups.*

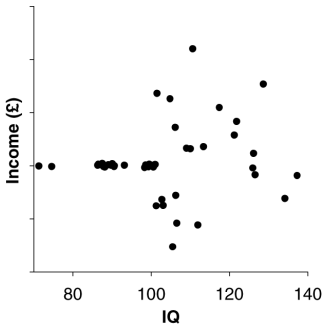
2.3. Is a correlation 'significant'?

Assumptions we must make

If all we want to do is to *describe the sample* that we have (e.g. with correlation and / or regression), we don't have to make any assumptions — although correlation and regression both aim to describe a **linear relationship** between two variables, so if the relationship isn't linear, then the answers we get from correlation and regression won't *mean* very much.

But if we want to perform statistical tests with the data (e.g. 'is this correlation coefficient significantly different from zero?'), we will effectively be asking questions to do with the *underlying population* that our sample was drawn from (i.e. 'what is the chance that a sample with correlation r came from an underlying population with correlation $\rho = 0$?'). This requires making some assumptions, or our statistical tests won't be meaningful. Basically, the data shouldn't look too weird:

- The variance of *Y* should be roughly the same for all values of *X*. This is often called *homogeneity of variance;* its opposite, what you don't want, is called *heteroscedasticity* (Greek *homo* same, *hetero* other, *skedastos* able to be scattered).
- If we are asking questions about ρ, we must assume that both *X* and *Y* are normally distributed.
- For all values of *X*, the corresponding values of *Y* should be normally distributed, and vice versa. [You may see the last two assumptions referred to together as the assumption of 'bivariate normality'.]



Heteroscedasticity: a Bad Thing. The variance in income is very different for low-IQ and high-IQ data.

Testing the 'significance' of r — is r significantly different from zero?

Let's suppose we take a sample of people, measure their IQs and incomes, and correlate them to find *r*. That's the correlation in the sample; but is there a correlation in the whole population? Our null hypothesis is that the population correlation coefficient (ρ) is zero. Without going into the details, we can compute a number called a ***t* statistic:**

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

We can use this number, *t*, to perform a ***t* test** with *n*–2 degrees of freedom. (The statistical catchphrase is that the number we have just calculated is 'distributed as *t* with *n*–2 degrees of freedom'; the *t* distribution is much like the normal distribution that we've mentioned before, so we need to look up the probability corresponding to our *t* statistic just like we might look up a probability corresponding to a *Z* score.) To interpret this using tables, we can look up the **critical value of *t*** for our particular value of α and the number of degrees of freedom (see p. 125); if our *t* statistic is bigger than the critical value, it's 'significant' and we reject the null hypothesis that there's no correlation in our underlying population.

Note that we use *r*, not r_{adj}, for this test.

This is an example of a *t* test; we'll cover these properly in Section 3.

2.4. Spearman's correlation coefficient for ranked data (r_s)

If our X and Y data are both **ranked** (see below for how to rank data), we can calculate the correlation coefficient *r* just as normal, except that we'll call it $\boldsymbol{r_s}$ (sometimes called Spearman's rho). However, when we want to test the significance of r_s, we have a problem, because we cannot make our assumption that the data are normally distributed. Some argue that there are substantial problems inherent in computing the significance of r_s (see Howell, 1997, p. 290). Anyway, with these caveats, what we'll do is to look up **critical values of $\boldsymbol{r_s}$** (see p. 124) if $n \leq 30$, and if $n > 30$ we'll calculate *t* and test that, just as before:

$$t_{n-2} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \text{ where } n > 30$$

To answer the question **'are these values in a particular order?'** you can correlate the rank of the data with the rank of their position. For example, suppose you take large spoonfuls of bran flakes from the top of a cereal packet, one by one, and find

the mean weight of individual bran flakes in each spoonful. These weights, in milligrams and in order, are 70, 84, 45, 50, 48, 40, 38, 40, 25, 30. If you want to establish whether it's true that big bran flakes come out of the packet first, you can correlate the set of positional ranks {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} with the corresponding ranks of the data {9, 10, 6, 8, 7, 4.5, 3, 4.5, 1, 2} to get $r_s = -0.918$ ($p < .001$).

How to rank data

Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

5	8	9	12	12	15	16	16	16	17

Then start assigning them ranks. When you come to a tie, give each value the mean of the ranks they're tied for — for example, the 12s are tied for ranks 4 and 5, so they get the rank 4.5; the 16s are tied for ranks 7, 8, and 9, so they get the rank 8:

X:	5	8	9	12	12	15	16	16	16	17
rank:	1	2	3	4.5	4.5	6	8	8	8	10

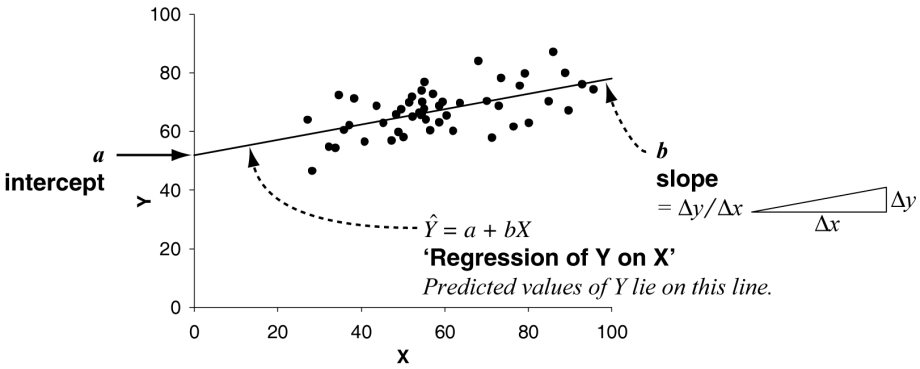
2.5. Regression

We've used correlation to measure *how much* of a relationship there is between two variables. We can use a related technique, **regression,** to establish exactly *what* that relationship is — specifically, to make predictions about one variable using the other. Suppose there's a positive correlation between serum cholesterol in 50-year-old men and their chance of having a heart attack in the next five years. If Mr Blobby has a serum cholesterol twice that of Mr Slim, are his chances of having a heart attack doubled? Increased by a factor of 1.5? Tripled? Let's find out.

If we call our two variables *X* (cholesterol) and *Y* (chance of having a heart attack), we can write an **regression equation** that describes the **linear relationship** between *X* and *Y*. It's just the equation of a straight line:

$$\hat{Y} = a + bX$$

We call this the **regression of Y on X,** meaning that we're predicting *Y* from *X*, not the reverse. The *Y* with a 'hat' ($\hat{Y}$) just means 'the predicted value of *Y*'. This is the picture that this equation represents:



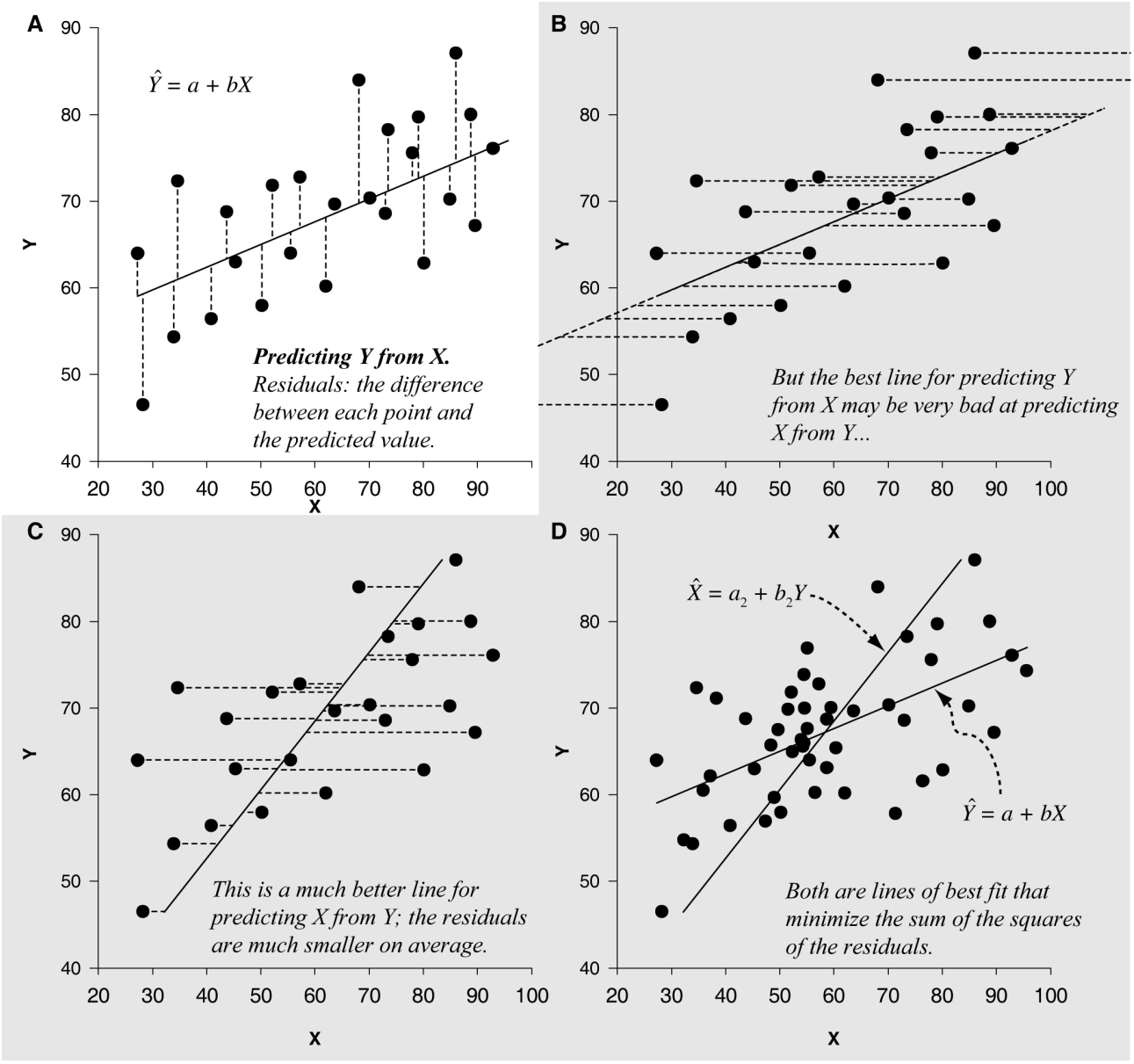
The regression equation and what it means. You might also see it written $y = bx + a$, *or* $y = y_0 + ax$, *or* $y = mx + c$, *or some other equivalent.*

We could draw thousands of lines like this. So which one is the **best fit** to our data? If we take a particular line $\hat{Y} = a + bX$, then for each {*x*, *y*} point, we can calulate a predicted value $\hat{y} = a + bx$. From this, we can calculate how wrong our prediction was: the prediction error is $y - \hat{y}$. This error is often called the **residual,** because it's what you have left after you've made your prediction. Since this will sometimes be positive and sometimes be negative, we can square it to get rid of the +/– sign, giving us the *squared error:* $(y - \hat{y})^2$. So we should aim to find a line that gives us the minimum possible total prediction error, or *sum squared error,* $\sum (y - \hat{y})^2$. (This procedure is called **least squares regression.**) As it happens, this is when

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\mathrm{cov}_{XY}}{s_X^2} \quad \text{(or } b = r\frac{s_Y}{s_X} \text{ if that's easier with your calculator)}$$

Note that regression is not a symmetrical process: the best-fit line for predicting *Y* from *X* is probably not the same as the best-fit line for predicting *X* from *Y* (illustrated in the figure below). This is **different from correlation,** which doesn't 'care' which way round *X* and *Y* are.



Residuals and lines of best fit. ***(A)*** *What's a residual?* ***(B–D), which are LESS IMPORTANT,*** *show why predicting Y from X is different from predicting X from Y.*

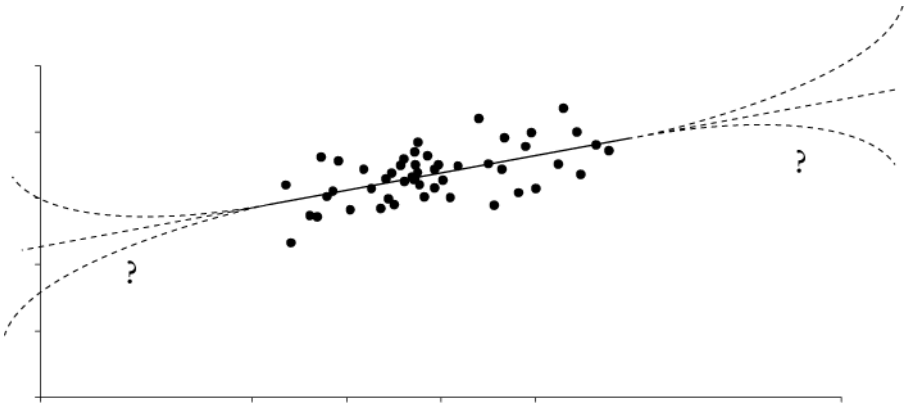
To save you the bother of doing this by hand, **your calculator should give you A and B (regression) and *r* (correlation) directly.** Learn how to use it for the exam! Typically, you put it into 'statistics mode' or 'linear regression' mode, clear the stats memory, then enter each data point as an {*x*, *y*} pair — then you can read out the answers.

Plotting and interpreting the regression line

To plot the line, you just need any two {*x*, $\hat{y}$ } pairs — though it helps if they're far apart, because this makes your line more accurate, and it's often wise to plot a third

point somewhere in the middle to make sure it lies on the same line! The line will also pass through the points $\{0, a\}$ and $\{\bar{x}, \bar{y}\}$.

If you actually need to predict a *y* value from some *x* value — say your father's got a particular cholesterol level and you wanted to predict his risk of a heart attack — then you can just use the regression equation directly. **Beware of extrapolating beyond the original data, though** (figure below). If you've based your regression equation on 50-year-old men with a cholesterol level of 4–8 mM, they may be pretty useless at predicting heart attack risks in 50-year-old men with a cholesterol level of 12 mM, or 100-year-old men, or 50-year-old women. Within the range of your data, though, you can also make statements like 'for every 1 mM drop in cholesterol, one would expect a 10% reduction in the risk of a heart attack' (or whatever it is); this information is based on the **slope** of the regression line.



Beware extrapolating beyond the original data.

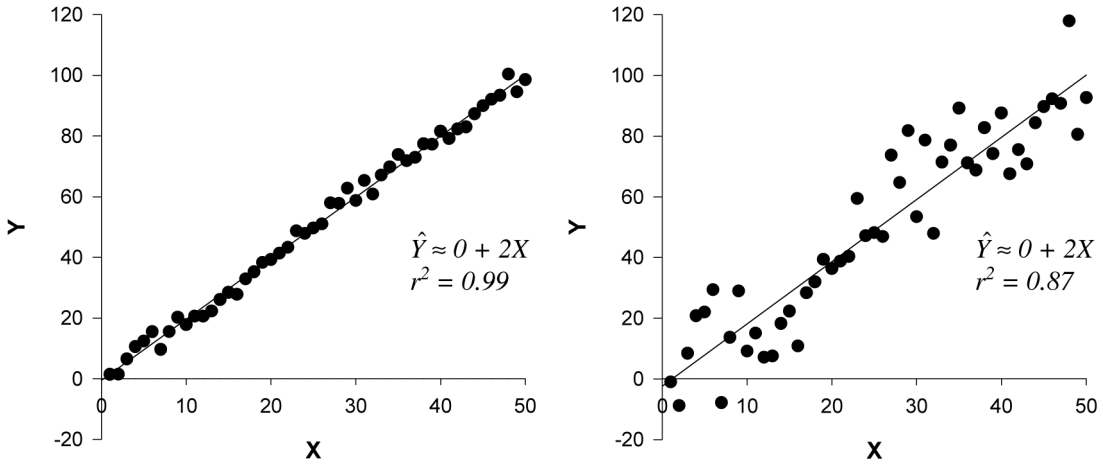
Finally, remember that **correlation and regression do not necessarily represent causation** (see above).

r^2 as a measure of how good a correlation or regression is

So far, we've drawn a regression line. But how good it is at predicting *Y* from *X* depends on how much of a relationship there is between *Y* and *X* — we could draw a regression line where *Y* was the chance of having a heart attack and *X* was shoe size, but it wouldn't be a very good one. How can we quantify 'how good' our best fit is?

r^2 represents the proportion of the variability in *Y* that's predictable from the variability in *X*, or (equivalently) the proportion by which the error in your prediction would be reduced if you used *X* as a predictor. Let's say the correlation between cholesterol levels and heart attack risk were ridiculously high, at $r = 0.8$; then $0.8^2 = 0.64 = 64\%$ of the variability in the risk of heart attacks would be attributable to variations in cholesterol. If $r = 0.1$, then $0.1^2 = 0.01 = 1\%$ of the variability in the risks of heart attacks would be attributable to differences in cholesterol levels.

Note, once again, that this doesn't tell you anything about causality. If rainfall is predictable from twinges in your gammy knee, that doesn't necessarily mean that twinges cause rain, or that rain causes twinges.



Two regressions with nearly identical equations ($\hat{Y} = 2X$) but different values of r^2.

Mathematical statement of this property of r^2

Let's start by taking the worst-case scenario. If you knew *nothing* about your subject's cholesterol level (*X*), how accurately could you predict his risk of a heart attack (*Y*)? Your best guess would be the mean risk of a heart attack, $\bar{y}$, and your error would be described in some way by the standard deviation of *Y*, s_Y , or the variance s_Y^2 . The variance, remember, is

$$s_Y^2 = \frac{\sum (y - \bar{y})^2}{n-1}$$

Now the bottom part of that, $n - 1$, is the number of **degrees of freedom (*df*)** our estimate of the variance was based on. (See page 14 — if you have *n* numbers, and you use them to calculate the sample mean, $\bar{y}$, then you can subsequently only alter $n - 1$ of the numbers freely without altering the mean. This is called the number of *degrees of freedom* you have left — it is the number of *independent* observations on which a given estimate is based.) The top part is the sum of the squares of the deviations of *Y* from the mean of *Y*, which we shorten to the **sum of squares of *Y* (SS_Y).** So we can write the variance as

$$s_Y^2 = \frac{SS_Y}{df_Y}$$

Let's now suppose that we do know our subject's cholesterol. We have a whole set of *n* observations with which to calculate a regression line, i.e. *a* and *b*. (Since we calculate two numbers, we're left with $n - 2$ degrees of freedom in our data.) But now we can estimate our subject's heart attack risk rather better, we hope — and the error in doing so will be related to the *residuals* (error) of our regression's prediction. This thing is called the **residual variance,** or **error variance,** also known as the '*mean square* (MS)' of the residuals:

$$s_{residual}^2 = \frac{\sum (y - \hat{y})^2}{n-2} = \frac{SS_{residual}}{df_{residual}} = MS_{residual}$$

and its square root, $s_{residual}$ (sometimes written $s_{Y \cdot X}$ to show that *Y* has been predicted from *X*), is called the **standard error of the estimate** (it's like a standard deviation — the square root of the variance of the errors is the standard deviation of the errors, abbreviated to the **standard error**). We can also express the residual variance and its square root like this:

$$s_{residual}^2 = MS_{residual} = \frac{SS_{residual}}{df_{residual}} = \frac{SS_Y(1-r^2)}{n-2} = \frac{s_Y^2(n-1)(1-r^2)}{n-2} = s_Y^2(1-r_{adj}^2)$$

$$s_{residual} = s_Y\sqrt{(1-r^2)\frac{n-1}{n-2}} = s_Y\sqrt{1-r_{adj}^2}$$

Actually, it's generally easiest to do the calculations in terms of the sums of squares, not variances, because then we don't have to worry about all these degree-of-

freedom corrections (r and r_{adj} and this $n - 1$, $n - 2$ business) — you can't add two variances together unless they're based on the same number of degrees of freedom, but **you can add sums of squares together** any way you like — and we find that

$$SS_{residual} = SS_Y(1-r^2)$$

$$r^2 = \frac{SS_Y - SS_{residual}}{SS_Y}$$

$$SS_Y = SS_Y(r^2) + SS_{residual}$$

In other words, the total variability in Y is made up of a component that's related to X ($SS_Y \cdot r^2 = SS_Y - SS_{residual}$, which we can also write as $SS_{\hat{Y}}$, the variability in the *predicted* value of Y) and a component that's residual error ($SS_{residual}$). Translated to our cholesterol example, people vary in their cholesterol levels (SS_X), they vary in their heart attack risk (SS_Y), a certain amount of the variability in their heart attack risk is predictable from their cholesterol ($SS_{\hat{Y}}$), and a certain amount of variability is left over after you've made that prediction ($SS_{residual}$). Or,

$$SS_Y = S_{\hat{Y}} + SS_{residual}$$

where

$$r^2 = \frac{SS_{\hat{Y}}}{SS_Y}$$

2.6. Advanced real-world topics

As with all the wavy-line sections, this section certainly isn't intended to be learned! It's for use with real-world problems that you may encounter. You will not be tested on any of this in the exam.

What's 'regression to the mean'?

Something related to regression, but quite interesting. It was discovered by Galton in 1886. He measured the heights of lots of families, and calculated the 'mid-parent height' (the average of the mother's and the father's height) — call it X — and the heights of their adult children — call it Y. He found that the average mid-parent height was $\bar{x}$ = 68.2 inches; so was the average height of the children ($\bar{y}$ = 68.2 inches). Now, consider those parents with a mid-parent height of 70–71 inches: the mean height of their children was 69.5 inches. That is, the height of these children (69.5) was closer to the mean of *all* the children ($\bar{y}$ = 68.2) than the height of the parents (70–71) was to the mean of all the parents ($\bar{x}$ = 68.2). But this wasn't a genetic phenomenon, it was a statistical phenomenon, and it worked backwards: if you took children with a height of 70–71 inches, the mean mid-parent height of their parents was 69.0 inches. This is called **regression to the mean.**

Why does it happen? Suppose we have the variables X and Y, with standard deviations s_X and s_Y, and the correlation between them is r. We've previously seen that

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y}$$

and the regression slope b is

$$b = \frac{\text{cov}_{XY}}{s_X^2}$$

Therefore,

$$slope = b = r\frac{s_Y}{s_X}$$

So a change of one standard deviation in X is associated with a change of r standard deviations in Y. And we know the regression line always goes through the point at the means of both X and Y — that is, the point { $\bar{x}$, $\bar{y}$ }. Therefore, unless there is perfect correlation (unless r = 1), the predicted value of Y is always fewer standard deviations from its mean than X is from its mean. Remember that predicting Y from

X is different from predicting *X* from *Y*, unless the two are perfectly correlated? This is another way of saying the same thing.

Examples of regression to the mean (from Bland & Altman, 1994)

- If we are trying to treat high blood pressure, we might measure blood pressure at time 1, then treated, and then measured again at time 2. We might see that blood pressure goes down *most* in those who had the *highest* blood pressure at time 1, and we might interpret this as an effect of the treatment. We'd be wrong; this is regression to the mean. It would happen even if the treatment had no effect. The two sets of observations (time 1, time 2) will never be perfectly correlated (because of measurement error and biological variation); $r < 1$. So if the difference between our 'high blood pressure' subgroup and the whole population was *q* at time 1, it will be *rq* at time 2 — i.e. the difference from the population mean will have shrunk. We should have compared our treated group to a randomized control group.
- In one study, people reported their own weight and had their weight measured objectively. A regression was used to predict reported weight from measured weight; the regression slope was less than 1. This might lead to you interpret that very fat people underestimate their weight when they report it, and very thin people overestimate it. But we'd never have expected *perfect* correlation. All this might be is regression to the mean — and if we'd predicted measured weight from reported weight, we'd also have a slope less than one, from which we might have concluded the opposite: that very fat people overestimated their weights and very thin people underestimated them.
- When scientific papers are submitted to journals, referees criticize them and editors select the 'best' ones to publish on the basis of the referees' reports. Because referees' judgements always contain some error, they cannot be perfectly correlated with any measure of the true quality of the paper. Therefore, because of regression to the mean, the average quality of the papers that the editor accepts will be less than he thinks, and the average quality of those rejected will be higher than he thinks.

Partial correlation — dealing with the effects of a third variable

Sometimes we are interested in the relationship between two variables and know that a third variable is also influencing the situation. Imagine we examine the correlation between IQ (*X*) and income (*Y*), and find it to be positive, but we suspect that one reason that higher IQ predicts higher income is because people with higher IQs are more likely to get into university, stay for higher degrees, and so on — and it's the degree that gets you the higher income, not your IQ itself. So is there any *further* relationship between IQ and income once you've taken into account this effect of studying for longer? One way of investigating this is to look at the correlation between IQ (*X*) and (say) number of years of study (*Z*), and the correlation between income (*Y*) and number of years of study (*Z*). We can then calculate the **partial correlation** between IQ (*X*) and income (*Y*) *having taken account of the relationship of each of these to number of years of study.* We call this 'partialling out' the effects of number of years of study. We term the partial correlation coefficient between *X* and *Y* with the effects of *Z* partialled out $r_{xy.z}$, and calculate it like this:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

Let's use some fictional numbers to illustrate this: suppose that the correlation between IQ and income is $r_{xy} = 0.6$, the correlation between IQ and years of study is $r_{xz} = 0.8$, and the correlation between income and years of study is $r_{yz} = 0.7$. Then the correlation between IQ and income having partialled out the effect of years of study would be only $r_{xy.z} = 0.09$. This would mean that $r_{xy.z}^2 = 0.0081$, so only 0.8% of the variability in income is predictable from IQ once you've taken account of the number of years of study, even though $r_{xy}^2 = 0.36 = 36\%$ of the variability in income is

predictable from IQ. This would suggest that nearly all the ability of IQ to predict income was due to the fact that high IQs predict more years of study.

The point-biserial correlation (r_{pb}) for a dichotomous variable

If we ask the question 'is body weight correlated with sex?', we have a bit of a problem with assuming that 'sex' is normally distributed; it clearly isn't. Body weight probably is, but mammals are either male or female; the sex variable is **dichotomous** (Greek *dikhotomos*, from *dikho-*, in two; *temnein*, to cut).

No problem: simply assign two values to the dichotomous variable as you see fit — e.g. male = 0, female = 1 (or male = 56, female = 98; it doesn't matter at all). Then calculate r as normal. Officially this r is called r_{pb}, the point-biserial correlation coefficient, but you can treat it like any r, and test it for significance in the same way (a t test on $n - 2$ df) as we saw before:

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

You might think that asking 'does weight vary with sex' and calculating a correlation is a bit daft here, and the more natural question is 'do the sexes differ in body weight?' You'd be right, really, but it is actually the same question. Since it's the same question, there must be a simple relationship between r_{pb} and the t statistic:

$$r_{pb}^2 = \frac{t^2}{t^2 + df}$$

The use of this is that if you test the difference between two groups (e.g. male body weights and female body weights) using a t test (which we'll cover in Practical 2), you can calculate r^2, and therefore the proportion of the variability in body weight explained by sex. And if you read the results of a t test in a research article, you can interpret them in terms of r^2 using this technique.

Correlations when the dichotomous variable is 'artificial'…

The male / female dichotomy is natural; all subjects are either one or the other. Sometimes a dichotomy is arbitrary, such as 'pass / fail' in an exam with a 60% pass mark; this dichotomy classifies people who scored 59% and people who scored 1% in the same category, but classifies people who scored 59% and people who scored 60% in different categories. If you have data like these and want to calculate a correlation, you have to use a slightly different technique; this is described by Howell (1997, p. 286).

Correlations with two dichotomous variables

If you want to calculate the correlation between two variables when *both* are dichotomous, again, you can do it. All you do is calculate r in the normal way; this time, its special name is ϕ (phi). And it's exactly equivalent to doing a χ^2 test (which we'll cover in Practical 4). And there's a relationship between the two:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Why is this useful? Again, because r^2 is a measure of the proportion in the variability in one variable that's explained by a variable — the practical significance of the relationship — and therefore so is ϕ^2. So if you see a χ^2 test reported in an article, you can calculate ϕ^2 to see whether the relationship is *important* (large) as well as *significant*; see Howell (1997, p. 285).

Is a regression slope (b) significantly different from zero?

We saw earlier how to test if a correlation (r) was significantly different from 0. Since correlation and regression are much the same thing, we can also calculate the t statistic from the regression parameter b (from $\hat{y} = a + bx$) using a different fornula. For this it helps to use the notation $s_{Y \cdot X}$ rather than $s_{residual}$ for the standard devia-

tion of the residuals left over when we have predicted *Y* from *X*. We would find that the *t* statistic we've just worked out could also be found like this:

$$t_{n-2} = \frac{b \cdot s_X \sqrt{n-1}}{s_{Y \cdot X}}$$

As before, this *t* statistic is distributed with $n - 2$ degrees of freedom (which is why the subscript on the *t* is $n - 2$).

If we calculate two regressions, are they significantly different?

Suppose we calculate the relationship between smoking and life expectancy in males and females. We'd probably find that the more you smoke, the shorter you live ($b < 0$). Let's suppose we find that this relationship is stronger in males (e.g. $b_{\text{male}} < b_{\text{female}} < 0$), suggesting that males decrease their life expectancy more than females for a given increment in the amount they smoke (though, of course, the regression by itself doesn't tell you anything about causality). Is this difference between males and females significant? If we have two variables X_1 and X_2 that both predict a third variable *Y*, and two sample regression coefficients b_1 and b_2, then we can calculate a *t* statistic (with $n - 4$ *df*) for the null hypothesis that the two underlying population regression coefficients are the same:

$$t_{n-4} = \frac{b_1 - b_2}{\sqrt{\frac{s^2_{Y \cdot X_1}}{s^2_{X_1}(n_1 - 1)} + \frac{s^2_{Y \cdot X_2}}{s^2_{X_2}(n_2 - 1)}}}$$

'I have two values of r from different (independent) groups. Are they different?'

If we want to do the same with correlations rather than regressions ('is the correlation r_1 between male smoking and male life expectancy significantly different from the correlation r_2 between female smoking and female life expectancy?') we have to use a slightly different test. We convert *r* to a related number *r´* and work out a *Z* score from those:

$$r' = 0.5 \ln \left| \frac{1 + r}{1 - r} \right|$$

$$z = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Then we look up our value of *z* in a table of the standard normal distribution to get our *p* value.

'I have two values of r, but they are not independent; are they different?'

Suppose we measured the number of GCSE points acquired by a group of 16-year-olds, then measure the number of A-Level points acquired by the same people aged 18, then measure their annual income when they are 30. We could calculate a correlation between any two of these variables. We could also ask whether the correlation between GCSE scores and income was better / worse than the correlation between A-Level scores and income. But these are clearly not independent correlations, because they were all based on the same people, and so there will probably be a correlation between GCSE scores and A-Level scores that we must take into account. If our three correlations are r_A, r_B, and r_C, and we want to know if the difference between r_A and r_B is significant, then the null hypothesis is that r_A and r_B are the same, and we can calculate a *t* statistic (with $n - 3$ df) like this:

$$|R| = (1 - r^2_{AB} - r^2_{AC} - r^2_{BC}) + 2 r_{AB} r_{AC} r_{BC}$$

$$t_{n-3} = (r_{AB} - r_{AC}) \sqrt{\frac{(n-1)(1 + r_{BC})}{2\left(\frac{n-1}{n-3}\right)|R| + \frac{(r_{AB} + r_{AC})^2}{4}(1 - r_{BC})^3}}$$

This is effectively a statistical test for partial correlations. The partial correlation coefficient will answer the question 'what is the correlation between *X* and *Y*, taking account of *Z*?' This test will answer the question 'is the correlation r_{xy} significantly different from the correlation r_{xz}, taking into account the fact that these two correlations are themselves related (non-independent)?'

'Is my value of r different from (a particular value)?'

Suppose we have a sample with a correlation of $r = 0.3$, and we want to know if this differs from a correlation of 0.5. The null hypothesis is that the sample $r = 0.3$ came from a population with $\rho = 0.5$. We can calculate a *Z* score like this:

$$r' = 0.5 \ln\left|\frac{1+r}{1-r}\right|$$

$$z = \frac{r' - \rho'}{\sqrt{\frac{1}{n-3}}}$$

If I calculate r, what are the confidence limits on ρ?

The expression for *z* above tells us that

$$\rho' = r' + \frac{z}{\sqrt{n-3}}$$

so we can calculate confidence intervals from appropriate critical values of *z* for a two-tailed α:

$$CI(\rho') = r' \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}}$$

If you want 95% confidence intervals, $z_{\alpha/2}$ would be 1.96. Once you've worked out confidence intervals for ρ', we can convert them back to ρ to get our final answer:

$$\rho' = 0.5 \ln\left|\frac{1+\rho}{1-\rho}\right| \Rightarrow \rho = \frac{e^{2\rho'} - 1}{e^{2\rho'} + 1}$$

I have a regression equation; what are the confidence intervals on predictions?

Suppose you calculate the regression equation $\hat{Y} = a + bX$. For a given value of *X*, (call it *x*) what are the confidence intervals on the predicted value of *Y* (call it $\hat{y}$)? Clearly, predictions where *x* is *near to* $\bar{x}$ are more likely to be accurate than predictions where it's very far away. First, we obtain the standard error of the estimate (as on p. 36→):

$$s_{residual} = s_Y \sqrt{(1-r^2)\frac{n-1}{n-2}}$$

This standard error of the estimate is useful as an *overall* measure of prediction error (for the data set), but it isn't specific to our value of *x* — we know that predictions are less accurate when *x* is far from $\bar{x}$. So if we want to know how accurate a prediction is for a given new value of *x*, we calculate something called the *standard error of prediction*, $s'_{residual}$ (Howell, 1997, p. 253):

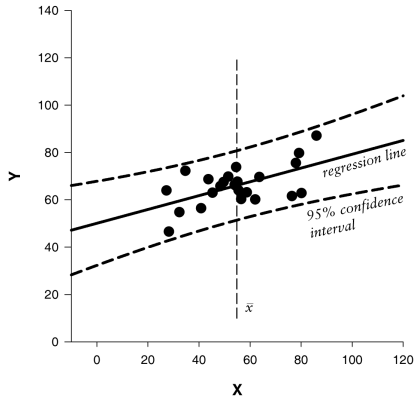
$$s'_{residual} = s_{residual} \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_X}}$$

$$= s_{residual} \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_X^2 (n-1)}}$$

Finally, we obtain the desired confidence interval (as on pp. 16 and 54):

$$\text{CI} = \hat{y} \pm t_{\text{critical for } n\text{-2 df}} \times s'_{residual}$$

The end result is a set of confidence intervals like those shown in the figure below (which plots some fictitious data with $r = 0.58$, the regression line $\hat{Y} = a + bX$, and the 95% CI). Note that the confidence interval widens the further x is from $\bar{x}$.



I have a group of subjects and have worked out a correlation for each one. Is this correlation significant for my whole group?

Stop and go back a stage. Suppose you have a group of 20 rats and you measure their performance on a test of attention and, simultaneously, the levels of the neurotransmitter acetylcholine in parts of their brain. Is there a relationship between acetylcholine and attentional performance? If you only make one measurement per rat, the problem is easy; you have 20 measurements of two variables, and can correlate them as usual. If you've made 100 measurements for *each* rat, the problem is harder. What can you do?

- You **must not** lump all the measurements together to give 2000 different {*x*, *y*} pairs — because these observations are definitely not equally independent, since subsets of observations are likely to be related by virtue of having come from the same rat.

- To ask whether subjects with high levels of acetylcholine have high levels of performance — a *between-subjects* question — you could take each rat's *mean* performance and *mean* acetylcholine and conduct a correlation as normal (n = 20). (And if you'd made 60 observations on some rats and 105 observations on others, it wouldn't matter, because you'd take the mean across all these subjects. If you really felt that it was worth placing more weight on data from subjects that you obtained more measurements from, you could conduct a *weighted analysis*, weighting for the number of observations per subject.)

If different rats have very different levels of acetylcholine, then we could end up with something like our wild-boar-and-runner-bean effect — for example, you might find a negative correlation across the group (rats with lots of acetylcholine do worse than rats with less acetylcholine), even though if you looked for it, you might find a positive correlation *within* each rat (when any given rat has what is a high level of acetylcholine *for that rat*, it performs better). So…

- If we want to know whether changes in one variable (acetylcholine) are paralleled by changes in the other variable (performance) in the same subject, and that this is consistent across subjects — a *within-subjects* question using data from multiple subjects — we can estimate the relationship within subjects using a very general technique, called general linear modelling. This particular way of using a general linear model (GLM) is called multiple regression or analysis of covariance (ANCOVA). The GLM technique will handle even more complicated problems, such as when we have two groups of rats (a control group and one that has had part of their brain destroyed) and we want to know whether the relationship between acetylcholine and performance is different in the two groups. We will not cover these advanced techniques in the IB course.

I want to predict a variable on the basis of many other variables, not just one.

Then you need multiple regression, which we're not going to cover.

Correlation/regression in Excel — relevant functions (see Excel help for full details)

COVAR(…)	*Population* covariance (i.e. divide-by-*n* formula). So to calculate *r* using this number, you need to divide this by the product of the 'population SDs' of *X* and *Y*, calculated using STDEVP(…) — or multiply COVAR(…) by *n* and then divide it by *n*–1, before dividing the result by the product of the sample SDs, calculated using STDEV(…).
CORREL(…)	Calculates *r*.
SLOPE(…)	Calculates *b*, the slope of a regression line $\hat{Y} = a + bX$.
INTERCEPT(…)	Calculates *a*, the intercept of a regression line $\hat{Y} = a + bX$.
RANK(…)	Don't use it — it gets the ranks wrong when there are ties.
Tools → Data Analysis → Covariance	Calculates sample covariances (i.e. divide-by-*n*–1 formula).
Tools → Data Analysis → Correlation	Calculates *r*.
Tools → Data Analysis → Regression	Calculates *r*, *a*, *b*, *p*.

2.7. Examples 2: correlation and regression

For questions 1–3,

(a) calculate Pearson's correlation coefficient r and test its significance
(b) find the equation of the best-fitting regression line (predicting the second variable from the first)
(c) plot a graph showing the data points and the regression line (using the first variable as the x axis and the second variable as the y axis)
(d) calculate r^2 to find the proportion of the variance in the second variable that is accountable for by predicting it from the first variable

Q1. The decay of a visual after-effect is believed to be slowed by blinking. The time for the after-effect to decay to half strength (in seconds) was measured for twelve subjects, and their blink rate (average number of blinks per minute) was also measured. Do the data show a correlation supporting the hypothesis?

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Blink rate	2.1	10.3	5.9	10.0	0.5	4.5	3.1	8.2	5.2	9.7	4.6	9.7
Decay time	24.5	29.8	27.9	32.9	23.0	21.0	23.2	25.3	24.7	30.7	26.5	28.9

Q2. The average number of alternations per minute seen in a Necker cube was measured for 15 subjects who were also tested on the Indecision subtest of the Kentucky Personal Effectiveness Test (KPET). Is there an association between the two scores?

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Alternations per minute	20	20	21	35	15	17	19	9	4	5	3	40	22	22	17
KPET	175	130	101	200	118	120	137	120	117	120	118	222	118	191	190

Q3. Ten frog retinal ganglion cells were repeatedly stimulated by a flash of light of constant intensity. The average latency of the first nerve impulse and the average number of spikes produced in the 10 ms after the first impulse are given below. Is there evidence that the shorter the latency, the more spikes are produced?

Cell	A	B	C	D	E	F	G	H	I	J
Latency (ms)	102	110	120	80	65	73	97	150	111	74
No. of impulses	1.7	3.1	4.6	7.2	6.3	6.0	4.8	2.1	5.3	5.9

Q4. I walk in a straight line away from St Peter's Basilica in Rome. Every time I encounter an ice-cream stall, I buy a vanilla cone. In order, the prices (in €) were as follows:

1.8 1.4 2.0 1.0 1.0 1.4 0.8 0.6 1.0 0.8

Do Roman street vendors charge more for their proximity to the Vatican?

Q5. Ten people are repeatedly offered the choice between a small, immediate monetary reward (e.g. £10 now) and a large, delayed monetary reward (e.g. £100 next week). From their preferences, a number is computed that represents their impulsivity when choosing. At the end of the test, they lie down and the experimenter inserts a needle into their spinal space to withdraw a sample of cerebrospinal fluid, which he tests for levels of 5HIAA, a metabolite of the neurotransmitter 5HT (serotonin). The results are as follows:

Subject	A	B	C	D	E	F	G	H	I	J
Impulsivity score	85	51	76	23	55	90	42	56	21	61
5HIAA level (ng / ml)	25	28	40	37	22	35	31	18	32	29

Do these data suggest a relationship between CSF 5HIAA and choice impulsivity?

3. Difference tests — parametric

Objectives

We will go through the various types of tests for asking the question 'is the mean of this sample significantly different from… (something)?' We will then look at the ***t* test,** a very popular 'parametric' test. This has various forms, depending on the kind of data you want to analyse. We will look at nonparametric tests in Practical 4.

Stuff with a solid edge, like this, is important.

But remember — you can totally ignore stuff with single / double wavy borders.

3.1. Background

Reminders

We've already discussed the differences between **one- and two-tailed tests** (p. 23). We've already talked about making **multiple comparisons** between groups (p. 24).

Paired and unpaired tests (related and unrelated data)

When we come to look at the difference between two samples of data, the samples can be *related* or *unrelated*. Suppose we want to compare the speed with which people can rotate figures mentally in two conditions: on land and underwater. (1) We could take a group of landlubbers and a group of divers, and compare them. There would be no particular relationship between individual data points from the land sample and the underwater sample. We would use statistical methods that are described as *unrelated*, *unpaired*, or *between-subjects*. (2) Alternatively, we could measure the *same* group of people in two conditions, on land and underwater. In this situation, there is a relationship between one subject's score on land and the same subject's score underwater — they are likely to be more similar than they would be by chance alone, because they come from the same person. Our statistical methods must reflect this fact; the techniques we would use are described as *related*, *paired*, or *within-subjects*.

It is absolutely **not** acceptable to fail to take account of relationships like this between data. A classic example of this sort of error is something called **pseudoreplication.** Suppose you test Alice, Bob, and Celia on land, and Eric, Frankie, and Greg underwater. You obtain 6 observations, $n = 3$ for each group. Your groups are not related. So far, so good. But suppose you want more than 6 observations; you might measure each subject three times. This would give you observations A_1, A_2, A_3, B_1… on land, and E_1, E_2, E_3, D_1… underwater. The error is to analyse this as if you had 18 observations ($n = 9$ for each group). This is wrong, because A_1, A_2 and A_3 are all *related* — more so than A_1 and B_1, or A_1 and E_1. We will not cover the analytical techniques required for this sort of situation, where we have multiple variables (in this case, land / underwater as a between-subjects variable, observation 1 / observation 2 / observation 3 as a within-subjects variable) — that's covered in the Part II course. If you have data like these, the simplest thing is to obtain some sort of 'overall' score for each subject (e.g. take Alice's overall score to be the mean of A_1, A_2, and A_3) and analyse those.

If you have data from *only* one subject, then you can consider the data to be 'unrelated' for the purposes of analysis, but your conclusions *only apply to that subject*. For example, if you measured my ability to remember sequences of digits (my digit span) ten times when I'm on dry land and ten times when I'm underwater, you could treat the data as unrelated — they have no relationship to each other *beyond* the fact that they come from the same subject, and that's part of your analytical 'context' anyway. You would have a sample ($n = 10$) of my dry-land digit span, and a sample ($n = 10$) of my underwater digit span. If the dry-land scores were significantly higher than the underwater scores, you could conclude that *my* digit span was better

on dry land than underwater — but this would tell you absolutely nothing about people in general, because I might not be a representative person. You would only know this by testing more people. (If you're wondering, the rather foolish situation in which you would need to deal with further 'relatedness' when you're only testing one subject might be something like this: you test me in a car on dry land, in a car underwater, drunk on dry land, drunk underwater, tired on dry land, tired underwater… then to ask the 'dry land versus underwater' question, you would treat the 'car' pair of observations as related, the 'drunk' pair as related, and so on.)

Parametric and non-parametric tests

In the tests we'll cover here, we analyse differences involving one or two samples by making *assumptions* about the populations they come from. Remember the jargon (p. 8): we estimate *parameters* of populations by using *statistics* of samples. The tests we'll cover in this section make assumptions about the parameters of the populations — for example, assuming that the underlying population is normally distributed. They are therefore called **parametric** tests.

If the assumptions of a parametric test are *not* justified — if our data are a bit odd — then we have two alternatives. (1) We can **transform** the data to make them fit the assumptions better. We won't cover this approach in the IB course, but it's important for 'real-life' data analysis. (2) We can use a test that does not make these assumptions about the distribution of the population — a **nonparametric** or **distribution-free** test.

If a test's assumptions are met, it should give an accurate value of p. We say that a test is **robust** if it gives a *good* estimate of p even if we violate its assumptions. (We may also say that it's **liberal** if it underestimates p when certain assumptions are violated — that is, says things are 'significant' more often than it should — or **conservative** if it does the opposite.)

In general, parametric tests have *more power*. If the assumptions of a parametric test are met, it's therefore better to use the parametric test. Many parametric tests are also quite *robust*, so people don't get too worried if the assumptions are not quite met, but not grossly violated. Parametric tests can also be used for *complex analyses* that can be quite hard to do with non-parametric tests. Transformations are a way of 'rescuing' the parametric test by making the data fit the test's assumptions better; this is why transformations are widely used. Non-parametric tests are sometimes viewed as a bit of a last resort, because they have lower power. (On the other hand, if you find a significant effect with a low-power test, you have no problem, and some statisticians argue that non-parametric tests are a generally Good Thing, though it's probably fair to say that most researchers prefer parametric tests.) Occasionally, if the data are 'odd', nonparametric tests have *more* power.

We'll cover some non-parametric tests in section 4.

3.2. The one-sample *t* test

Overview

Suppose we have **one group** of n men and want to know if they are unusually tall. We can measure their height, and ask the question 'does the mean of this sample differ significantly from μ metres?', where μ is the average height of our reference population (all the men in the UK, perhaps). To do this, we define the null hypothesis that the sample comes from a population with mean height μ metres. We calculate the sample mean $\bar{x}$ and the sample standard deviation s_X. From this, we can calculate the **standard error of the mean,** $s_{\bar{x}} = s_X / \sqrt{n}$. Then, we can calculate a ***t*** **statistic:**

$$t_{n-1} = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

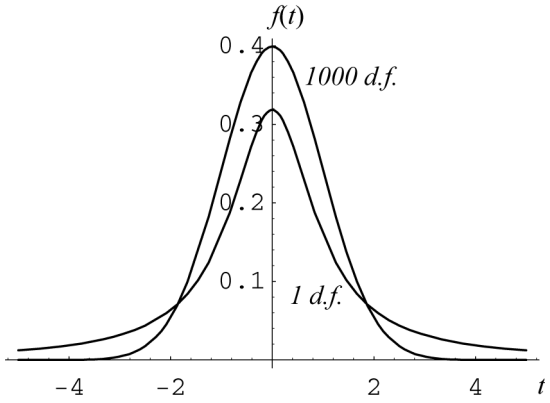
This is called a *t* statistic with ***n* – 1 degrees of freedom (*df*).** We look up our *t* statistic in our tables (see p. 125) to find a **critical value** of *t* for this many *df* and our desired level of α. (If we want a two-tailed test with a level of α, we have to allocate $\alpha/2$ to each tail.) If our value of *t* is bigger than this critical value, we reject the null hypothesis.

Significant values of *t* can be big positive numbers or big negative numbers. Non-significant values of *t* are close to zero.

The *t* test is always obtained by taking a **number,** subtracting from it a **test value,** and dividing the result by the **standard error of the number.** We'll see several different forms of the *t* test for different types of data (one sample, two samples, etc.), but they all have the same general format.

How did we arrive at this? You don't need to know, but if you're interested, see page 57.

Some people use the subscript on the *t* to refer to the number of degrees of freedom (e.g. '$t_6 = 2.5$, two-tailed $p < 0.05$'); others use it to denote critical values ('for $df = 6$ and two-tailed $\alpha = 0.05$, $t_{\alpha/2} = t_{0.025} = 2.447$; our $t = 2.5$, so $p < 0.05$'). I prefer the first of these, as you can probably tell.



The t distribution with different degrees of freedom. With many degrees of freedom (as df → ∞) it becomes just like the normal distribution, but with few degrees of freedom it is a different shape; critical values of t for few df are therefore larger than critical values of Z.

What is the standard error of the mean (SEM)?

Suppose we have a population with mean μ and variance σ^2, and we repeatedly take very many samples from it, with each sample containing *n* observations. We can say some things about the samples that we take. For each sample, we can calculate a sample mean $\bar{x}$. So we can collect lots of different sample means — many values of $\bar{x}$. Now we can ask what might at first appear to be an odd question: what will be the *distribution* of these sample means? The mean of all the sample means (the mean of all the values of $\bar{x}$), written $\mu_{\bar{x}}$, will be the same as the population mean, μ. The standard deviation of all these sample means (the standard deviation of all the values of $\bar{x}$), written $\sigma_{\bar{x}}$, is usually called the **standard error of the mean (SEM).** It's a measure of how much the value of the sample mean $\bar{x}$ may vary from sample to sample taken from the same population. It can be used to compare the observed mean to a hypothesized value — as we saw above, it's the basis of the *t* test. If we know the population standard deviation σ and the sample size *n*, we can calculate the SEM like this:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If we don't know σ, we can estimate the population SEM using the sample standard deviation *s*, to give us the sample SEM:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

3.3. The two-sample, paired *t* test

It's very easy to extend the one-sample *t* test to **two related groups.** Suppose you measure the heights of *n* girls when they're 10, and then again when they're 11, so you have two measurements for each girl. These two measurements are clearly related (more so than two measurements for two different girls). We want to know if our girls are growing normally. For each girl, we can therefore calculate the **difference** or **difference score** between the two related measurements — we just subtract one from the other. We will obtain *n* difference scores (the amount that each girl has grown). Suppose we know that the average girl grows 5 cm between the ages of 10 and 11 ($\mu = 5$). We can just run a *t* test on the difference scores, exactly as before:

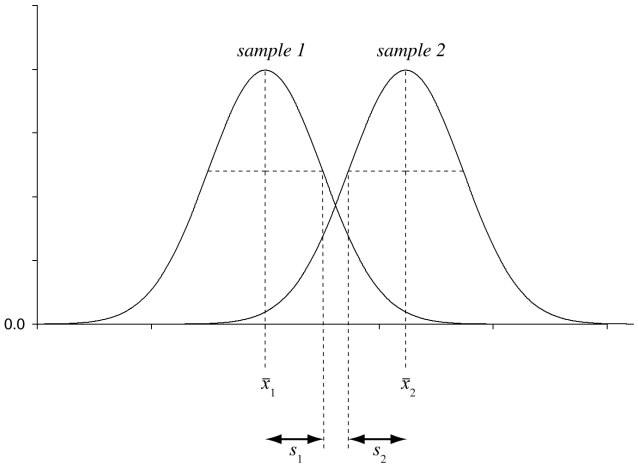
$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

If our value of *t* exceeds the relevant critical value for $n - 1$ *df* and an appropriate α, we reject the null hypothesis that our girls come from a normal-growing population.

The paired *t* test is used for **related (or matched) samples.** Two samples are related whenever you can use one sample to make better-than-chance predictions of the other. In this example, knowing one girl's height aged 10 allows you to make a better-than-chance prediction of the same girl's height aged 11, but doesn't allow you to predict another 11-year-old girl's height. In this example, the two samples come from the same subject, but sometimes related samples don't come from the same subject. For example, if you ask different couples to rate their satisfaction about their relationships, it is likely that if the man is very dissatisfied with the relationship, the woman is too, so their scores would be related (but would not be related to scores from a different couple).

Here's an example: suppose the initial heights of the girls in cm are {125, 148, 132, 135, 139, 129} and after a year they are {129, 153, 135, 140, 148, 136}. The difference scores (age 11 heights minus age 10 heights) are {4, 5, 3, 5, 9, 7}. The mean of this sample of difference scores is $\bar{x} = 5.5$; the sample SD is $s_X = 2.17$; $n = 5$. We want to know if our group differs from a population with mean $\mu = 5$. We can calculate that $t = (5.5 - 5) / (2.17 / \sqrt{5}) = 0.51$. This *t* statistic has $n - 1 = 4$ *df*. For a two-tailed $\alpha = 0.05$, the critical value of *t* is 2.776. Our *t* is less than this, so we do not reject the null hypothesis; the girls are growing normally.

3.4. The two-sample, unpaired *t* test, for equal sample variances



The essence of a two-sample t test. We have two samples with means $\bar{x}_1$ and $\bar{x}_2$. If the distance (difference) between means ($\bar{x}_2 - \bar{x}_1$) is big enough, we say that the two samples are significantly different (which is to say, the two samples come from underlying populations whose means are different). We measure the distance between the means — somehow — in terms of the standard deviations of the samples, s_1 and s_2.

Overview

If we have **two independent (unrelated) groups,** X_1 and X_2, with **equal variances** ($s_1^2 = s_2^2$), we can ask if they are significantly different from each other. The null hypothesis is that the two underlying populations have the same mean ($\mu_1 = \mu_2$). We can calculate a *t* statistic, which has the same general form as before: it's the **differ-**

ence between means divided by the **standard error** of that difference, and this time it has $(n_1 + n_2 - 2)$ degrees of freedom.

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

where s_p^2 (called the **pooled variance**) is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If the groups are of the same size ($n_1 = n_2 = n$), then the formula becomes simpler:

$$t_{2n-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

This test assumes that the two samples come from populations with equal variances ($\sigma_1^2 = \sigma_2^2$), whether or not $n_1 = n_2$. If this assumption is violated, we must use the *unequal variances* version of this test (see below).

Example

Suppose we collect young horses and assign them to one of two groups at random. We feed one group (n = 10) FastDope, a drug that we suspect of having performance-enhancing properties. The other group (n = 10) are given a placebo. They are then timed running along a 1 km racetrack and their speed is calculated in $m \cdot s^{-1}$. The null hypothesis is that the speeds of the drugged and undrugged groups do not differ. We find that the speeds of the drugged group (group 1) are {14.6, 12.6, 12.2, 15.0, 12.5, 12.1, 13.1, 12.2, 14.1, 14.2} and the speeds of the placebo group (group 2) are {10.8, 11.9, 9.7, 9.3, 12.0, 9.6, 10.7, 8.9, 12.5, 12.0}. Since $n_1 = n_2$, we can use the simpler of the two formulae for t, and can therefore calculate

$$t_{n_1+n_2-2} = t_{18} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} = \frac{13.26 - 10.74}{\sqrt{\frac{(1.11)^2 + (1.31)^2}{10}}} = \frac{2.52}{0.543} = 4.64$$

For 18 *df*, the critical value of t for a two-tailed $\alpha = 0.05$ is 2.101. Since our t statistic exceeds this critical value, we reject the null hypothesis; the drugged group ran faster.

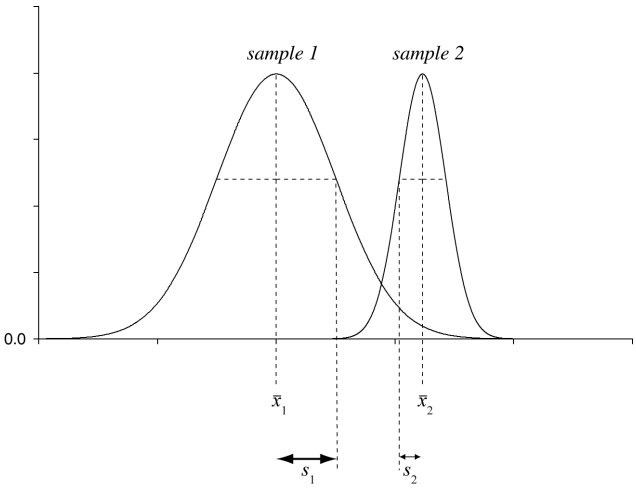
How did we derive this *t* test? If you're interested, see page 59.

If there is a significant difference, which way round is it?

If you calculate a two-tailed t test and find a significant result, then you may declare the group means to be significantly different — and by looking at the means, you can see which mean is bigger. So that mean is significantly bigger! Sometimes people calculate a t test, find a significant result, and then think they're not allowed to say which of the groups has the significantly larger mean, which is nonsense.

If, on the other hand, you run a one-tailed test, you specify *in advance* which direction of difference you are interested in (i.e. whether you're only interested if group 1 > group 2, or only interested if group 2 > group 1). Suppose you are only interested if group 1 > group 2. If you conduct a one-tailed t test and find a significant result, you may declare the mean of group 1 to be significantly bigger than that of group 2. If you do not find a significant result, you may merely say that the group 1 mean is 'not significantly bigger' than that of group 2. You should not then proceed to see if it is *smaller* instead, because then you would have broken the implicit promise of the one-tailed test — not to be interested in differences of the opposite kind — and your α and p values would be misleading (see p. 23).

3.5. The two-sample, unpaired *t* test, for unequal sample variances



Unequal sample variances.

If the two sample variances are not equal **(heterogeneous variances),** we have a bit of a problem. First, the number we calculate will not have a *t* distribution, so if we look it up using *t* tables we'll get the wrong answer. Second, it makes no sense to use s_p^2 in our formula (to 'pool' the variances of the two groups) since that procedure also assumes equal variances (as explained in section 3.12 if you're really interested). But we can still run a *t* test, although we'll lose a bit of power (the test is more conservative). We use this formula and call the result *t*′:

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

We then test it just as if it were a *t* score, but with a **different number of degrees of freedom.** If we're doing it by hand,

degrees of freedom = (n_1 – 1) or (n_2 – 1), whichever is smaller.

If you have a computer, you can get a slightly better answer, which will lie somewhere between the hand-calculated version above and the original, uncorrected formula (using $df = n_1 + n_2 - 2$). It's called the Welch–Satterthwaite approximation (see Howell, 1997, p. 197), and it'll give us slightly more power. But you'll be doing it by hand in the exam and the W–S technique is too laborious to do by hand.

3.6. So are the variances equal or not? The *F* test

If you want to know whether to use the *equal variances* or *unequal variances* version of the two-sample unpaired *t* test, you obviously need to know whether your population variances are equal or not, and the only way you can usually find that out is to test whether your sample variances are equal or not. Actually, what we do is to ask if our sample variances are *significantly different* from each other; if they are, we use the 'unequal variances' *t* test; if they're not, we use the 'equal variances' *t* test.

There are several methods available for testing differences between variances. Firstly, **look at the data;** it may be obvious. A good formal statistical test is Levene's test, provided by all good statistical packages, but it's a bit too much work to calculate by hand. Even the pen-and-paper test suggested by Howell (1997, p. 198), though good, would take a lot of time in the exam. So we'll use the ***F* test.** This may not be the 'best' test (it has problems — becomes liberal — if the data are not normally distributed, though if they are, it's the most powerful at detecting differences in variances). However, it's quick and good enough to decide whether the variances are too different for the 'equal variances' version of the *t* test.

The F test

The *F* statistic is a ratio of two variances. If the two variances are equal, $F = 1$. If they're not, $F \neq 1$. How much more / less than 1 does it need to be before we declare the difference 'significant'? We find that from tables of critical values of *F* (see p. 126). The *F* distribution is based on *two* numbers for the degrees of freedom: one for the numerator, and one for the denominator. We might write this as $F_{a,b}$ where *a* is the number of *df* for the numerator and *b* is the number of *df* for the denominator:

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2}$$

In practice, tables of *F* don't give critical values for $F < 1$; they only give critical values for $F > 1$ (if you had $F < 1$, you could always take the reciprocal, 1 / F, and test that). So to make sure that our $F > 1$, we always put the **biggest variance on the top** (numerator) of the ratio, and the smallest variance on the bottom (denominator). So if the variances are different, the *F* statistic will be bigger than 1. In other words:

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2} \text{ if } s_1^2 > s_2^2$$

$$F_{n_2-1,n_1-1} = \frac{s_2^2}{s_1^2} \text{ if } s_2^2 > s_1^2$$

So you can run an *F* test on your data before choosing a *t* test; if it's significant (especially if $n_1 \neq n_2$), use the unequal variances *t* test; if it isn't, use the equal variances *t* test.

One more thing, though — if you want to test whether the variances are *different* with $\alpha = 0.05$ (two-tailed), you must run the *F* test itself with $\alpha = 0.025$. If you run the test with tabled values for $\alpha = 0.05$ (one-tailed), your actual two-tailed α will be 0.1. Why? Well, asking whether the variances are different without specifying the direction of the difference is a two-tailed test. The critical values of *F*, however, are for a one-tailed test (because we only test significance when $F > 1$, rather than $F < 1$). You've forced it to become a two-tailed test by calculating *F* in such a way that that $F > 1$; you must therefore double the stated one-tailed α to get the two-tailed α. The Tables & Formulae (p. 126) give both the one-tailed and 'two-tailed equivalent' values of α; you should use the 'two-tailed equivalent' α for testing whether two variances are different in this context.

One-tailed, two-tailed… notes only for Part II students using this for revision

When using *F* tests as part of analysis of variance (ANOVA), covered in Part II, use the *one-tailed* critical values of *F*. Why? Because ANOVA compares a measure of 'effect size' (MS_{effect}) with a measure of variability (MS_{error}): $F = MS_{effect} / MS_{error}$. MS_{effect} gets bigger no matter what the direction of the effect. We are only interested in whether an effect is bigger than we'd expect by chance; given the assumptions of ANOVA, it would not be sensibly possible to get an effect 'smaller' than we'd expect by chance alone. So we use the one-tailed critical values of *F*. It is in this (one-tailed) sense that $F = t^2$ as discussed below.

Relationship between the F test and the t test

The *t* test is actually a special case of the *F* test:

$$F_{1,k} = t_k^2 \text{ and } t_k = \sqrt{F_{1,k}}$$

where *k* is the number of degrees of freedom. In other words, a *t* test on *k df* is directly equivalent to an *F* test on 1 and *k df*. The difference that the *t* distribution is symmetrical about zero, since it deals with the differences between things, so values of *t* can be positive or negative. The *F* test deals with squared values, which are always positive, so *F* ratios are always positive (see Keppel, 1991, p. 121).

3.7. Assumptions of the *t* test

For any *t* test:

- You're testing hypotheses about the mean, which only makes sense if the mean is meaningful (it may not be if the measurement scale you used wasn't an interval or ratio scale — see p. 7).
- The maths behind the *t* test assumes that the underlying populations of the scores — or difference scores, for the paired *t* test — are **normally distributed.** (Large samples help to make up for lack of normality, but see below for more explanation. Casual rule of thumb: if $n > 15$ and the data don't look too weird, it's probably OK to use a *t* test; if $n > 30$, it's usually fine.) If this assumption is violated, you can't use *any* form of *t* test.

For two independent samples, to use the equal-variance *t* test, we assume

- The two samples come from populations with equal variances ($\sigma_1^2 = \sigma_2^2$), whether or not $n_1 = n_2$.

The *t* test is fairly robust to violations of this assumption if $n_1 = n_2$, but not if $n_1 \neq n_2$.

The case where the variances are unequal and so are the *n*s is the only situation in which the *t* test becomes liberal (see p. 46) — particularly if the sample with the smaller *n* has the larger variance (Boneau, 1960).

More detailed explanation of the normality assumption

The assumption about normally-distributed data was stated above casually. This is the full explanation; I'll use 'scores' to refer to the numbers being analysed.

1. The logic behind the *t* test doesn't make assumptions about the distribution of the scores *per se*, but it does assume that the *means* taken from samples of size *n* are themselves normally distributed (see p. 57). This is always true if the scores themselves are normally distributed, but is also true if the scores are not themselves normally distributed but the sample size is large (e.g. >15 if the scores are not too far from a normal distribution; >30 if the scores are very non-normal) — this is a consequence of something called the Central Limit Theorem (see p. 57). (See also Frank & Althoen, 1994, pp. 388-390, 401-406.) For the two-sample *t* test, simply read 'difference scores' instead of 'scores'.
2. The *t* test also makes assumptions about the distribution of the *variance* of the samples — it assumes that they have a χ^2 distribution (see pp. 58 and 81), which is true if the underlying scores are normally distributed, but may not be true if they're not (e.g. highly skewed; see Howell, 1997, p. 177, and core.ecu.edu / psyc / wuenschk / StatHelp / t-CLM.txt).
3. There is an additional reason for wanting the scores themselves to be normally distributed. If they aren't, the sample mean and the sample variance (or standard deviation, if you prefer) are not independent. For example, consider a positively skewed set of scores (see p. 17). Because low scores are generally closer together than high ones, samples with low means will tend to have lower variances than samples with high means. This skew can make the *t* test less powerful. The two-sample *t* test can also give distorted results if the two samples have *different* skew (see also Howell, 1997, p. 201).

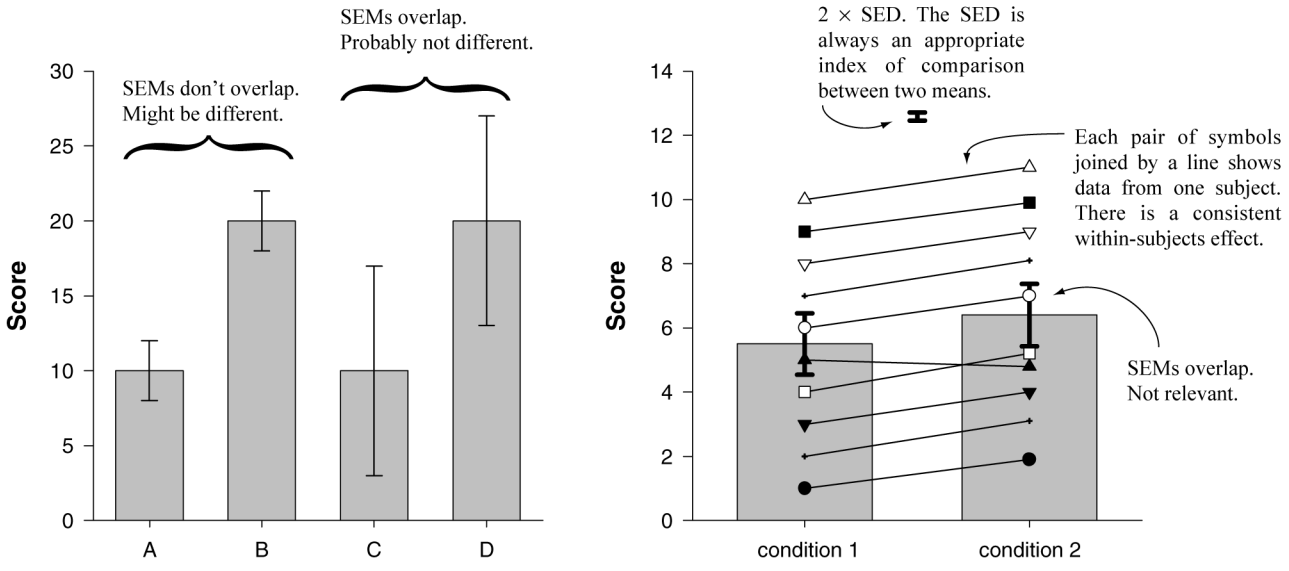
3.8. Graphical representation of between- and within-subject changes

'Error bars' (or 'mean ± variation') — the SEM is commonly used

The SEM is frequently used when people publish data. They may quote a measurement of '25.4 ± 1.2 g', or display a datum on a graph with a value of 25.4 units and error bars that are each 1.2 units long. These 'variation' indices could be one of several things — mean ± SD, mean ± 95% CI, mean ± SEM… The paper should state somewhere which one is being used, but usually it's the SEM. Why? First, it's smaller than the SD, so it conveys an impression of improved precision (remember that **accuracy** is how close a measurement is to a 'true' value and **precision** is how well it is defined; thus, 2.5000000003×10^8 m·s^{-1} is a more precise but far less accurate measurement of the speed of light than 3.0×10^8 m·s^{-1}). In fact, using the SEM is perfectly fair and correct: the precision of an estimator is generally measured by the standard error of its sampling distribution (Winer, 1971, p. 7). Secondly — more importantly — if the SEM error bars of two groups overlap, it's very unlikely that the two groups are significantly different. (This is explained somewhat in the figure.) The opposite isn't necessarily true, though — **just because two sets of error bars don't overlap doesn't mean they are significantly different** (they have to 'not overlap' by a certain amount, and that depends on the sample size, and so on).

Within-subjects comparisons and the SED

For **within-subjects** comparisons, SEMs calculated for each condition are highly misleading (see figure). For this comparison — indeed, for any comparison — the SED is an appropriate index of comparison, because that's what the *t* test is based on (*t* = difference between means / SED). So **if the difference between two means is 2 × SED. The SED is always an appropriate index of comparison between two means.**



Height of bar = mean
Error bar = ±1 SEM (1 SEM above, 1 SEM below the mean).

If the *n*s and SEMs of two groups are the same, then *t* = (difference between means) divided by ($\sqrt{2} \times$ SEM). And if the SEMs of the two groups are the same and the SEMs overlap, then the means differ by $<2 \times$ SEM, so $t < 2 / \sqrt{2} = 1.4$. And $t < 1.4$ is never significant even at the 0.1 level.

So for independent groups, if the SEM error bars overlap, there's probably not a significant difference.

For within-subject comparisons, the SEM of each condition is not helpful. The vertical bars show group means; their error bars show ±1 SEM. You would think that the groups don't differ. But in fact, the same subjects were tested in condition 1 and condition 2. The subjects all scored very differently, but there is a consistent improvement from condition 1 to condition 2. If we ran a paired-sample *t* test on the difference scores, we would find a highly significant difference between the two conditions. The appropriate index of variation to compare the two conditions is the standard error of the difference between means (SED), shown at the top.

Another way of plotting these data would just be to plot the difference scores, with their SEM; readers could then visually compare that mean to zero. However, that would not show the baseline scores.

The SED is always an appropriate index of comparison; a *t* test is calculated as (difference between means) divided by (appropriate SED). But different comparisons require different SEDs. If your error bars don't convey the right impression, consider using SEDs (as in the top-right example; you could say "the error bar is 2 × the standard error of the difference for the comparison between …").

Graphical presentation and interpretation of between- and within-subject changes.

greater than twice the SED, $t > 2$. And for a healthy n, $t > 2$ is significant at the two-tailed $\alpha = 0.05$ level (have a quick glance at your tables of critical values of t — p. 125).

The SED is therefore a very good index of variation that can be used to make visual comparisons directly, particularly if you draw error bars that are 2SED long — if the means to be compared are further apart than the length of this bar, there's a good chance the difference is significant. However, it's a bit more work to calculate the SED, which is why you don't see it very often.

If you want to work out an SED, just choose the appropriate t test and calculate the denominator of the t test. For between-group comparisons where the group SEMs are SEM_1 and SEM_2, you'll see that $SED = \sqrt{(SEM_1^2 + SEM_2^2)}$.

To summarize, for within-subject changes:
1. The mean within-subject change equals the difference of the group means.
2. The variance of the within-subject change may differ greatly from the variance of any one condition (group).
3. Present within-subject changes when the baseline varies a lot, or you want to show variance of the within-subject measure.
4. Present group means when the baseline matters.

3.9. Confidence intervals

One sample — confidence intervals on the population mean, μ

We can use the t formula to establish confidence intervals for particular measurements, just as we did for Z scores (p. 16). Suppose when we measured the heights of a group of $n = 10$ UK men and found $\bar{x} = 1.82$ m, $s = 0.08$ m. We could calculate the 95% confidence interval like this. Since

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

we can work out 95% critical values for t (i.e. $\alpha = 0.025$ each tail) with $n - 1 = 9$ *df*. From our tables (p. 125), these critical values are ±2.262. We can plug these into the formula above to find an expression for μ as a 95% confidence interval:

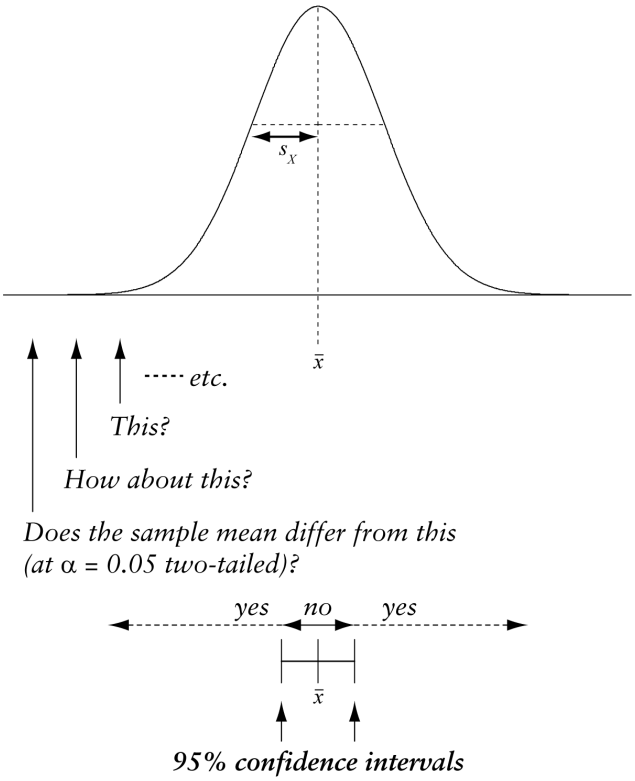
$$\pm 2.262 = \frac{1.82 - \mu}{\frac{0.08}{\sqrt{10}}}$$

$$\mu = 1.82 \pm 0.06$$

What would this mean? That there is a **95% chance that the true mean** height of UK men is in the range 1.76 to 1.88 m. We could also write this as a general formula:

$$\mu = \bar{x} \pm t_{critical\,(n-1)df} \frac{s_X}{\sqrt{n}}$$

Any value outside the confidence interval is significantly different from the sample mean at the specified level (e.g. any value outside the 95% CI is significantly different from the sample mean at $\alpha = 0.05$, two-tailed). Any value *inside* the CI is *not* significantly different from the sample mean at the specified level of α.



Confidence intervals.

Two samples — confidence intervals for a difference between means, $\mu_1 - \mu_2$

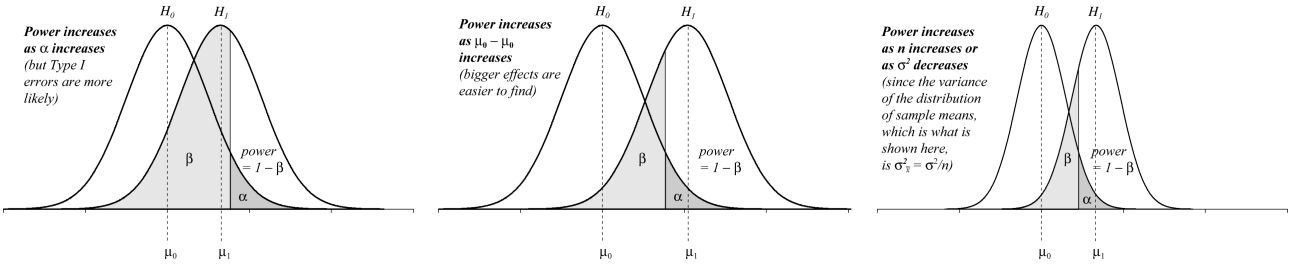
Similarly, if we have two samples whose mean difference is $\bar{x}_1 - \bar{x}_2$, we can use the formula for a two-sample *t* test to find the interval within which there is a 95% chance of finding the underlying population difference, $\mu_1 - \mu_2$.

3.10. Power and things that affect it

We won't talk about power in any great detail; certainly, you're not expected to calculate power. But it is helpful to understand what power is. Remember (see page 22) that α is the probability of rejecting the null hypothesis H_0 when it is in fact true (a Type I error); β is the probability of not rejecting H_0 when it is in fact false (a Type II error); power is $(1 - \beta)$, or the probability of rejecting H_0 when it is in fact false. If your power is 0.8, it means that you will detect 'genuine' effects with $p = 0.8$.

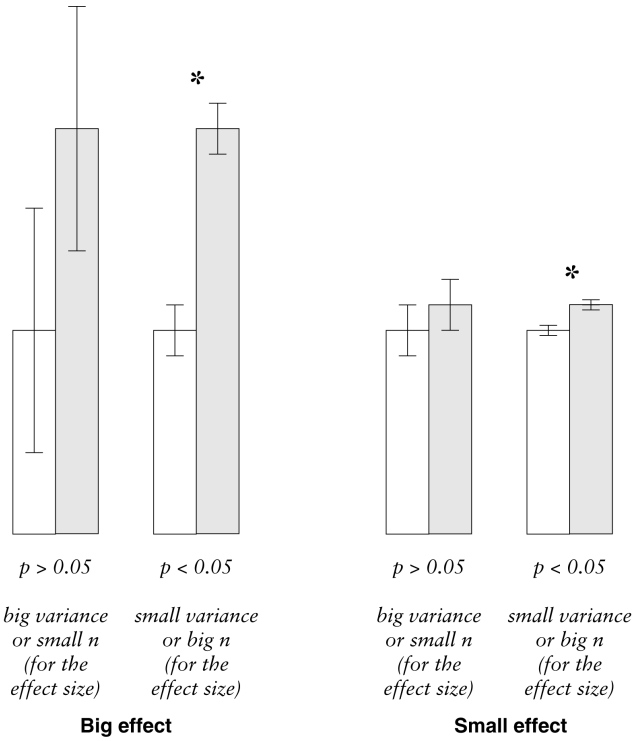
The consequences of Type II errors can be just as serious as those of Type I errors. If you run an expensive experiment with a very low power, you have a very small chance of finding the effect that you're looking for even if it does exist; if you then *don't* find it, you've probably wasted your time and money. (If you ever plan to run a seriously expensive experiment, make sure you understand how to do power calculations to work out how big your sample size should be, or ask a statistician to do it for you!)

Several things affect power: the size of the effect you're looking for (the difference between μ_0 and μ_1 — bigger effects give higher power), the sample size (n — the more observations you have, the higher the power), the variance of the sample (σ^2 — smaller variances give higher power), and of course your chosen level of α (higher α means lower β and therefore higher power, although higher α increases the chance of a Type I error). Have a look at the piccie (below).



Factors affecting power. If H_0 is true, and we take a set of samples each with mean $\bar{x}$, the mean of all the values of $\bar{x}$ will be μ_0. If H_1 is true, the mean of $\bar{x}$ will be μ_1. The distribution of all the values of $\bar{x}$ — the so-called 'sampling distribution of the mean' will be the curve labelled H_0 (if H_0 is true) or H_1 (if H_1 is true instead). The area under each curve is 1. Our job is to try to distinguish whether H_0 or H_1 is true on the basis of a single sample mean $\bar{x}$. We do this by setting α, the proportion of times that we reject H_0 when it is true. Setting α creates a criterion and thereby determines β, the chance of rejecting H_1 when it is true. In turn, this determines power, since this is $1 - \beta$ (the rest of the area under the H_1 curve). However, things other than α also affect power (middle and right-hand figures).

One thing that you should remember from this is that **significance levels do not indicate effect size.** Extremely large samples have power to detect very small effects with very small *p* values. Suppose a carefully-controlled study of a million people finds that running two miles a day decreases the risk of puffy ankles by 1% ($p < 0.001$). This is a study with high power finding a small effect that probably isn't important. On the other hand, **absence of evidence is not evidence of absence** — underpowered studies may fail to find large effects. A study of twenty 50-year-old men with heart disease might find no evidence that aspirin decreases the risk of a heart attack over the next five years ($p > 0.1$). This is a study with very low power failing to detect quite a substantial and important effect (aspirin does indeed reduce this risk).



Significance is not the same as effect size.

Reporting both may be useful (for example, giving the effect size with its 95% confidence interval; if the confidence interval includes 0, then the effect size is not significantly different from 0).

How big an effect needs to be to be important *depends on the experiment.*

3.11. Supplementary material: deriving the one-sample *t* test

The sampling distribution of the mean and the central limit theorem

Suppose we have a population with mean μ and variance σ^2. If we repeatedly take samples of n observations, we can say some things about the samples that we take. For each sample, we can calculate a sample mean $\bar{x}$. So we can collect lots of different sample means — many values of $\bar{x}$. Now we can ask what might at first appear to be an odd question: what will be the distribution of the sample means (also known as the **sampling distribution of the mean**)? What will be the mean of all the sample means (the mean of all the values of $\bar{x}$, written $\mu_{\bar{x}}$)? What will be the standard deviation of all these sample means (the standard deviation of all the values of $\bar{x}$, written $\sigma_{\bar{x}}$)? What we need to know is contained in a fact called the **central limit theorem.** There are various ways of stating this. The simplest is that if W_1, W_2, … W_n are independent, identically distributed random variables and $Y = W_1 + W_2 + … + W_n$, then the probability density function of Y approaches the normal distribution as $n \to \infty$. (This explains why the normal distribution so closely approximates so many biological, sociological, economic, and other variables that are themselves the sum of the effects of many other variables.) A more thorough version of the central limit theorem applicable to our present needs is this:

> Given a population with mean μ and variance σ^2, from which we take samples of size n, the distribution of sample means will have a mean $\mu_{\bar{x}} = \mu$, a variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$, and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. As the sample size n increases, the distribution of the sample means will approach the normal distribution.

This is very important. It doesn't matter whether or not the population is normally distributed; if you sample from it, the distribution *of* the sample means always approaches the normal distribution. (If the population is normally distributed and unimodal, the sample means will be normally distributed even if n is small; if the population is very skewed, n may have to be quite large — e.g. >30 — before the distribution of the means starts to become normally distributed.)

Very hard bit: why $\sigma_{\bar{x}} = \sigma/\sqrt{n}$? Well, the variance of individual observations is σ^2. So the variance of means of sample size 1 is σ^2. The variance of means of sample size n is the variance of (a sum of n individual observations, divided by n), by the definition of a mean, which is the variance of (a sum of n individual observations) divided by n^2. This is a consequence of the variance law $V(cX) = c^2V(X)$, or $V(X/c) = V(X)/c^2$. And the variance of (a sum of n individual independent observations) is $n\sigma^2$. This is a consequence of the variance sum law $V(X+Y) = V(X) + V(Y)$, which is true so long as X and Y are independent. So the variance of the samples means is $n\sigma^2/n^2 = \sigma^2/n$, and the standard deviation is the square root of this.

If we know the population SD, σ, we can test hypotheses very simply with a Z test

It's unusual for us to know the population standard deviation, σ. But sometimes we do. For example, we know that IQ in the general population has a mean of 100 and a standard deviation of 15. In this case, we saw (see p. 15) that we could calculate the probability that a single individual with an IQ of 89 came from the general population. We could calculate a Z score:

$$z = \frac{x - \mu}{\sigma}$$

which in this case would be $z = (83 - 100)/15 = -1.13$; we could look this up in our tables (p. 123) and find that the probability that a single IQ score of 83 or less could come from the general population is 0.129. We would not reject the null hypothesis that this subject was drawn from the general population.

But suppose that we have five subjects, and their IQs are 89, 94, 73, 82, and 77. Are these *five* subjects drawn from a healthy population (mean 100, SD 15)? The null hypothesis is that they are (null hypothesis: population mean $\mu = 100$). So what we do is this. We calculate our sample mean $\bar{x} = 83$ and sample size $n = 5$. We know from the central limit theorem that if we repeatedly took samples of size 5 from a population with $\mu = 100$ and $\sigma = 15$, that these sample means ($\bar{x}$) themselves would have a mean of $\mu_{\bar{x}} = 100$ and a standard deviation $\sigma_{\bar{x}} = 15/\sqrt{5} = 6.71$. We also know from the central limit theorem that the distribution of the sample means ($\bar{x}$) approaches a normal distribution. So we could obtain a *Z* score again:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{83 - 100}{6.71} = -2.53$$

Using our tables of *Z* scores (p. 123), we'd find that the probability of obtaining a *Z* score of –2.53 or more extreme is 0.0057. If we set our α to be 0.05 with a two-tailed test ($\alpha = 0.025$ each tail), we'd reject the null hypothesis, and conclude that our group of five subjects were not drawn from the general healthy population; the group mean of 83 was *significantly different* from 100. (It should be fairly obvious that our likelihood of finding a significant difference depends on the sample size *n*; larger samples have more **power** to detect a significant difference.)

More often, we do not know the population SD, σ, and can't use a Z test…

It's much more common that we don't know the population SD, σ, or the population variance, σ^2, so we have to estimate it from the sample SD, *s*, or the sample variance, s^2. Unfortunately, this complicates matters a bit. Although in the long run, the average value of the sample variance s^2 is equal to σ^2 (it's an *unbiased estimator*; see p. 13→ if you're interested), the distribution of s^2 is *positively skewed*. That means that although the average value of s^2 equals σ^2, more than half the values of s^2 are less than σ^2 (and less than half are more than σ^2 — though the values that are more than σ^2 are *much* more than σ^2, to balance things out). So any *individual* value of s^2 is likely to underestimate σ^2. (In fact, s^2 has a χ^2 distribution; see p. 81.)

What we have to do to compensate is to change from a *Z* test to something called a *t* test. Instead of calculating a *Z* score based on σ:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

we calculate a *t* score based on *s*:

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

So just as a *Z* score tells you how far a score (in this case $\bar{x}$) is from the mean (in this case $\mu_{\bar{x}}$) in terms the number of standard deviations (in this case $\sigma_{\bar{x}}$), a *t* score tells you how far a score (in this case $\bar{x}$) is from the mean (in this case $\mu_{\bar{x}}$) in terms of the number of *estimated* standard deviations (in this case $s_{\bar{x}}$).

But since s^2 is more likely than not to be smaller than σ^2, *t* is more likely than not to be bigger than *z*. (This also means that if you tried to calculate a *Z* score but incorrectly used s^2 rather than σ^2, your test would be too liberal.) The *t* score is not normally distributed; it has its own distribution. This distribution was worked out by William Gossett in 1908. Gossett worked for Guinness and they wouldn't let him publish under his own name, so he published under the pseudonym of Student. The distribution is therefore called **Student's *t* distribution.** There are in fact infinitely many *t* distributions, one for each *degree of freedom* (*df*; see below). For a one-sample *t* test, the number of degrees of freedom is $n - 1$, where *n* is the number of observations in the sample. As $n \to \infty$, $df \to \infty$, the distribution of s^2 becomes less and less skewed, and the *t* distribution becomes more and more like the normal dis-

tribution, *Z*. Anyway, we don't routinely need to calculate the distribution of *t* because we have it in the form of pre-calculated tables (p. 125). If our calculated value of *t* exceeds the relevant critical value for the appropriate number of degrees of freedom and α, we reject the null hypothesis.

More formally, when we calculate a *Z* test using $z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$, we assume that $\bar{x}$ is normally distributed; μ is a constant, and so is σ^2; therefore, the *z* score we calculate is also normally distributed. When we calculate $t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$, we assume that $\bar{x}$ is normally distributed (either as a consequence of the underlying scores *x* being normally distributed, or as a consequence of the Central Limit Theorem as the sample size becomes large; see p. 57). Again, μ is a constant, and so is *n*, but this time, assuming the underlying scores are normally distributed, s^2 has a χ^2 distribution (see p. 81). Therefore, we obtain something (*t*) that is a normally-distributed variable divided by the square root of a χ^2-distributed variable; that's what the *t* distribution really is.

Degrees of freedom (df)

When we begin, we have *n* observations, and all of them are free to vary. When we obtained the sample variance, s^2, we calculated the deviations of each observation from the sample mean ($x - \bar{x}$), rather than from the population mean ($x - \mu$). Because the sum of the deviations about the mean, $\sum(x - \bar{x})$, is always zero, only $n - 1$ of the deviations are free to vary. We've 'used up' one of our degrees of freedom by calculating $\bar{x}$ using data from our sample. So s^2 is based on $n - 1$ degrees of freedom, and so is our *t* statistic.

3.12. Supplementary material: deriving the two-sample *t* test

The distribution of differences between means; deriving the two-sample t test

When we want to compare two groups, what we do is take two samples from two different populations, X_1 and X_2, and ask if the two populations have the same mean ($\mu_1 = \mu_2$) or not ($\mu_1 \neq \mu_2$). Suppose the populations have means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. If we draw pairs of samples, of size n_1 from population X_1 and of size n_2 from population X_2, we can calculate the *difference* between each pair of sample means $\bar{x}_1$ and $\bar{x}_2$, or $\bar{x}_1 - \bar{x}_2$. If we draw many pairs of samples, we can calculate the **distribution of the differences between sample means,** also called the sampling distribution of differences between means. The mean difference between sample means (μ_d) will be given by

$$\mu_d = \mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

From the central limit theorem, we know that the variance of sample means from X_1 will be $\sigma^2_{\bar{x}_1} = \frac{\sigma_1^2}{n_1}$, and similarly $\sigma^2_{\bar{x}_2} = \frac{\sigma_2^2}{n_2}$. The **variance sum law** states that the variance of a sum or difference of two variables is:

$$\sigma^2_{X_1 + X_2} = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$$

$$\sigma^2_{X_1 - X_2} = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

where ρ is the correlation between them; therefore, for two *independent* variables ($\rho = 0$), the variance of the sum or difference of the variables is the sum of their variances ($\sigma^2_{X_1 \pm X_2} = \sigma_1^2 + \sigma_2^2$). Therefore, the variance of the difference between our two means will be

$$\sigma_d^2 = \sigma_{\bar{x}_1 - \bar{x}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

so the corresponding standard deviation is

$$\sigma_d = \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This is called the **standard error of the difference between means (SED).** We now know the mean and SD of the distribution of the differences between sample means; all that's left is to determine the shape of this distribution. Another theorem tells us that the sum or difference of two independent normally-distributed variables is itself normally distributed; we're basically done. If we knew the population SDs σ_1 and σ_2 — which is very unusual! — we could perform a *Z* test:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

… but that's very unlikely. So just as we used a *t* test in place of a *Z* test earlier, when we had to estimate σ based on the sample SD, *s*, we want to do the same thing now. Only there's a problem. Remember that the shape of the *t* distribution depends on how skewed our estimate of the standard error is — that is, on the number of degrees of freedom of our estimate (*s*). But in this two-sample *t* test, we have *two different* variances in the denominator. What do we do?

The simplest thing to do is to assume that the two populations have equal variances ($\sigma_1^2 = \sigma_2^2$). We can denote this variance simply σ^2, and its estimate s_p^2 (the 'pooled' variance estimate, explained in more detail below). Then we only need to worry about a single estimate, and its skewness (*df*), so we can use this formula in place of the *Z* test shown above:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Generally, the null hypothesis is that the means are the same ($\mu_d = \mu_1 - \mu_2 = 0$), so we can simplify this a bit further:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

and there we have the formula that I stated earlier. What about the **degrees of freedom?** Well, we started with $n_1 + n_2$ degrees of freedom. We've calculated two sample variances, so we've lost 2 *df*; we're left with $(n_1 + n_2 - 2)$ *df*.

Pooling variances when $n_1 \neq n_2$

We've seen that this use of the *t* test for two independent samples requires the assumption that the two samples come from populations with equal variances ($\sigma_1^2 = \sigma_2^2$). We denoted this variance simply σ^2. This is often a reasonable assumption, particularly if we start with two groups of equivalent subjects ($\Rightarrow$ equal variances) and then do something to one or both groups that affects the mean of those groups; the variances will often be relatively unaffected. Anyway, when we use the *t* test, we are using the sample variances s_1^2 and s_2^2 to estimate σ^2. If our sample sizes are not equal ($n_1 \neq n_2$), then the larger sample will probably give us a *better* estimate of σ^2 (both s_1^2 and s_2^2 are meant to be estimating the same thing, since we're assuming $\sigma_1^2 = \sigma_2^2$, and the larger sample contains more information). Accordingly, we would be better off with a **weighted average,** in which the sample variances are weighted by their degrees of freedom ($n - 1$), the number of observations on which they are based. This weighted average is usually called the **pooled variance estimate,** s_p^2:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If we use that in our *t* test, we get the general formula for the two-sample unpaired *t* test (equal variance version) that we've just seen:

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1-\bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

As always, this formula for *t* involves dividing the difference between means by the standard error of the difference between means (SED).

Another way of thinking about the pooled variance is in terms of *sums of squares* (we mentioned this on p. 36→ in some of the wavy-line bits to do with what r^2 means in correlation). A variance is a 'sum of squares' (the sum of squared deviations from the mean) divided by the degrees of freedom. So when we multiple each sample variance by its own *df* we get the sample sums-of-squares. We also said that you could only add sample variances meaningfully when they were based on the same *df*, but you can add sums of squares any way you like — so to calculate the pooled variances, we convert the sample variances to the sample sums-of-squares, add them together, and divide by the overall number of *df* to get the overall (pooled) variance.

If the sample sizes are equal ($n_1 = n_2 = n$), then this formula can be simplified like this:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(n - 1)s_1^2 + (n - 1)s_2^2}{2n - 2} = \frac{s_1^2 + s_2^2}{2}$$

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{n}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2}{n}s_p^2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2}{n}\left(\frac{s_1^2 + s_2^2}{2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

The two-sample t test with unequal sample variances

If the two sample variances are *not* equal, then we can't make our simplifying assumption that $\sigma_1^2 = \sigma_2^2$. Unfortunately, this means we have *two different variance estimates* in the denominator for our formula — both of which have their own χ^2 distribution — whereas the *t* distribution is predicated on dividing by something involving a *single* variance estimate. As a result, the resulting statistic will not necessarily have a *t* distribution this time. It is therefore written *t*′:

$$t' = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

However, we know that the sum of our two estimated standard errors can, at most, be only as skewed (lopsided) as the more lopsided of our two estimates, which is the one with the lower *df*. In other words, the probability of getting extreme values of *t*′ must be no more than the probability of getting extreme values of a *t* score *with the lower of the two dfs*. Consequently, the usual procedure for calculating this test by hand is to look up the value of *t*′ in tables of *t*, but using the **smaller of ($n_1 - 1$) and ($n_2 - 1$)** as the number of degrees of freedom. A computer would calculate something more accurate for the *df*, which lies somewhere between the smaller of ($n_1 - 1$) and ($n_2 - 1$) and their sum, ($n_1 + n_2 - 2$). This is called the Welch–Satterthwaite approximation to the *t*′ distribution (Welch, 1938; Satterthwaite, 1946), but it's too complicated to do by hand.

3.13. Examples 3: parametric difference tests

Interval estimation from samples (and a one-sample* t *test)

Q1. The following ten measurements were made of the light intensity at which a glare source impaired reading (in log trolands). Within what interval can we infer that the true mean (i.e. *population* mean, or mean of a very large number of such measurements) lies, with a 90% probability of being right?

4.32 5.07 4.29 6.02 5.11 4.93 3.98 4.83 5.50 6.10

Q2. In an experiment to measure the speed of discriminating words from non-words, the following 12 discrimination times (in ms) were recorded:

605 460 752 321 550 612 700 680 800 491 523 594

Within what interval is there a 95% probability that the true (population) mean lies?

Q3. In an experiment on judging the equality of weights the following ten values were set by a subject as being equal to 100 g. What is the 95% confidence interval within which the true (population) mean lies? Is the mean significantly different from 100 at the 5% level?

100.2 96.3 110.9 89.3 95.0 98.5 105.6 99.8 102.4 97.6

Two-sample* t *test

These examples also appear in the 'Examples 4' section for nonparametric difference tests — and you might like to try a few more of the examples from 'Examples 4' using t tests as practice.

Q4. A traffic survey measures the speed of 15 cars chosen randomly each morning over a quarter-mile stretch of road. One ordinary Monday these were (in m.p.h.):

32 45 37 41 28 36 40 49 34 36 33 30 40 38 39

On the next Monday in another ordinary working week on which there were similar weather conditions, a 'simulated accident scene' was placed 50 yards before the start of the measurement area, and the speeds of fifteen cars measured were:

33 27 38 35 30 32 29 20 37 44 31 36 30 34 32

Did the simulated accident significantly reduce drivers' speeds?

Q5. Twelve student volunteers performed a card-sorting task: they sorted 250 cards on one day, 500 on the next day starting 20 min after having ingested a pharmacologically-active substance, and 250 on a third day. The table gives the number of errors in sorting they made on the second day, and the total errors on the first and third days. Does the substance have any effect on card-sorting accuracy?

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Day 2	12	17	9	3	16	10	28	14	5	19	20	8
Days 1 & 3	16	16	11	5	10	13	36	11	8	11	20	14

Q6. The short-term memory span for digits was measured for a number of students specializing in arts (A) and science (S) subjects. The table gives each student's mean span with his subject group:

A	A	S	A	S	S	S	S	A	S	A	A	S	S	A
5.8	7.3	7.1	6.9	8.2	5.9	6.4	6.8	7.7	6.0	6.3	5.2	6.2	6.6	7.4

A	S	S	S	A	S	A	A	S	S
6.5	7.0	7.2	6.1	7.9	7.4	7.0	6.2	6.4	8.0

Is there a significant difference between the digit spans of arts and science students?

Q7. Two groups of subjects are shown an ambiguous figure, and the time taken until the first reversal of its appearance is measured for each subject. One group had previously seen the figure in a form strongly biased to show one of its alternative appearances; the other had no such pre-exposure (control group). The times to first reversal (in s) were:

Pre-exposure group	7.4	7.0	6.8	8.2	6.5	7.5	5.8	6.3	7.1	6.6
Control group	6.2	7.3	5.6	5.9	6.0	6.9	6.1	5.4		

Does pre-exposure to the biased figure lengthen the time to first reversal?

Q8. In an experiment in which briefly-flashed letters were superimposed on either a random or a checkerboard black-and-white pattern, one subject gave the following results:

Letter	a	c	e	n	o	s	u	v	x	z
% correct recognitions:										
On random field	67	43	49	31	40	52	35	74	83	77
On checkerboard	79	51	58	28	44	52	28	87	90	81

Do the checkerboard and random fields have significantly different effects on the visibility of the letter?

F test

Do the variances of the following pairs of groups differ at the .10 level?

Q9.	Group A:	–10	4	3	–5	12	6	7	1	–8
	Group B:	1	4	2	–1	5	–2	0	3	
Q10.	Group A:	20	47	150	10	60	120			
	Group B:	60	65	45	30	25				

4. Difference tests — nonparametric

Objectives

This time, we'll discuss some nonparametric difference tests. If you recall, nonparametric tests generally have lower power than parametric tests, but make fewer assumptions about the distribution of the data, so they may be valid when parametric tests are not. These are the rough equivalents of the parametric and nonparametric tests we cover:

Parametric test	Equivalent nonparametric test
Two-sample unpaired *t* test	Mann–Whitney *U* test
Two-sample paired *t* test	Wilcoxon signed-rank test with matched pairs
One-sample *t* test	Wilcoxon signed-rank test, pairing data with a fixed value

They assume that the variable is measured on at least an ordinal scale. (That's it.)

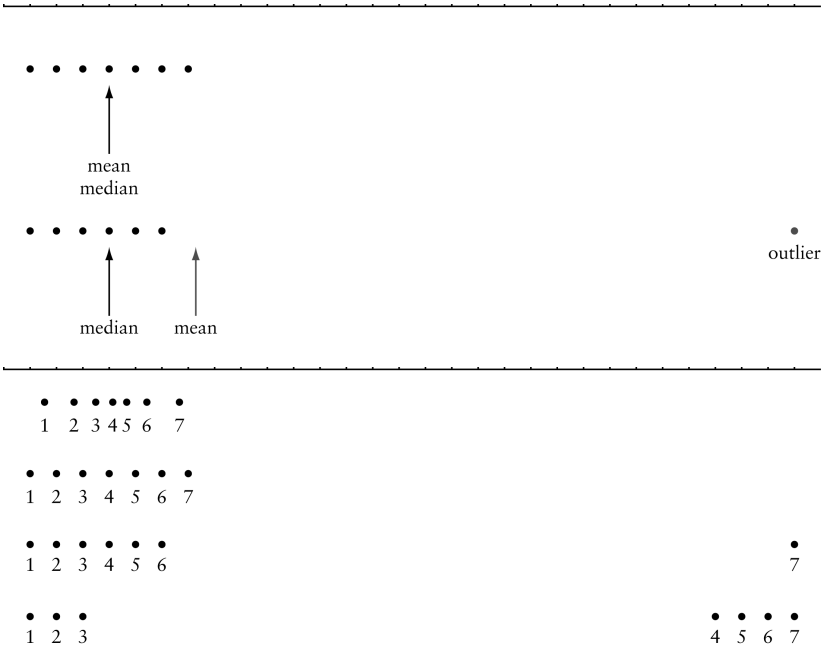
Stuff with a solid edge, like this, is important.

But remember — you can totally ignore stuff with single / double wavy borders.

4.1. Background

Nonparametric tests often operate on the **rank** order of a set of numbers, rather than on the numbers themselves. This also means that nonparametric tests are less affected by **outliers** (a few extreme scores) than parametric tests. Outliers may make parametric tests *less* powerful (they increase the variance as well as distorting the mean), sometimes less powerful than the nonparametric equivalent.

It should be obvious how ranking 'removes' information about the distribution. The scores {2,8,10,12,14,24} might have come from a normal distribution and the scores {1,2,3,100,101,102} might have come from a bimodal distribution, but both reduce to the ranks {1,2,3,4,5,6}.



Medians are less sensitive to outliers than means.

Ranking removes information about the distribution of scores.

Ranking removes information about the distribution.
Whatever the distribution (normal, flat, skewed, bimodal…), the ranks are the same:
1, 2, 3, 4, 5, 6, 7.

How to rank data (repeated from p. 33)

Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

5	8	9	12	12	15	16	16	16	17

Then start assigning them ranks. When you come to a tie, give each value the mean of the ranks they're tied for — for example, the 12s are tied for ranks 4 and 5, so they get the rank 4.5; the 16s are tied for ranks 7, 8, and 9, so they get the rank 8:

X:	5	8	9	12	12	15	16	16	16	17
rank:	1	2	3	4.5	4.5	6	8	8	8	10

4.2. The Mann–Whitney *U* test (for two independent samples)

This is a nonparametric analogue of a two-sample unpaired *t* test. Its null hypothesis is that the two samples were drawn from identical populations (rather than the *t* test's null hypothesis that the two samples were drawn from populations with the same means). So a 'significant' Mann–Whitney result might be due to a difference between the *central tendency* of the two populations (like a 'significant' *t* test) but it might also have been due to some other difference, such as a difference in the *distributions* of the populations. If we assume the distributions are similar, a significant Mann–Whitney test suggests that the **medians** of the two populations are different.

Basic logic of the test

Let's suppose we have two samples with n_1 and n_2 observations in each ($n_1 + n_2 = N$ observations in total). We can rank them, lowest to highest, from 1 to N. If the two samples come from identical populations, the sum of the ranks of 'sample 1' scores is likely to be about the same as the sum of the ranks of 'sample 2' scores. If, on the other hand, sample 1 comes from a population with generally much lower values then sample 2, then the sum of the ranks of 'sample 1' scores will be lower than the sum of the ranks of 'sample 2' scores.

In the test, shown below, the rank sums are first 'corrected' for the fact that larger groups tend to have larger rank sums simply because they have more observations; the resulting 'corrected' numbers are called U_1 and U_2. So if the groups are very different, one of these numbers will be very high and the other will be very low; if the groups are very similar, both U_1 and U_2 will be close to each other. Since they are related, though, we simply pick the smaller of the two, call it U, and look it up to see if it's smaller than some critical value. (We could equally have picked the larger of the two and looked it up in a different set of tables to see if it was greater than a critical value, but we haven't supplied those tables!)

Calculating the Mann–Whitney U statistic

1. Call the smaller group 'group 1', and the larger group 'group 2', so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 (= R_1) and group 2 (= R_2).
3. $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$
4. $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$
5. The Mann–Whitney statistic U is the smaller of U_1 and U_2.

Check your sums: verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$.

It doesn't matter which numbers you call U_1 and U_2, since all you do is take the smaller. Incidentally, why this formula for $R_1 + R_2$? Because if you add consecutive numbers from 1 to x, the total is $\sum_{i=1}^{x} i = \frac{x(x+1)}{2}$.

Handy hint: if the ranks don't overlap at all

If the ranks of the two groups **do not overlap,** then $U = 0$. This can save you some time in calculation. For example, if the ranks of group 1 are {1, 2, 3, 4, 5} and the ranks of group 2 are {6, 7, 8, 9}, then $U = 0$.

Determining a significance level from U

If n_2 is small, look up the critical value of U in tables (see p. 128) — **values of U smaller than the critical value are significant. If $n_2 > 20$,** the U statistic is approximately normally distributed; mean $\mu = \frac{n_1 n_2}{2}$, variance $\sigma^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$. So we can calculate a ***Z* score:**

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

and test that in the usual way (see p. 15).

If there is a significant difference, which way round is it?

When you calculate a t test, if you find a significant difference, then it's obvious 'which way round' the difference — you've already calculated the group means, so you can instantly see which group has the larger means (and the t test has told you that this is a significant difference). But with the Mann–Whitney U test, rejection of the null hypothesis tells you that the samples are unlikely to have come from the same population — and if you assume that the samples have the same distribution (see above), then a significant U tells you that the **medians** are different. But to establish which group actually has the larger median, **you have to calculate the medians** of each group. You can't rely on the rank sums — if the group sizes are unequal, it's *not* always the case that the group with the larger rank sum has the larger median.

Example

Borrowing an example from Howell (1997, p. 651), suppose we imagine that we collect information on birth weights of babies whose mothers received prenatal care either from the first trimester onwards or from the third trimester onwards. Suppose these birthweights, in kg, were {1.68, 3.83, 3.11, 2.76, 1.70, 2.79, 3.05, 2.66, 1.40, 2.775} for the first trimester group and {2.94, 3.38, 4.90, 2.81, 2.80, 3.21, 3.08, 2.95} for the third trimester group. If we chose to calculate a Mann–Whitney test on these data, we would calculate the ranks as {2, 17, 14, 5, 3, 7, 12, 4, 1, 6} for the first trimester group (n = 10, rank sum = 2 + 17 + 14 + … = 71) and {10, 16, 18, 9, 8, 15, 13, 11} for the third trimester group (n = 8, rank sum = 100). We'd therefore call the third trimester group 'group 1', because it's the smaller, and the first trimester group 'group 2'. So we have $n_1 = 8$, $n_2 = 10$, $R_1 = 100$, $R_2 = 71$. From this we can calculate $U_1 = 64$, $U_2 = 16$. The Mann–Whitney U is the smaller of these, i.e. 16.

From our tables (see p. 128) we can find that the critical value of U for these values of n and a two-tailed test at $\alpha = 0.05$ is 18. Our U is less than this, so it's significant; we reject the null hypothesis, and say that there's a difference between the birthweights of our two sets of babies, $p < 0.05$.

If our ns had been larger, we could have calculated a Z score. Pretending for a moment that our ns were larger, for these data $z = -2.13$, corresponding to $p = 0.033$.

For the same data, a two-sample unequal-variance *t* test would have given $p = 0.063$, and a two-sample equal-variance *t* test would have given $p = 0.066$. This is an example when a nonparametric test has more power because the assumptions of the parametric test — in this case normality of the underlying distribution — were not met.

4.3. The Wilcoxon matched-pairs signed-rank test (for two related samples)

This is a nonparametric test for **paired** scores. It's the nonparametric analogue of the *t* test for related samples (the paired *t* test). The null hypothesis is that the distribution of differences between the pairs of scores is symmetric about zero. (Since the median and the mean of a symmetric population are the same, the null hypothesis can be restated either as 'the differences between the pairs of scores are symmetric with a mean and a median of zero'.)

Let's do this as a worked example (borrowed from Howell, 1997, p. 653). Suppose 10 subjects have their systolic blood pressure measured (BP_1), engage in a running program for 6 months, and then have their systolic blod pressure measured again (BP_2). We can calculate the difference for each subject as $BP_2 - BP_1$. If there's no difference between the 'before' and 'after' scores, there should be about as many differences that are positive as there are differences that are negative…

Calculating the Wilcoxon matched-pairs signed-rank statistic, T

The procedure is:

1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or –).
4. Add up all the ranks for difference scores that were positive; call this T^+.
5. Add up all the ranks for difference scores that were negative; call this T^-.
6. The Wilcoxon matched-pairs statistic *T* is the smaller of T^+ and T^-.

Check your sums: verify that $T^+ + T^- = \frac{n(n+1)}{2}$.

Here's a worked example:

Before (BP_1):	130	148	170	125	170	130	130	145	119	160
After (BP_2):	120	148	163	120	135	143	136	144	119	120
Difference ($BP_1 - BP_2$):	10	0	7	5	35	–13	–6	1	0	40
Rank of difference (ignoring zero differences and sign):	5		4	2	7	6	3	1		8
'Signed rank'	*5*		*4*	*2*	*7*	*–6*	*–3*	*1*		*8*
Ranks of positive differences:	5		4	2	7			1		8
Ranks of negative differences:						6	3			

The 'signed rank' row is what gives the test its name; it's what you get when you put the signs (+ or –) from the difference scores back on the ranks you calculated by ignoring those signs. But you don't need to do this to calculate *T*.

The difference scores don't appear to be anything like normally distributed, so we want to use a distribution-free (nonparametric) test. We can calculate $n = 8$ (since we ignore zero differences), $T^+ = 5 + 4 + 2 + 7 + 1 + 8 = 27$, and $T^- = 6 + 3 = 9$. Therefore the Wilcoxon statistic $T = 9$.

Determining a significance level from T

For small *n*, look up the critical value of *T* in tables (see p. 129) — **values of *T* smaller than the critical value are significant. If *n* > 20,** the *T* statistic is approximately normally distributed; mean $\mu = \frac{n(n+1)}{4}$, variance $\sigma^2 = \frac{n(n+1)(2n+1)}{24}$. So we can calculate a ***Z* score:**

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

and test that in the usual way (see p. 15).

In our example, since $n = 8$, our tables (p. 129) tell us that the critical value of T is 6 (for $\alpha = 0.05$ two-tailed). Our T was 9. Therefore, the difference in blood pressure was not significant ($p > 0.05$).

If there is a significant difference, which way round is it?

In the case of the Wilcoxon test, it is always true that the set of scores with the larger rank sum has the larger median. So if the Wilcoxon test gives you a significant result and $T^+ > T^-$, then the difference between scores is significantly *greater* than zero; if $T^+ < T^-$, then the difference between scores is significantly *less* than zero.

4.4. Using the Wilcoxon signed-rank test as a one-sample test

The Wilcoxon signed-rank test may also be used to test whether the median of one group of scores is significantly difference from some expected value *M*. In this case, the null hypothesis is that the median is equal to *M*. Calculate a difference score **(*x* – *M*)** for each score *x*, and proceed as above.

4.5. Supplementary and / or advanced material

Tied ranks

The formulae given here for the Mann–Whitney and Wilcoxon tests should actually be modified if there are *tied* ranks (i.e. if two observations have the same value), and a computer would do this for us, but since there are no major problems if there aren't very many tied ranks and the formulae are complex that way, we simply ignore the problem when calculating *U* or *T* by hand.

The Wilcoxon rank-sum test (not the same as the Wilcoxon signed-rank test!)

There are actually *two* tests based on the logic used for the Mann–Whitney *U* test: they are the Mann–Whitney *U* test itself and the *Wilcoxon rank-sum test*. They're directly equivalent: both will give the same *p* value. (Some people even mix the names up, calling *U* a Wilcoxon rank-sum statistic, which confuses everybody.) The Mann–Whitney *U* test is more popular and has a name that's not so easily confused with the Wilcoxon signed-rank test. However, the Wilcoxon rank-sum calculations make it a bit clearer how we get a statistic out of the sums of a set of ranks, so I've included it here **only** in case you want to understand how the two tests work.

1. Call the smaller group 'group 1', and the larger group 'group 2', so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 (= R_1) and group 2 (= R_2).
3. If $n_1 < n_2$, then $W_S = R_1$. If $n_1 = n_2$, then W_S = whichever of R_1 and R_2 is smaller.
4. Calculate $W'_S = n_1(n_1 + n_2 + 1) - W_S$.

Now we evaluate W_S and W'_S using tables (not supplied here). The smaller W_S is, the more likely it is to be significant. W_S will be significant (small) if the smaller group (group 1) contains significantly smaller-than-average ranks, or if the larger group (group 2) contains significantly larger-than-average ranks, i.e. if group 1 < group 2. W'_S is the sum of the ranks we would have found if we reversed our ranking and ranked from high to low; it will be significant (small) if group 2 > group 1. Normally we want to test for a two-tailed difference between groups; we'd then pick

whichever of W_S and W_S' is the smaller and look up the critical values in tables (doubling α if the table gives one-tailed values).

Two other ways of calculating the Mann–Whitney U statistic

This shows the equivalence of the Wilcoxon rank-sum and Mann–Whitney tests:

1. Compute W_S and W_S' as above. Let W_S'' be whichever of the two is larger.
2. The Mann–Whitney statistic $U = \frac{n_1(n_1 + 2n_2 + 1)}{2} - W_S''$

A third method is this:

1. For each observation in group 1, count the number of observations in group 2 that exceed it (score 0.5 for equality). Sum these values to obtain U_1.
2. Do the same for group 2 to obtain U_2.
3. The Mann–Whitney statistic U is the smaller of U_1 and U_2.

4.6. Examples 4: nonparametric difference tests

Many of these examples are also suitable for further practice with *t* tests.

Mann–Whitney U test

Find the value of *U* for each of the following pairs of groups of observations, and discover whether the difference between the groups is significant at the 0.05 level, two-tailed.

Q1.	Group A:	43	39	57	62						
	Group B:	51	63	70	55	59	66				
Q2.	Group A:	4.5	2.3	7.9	3.4	4.8	2.7	5.6	6.1	3.5	
	Group B:	3.5	4.9	1.1	2.5	2.3	4.1	0.7			
Q3.	Group A:	650	710	437	520	583	492	555			
	Group B:	573	617	648	861	732	689	741			
Q4.	Group A:	43	70	51	35	60	77	48	62	57	75
	Group B:	90	45	73	64	86	59	88	72	89	
Q5.	Group A:	48	60	75	86	79	39	52	75	93	57
		62	71	69	80	69	62	70			
	Group B:	54	93	82	67	81	77	91	79	63	74
		99	84	76	68	71	90				

Wilcoxon matched-pairs signed-rank test

In the following examples, find the significance level of the differences between the groups in (*a*) a one-tailed and (*b*) a two-tailed test. The groups are arranged in matched pairs, the members of each pair being shown one above the other.

Q6.	Group A:	4.5	2.3	7.9	6.8	5.3	6.2	5.7		
	Group B:	4.3	2.7	9.0	6.7	5.6	10.1	6.9		
Q7.	Group A:	127	163	149	101	137	125	141	142	133
	Group B:	135	170	181	111	151	120	138	153	140
Q8.	Group A:	5	3	7	11	9	4	3	2	
	Group B:	7	4	6	12	6	10	9	3	
Q9.	Group A:	14	17	19	25	33	15	17	19	23
	Group B:	11	17	15	26	19	14	13	20	18

Mixed examples

(These are all fictitious experiments!)

Q10. A traffic survey measures the speed of 15 cars chosen randomly each morning over a quarter-mile stretch of road. One ordinary Monday these were (in m.p.h.):

32 45 37 41 28 36 40 49 34 36 33 30 40 38 39

On the next Monday in another ordinary working week on which there were similar weather conditions, a ‘simulated accident scene’ was placed 50 yards before the start of the measurement area, and the speeds of fifteen cars measured were:

33 27 38 35 30 32 29 20 37 44 31 36 30 34 32

Did the simulated accident significantly reduce drivers’ speeds?

Q11. In a reaction-time experiment, the stimulus to react to was a recorded voice, sometimes the same voice that had just given a 'ready' signal, and sometimes a different one. Twelve subjects' results were as follows (RTs in ms):

Subject	1	2	3	4	5	6	7	8	9	10	11	12
RT to same voice	302	287	350	296	411	337	326	343	315	371	299	316
RT to different voice	340	302	359	352	408	361	328	340	347	392	326	333

Is there a significant difference in RT between the two conditions?

Q12. Twelve student volunteers performed a card-sorting task: they sorted 250 cards on one day, 500 on the next day starting 20 min after having ingested a pharmacologically-active substance, and 250 on a third day. The table gives the number of errors in sorting they made on the second day, and the total errors on the first and third days. Does the substance have any effect on card-sorting accuracy?

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Day 2	12	17	9	3	16	10	28	14	5	19	20	8
Days 1 & 3	16	16	11	5	10	13	36	11	8	11	20	14

Q13. A survey was conducted to determine people's opinions of selected foreign countries. The overall order of preference among the sample interviewed (starting with the most preferred) was Australia, Canada, Denmark, New Zealand, Holland, Germany, France, Zimbabwe, Spain, South Africa, Italy.

Was a significant preference shown for Commonwealth and ex-Commonwealth countries on the one hand over European countries on the other?

Q14. Twelve cod-graders grade the following numbers of cod per hour:

1382	1545	1106	1761	1560	1669	1292	1418	1477	1351	1523	1618

After a number of sessions working through the teaching programme *Defect Detection in White Fish Processing: Intermediate Level*, their cod grading rates were (taking the graders in the same order as above):

1390	1422	1119	1578	1553	1682	1101	1376	1468	1099	1478	1564

Has the teaching programme had any effect on their grading rates?

Q15. The short-term memory span for digits was measured for a number of students specializing in arts (A) and science (S) subjects. The table gives each student's mean span with his subject group:

A	A	S	A	S	S	S	S	A	S	A	A	S	S	A
5.8	7.3	7.1	6.9	8.2	5.9	6.4	6.8	7.7	6.0	6.3	5.2	6.2	6.6	7.4

A	S	S	S	A	S	A	A	S	S
6.5	7.0	7.2	6.1	7.9	7.4	7.0	6.2	6.4	8.0

Is there a significant difference between the digit spans of arts and science students?

Q16. Two groups of subjects are shown an ambiguous figure, and the time taken until the first reversal of its appearance is measured for each subject. One group had previously seen the figure in a form strongly biased to show one of its alternative appearances; the other had no such pre-exposure (control group). The times to first reversal (in s) were:

Pre-exposure group	7.4	7.0	6.8	8.2	6.5	7.5	5.8	6.3	7.1	6.6
Control group	6.2	7.3	5.6	5.9	6.0	6.9	6.1	5.4		

Does pre-exposure to the biased figure lengthen the time to first reversal?

Q17. Twelve people are engaged in 'experimental conversation'. In the 'positive' condition they are 'reinforced' by an approving 'uh-huh' from the experimenter whenever they use the personal pronoun 'I'. In the 'negative' condition they are 'punished' by a disapproving 'huh' when they say 'I'. The rates of 'I' emission in the experiment are as follows (responses in a 10-min interval):

Subject	A	B	C	D	E	F	G	H	I	J	K	L
Pre-exposure group	17	62	20	11	31	25	15	38	47	22	26	8
Control group	14	68	19	3	27	26	9	22	40	19	20	11

Does reinforcement have the effect you would expect?

Q18. Two new-born bats are taken from each of a number of litters. One of each pair is kept in a cage, the other being allowed to live freely in the experimenter's office (despite protests from the occupants of nearby offices). After one month, their moth-catching abilities are tested in a standard Batman™ experimental chamber. The number of moths caught (out of a possible total of 25) are given below. Does experience in the first month of life have any effect on moth-catching ability in bats?

Litter number	1	2	3	4	5	6	7	8	9	10
Caged bat	8	16	0	10	6	12	8	2	15	9
Free-living bat	18	25	17	6	11	11	12	10	15	14

Q19. The following are the scores on the Seashore Test of Musical Aptitude of a number of 10-year-olds:

Right-handed children	28	54	37	102	66	30	41	56	34	72
Left-handed children	46	50	83	27	40	39	61	33	59	87

Do these data reveal a relationship between handedness and musical aptitude as measured by the Seashore Test?

Q20. In an experiment in which briefly-flashed letters were superimposed on either a random or a checkerboard black-and-white pattern, one subject gave the following results:

Letter	a	c	e	n	o	s	u	v	x	z
% correct recognitions:										
On random field	67	43	49	31	40	52	35	74	83	77
On checkerboard	79	51	58	28	44	52	28	87	90	81

Do the checkerboard and random fields have significantly different effects on the visibility of the letter?

Q21. The following were all Republican candidates in the electoral contests for various local offices in the city of Meltingpot, Ohio. **Elected:** Aaronson, Blomberg, Evans, Horsley, Jaspers, McTavish, O'Shaughnessy, Scorbini. **Defeated:** Neuhaus, Pickford, Rodsky, Toft, Verploot, Wilhelm, Young, Zotterman.

The ballot papers were organized alphabetically. Do these results show a relation between position on the ballot paper and electoral success?

Q22. A number of rats were assessed on the Nebraska Rodent Personality Scale, and the ten most introverted and the ten most extroverted were selected. They were trained to criterion on a discrimination task, and the number of trials required for extinction was then counted for each rat:

Introverted rats	23	18	107	16	35	40	28	21	46	21
Extroverted rats	62	17	33	25	38	19	44	29	80	36

Is there any connection shown between rate of extinction and the extroversion scale of the NRPS?

Q23. The crew of a radar station work four-hour shifts. The following are the numbers of guided missiles falsely reported by each operator in the first and last half-hours of her shift:

Operator	A	B	C	D	E	F	G	H	I	J	K	L
First half hour	6	3	0	2	4	3	8	5	0	1	7	2
Last half hour	5	8	3	4	2	7	12	9	2	0	5	8

Are operators significantly more prone to make false reports at either end of their shifts?

Q24. Sixteen subjects made settings of the same colour discrimination threshold on two successive days. The differences between the two settings made were as follows (in nanometres). Is there evidence of improvement (improvement = positive difference score)?

0.3 –0.6 1.2 2.3 –1.0 3.5 –2.0 1.1 0.8 1.4 2.7 –1.5 –2.6 2.4 3.1 1.9

Mann–Whitney test using a normal approximation

Q25. In a Mann–Whitney U test with $n_1 = 20$ and $n_2 = 60$ we find $U = 400$. What is (*a*) the one-tailed probability, and (*b*) the two-tailed probability of getting a value of U as extreme as this? (See instructions on page 128 in the *Tables and Formulae* section giving critical values of U.)

5. χ^2 test

Objectives

We'll cover the chi-square (χ^2) test for categorical data (goodness-of-fit test) and extend it to examine whether two categorical variables are related (contingency test). By the way, *chi* is pronounced *k*ai, not *ch*ai. Related supplementary material is presented for those who are interested.

Stuff with a solid edge, like this, is important.

But remember — you can totally ignore stuff with single / double wavy borders.

5.1. The chi-square (χ^2) test

One categorical variable, two categories

The χ^2 test, sometimes called Pearson's χ^2 test, is all about analysing **categorical data.** Suppose we ask 100 people to choose between chocolate and garibaldi biscuits (so every person falls into one of two categories); 65 choose chocolate and 35 choose garibaldi. Does this differ from chance, i.e. a 50:50 split? The **expected values** based on the null hypothesis are 50 chocolate and 50 garibaldi. The **observed values** are 65 and 35. From this, we can calculate the **χ^2 statistic:**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is the observed frequency in each category, and E is the expected frequency. We sum over all the categories. **A big χ^2 means that the observed frequencies differ considerably from the expected frequencies. (Significant values of χ^2 are big. Non-significant values of χ^2 are close to zero.)** If we have c categories, we have ***c* – 1 degrees of freedom.**

This is called a **goodness-of-fit** test. It asks whether the data (observed values, O) are a good fit to some model (expected values, E).

So in this example, $\chi^2 = \frac{(65-50)^2}{50} + \frac{(35-50)^2}{50} = 9$. We have two categories, but since we know n (100), then as soon as we know the frequency of one category (chocolate) we automatically know the frequency of the other (garibaldi). So we have only 1 *df*. All we need now is to know the **critical value** of χ^2_1 for our chosen value of α (say 0.05); our handy statistical tables (see p. 130) will tell us that this is 3.84. Since our χ^2 value was 9, we can reject the null hypothesis and say that people's preferences differed from chance ($\chi^2_1 = 9.0$, $p < .05$). If we were using a computer, we could derive an exact p value for our χ^2 value of 9 — it's 0.0027 — so we could report our biscuit analysis like this: 'The group's preference differed from chance ($\chi^2_1 = 9.0$, $p = .0027$).'

Note that although the process of testing χ^2 involved a one-tailed test (was χ^2 *bigger* than a critical value?), the process of *obtaining* the value of χ^2 was inherently two-tailed (the way we calculate χ^2 detects observed values that are bigger *or* smaller than the expected value). So the α we use to obtain a critical value of χ^2 is effectively a two-tailed α. For more details on this, see Howell (1997, p. 144).

One categorical variable, more than two categories

This approach can be used for any number of categories, and any expected values. So if a furniture warehouse stocks a vast number of chair backs, chair seats, and chair legs, then we could take random samples of items, classify each item in the

sample into one of these three categories ($c = 3$), and test the hypothesis that in the total stock (the population) these items were in the correct chair-building ratio 1:1:4 using a χ^2_2 test (note 2 degrees of freedom = $c - 1$).

More than one categorical variable (contingency tests)

We're often interested in data that's classified by more than one variable, and in asking whether these variables are *independent* of each other or in some way *contingent* upon each other. Here's an example (see Howell, 1997, p. 144), based on a 1983 study of jury decisions in rape cases. Decisions were classified on two variables: (1) guilty or not guilty; (2) whether the defence alleged that the victim was somehow partially at fault for the rape. The researcher analysed 358 cases:

Obtained values	Guilty verdict	Not guilty verdict	Total
Victim portrayed as low-fault	153 (*a*)	24 (*b*)	*177*
Victim portrayed as high-fault	105 (*c*)	76 (*d*)	*181*
Total	*258*	*100*	*358*

Now if these two variables (verdict and victim portrayal) are *independent*, then we would expect that $^a/_b = {}^c/_d$ and that $^a/_c = {}^b/_d$. But if they are not independent, we might expect a different picture. We can use a χ^2 test to answer this question. This is called a **contingency test,** because it asks whether one variable is in some way contingent upon the other. The null hypothesis is that the two variables are independent. We can calculate the expected value of each cell as follows:

$$E(row_i, column_j) = \frac{R_i C_j}{n}$$

where $E(row_i, column_j)$ is the expected value of the cell in row i and column j, R_i is the **row total** for row i, C_j is the **column total** for column j, and n is the overall total number of observations.

For our example, we can calculate that $E(1,1) = (177 \times 258) / 358 = 127.559$. We can fill in all the other expected values like this:

Expected values	Guilty verdict	Not guilty verdict	Total
Victim portrayed as low-fault	127.559	49.441	*177*
Victim portrayed as high-fault	130.441	50.559	*181*
Total	*258*	*100*	*358*

Then we can calculate χ^2 in the usual way:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

In general, if we have a table with R rows and C columns, we have **$(R - 1)(C - 1)$ degrees of freedom.** This method extends to any $R \times C$ table.

So in our example, there are four numbers to sum over (you should obtain the answer $\chi^2 = 35.93$), and we have (2–1)(2–1) = 1 *df*. This should make sense: once you know the row and column totals, you need to know only one cell frequency to be able to work out all the others. The critical value of χ^2_1 for $\alpha = 0.05$ is 3.84, so we reject the null hypothesis. When the victim was portrayed as low-fault, the defendant was found guilty 86% of the time, but when the victim was portrayed as high-fault, the defendant was convicted only 58% of the time, and this is a significant difference ($\chi^2_1 = 35.93$, $p < 0.001$).

Assumptions of the χ2 test

All statistical tests have assumptions. If they are violated, using the test is pointless: the results of the test will not be the probabilities we're interested in, and therefore our conclusions will be meaningless. This is what the χ^2 test assumes:

- **Independence of observations.** In all the examples given so far, each observation has been independent. One person didn't affect another's biscuit choice, and one court case didn't affect another. If this is not the case, you can't use a χ^2 test. In particular, one thing you mustn't do is to analyse data from several subjects when there are multiple observations from one subject, because they won't be independent. (It's possible to analyse data from *only* one subject, because the observations are then *equally* independent, but your conclusion will only tell you something about that one subject.)
- **Normality.** There shouldn't be any very small expected frequencies **(none less than 5),** otherwise the data won't approximate a normal distribution.
  [Actually, the "none <5" rule is a bit conservative; it's probably OK to use the test with even smaller expected frequencies if the row totals aren't too dissimilar and neither are the column totals (see Howell, 1997, p. 152) — but no expected value can be zero!]
- **Inclusion of non-occurrences.** To see what this means, let's take an example. Suppose that 17 out of 20 men supported the sale of alcohol in petrol stations, and 11 out of 20 women did. We want to know if significantly more men than women support this idea. This would be **wrong:**

Obtained values	Men	Women
Support booze	17	11

  This would give us expected values of 14 and 14 under the null hypothesis of 'no difference', and therefore χ^2_1 = 1.29 (not significant). But this is wrong because we've *lost information* about the total number of responders. We should be doing this:

Obtained values	Men	Women
Yes to booze	17	11
No	3	9

  This would give us χ^2_1 = 4.29 (p = 0.038). Including information on non-occurrences is vital — suppose we'd interviewed 2000 men and 17 said yes:

Obtained values	Men	Women
Yes to booze	17	11
No	1983	9

  We'd have a totally different picture, which the first table would have missed completely.

5.2. Supplementary material: odds ratios and relative risk

Although a χ^2 test may tell you that two variables are associated, it won't tell you by how much. One way of doing this is by using the **odds ratio.** Here's some 1998 data in which 20,000 male physicians were given daily aspirin or placebo for some time, and the incidence of heart attacks monitored.

	Heart attack	No heart attack	*Total*
Aspirin	104 (*a*)	10,933 (*b*)	*11,037*
Placebo	189 (*c*)	10,845 (*d*)	*11,034*
Total	*293*	*21,778*	*22,071*

The probability of someone in the aspirin group having a heart attack was $^a/_{a+b}$ = 0.94%. The probability of someone in the placebo group having a heart attack was $^c/_{c+d}$ = 1.7%. The **probability ratio** or **relative risk** is therefore $^a/_{a+b} \div {}^c/_{c+d}$ = 0.55 (or, taking the reciprocal of this, 1.82). The **odds** of someone in the aspirin group having a heart attack were $^a/_b$ = 0.0095 (see p. 18 for definition of odds). The odds of someone in the placebo group having a heart attack were $^c/_d$ = 0.0174. The **odds ratio** is $^a/_b \div {}^c/_d = {}^{ad}/_{bc}$ = 0.54 (and its reciprocal, $^{bc}/_{ad}$, is 1.83). So these men were about half as likely to have a heart attack if they were on aspirin.

Probability versus odds: be careful

Applying this technique to the rape jury data above might lead you to the conclusion that the jury were five times as likely to acquit if the defendant was portrayed as

being at fault. The probability of conviction in the low-fault condition was 0.86, equivalent to odds of 6.40. The probability of conviction in the high-fault condition was 0.58, equivalent to odds of 1.38. The **odds ratio** is therefore 4.6 (or 0.22 depending on which way round you view it). However, the probability ratio (**relative risk**) is only 1.49 (or 0.67) and the **absolute risk** increased by 0.86 – 0.58 = 0.28. Were the jury 4.6 times as likely to convict if the defendant was portrayed as being at fault, or 1.5 times? This depends on what you mean by 'as likely'! Remember that **probability = odds/(1+odds).** The odds on them acquitting were increased 4.6 times; the probability was increased 1.5 times.

To get a feeling for these counter-intuitive numbers, consider a couple of examples. Take a 100-kg sack of tomatoes that are 99% water. If you dried out the tomatoes completely, they'd have a mass of 1 kg. What would their mass be if you dried them out partially, until they were 98% water? The answer is 50 kg. So consider a group of patients that has a 99% chance of dying from the disease. If you give them a drug that reduces their probability of dying to 98% (so relative risk of dying: 0.98 / 0.99 = .9899), you have halved their odds of dying from 100:1 to 50:1 (odds ratio 0.5). But beware another property of relative risk: it matters which way round you view things. The patients' chance of *survival* has increased from 1% to 2% (relative risk of surviving: 0.02 / 0.01 = 2, which is nothing like the reciprocal of the relative risk of dying) but their odds of survival have increased from 1:100 to 1:50 (odds ratio 2, which is exactly the reciprocal of the odds ratio of dying).

Be careful not to be misled by papers that report odds ratios. If the overall event rate is low, odds ratios and relative risk are very similar; if high, they can be very different. The mathematical properties of odds ratios encourage their use (you can't double a probability of 0.6, for example), and they can be used in studies where you do not know the absolute probabilities (risks) of something happening (e.g. clinical case–control studies). However, they don't reflect our intuitive view of probability very well. Perhaps the clearest way to report these things is to give absolute probabilities, if you can, and then readers can work out all the other measures.

5.3. Supplementary material: the binomial distribution

Where does the χ^2 test come from? Read on if you're interested…

Imagine you have a coin that you flip a number of times. Each time, there are only two possible outcomes (heads or tails). If it's a fair coin, the probability of a head on each trial, call it p, is 0.5. Let's call the probability of a tail q; also 0.5. If you flip the coin five times, what is the probability that you get five heads? There's only one way to do this — HHHHH. So the probability is $0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 = (0.5)^5 = p^5 = 0.03125$. Similarly, the probability of zero heads, i.e. five tails (TTTTT), is $q^5 = 0.03125$ as well. But if you flip the coin five times, what's the probability that you get *three* heads? This is trickier, because there are several ways to do it. You might throw HHHTT, or HHTTH, or TTHHH… the probability of each pattern is $(0.5)^5$, but we'd like an easy way to work out the number of ways of getting three heads.

Permutations and combinations

We might as well make this general. If we have n lottery balls in a lottery ball machine, and we draw out r of them in a particular order, the number of ways we could draw them is called the number of **permutations,** written P_r^n. For example, if there are 50 balls in the National Lottery and we draw out 6 of them, then one permutation is {1,2,3,4,5,6}; another is {6,5,4,3,2,1}; another is {17;42;22;5;38;9}. Since we don't care about the order of the balls in the lottery, we can also talk about the number of **combinations** of drawing r balls out of n balls, or C_r^n — combinations are the same as permutations except that they don't care about the *order*, so {1,2,3,4,5,6} and {6,5,4,3,2,1} count as two separate permutations but are just two ways of writing the same combination. We can calculate P_r^n and C_r^n very simply once we know what **factorial** means: 6 factorial, written **6!,** is $6 \times 5 \times 4 \times 3 \times 2 \times 1$. So, written mathematically, here's what we need to know:

$x! = x(x-1)(x-2)\ldots 1$ Note that $0! = 1$ (a special case)

$$P_r^n = \frac{n!}{(n-r)!};\ P_r^n \text{ is sometimes written } {}_nP_r$$

$$C_r^n = \frac{n!}{r!(n-r)!};\ C_r^n \text{ is sometimes written } {}_nC_r \text{ or } \binom{n}{r}$$

We can use this to find out that there are $C_6^{50} = \frac{50!}{6!(50-6)!} = \frac{50!}{6!\times 44!} = 15{,}890{,}700$ possible outcomes in the National Lottery. But we can also use it to find out that there are $C_3^5 = \frac{5!}{3!\times 2!} = 10$ ways of flipping three heads in five coin flips.

The binomial distribution

Since we know that the probability of any particular sequence of five coin flips is $(0.5)^5 = 0.03125$, we now know that the probability of flipping three heads is $10 \times (0.5)^5 = 0.31$. In general, if we have ***n* independent trials,** each of which has **two outcomes,** one of which we'll call 'success' and one of which we'll call 'failure', where the probability of success is ***p*** and the probability of failure is ***q* = 1 – *p*,** and ***X*** is a discrete random variable representing the number of successes, then the probability of ***r* successes,** written $P(X = r)$, is given by the **binomial distribution:**

$$P(X = r) = C_r^n p^r q^{n-r}$$

We would call this distribution ***B*(*n*, *p*).** We can calculate the mean (the expected value) and the variance of $B(n, p)$:

$$E[B(n, p)] = np$$

$$Var[B(n, p)] = npq$$

In other words, the mean number of heads in five coin flips is $5 \times 0.5 = 2.5$, and the variance of this is $5 \times 0.5 \times 0.5 = 1.25$ (so the standard deviation is $\sqrt{1.25} = 1.12$).

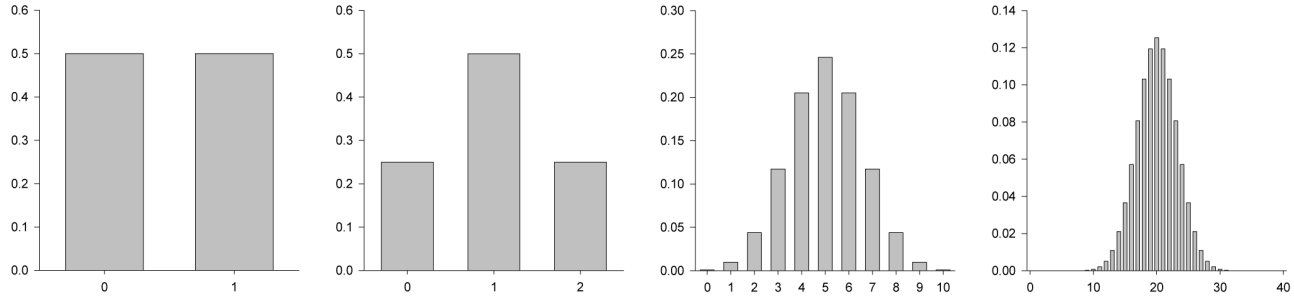
Using the binomial distribution as a statistical test

If a gambler inveigles us into a betting game, flips a coin 100 times and obtains 90 heads, is the coin fair? The null hypothesis is that the coin is fair ($p = q = 0.5$), and the observed number of heads was observed by chance. If the null hypothesis is true, then the number of heads in 100 flips should obey the binomial distribution $B(100, 0.5)$. The probability of obtaining 90 heads is therefore $P(X = 90) = C_{90}^{100} 0.5^{90} 0.5^{10}$. But we're actually interested in the probability of obtaining *90 or more* heads. We therefore want to know $P(X \geq 90) = P(X = 90) + P(X = 91) + P(X = 92) + \ldots + P(X = 100)$; a bit of calculation gives the answer $P(X \geq 90) = 1.53 \times 10^{-17}$. This is considerably less than our conventional α of 0.025 (we'd be using a two-tailed test here, since we'd want to detect a bias in either direction, so $\alpha = 0.025$ for each tail); we would therefore reject the null hypothesis and accuse the gambler of fraud. The clever fraudster would do better to use a very slightly biased coin: if he flipped 60 heads, then as $P(X \geq 60) = 0.028$, a two-tailed test with overall $\alpha = 0.05$ would not reject the null hypothesis of a fair coin. We'd need to observe the slightly biased coin for longer (more trials) to be able to detect its bias. This is a general principle of statistics: **more observations help you detect smaller effects.**

The normal distribution as an approximation to the binomial distribution

For large sample sizes (e.g. $np > 5$ and $nq > 5$), the binomial distribution $B(n, p)$ approximates the normal distribution $N(np, npq)$ — that is, a normal distribution with mean np and variance npq (see figure).

For very small p (or very large p), the binomial distribution does *not* approximate a normal distribution (it approximates a Poisson distribution instead; see www.mathworld.com/BinomialDistribution.html). Since the χ^2 test assumes normal distributions, this is why the χ^2 test is not valid with very small expected frequencies; see below.



Probability (y axis) of all possible total numbers of heads observed (x axis) when you flip a coin 1, 2, 10, or 40 times (from left to right). The binomial distribution approximates a normal distribution as n increases.

5.4. Supplementary material: the sign test

The sign test (sometimes called the Fisher sign test) evolves from the binomial test and is very simple indeed. Using an example borrowed from Howell (1997, p. 127), suppose we want to test whether people that know each other are more tolerant of individual differences. We might ask a dozen male first-year students to rate the physical attractiveness of a dozen other first-years (of the same sex) at the start and the end of the year. Suppose the median ratings (high = attractive) are as follows:

Target	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Start	12	21	10	8	14	18	25	7	16	13	20	15
End	15	22	16	14	17	16	24	8	19	14	28	18
Gain	3	1	6	6	3	–2	–1	1	3	1	8	3

The sign test looks at the sign (direction), but not the magnitude (size) of each difference. The null hypothesis is that there is no change in rating. Ignoring gains of 0 (which we don't have here anyway), the null hypothesis would therefore predict that by chance, about half the ratings would improve and about half would worsen, i.e. p(higher) = p(lower) = 0.5. In our hypothetical data set, we have 10 improvements out of 12 targets. We want to calculate $P(X \geq 10) = P(X = 10) + P(X = 11) + P(X = 12)$. Using the binomial distribution $B(12, 0.5)$, we know that $P(X = 10) = C_{10}^{12} 0.5^{10} 0.5^{2}$, and so on; the total $P(X \geq 10)$ is 0.0192. As this is less than our traditional $\alpha = 0.05$, we would reject the null hypothesis and say that there was a significant change in rating over the year.

The sign test using the normal approximation to the binomial distribution

For the null hypothesis, p(positive sign) = p(negative sign) = 0.5. So if the number of non-zero difference scores $n > 10$, and x is the number of difference scores of one sign (e.g. positive), we can use the normal approximation to the binomial distribution to get a quick answer. The mean of this distribution is $np = n/2$, and the variance is $npq = n/4$. So we can calculate a ***Z* score:**

$$z = \frac{x - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{2\left(x - \frac{n}{2}\right)}{\sqrt{n}}$$

and test that Z score in the usual way (see p. 15).

Comparing the sign test to the Wilcoxon matched-pairs signed-rank test

From our discussion of the Wilcoxon matched-pairs signed-rank test (p. 67), you'll see that the sign test is pretty similar in overall logic — except that the sign test throws away *even more* information about the distribution (it doesn't care about the magnitudes of the difference scores at all, just their signs). You pay a price in power,

but gain generality; the sign test is a nonparametric test that can be used with ordinal or even categorical data.

5.5. Supplementary material: the multinomial distribution

If we want to consider more than two alternatives for each trial, we need to use the **multinomial distribution.** Let there be n trials and k alternatives for each trial, numbered from 1 to k, each with the probabilities p_1, p_2, … p_k. Then the probability of obtaining exactly X_1 outcomes of event$_1$, X_2 outcomes of event$_2$, … and X_k outcomes of event$_k$ is given by

$$p(X_1, X_2, \ldots, X_k) = \frac{n!}{X_1! X_2! \cdots X_k!} p_1^{X_1} p_2^{X_2} \cdots p_k^{X_k}$$

An example: if we had a die with two black sides, three red sides, and one white side, then for each trial p(black) = $^2/_6$, p(red) = $^3/_6$, and p(white) = $^1/_6$. So if we roll the die 10 times, then the probability of obtaining exactly 4 blacks, 5 reds, and 1 white is

$$p(4,5,1) = \frac{10!}{4! \times 5! \times 1!} \left(\frac{2}{6}\right)^4 \left(\frac{3}{6}\right)^5 \left(\frac{1}{6}\right)^1 = 0.081$$

5.6. Supplementary material: the χ^2 distribution; an outline of deriving the χ^2 test; other points

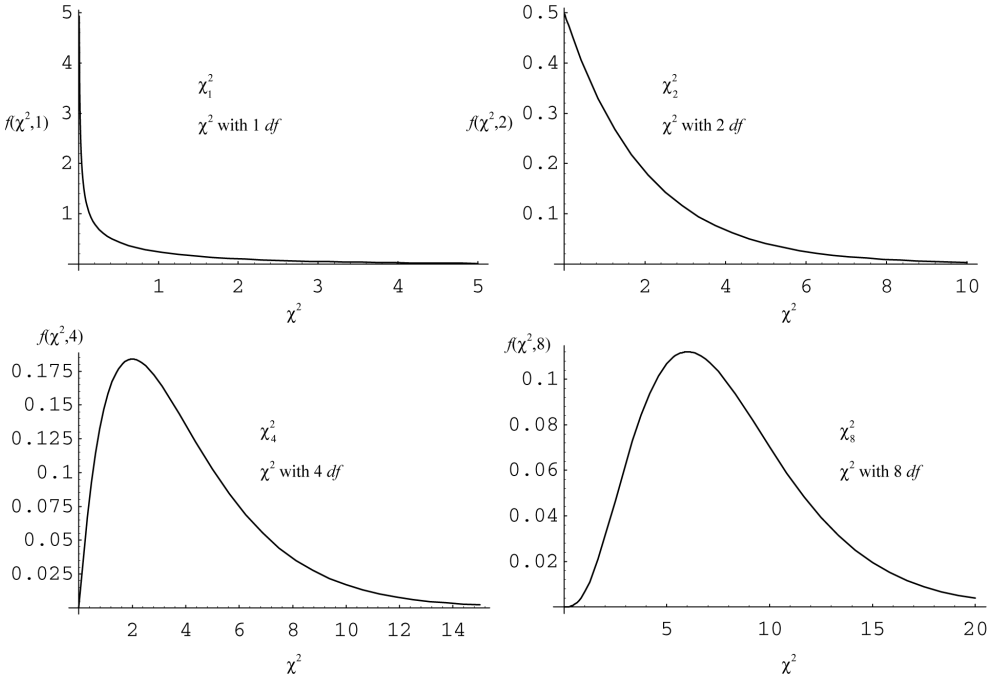
The χ^2 distribution

The χ^2 probability density functions are shown in the figure below; you can see that the shape of the distribution depends on the number of degrees of freedom, k. It is a *positively skewed* distribution, especially when k is small. The distribution is often written as χ^2_{df}, or sometimes $\chi^2(df)$. To obtain critical values of χ^2, we need to know the value of χ^2 above which (say) 5% of the area falls. In practice, we'll get this from tables (p. 130) or a computer.

Relationship between χ^2 and the normal distribution

If we have a normal random variable $N(\mu, \sigma^2)$, we can sample one value x from it, convert it to a standard normal variable z, and square it:

$$z^2 = \frac{(x-\mu)^2}{\sigma^2}$$



The χ^2 distribution, shown with 1, 2, 4, and 8 degrees of freedom. You can see that the distribution is positively skewed, but that as the number of degrees of freedom increases, it becomes more like a normal distribution.

If we repeated this *ad infinitum*, sampling independently each time, we would have a great number of values of z^2. We could therefore plot the distribution of z^2. We would find that this distribution is the same as χ_1^2 (χ^2 with 1 *df*):

$$\chi_1^2 = z^2$$

Now suppose that instead of sampling one number at a time, we sample *n* numbers at a time. For each observation within each sample we calculate z^2; for each sample, we calculate Σz^2. So each sample produces one value of Σz^2. Now we plot the distribution of these values of Σz^2. We find that the distribution is the same as χ_n^2:

$$\chi_n^2 = \sum_{i=1}^{n} z_i^2 = \sum \frac{(X_i - \mu)^2}{\sigma^2}$$

In other words, then if *Y* is the sum of squares of *n* independent standard normal variables, then *Y* is distributed as χ^2 with *n* degrees of freedom. (Since z_i^2 has a χ^2 distribution, this result also shows that the sum of a set of independent values of χ^2 itself has a χ^2 distribution, given the restrictions of independent sampling and an underlying population with a normal distribution.)

χ^2 tells us something about the distribution of sample variances

If we have a normal random variable $N(\mu, \sigma^2)$, we can draw an infinite number of samples from it. From each sample, we can calculate the sample variance s^2. We could then plot the distribution of these sample variances. We would find that it is related to the χ^2 distribution:

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2} \text{ and therefore } s^2 = \frac{\chi_{n-1}^2 \sigma^2}{n-1}$$

Since $\sigma^2/(n–1)$ is constant for a given σ^2 and sample size (*n*), the sampling distribution of the variance (the distribution of a set of sample variances) has a χ_{n-1}^2 distribution. Since the χ^2 distribution is skewed, this tells us that the distribution of s^2 is too — although the average value of a lot of s^2 measurements will equal σ^2, more than half the time s^2 will be less than σ^2.

Remember, it is because the distribution of s^2 is skewed (because it has a χ^2 distribution) that we use the *t* test rather than the *Z* test when we use s^2 as an estimator of σ^2 (p. 58). Of course, the bigger the sample, the more *df* the χ^2 distribution has, so the less skewed it becomes, and the more it and the resulting *t* distribution become like the normal distribution.

Deriving the χ^2 test from the binomial distribution (via the normal distribution)

Suppose we ask 100 people to choose between chocolate and garibaldi biscuits. Let's say that 65 choose chocolate and 35 choose garibaldi. Does this differ from chance, i.e. a 50:50 split? We could answer this with the binomial distribution, $B(100, 0.5)$, but there'd be a lot of adding up to find $P(X \geq 65)$. So let's do it a different way. **For large sample sizes** (e.g. $np > 5$ and $nq > 5$), **the binomial distribution $B(n, p)$ approximates the normal distribution $N(np, npq)$.** We've seen that

$$\chi_1^2 = z^2 = \frac{(x-\mu)^2}{\sigma^2}$$

where *x* is sampled from a normal distribution $N(\mu, \sigma^2)$. As we know the mean of a binomial distribution is *np* and the variance (σ^2) is *npq*, we can derive this approximation:

$$\chi_1^2 = z^2 = \frac{(x-np)^2}{npq}$$

To make things easier for later, we'll call the **observed frequencies** O_1 and O_2, and the **expected frequencies** E_1 and E_2. Specifically, $E_1 = np$ and $E_2 = nq$, and $O_1 + O_2 = E_1 + E_2 = n$. In our biscuit example, $O_1 = 65$, $O_2 = 35$, $E_1 = 50$, and $E_2 = 50$. Expanding and substituting these in to the previous formula, we would eventually get

$$\chi_1^2 = \frac{(x-np)^2}{np} + \frac{(n-x-nq)^2}{nq}$$

$$\chi_1^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

or, more generally,

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

which is the general formula for χ^2 that we've been using. This formula also extends to more than two categories, using the multinomial distribution.

5.7. Supplementary material: other points about χ^2

'Too good a fit' — the other tail of the χ^2 distribution

Very advanced stuff, this. But if you look at the χ^2 distribution (especially for higher *df*), you'll see that it's also possible to have an unusually *low* χ^2. Obviously, χ^2 can never be negative, but it can be lower than you'd expect by chance — in the far left tail of the distribution (whereas conventional χ^2 testing always asks whether χ^2 is greater — further right — than a critical value). What does it mean to say that χ^2 is unusually close to zero? That there is *too good a fit* to a proposed model. For example, if we flip a coin 10,000 times, then if the coin is unbiased it would be unlikely that we get 5100 heads and 4900 tails ($\chi_1^2 = 4$, $p < 0.05$). But if the coin is unbiased it is also quite unlikely that we get exactly 5000 heads and 5000 tails.

A famous example of this sort of criticism was Fisher's (1936) analysis of Gregor Mendel's (1866) experiments on plant breeding that established the field of genetics. Fisher argued that Mendel's data were 'too good a fit' to his model; for example, one study predicted a 3:1 ratio of plants with yellow and green seeds and obtained a ratio of 6022:2001 ($\chi_1^2 = 0.015$, p = 0.903). This is a pretty good fit, but nothing improbably good. However, Fisher looked at several such examples, and totalled up their χ^2 values. (This is perfectly valid mathematically; remember, we saw above that the sum of a set of independent χ^2 variables itself has a χ^2 distribution. The expected value [mean] of χ_1^2 is 1; the expected value of χ_n^2 is *n*.) Fisher obtained a cumulative $\chi_{84}^2 = 41.61$, p = 0.99997 — far *smaller* than the expected χ^2 value of 84. Now if the probability of obtaining data that deviate from Mendel's model by this much or more, given that the model is correct, is p = 0.99993, then the probability of obtaining data that deviate from the model by this much or *less*, given that the model is correct, is approximately $1 - p$ = 0.00003. Therefore, the data were actually unlikely (given that Mendel's model was true and that plants were randomly sampled) because they were too good a fit to that model. Fisher argued on this basis that Mendel or his assistant falsified or biased experiments so as to agree with Mendel's expectations — though perhaps only because he stopped when he was satisfied that his theory had been demonstrated (see Abelson, 1995, p. 96). There has been some vigorous and often fallacious debate on the validity of this approach (Pilgrim, 1984; Edwards, 1986; Pilgrim, 1986a; Pilgrim, 1986b) but Fisher's methods were sound (Edwards, 1986). But in any case, Mendel's experimental hypothesis was correct!

Note that there are other reasons that statistics can come out 'too small', such as when the assumptions of the test are not met (see Abelson, 1995, pp. 93-97). For example, it would be possible to obtain a 'too-good-to-be-true' situation in a χ^2 test if the observations were not independent.

What's Yates' correction to χ^2?

Something we won't use. In full, *Yates' correction for continuity* is a procedure applied for the analysis of 2 × 2 contingency tables (only) using the χ^2 test. It is intended to correct for the fact that the theoretical χ^2 distribution is continuous, but the χ^2 statistic can only take certain values in real situations (is discontinuous). To use the correction, subtract 0.5 from the absolute value of (O – E) before squaring it. That is, the correction applies the formula

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

For example, for a category where O = 6 and E = 4, or where O = 4 and E = 6, you would add $1.5^2 / 4$ to your χ^2. It is popularly applied when the expected values are small (typically, less than 5). However, Howell (1997, p. 146, q.v.) argues that there's almost never a good reason to use it — there are specific reasons when it's a good thing, but they are rare situations. We will ignore it.

Chi-square, or chi squared?

Note that χ^2 is usually written in full as 'chi-square', not 'chi squared', on the basis that it is a single statistical variable, not the square of some quantity χ (e.g. Howell, 1997 and see physics.ucsc.edu/~drip/133/ch4.pdf) — but other mathematicians disagree (see www.mathworld.com/Chi-SquaredDistribution.html and its links).

Relevant functions in Excel (see Excel help for full details)

BINOMDIST()	Gives you the binomial p.d.f., $P(X = x)$, or c.d.f., $P(X \leq x)$, where X has a binomial distribution.
CHIDIST()	From χ^2 and the *df*, gives you the probability that $P(X > \chi^2)$, where X has a χ^2 distribution.
CHIINV()	From p and the *df*, gives you the critical value of χ^2 such that $P(X > \chi^2) = p$.
CHITEST()	Does a χ^2 test for you, working out the *df* automatically and returning the p value. But it's often difficult to be sure that what it's doing is what you hope it's doing, so it's much safer to compute χ^2 step by step and then use CHIDIST() to get the p value.

5.8. Examples 5: χ^2

Q1. A coin is tossed 100 times. It comes down heads 40 times, tails 60 times. Is it biased?

Q2. 100 male rats and 100 female rats are tested on an up / down jumping stand. 40 of the males jump up, and 60 down. Only 16 female jump up, and 84 down. Is this good evidence that females jump down more than males?

Q3. In January 2007 there will be 7,000 road accidents in Cambridge. 1,643 of them will happen on Sundays. Will this be good evidence that accidents tend to happen on Sundays more than on other days?

Q4. A die is rolled 342 times. The various sides appear with the following frequencies. Is it biased?

1	2	3	4	5	6
53	48	75	49	60	57

Q5. Samples from three strains of giraffe, with 80 giraffes in each sample, were divided into four groups on the basis of ophthalmoscopic measurements of their refractive errors. The numbers in each group are shown below. Is there any evidence that the strains differ in the relative frequencies of different refractive errors (including Normal)?

	Short-sighted	Long-sighted	Astigmatic	Normal
Strain A	5	7	23	45
Strain B	13	17	30	20
Strain C	10	5	15	50

6. Past exam questions

About the NST 1B Psychology exam

'Section B contains a single question. In Paper 1, this question requires a statistical analysis of a set of data; in Paper 2, the design of an experiment or series of experiments is required… Section B [is] assigned 25% of the total maximum marks for a paper.'

Q1 (2000, Paper 1).

In a treatment trial for depression all patients received treatment with imipramine, an antidepressant drug. In addition, half received cognitive therapy (Cogth) while half received counselling (Couns). Patients were assessed on the Beck Depression Inventory prior to (Pre) and following (Post) the treatment programme. The results are shown overleaf.

a) Which treatment is more effective?
b) Are there any differences between the levels of depression in men and women prior to treatment?
c) Is there a relationship between depression before and after treatment in the cognitive therapy group?

Treatment	Gender	Pre	Post
Cogth	F	20	10
Cogth	F	18	11
Cogth	F	17	6
Cogth	F	19	10
Cogth	F	21	8
Cogth	M	42	20
Cogth	M	35	17
Cogth	M	32	18
Cogth	F	15	3
Cogth	M	28	18
Cogth	F	22	11
Cogth	F	21	7
Cogth	M	26	17
Cogth	F	27	14
Cogth	F	19	8
Couns	F	19	14
Couns	F	17	13
Couns	F	18	19
Couns	F	20	14
Couns	F	23	19
Couns	M	38	30
Couns	M	33	29
Couns	M	34	27
Couns	F	13	10
Couns	M	29	20
Couns	M	23	16
Couns	F	24	16
Couns	F	28	25
Couns	F	24	16
Couns	F	17	13

Q2 (2000, Paper 2).

Answer **one** of the following three questions on Experimental Design.

(1) Design a research project to investigate the relative importance of different depth cues in determining size constancy.

(2) It has been shown that babies of 8 to 9 months can show "conditioned joint attention". The experimenter and baby face one another and then the experimenter turns his / her head to look at a movable toy either to the left or the right of the baby. If the baby turns and looks in the same direction, the toy is activated, thus acting as a reinforcer. Design an experiment to assess the kind of learning that underpins this behaviour.

(3) A pharmaceutical company has developed a new drug for alleviating anxiety. However, they are concerned that this drug may also have effects on working memory, and therefore they wish to assess the effects of the drug on this form of memory using an animal model before conducting trials with human volunteers. Design an experiment to assess this question using rats as subjects.

Q3 (2001, Paper 1).

In an incidental memory experiment, 10 subjects were presented with a series of preference judgement trials. On each trial, the subjects were asked to rate a picture for attractiveness. In two subsequent tasks, the subjects were tested first for their recall, and then for their recognition, of the fifty pictures used in the preference task. The number of pictures correctly recalled and recognised by each subject is given below:

Subject	Recalled	Recognised
1	19	12
2	27	38
3	24	34
4	40	47
5	29	39
6	50	50
7	17	17
8	25	43
9	30	31
10	38	35

Determine whether recognition performance is better than recall performance using an appropriate statistical test.

Construct a scatter plot of the number of pictures recognised against the number recalled. Did those subjects who recalled more pictures also recognise more of them?

Plot on your graph the line that best predicts recognition performance from recall performance.

Q4 (2001, Paper 2).

Answer **one** of the following three questions on Experimental Design.

(1) Subjects are able to adapt to the displacement of the visual field produced by a prism worn over one eye. Design a series of experiments to demonstrate this effect and to investigate what subjects learn during adaptation.

(2) Some words in English also happen to be words in another language. For example, the word "*gift*" means *poison* in German, whereas the word "*four*" means *oven* in French. Design an experiment to investigate:

(a) whether bilingual speakers automatically access meaning in both languages when they read a word like "*gift*", and

(b) whether it makes a difference if the words in the two languages have the same phonology as well as the same spelling.

(Note: You do *not* need to display any knowledge of a second language or use real words as examples.)

(3) Design an experiment to determine whether a therapy that is aimed at reducing expressed emotion in the families of individuals with schizophrenia has an effect on relapse rate.

Q5 (2002, Paper 1).

The results below were obtained in a recent practical class on mental rotation. The subject was asked to indicate as quickly as possible whether a letter was presented in its normal form or as a mirror-image. The letter was presented in different orientations on different trials. The first row of the table shows the number of degrees by which the letter was rotated from the upright position. The second row shows the corresponding mean reaction time for those trials in which the target was presented in its normal form. The final row shows the average error rate for each orientation.

Rotation (deg)	0	45	90	135	180	225	270	315
RT (msecs)	518	563	638	781	896	738	625	552
Error rate (%)	0.87	0.84	1.47	2.44	4.20	2.29	1.33	0.53

A standard theory holds that the subject performs a 'mental rotation' of the target before judging whether it is in its normal form. The transformation is thought to be carried out over the most direct route. On the assumption that this theory is correct, estimate the rate of 'mental rotation' from the data. What is your best estimate of the time occupied by the remaining components of the reaction time?

Is there a significant relationship between reaction time and error rate?

Q6 (2002, Paper 2).

Answer **one** of the following three questions on Experimental Design. Say what data you would collect and what statistical procedure(s) you would use.

a) Young infants from about 12 months have been observed to look at their caregiver before they approach or reach out for an unfamiliar object. If their caregiver looks anxious, the infant will withdraw from the unfamiliar object, but they will approach if their caregiver looks happy. This phenomenon is called 'social referencing'. Some have argued it reflects the infant's new capacity to understand that their caregiver has mental states, which the infant infers from the adult's facial expression. Design an experiment to assess this claim against alternative explanations for 'social referencing'.

b) Design an experiment to determine whether the left or the right ear is more accurate in recognising words presented, one to each ear, simultaneously. What particular problems does such an experiment encounter?

c) Suppose you are interested in determining the mechanisms that are responsible for retroactive and proactive interference. Design an experiment that might help elucidate these mechanisms.

Q7 (2003, Paper 1).

In an initial experiment to measure the reaction times for discriminating 'positive affect' faces (expressing 'happiness') from 'negative affect' faces (expressing 'sadness') the following ten reaction times from ten subjects were recorded in milliseconds (msec):

630 580 604 596 720 549 613 660 578 618

Within what interval is there a 95% probability that the true population mean lies (assuming that the 10 observations have been sampled randomly from a normally distributed population)?

In a subsequent experiment, 12 subjects were randomly assigned to two groups. One group was given a caffeine tablet (condition A) while the other group was given a placebo — a 'sugar pill' with no physiological effect (condition B). Reaction times were then taken for subjects in both groups on the 'positive affect' versus 'negative affect' face discrimination test. These are given below.

	Reaction time score (msec)					
Condition A:	643	497	567	521	596	507
Condition B:	586	601	547	630	654	593

Is there a significant difference between the two groups?

Q8 (2003, Paper 2).

Answer **one** of the following three questions on Experimental Design. State what data you would collect and what statistical procedure(s) you would use, and give the reasons for your choices.

(1) Design an experiment to demonstrate blocking in humans. Indicate how you would investigate the processes responsible for the effect in further experiments.

(2) It has been claimed that aspirin adversely affects the operation of the active mechanism in the cochlea. Design an experiment to test this claim.

(3) Design an experiment to investigate whether people who are blind perform a mental imagery task. Be sure to choose a task that requires visual mental imagery and cannot be performed using semantic knowledge.

Q9 (2004, Paper 1).

An experimenter is interested in schizotypy, a personality type thought to predispose to the development of schizophrenia. One aspect of schizotypy is 'magical ideation', the tendency of a person to believe in magical phenomena. Schizophrenia has been suggested to involve abnormal development of the left / right asymmetry of the cerebral hemispheres, and abnormalities of the dopaminergic neurotransmitter system. High dopamine levels in one side of the brain are known to make animals turn to the opposite side, and consequently, the experimenter wonders if people with magical ideation might have higher dopamine levels in the right-hand side of their brain, which might result in a tendency to turn left more often than would be expected by chance.

She selects a group of 16 subjects at random and measures their Magical Ideation scores using standard techniques. She then equips her subjects with devices to measure their turning behaviour. She gives each subject one of either dopamine agonist (increases levels of dopamine [actually, mimics dopamine!]) or placebo (in a

randomised double-blind design), records the number of left and right whole-body turns they make in their everyday activities over a certain time period, and then calculates the percentage of left turns made. Her data are shown below.

Subject	Magical Ideation score	Dopamine agonist or placebo	Percentage of turns made to the left
1	8.1	Placebo	44.7
2	5.2	Placebo	41.6
3	9.7	Placebo	49.6
4	12.8	Placebo	55.3
5	12.6	Placebo	50.9
6	14.2	Placebo	60.8
7	2.4	Placebo	41.7
8	8.3	Placebo	51.4
9	12.3	Dopamine agonist	61.9
10	5.7	Dopamine agonist	57.2
11	6.9	Dopamine agonist	63.2
12	3.9	Dopamine agonist	53.1
13	3.5	Dopamine agonist	52.3
14	6.1	Dopamine agonist	47.3
15	6.7	Dopamine agonist	54.2
16	2.6	Dopamine agonist	48.3

Assume that the percentage of turns made to the left by humans is a normally-distributed variable.

a) Considering only those subjects who received the placebo, sketch a scatterplot showing the relationship between magical ideation scores and percentage of left turns. Is there a significant linear relationship between magical ideation and left-turning behaviour in these subjects? What proportion of the variability in left-turning is predictable from the magical ideation scores in these subjects?

b) Did the dopamine agonist affect turning behaviour when compared to placebo?

Q10 (2004, Paper 2).

Answer **one** of the following three questions on Experimental Design. State what data you would collect and what statistical procedure(s) you would use, and give the reasons for your choices.

a) Design an experiment to demonstrate that unseen visual stimuli can affect behaviour.

b) It has been claimed that moderate intake of alcohol enhances subjects' ability to solve anagrams. Design an experiment to test this claim.

c) Design an experiment to investigate whether young children are able to rehearse items in memory.

7. Further mixed examples

Q1. The popliteal height (distance from underside of thigh to sole of foot when seated) of adult males is normally distributed with a mean of 17.0" and a standard deviation of 0.8". What percentage of men will be unable to rest their feet on the floor when sitting on a chair whose seat is 15" from the ground? Assume they wear no shoes or socks.

Q2. An experimenter has reason to believe that subjects' reports of the difference between the lengths of the lines in a Müller–Lyer illusion will be affected by the presence of a stooge who gives a pre-arranged false report. 30 subjects were each presented with the illusion and had to judge, orally, the apparent difference in length of the two lines, in inches. The experimental group were each accompanied by a stooge partner, whereas the control group worked on their own. Each subject's judgement is given below. Is there a significant difference (at the 5% level) between the two groups?

Experimental	–0.41	0.95	0.82	0.44	–0.64	0.76	–0.12	0.41	0.34	0.43	0.08
	–0.57	–0.06	–0.05	0.09							
Control	0.22	–0.17	–0.11	0.19	0.32	0.97	–0.17	–0.79	–0.11	0.05	–0.49
	0.23	0.22	–0.45	–0.28							

Q3. Suppose that on a final exam in statistics the mean score was 50 and the standard deviation (σ) was 10. Find the following:

(a) The standardized (Z) scores of students receiving the following grades: 50, 25, 0, 100, 64
(b) The raw grades corresponding to Z scores of –2, 2, 1.95, –2.58, 1.65, –0.33.

The instructor had reason to believe that the scores were normally distributed. Our of 200 students, how many should he have expected to achieve scores:

(c) within $1 \times \sigma$ of the mean?
(d) 3σ or more above the mean?
(e) between -1.96σ and -0.5σ away from the mean?

Q4. In an experiment to see whether position preferences are heritable in mice, two strains of mice were bred, one selected for left-turning behaviour and the other for right-turning. After ten generations, members of each strain and selected controls are observed in a maze. Their first turns are as follows. Does the experiment demonstrated inherited position preference?

	Right	Left	Total
Bred for R turn	17	9	26
Bred for L turn	13	15	28
Unselected	18	12	30

Q5. After the TV appeal on behalf of the Ski Slopes for the Disabled Fund, cheques were received for the following amounts (in £), and in the order given. Is there a relation between the size of the gift and the promptness with which it was sent off?

1000	120	5	15	10	6.30	10	25	2.50	2	4	0.12	1	1	8

Q6. Suppose that the lecturer in your History of Babylonia course informs you that on the final examination two students received grades of 60 and 30, and that the standardized scores corresponding to those grades were 1 and –1 respectively.

(a) What are the mean and standard deviation of the scores?
(b) If the scores are normally distributed, what proportion of them should lie between 25 and 65?
(c) On the same assumption, between what two scores should the middle 50% of cases lie?

Q7. In a study of reaction time (RT) to an auditory as opposed to a visual stimulus, 20 pairs of Basic Airmen were selected at random. Each pair was matched in physical stature, age, intelligence, and so forth. The members of the pairs were then assigned at random to two experimental conditions. In one, the man was to touch a button as soon as possible after the appearance of a visual stimulus. The other condition was the same except that the stimulus was a buzzer. The average RTs are shown below, in ms. Is there a significant (at 5%) difference in RT to the different types of stimuli?

Pair	1	2	3	4	5	6	7	8	9	10
Auditory RT	130	200	150	140	230	160	180	150	200	170
Visual RT	160	130	120	150	120	160	110	210	170	140
Pair	11	12	13	14	15	16	17	18	19	20
Auditory RT	220	140	220	230	180	190	180	210	230	220
Visual RT	170	160	110	290	100	190	210	180	170	160

Q8. A subject has to set the two sides of a rectangle to be visually equal on a number of trials. The table gives the time taken for each adjustment (in sec) and the error (in mm). Is there any relation between the time taken in adjustment and size of error?

Trial	1	2	3	4	5	6	7	8	9	10	11	12
Time	17	8	11	24	9	15	20	12	35	9	14	17
Error	4.4	5.7	4.0	4.2	3.6	1.9	2.9	5.3	4.1	6.3	0.8	2.0

Q9. The following are the wavelenghts (in nm) for maximum visual sensitivity of fiteen colour-normal observers and eight deuteranopes (colour-blind individuals lacking the 'green' pigment). Do these data show any difference between the two groups?

Colour-normals	560	558	563	561	552	557	562	560	569	559
	564	554	559	560	556					
Deuteranopes	561	567	559	570	564	555	566	561		

Q10. IQ is defined to be normally distributed with mean 100 and standard deviation 15. What IQ is high enough for only 100,000 people in England and Wales (take population to be 50,000,000) to exceed it? How many people lie between 95 and 105? Between 60 and 70?

Q11. A psychologist was interested in the verb–adjective ratio as an index of individual speech habits. 10 science and 12 arts students were chosen at random, and a sample of free speech of each subject obtained. Each sample was scored according to the number of verbs used divided by the number of adjectives. The data are shown below. Do science students have a significantly higher verb–adjective ratio?

Sci	1.32	2.30	1.98	0.59	1.02	1.88	0.92	1.39	1.95	1.25		
Arts	1.04	0.93	0.75	0.33	1.62	0.76	0.97	1.21	0.80	1.16	0.71	0.96

Q12. Four large US midwestern universities were compared with respect to the fields in which graduate degrees were given. The graduation rolls for last year from each university were taken and the results put into the contingency table below. Is there a significant association (at the 5% level) between the university and the fields in which it awards degrees? What are we assuming when we carry out this test?

University	Law	Medicine	Sciences	Humanities	Other
A	29	43	81	87	73
B	31	59	128	100	87
C	35	51	167	112	252
D	30	49	152	98	215

Q13. In the Tripos examination for Part II Neurobotany, male (M) and female (F) candidates are placed in the order shown below (highest marks first). Do either men or women make better neurobotanists, insofar as this elusive quality is measured by the examination?

M M F M M M F F M F M F F M M M M M F M

Q14. A number of 30-second samples were taken from each of three TV channels, and classified according to whether the subject matter was primarily sex, violence, or general interest. The table below shows how many were in each category. Do these data show a significant difference in the contents of the three channels?

	Sex	Violence	General interest	Total
BBC1	26	17	57	100
BBC2	17	5	38	60
ITV	19	33	48	100

Q15. On each trial of a discrimination experiment, a monkey was watched for neck-scratching and tooth-baring behaviour. These occurred on the numbers of trials shown below. Do these data show any connection between neck-scratching and tooth-baring?

Neck-scratching alone	46
Tooth-baring alone	22
Neck-scratching and tooth-baring	5
Neither	53

Q16. In a study of demographic trends, 26 newly-married couples were asked, one individual at a time, how many children they would like to have. Responses are listed below. What can be concluded from these data? (Note that at least four different tests can be applied. To what question is each appropriate?)

Couple	1	2	3	4	5	6	7	8	9	10	11	12	13
Husband	3	0	1	2	0	0	1	2	2	1	8	0	3
Wife	2	1	0	2	3	3	2	3	3	2	4	1	4

Couple	14	15	16	17	18	19	20	21	22	23	24	25	26
Husband	5	7	1	0	4	10	5	2	0	1	3	5	2
Wife	2	2	2	3	4	3	3	4	2	3	2	2	1

Q17. Ten subjects were tested on a high-speed sorting task while subjected to 100 dB SPL white noise through earphones. On session 1 and 4 the test was preceded by a period of 115 dB SPL noise and on sessions 2 and 3 by 85 dB SPL noise. Their error scores are shown below. Does the intensity of the preceding noise have a significant effect on task performance?

Subject	1	2	3	4	5	6	7	8	9	10
Errors in 2 & 3	37	29	60	44	21	47	46	38	28	66
Errors in 1 & 4	24	24	31	30	26	42	33	19	32	45

Q18. The following are percentages of a group of 15 subjects on (a) a motor-tracking test and (b) the Body Image Awareness scale. Is there any relationship between the two scores?

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Tracking	74	31	80	66	41	53	77	39	46	19	55	62	38	49	59
BIAS	43	60	51	53	57	70	39	85	68	73	54	48	59	39	87

Q19. A die is thrown 300 times, each face appearing uppermost on the number of times shown below. Would you accuse the owner of having a loaded die?

Face	1	2	3	4	5	6
No. of throws	42	37	58	44	61	58

Q20. Fifteen sea slugs are placed on a line perpendicular to a gradient of light intensity. After two minutes their distances from the line are as shown below (cm; +ve distances are towards the light). Do the slugs show a significant tendency to be phototropic?

+13	+21	+5	–1	+3	–7	+12	+9	+24	–19	+21	+11	+48	–8	+15

Q21. In an experiment on the effects of context on letter perception in reading, a subject has to cross out every letter ‘e’ she spots in a page of the *Times*, working as fast as possible. For *e*s in contexts where they have the sounds as in ‘evil’, ‘belt’, and unstressed ‘the’, she misses 17 out of 243, 64 out of 409, and 99 out of 688, respectively. For silent *e*s she misses 108 out of 595. Are there significant differences in the proportions missed in the four different pronunciations?

Q22. In a sample of 1000 people in a mass eye-testing program, the numbers needing various powers of corrective lenses are as follows (power in dioptres, for right eye only). Do these powers deviate significantly from a normal distribution? ***This question is well beyond the standard required for the exam!***

–5D	–4D	–3D	–2D	–1D	0	+1D	+2D	+3D	+4D	+5D
11	29	41	67	106	532	133	40	22	14	5

8. Answers to examples

Answers calculated by RNC — *caveat emptor*. Thanks to MRFA for checking them.

In some of these examples I'll quote exact *p* values, rather than just saying '$p < 0.05$'. Don't worry about this — since you're operating from tables and I'm doing some of these questions on a computer to save time, I can quote exact *p* values when you can't. If I say '$p = .03$', your tables would show that $p < .05$, but not that $p < .01$. If I say '$p = .125$', your tables would show that the answer is not significant at $p = .1$ (i.e. $p > .1$)… and so on.

8.1. Answers to Examples 1: background and normal distribution

Q1 RV (a) 0.933; (b) 0.015; (c) 0.015; (d) 0.136; (e) 0.046

Explanations…
If X is a normally-distributed random variable with mean 23.5 and SD 3.0, then we can invent a variable Z that has a mean of 0 and a SD of 1 — a 'standard' normal variable' by calculating Z = (X – mean)/SD = (X – 23.5)/3.0.

So when we want to ask 'what's the probability that X is less than 28', we can ask instead 'what's the probability that Z is less than (28 – 23.5)/3.0 = 1.5'. We can look up the probability of Z being less than 1.5 from tables of Z (see p. 123); it's 0.933. So this is also the probability that X < 28.

If you want to find the probability that X > 30, that's equivalent to asking 'what's the probability that Z > (30 – 23.5)/3.0 = 2.167'. From tables, the probability that Z is ***less*** *than 2.167 is 0.985, so the probability that Z is* ***bigger*** *than 2.167 (or X is bigger than 30) is 1 – 0.985 = 0.015.*

This logic applies to all these examples. When you want to find the probability that 26 < X < 28, find the probability that X < 28, and take away from it the probability that X < 26.

Q2 IQ (Since you're multiplying probabilities by a large number — 60,000,000 — you will notice differences between the answers you'd get from your tables and those you'd get with a computer. I'd expect you to use the tables — you'll have to in the exam — but have quoted both answers here.)

(a) 78,000. The probability $P(\text{IQ} > 145)$ is the same as the probability $P(Z > 3)$, which is 0.0013 from your tables. So this corresponds to $0.0013 \times 60{,}000{,}000 = 78{,}000$ people. (If you calculate this more precisely with a computer, you get a probability of 0.001349967… and the answer 80,988.)
(b) $P(Z < -1.33) = 0.0918$, so the answer's 5,508,000 (or, with a computer, 5,472,677).
(c) $P(-1 < Z < 1) = 0.6826$, so the answer's 40,956,000 (or, with a computer, 40,961,369).

Q3 meerk (a) SD = 2 cm; (b) $P(0 < Z < 0.5) = 0.192$; (c) 227; (d) 26.08 to 33.92 cm; (e) 0.067; (f) 0.933; (g) zero.

To explain (g) a little… the probability of finding a meerkat whose height is the same as a particular value depends on what we mean by 'the same as'! As we become more and more restrictive (the meerkat has to be within a centimetre… millimetre… micron… of the specified height) the probability of finding such a meerkat becomes smaller and smaller. As the range of acceptable heights shrinks to zero, so does the probability, so the probability of finding a meerkat of 'exactly' a given height is zero.

Q4 RCBF (a) 0.0082; (b) 0.0164; (c) 38 ml / min; (d) 0.809.

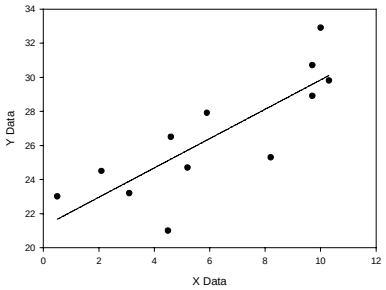
Explanation of (d): a Z score of ±2.4 is equivalent to p = 0.0164, so P(make at least one Type I error at p = 0.0164) = 1 – P(never make any Type I errors at p = 0.0164 in 100 comparisons) = 1 – (1–0.0164)^{100} = 1 – 0.191 = 0.809.

Caveat: the method for finding (d) doesn't take into account the fact that nearby areas are likely to have related blood flows — but then this is a statistics example, not an functional imaging tutorial.

Q5 poem Depends on the food, accommodation and risk aversion; both $P(\text{lose money}) = 0.18$. (From Frank & Althoen, 1994.)

8.2. Answers to Examples 2: correlation and regression

Q1 *visual decay*



sample covariance = 9.789
$s_X = 3.373$
$s_Y = 3.558$

$r = .816$, $p = .001$ two-tailed, $n = 12$
regression $Y = 21.24 + 0.86\,X$

Full working for Q1:

Call blink rate X and decay time Y. Plot your scatterplot as above. There's no obvious non-linear relationship, so doing a linear correlation makes sense. Data points, written as {*x*,*y*} pairs, are {2.1, 24.5}, {10.3, 29.8}, {5.9, 27.9}, {10, 32.9}, {0.5, 23}, {4.5, 21}, {3.1, 23.2}, {8.2, 25.3}, {5.2, 24.7}, {9.7, 30.7}, {4.6, 26.5}, {9.7, 28.9}. **You should be able to enter these into your calculator and get *r* directly.** If you were to do it by hand, you'd calculate these:

$$\sum xy = (2.1\times 24.5) + (10.3\times 29.8) + \ldots + (9.7\times 28.9) = 2065.84$$
$$\sum x = 2.1 + 10.3 + \ldots + 9.7 = 73.8$$
$$\sum x^2 = 2.1^2 + 10.3^2 + \ldots + 9.7^2 = 579.04$$
$$\sum y = 24.5 + 29.8 + \ldots + 28.9 = 318.4$$
$$\sum y^2 = 24.5^2 + 29.8^2 + \ldots + 28.9^2 = 8587.48$$
$$n = 12$$

OK… now for the sample covariance and sample standard deviations. Using the formula sheet:

$$\mathrm{cov}_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{2065.84 - \frac{73.8\times 318.4}{12}}{11} = 9.789$$

$$s_X = \sqrt{s_X^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{579.04 - \frac{(73.8)^2}{12}}{11}} = \sqrt{11.379} = 3.373$$

$$s_Y = \sqrt{s_Y^2} = \sqrt{\frac{\sum (y-\bar{y})^2}{n-1}} = \sqrt{\frac{\sum y^2 - \frac{(\sum y)^2}{n}}{n-1}} = \sqrt{\frac{8587.48 - \frac{(318.4)^2}{12}}{11}} = \sqrt{12.661} = 3.558$$

Now we can calculate r (and r^2):

$$r_{XY} = \frac{\mathrm{cov}_{XY}}{s_X s_Y} = \frac{9.789}{3.373\times 3.558} = 0.816$$
$$r^2 = 0.666$$

… and a t statistic:

$$t_{n-2} = t_{10} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.816\sqrt{10}}{\sqrt{1-0.816^2}} = 4.464$$

With 10 *df*, $t = 4.464$ is significant at the $\alpha = 0.01$ two-tailed level (i.e. $p < .01$ two-tailed). (A computer would tell you that $p = .001$.) Next, to calculate the **regression** of Y on X (predicting Y from X), we aim to calculate the equation

$$\hat{Y} = a + bX$$

Your calculator should be able to give you *a* and *b* directly (and you've already entered the data to calculate *r*, so you should be able to retrieve *a* and *b* very quickly). But if you had to calculate them by hand, you'd do it like this… First, we need the means of x and y:

$$\bar{x} = \frac{\sum x}{n} = \frac{2.1 + 10.3 + \ldots + 9.7}{12} = 6.15$$
$$\bar{y} = \frac{\sum y}{n} = \frac{29.8 + 27.9 + \ldots + 28.9}{12} = 26.533$$

Now we have all the information to calculate *a* and *b*:

$$b = \frac{\text{cov}_{XY}}{s_X^2} = r\frac{s_Y}{s_X} = 0.816 \times \frac{3.558}{3.373} = 0.86$$

$$a = \bar{y} - b\bar{x} = 26.533 - 0.86 \times 6.15 = 21.24$$

So our regression equation, which you can add to your scatterplot, is

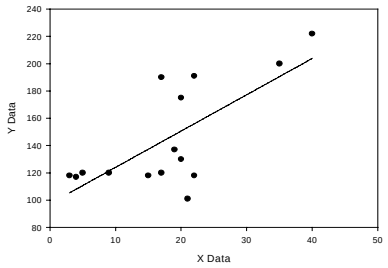
$$Y = a + bX = 21.24 + 0.86X$$

You can plot it by taking two or more *x* values that are reasonably far apart and calculating predicted values of *Y*, giving you $\{x,\ \hat{y}\}$ pairs. You should also find that the line passes through {0, *a*}, and $\{\bar{x},\ \bar{y}\}$, i.e. through {0, 21.24} and {6.15, 26.533}.

(This calculation — not hard, but time-consuming — should emphasize the importance of having a calculator that does the hard work for you in the exam!)

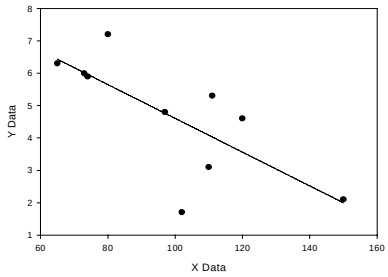
Q2 *Necker*

r = .711, *p* = .003 two-tailed, *n* = 15
regression Y = 97.431 + 2.66 X



Q3 *frog RGC*

r = –.738, *p* = .015 two-tailed, *n* = 10
regression Y = 9.825 – 0.0522 X

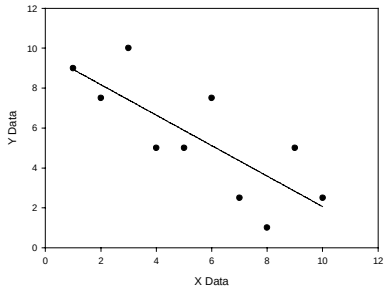


Q4 *Vatican / ice cream*

Correlate location rank with price rank (calling the result Spearman's correlation coefficient r_s):

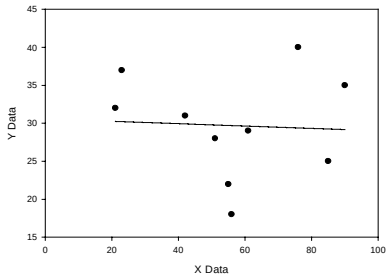
Location rank: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Price rank: 9, 7.5, 10, 5, 5, 7.5, 2.5, 1, 5, 2.5

This gives you r_s = –.778. Since *n* = 10, *p* = .008 two-tailed (as calculated by the program SPSS). However, different calculation techniques will give slightly different answers for *p*; for r_s = –.778 and *n* = 10 your tables (see p. 124) will show you that .01 < *p* < .02, two-tailed.



Q5 *impulsivity, CSF 5HIAA*

r = –.054, *p* = .883 two-tailed, *n* = 10
(regression Y = 30.57 – .0155 X — you'll often see published figures in which 'non-significant' regression lines are plotted, mainly so you can see the line is flat and useless as a predictor.)



8.3. Answers to Examples 3: parametric difference tests

Q1 glare

This question is asking for the 90% confidence intervals. From the formula sheet, we see that we don't know the population mean and SD, but we can work out the sample mean and SD, so we use this formula:

$$\text{Confidence intervals} = \bar{x} \pm \frac{s_X}{\sqrt{n}} t_{critical(n-1)df}$$

(For 95% confidence intervals, use t for $\alpha = 0.05$ two-tailed.)

We want the 90% confidence intervals, though, so we use t for $\alpha = 0.1$ two-tailed. Working as follows:

- sample mean $\bar{x} = 5.015$
- sample standard deviation (SD) = $s_X = 0.711$ (see Examples 2 for a worked example of calculating this by hand, but your calculator should give it to you directly)
- $n = 10$
- standard error of the mean $s_{\bar{x}} = \frac{s_X}{\sqrt{n}} = \frac{0.711}{\sqrt{10}} = 0.225$
- number of degrees of freedom $df = n - 1 = 9$
- with 9 *df*, critical values of t for 5% each tail = ±1.833 (if there's 5% in each tail, then 90% of values of t lie within ±1.833 of the mean). This value is listed in the *Tables and Formulae* sheet as the critical value of t for a two-tailed α of 0.1, or a one-tailed α of 0.05. All these are different ways of saying the same thing!
- the 90% confidence intervals are therefore 5.015 ± (1.833 × 0.225) = 5.015 ± 0.4124 = 4.6026 and 5.4274

So there's a 90% chance the true mean lies between 4.6026 and 5.4274.
(Do your working with maximum accuracy to avoid rounding errors, but when you've finished it's probably best to express the final answer to 3 significant figures — so we'd state that the 90% confidence intervals are 4.60 and 5.43 to 3 *sf*.)

Q2 nonwd

95% confidence interval: 506–676 ms (to 3 *sf*).
(Same technique as Q1, but with 95% confidence intervals.)
Intermediate steps: $\bar{x} = 590.6667$; $s_X = 133.6259$; $n = 12$; SEM = 38.57447; $df = 11$; $t_{critical}$ for $df = 11$ and two-tailed α of 0.05 is 2.201.

Q3 weight

(a) 95% confidence interval: 95.3–103.8 g (to 1 *dp*). [Working: $\bar{x} = 99.56$; $s_X = 5.918558$; $n = 10$; SEM = 1.871612; $df = 9$; $t_{critical} = 2.262$ for $df = 9$ and two-tailed α of 0.05.]
(b) Since the interval we calculated in (a) includes 100 g, the mean is not significantly different from 100 g at the 5% level. Alternatively, if you wanted to make more work for your good self by practising a one-sample t test in full, you could run a one-sample t test: t_9 = (mean – 100 g) / SEM = (99.56 – 100) / 1.872 = –0.24. This is not significant at the 5% level.

Q4 traffic

This question calls for an unpaired t test (they're not the same cars each week).

Step 1 – which t test?

Which of the following do we choose from the formula sheet?

- *'Two-sample t test for unrelated samples — where the variances of the two groups are equal'*
- *'Two-sample t test for unrelated samples — where the variances of the two groups are unequal'*

Well, we can run an F test to decide. We can calculate

$n_1 = n_2 = 15$; standard deviations $s_1 = 5.53173$; $s_2 = 5.46243$; variances $s_1^2 = 30.6$; $s_2^2 = 29.838095$

To get F, we put the bigger variance on top of the smaller: $F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2} = F_{14,14} = \frac{30.6}{29.838} = 1.0255$. This is not significant even at the $\alpha = 0.1$ level. So we want the 'equal variances assumed' formula.

Step 2 – the t test

Since $n_1 = n_2$, we use this formula (note that $df = 28$):

$$t_{2n-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} = t_{28} = \frac{37.2 - 32.53333}{\sqrt{\frac{30.6 + 29.838}{15}}} = \frac{4.66667}{\sqrt{4.0292}} = 2.325$$

This is significant at $\alpha = 0.05$ two-tailed. Using a computer, we would find that for $t_{28} = 2.325$, we obtain an exact value $p = 0.0276$ (two-tailed).

Comments

- You could argue for the use of a one-tailed *t* test (you'd find $p = 0.0138$ one-tailed) given the question is asking specifically about reductions, but I think a two-tailed test is more sensible — you wouldn't really ignore the result if it turned out that the simulated accident *increased* speeds — and in any case, it doesn't alter the conclusion here: the simulated accident *did* significantly reduce speeds.
- Note that when performing an *F* test for the purpose of deciding which *t* test to use, it is 'conventional' (though not obligatory) to use a decision criterion of $\alpha = 0.05$ for the *F* test (corresponding to the first *F* table in the formula sheet). Since you rigged the *F* ratio so that it's never less than 1, this criterion is equivalent to $\boldsymbol{\alpha = 0.1}$ for the two-tailed question 'are the variances different?' (see p. 51 for explanation).

Q5 cards — No.

This question calls for a paired (two-related-sample) *t* test; in other words using the formula

$$t_{n-1} = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

First we calculate the *difference* score for each subject (and subsequently ignore the raw scores completely and operate only with the difference scores). For the null hypothesis that there was no effect of the drug, $\mu = 0$. We note that $n = 12$, $df = 11$, mean = 0.833, SEM = 1.359, $t_{11} = (0.833 - 0) / 1.359 = 0.613$; exact two-tailed $p = 0.552$ (if you have access to a computer), NS [not significant].

Q6 digit — No.

Call Arts students group 1 and Science students group 2. We find that

$$n_1 = 11;\ n_2 = 14$$

standard deviations $s_1 = 0.8311$, $s_2 = 0.7184$

An *F* test (see Q4 for method) shows that the variances are not significantly different; $F_{10,13} = 0.8311^2 / 0.7184^2 = 1.338$ (not significant; NS). So we want a two-sample unpaired *t* test assuming equal variances ($df = 23$)… Group sizes are different, so we use this formula:

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \text{ where } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \text{ (the 'pooled variance estimate')}$$

We calculate

$$s_p^2 = \frac{(11-1)\times(0.8311)^2 + (14-1)\times(0.7184)^2}{11+14-2} = 0.592 \text{ and } t_{23} = \frac{6.745 - 6.807}{\sqrt{\frac{0.592}{11} + \frac{0.592}{14}}} = -0.199\text{, NS.}$$

Q7 revfig — Yes.

Methods just as Q6. Initial *F* test: no significant differences between variances ($F_{9,7} = 1.145$, *NS*). Two-sample unpaired *t* test assuming equal variances gives $t_{16} = 2.37$, two-tailed $p = 0.03$.

Q8 letters — Yes.

Paired *t* test; $t_9 = -2.299$ (or +2.299, depending on which way you calculate the differences), $p = .047$ two-tailed. (Methods just as Q5.)

Q9 — Yes.

F test; $F_{8,7} = 9.02$, $p = 0.0089$ two-tailed. (Methods as for the *F* test in Q4.)

Q10 — Yes.

F test; $F_{5,4} = 9.98$, $p = .045$ two-tailed. (Methods as for the *F* test in Q4.)

8.4. Answers to Examples 4: nonparametric difference tests

Q1 Short answer: $U_{4,6} = 5$. Critical value is 3, so not significant (NS).

Step by step:

- Group B is the larger, so group A is 'group 1' and group B is 'group 2'.
- Group A: $n_1 = 4$. Group B: $n_2 = 6$.

Original data:

group 1 (A):	43	39	57	62		
group 2 (B):	51	63	70	55	59	66

Corresponding ranks:							***Sum of ranks***
group 1 (A):	2	1	5	7			15 (= R_1)
group 2 (B):	3	8	10	4	6	9	40 (= R_2)

Then $U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 15 - \frac{4\times5}{2} = 5$ and $U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 40 - \frac{6\times7}{2} = 19$. So U is the smaller of the two, i.e. $U = 5$. We'd write $U_{4,6} = 5$ to indicate n_1 and n_2 as well.

(Just to check our sums: $U_1 + U_2 = 5 + 19 = 24$ and $n_1n_2 = 4 \times 6 = 24$, so they match, which they must do. Similarly $R_1 + R_2 = 15 + 40 = 55$ and $\frac{(n_1+n_2)(n_1+n_2+1)}{2} = \frac{10\times11}{2} = 55$ so they also match.)

Now we look up a critical value for $U_{4,6}$ (critical U with $n_1 = 4$ and $n_2 = 6$); we find that it's 3. Since our U is *not* smaller than this, it's *not* significant.

Q2 $U_{7,9} = 15$. Critical value is 13, so NS.

The method is exactly the same as in Q1. Just to make sure you get the ranks right when there are ties, here they are:

Original data:

group 2 (A):	4.5	2.3	7.9	3.4	4.8	2.7	5.6	6.1	3.5
group 1 (B):	3.5	4.9	1.1	2.5	2.3	4.1	0.7		

Corresponding ranks (in bold where ties have been split by taking the mean of the tied ranks):

group 2 (A):	11	**3.5**	16	7	12	6	14	15	**8.5**
group 1 (B):	**8.5**	13	2	5	**3.5**	10	1		

In this example, no more than two scores are tied for the same rank — but you may come across examples when more scores are tied. The principle is just the same; take the mean of the ranks for which they are tied. So the ranks of {10, 50, 50, 50, 60} are {1, 3, 3, 3, 5}. The ranks of {2.3, 2.3, 2.3, 2.3, 8.1, 8.9} are {2.5, 2.5, 2.5, 2.5, 5, 6}.

Q3 $U_{7,7} = 8$. Critical value is 9, so *significant*.

Q4 $U_{9,10} = 20$. Critical value is 21, so *significant*.

Q5 $U_{16,17} = 76.5$. Critical value is 82, so *significant*.

Q6 $T_7 = 3$. Significant at $\alpha = 0.05$ (one-tailed) or $\alpha = 0.1$ (two-tailed) (critical value 4).

Full working:

Group A	4.5	2.3	7.9	6.8	5.3	6.2	5.7	
Group B	4.3	2.7	9.0	6.7	5.6	10.1	6.9	
Difference (B–A)	–0.2	0.4	1.1	–0.1	0.3	3.9	1.2	
Non-zero differences	(as previous row)							
Ranks of non-zero differences (ignoring sign)	2	4	5	1	3	7	6	***n* = 7**
Ranks of + differences		4	5		3	7	6	***sum* = 25 = T^+**
Ranks of – differences	2			1				***sum* = 3 = T^-**

The T statistic is the smaller of T^+ and T^-, i.e. 3. We can write $T_7 = 3$ (to show that $n = 7$). This value, 3, is smaller than the critical value of T_7 for $\alpha = 0.05$ (one-tailed) or $\alpha = 0.1$ (two-tailed), which is 4. But our T

is not smaller than the critical value of T_7 for any smaller values of α shown in our tables. So we could say 'T_7 = 3, significant at α = 0.05 (one-tailed) or α = 0.1 (two-tailed)'.

(To check our sums, $T^+ + T^- = 25 + 3 = 28$ and $\frac{n(n+1)}{2} = \frac{7\times 8}{2} = 28$ so all's well with the world.)

Q7 T_9 = 3. Significant at α = 0.01 (one-tailed) or α = 0.02 (two-tailed) (critical value 4).

Q8 T_8 = 8.5. Not significant ($p > 0.05$ one-tailed; $p > 0.10$ two-tailed; critical value 6).

Q9 T_8 = 4. Significant at α = 0.05 (one-tailed) or α = 0.10 (two-tailed) (critical value 6).

	Nonparametric test (subscripts are n, probabilities are two-tailed unless stated):	*Parametric equivalent (two-tailed in all cases):*
Q10 traffic	Mann–Whitney $U_{15,15} = 59$, $p < .05$ (The question phrases a one-tailed question, but you could argue for a two-tailed test.)	F test for heterogeneity of variance: $F_{14,14} = 1.026$, NS Unpaired t test, equal variances: $t_{28} = 2.325$, $p = .027$
Q11 RT	Wilcoxon matched-pairs signed-rank $T_{12} = 5$, $p < .01$	Paired t test: $t_{11} = 3.879$, $p = .00257$
Q12 cards	Wilcoxon matched-pairs signed-rank $T_{11} = 25$, NS	Paired t test: $t_{11} = 0.613$, NS
Q13 xeno	Mann–Whitney $U_{5,6} = 10$, NS (The question phrases a one-tailed question, but you could argue for a two-tailed test.)	—
Q14 cod	Wilcoxon matched-pairs signed-rank $T_{12} = 11$, $p < .05$	Paired t test: $t_{10} = 2.872$, $p = .0166$
Q15 digits	Mann–Whitney $U_{11,14} = 76.5$, NS	F test for heterogeneity of variance: $F_{10,13} = 1.338$, NS Unpaired t test, equal variances: $t_{23} = 0.199$, NS
Q16 revfig	Mann–Whitney $U_{8,10} = 16$, $p < .05$ (The question phrases a one-tailed question, but you could argue for a two-tailed test.)	F test for heterogeneity of variance: $F_{9,7} = 1.146$, NS Unpaired t test, equal variances: $t_{16} = 2.278$, $p = .031$
Q17 conv	Wilcoxon matched-pairs signed-rank $T_{12} = 13.5$, $p < .05$	Paired t test: $t_{11} = 2.218$, $p = .0485$
Q18 bats	Wilcoxon matched-pairs signed-rank $T_9 = 3.5$, $p < .02$	Paired t test: $t_9 = 2.743$, $p = .0228$
Q19 music	Mann–Whitney $U_{10,10} = 48$, NS	F test for heterogeneity of variance: $F_{9,9} = 1.327$, NS Unpaired t test, equal variances: $t_{18} = 0.051$, NS
Q20 letters	Wilcoxon matched-pairs signed-rank $T_9 = 5.5$, $p < .05$	Paired t test: $t_9 = 2.299$, $p = .0471$
Q21 vote	Mann–Whitney $U_{8,8} = 4$, $p < .05$	—
Q22 rats	Mann–Whitney $U_{10,10} = 40$, NS	F test for heterogeneity of variance: $F_{9,9} = 1.899$, NS Unpaired t test, equal variances: $t_{18} = 0.051$, NS
Q23 radar	Wilcoxon matched-pairs signed-rank $T_{12} = 12$, $p < .05$	Paired t test: $t_{11} = 2.449$, $p = .0323$
Q24 col'r	Wilcoxon signed-rank $T_{16} = 37$, NS (The question phrases a one-tailed question, but you could argue for a two-tailed test.)	One-sample t test: $t_{15} = 1.730$, NS.

Q25

Use the normal approximation for U. If $U_{20,60} = 400$, then $z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{400 - 600}{\sqrt{\frac{1200\times 81}{12}}} = -2.22$

This Z score is associated with a p value of 0.0132 (one-tailed) or $2 \times 0.0132 = 0.0264$ (two-tailed).

8.5. Answers to Examples 5: χ^2

Q1 coin Yes: $\chi^2 = 4.00$, df = 1, $p < .05$.
This is a simple 'goodness-of-fit' χ^2 test with 2 categories, so 1 degree of freedom. It's simple:

category	*observed, O*	*expected, E (based on null hypothesis)*	$(O-E)^2/E$
heads	40	50	$10^2/50 = 2$
tails	60	50	$10^2/50 = 2$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2 + 2 = 4$$

With *df* = 1, critical value of χ^2 for $\alpha = .05$ is 3.84, so our test is significant at this level (but not at the .01 level, for which the critical value is 6.63). A computer would tell us that $p = .046$.

Q2 rat Yes: $\chi^2 = 14.29$, df = 1, $p < .001$

Jump up, jump up, and get down. This is a two-way 'contingency' χ^2 test. All the rats either jump up or down (beware — if the up / down numbers didn't add up to the total number of rats, we'd have to add a third category… 'white rats don't jump'.)

Observed values (O):

	females	*males*	
up	16	40	*row 1 total = 56*
down	84	60	*row 2 total = 144*
	column 1 total = 100	*column 2 total = 100*	*overall total (n) = 200*

To work out the expected values, we use the formula

$$E(row_i, column_j) = \frac{R_i C_j}{n}$$

For example, row 1 ('up') has a total of 56; row 2 ('down') has a total of 144; both columns have totals of 100. The total number of observations is 200. Therefore, the expected value for (row 1, column 1) is $56 \times 100 / 200 = 28$, and so on. So we obtain this:

Expected values (E) under the null hypothesis (no relationship between sex and jumping):

	females	*males*
up	28	28
down	72	72

$(O–E)^2 / E$

	females	*males*
up	$(16–28)^2/28 = 5.143$	$(40–28)^2/28 = 5.143$
down	$(84–72)^2/72 = 2$	$(60–72)^2/72 = 2$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 5.143 + 5.143 + 2 + 2 = 50.24 = 14.286$$

$$df = (\text{rows} - 1) \times (\text{columns} - 1) = (2-1) \times (2-1) = 1$$

Our χ^2 is therefore significant at the 0.001 level. (A computer would tell us that $p = 0.000157$.)

Q3 crash Yes: $\chi^2 = 482.36$, df = 1, $p < .001$ (exact $p = 6.56 \times 10^{-107}$).
Explanation: two categories. Expected values are 1000 (Sundays), 6000 (days other than Sundays).

Q4 die No: $\chi^2 = 8.67$, df = 5, NS (exact $p = .123$).

Q5 giraffe Yes: $\chi^2 = 30.5$, df = 6, $p < 0.001$ (exact $p = 3.2 \times 10^{-5}$).

8.6. Answers to Examples 6: past exam questions

Q1 (2000 Paper 1) **Hidden assumptions — or, waffle you can skip in order to get to the answer.**

Before we answer the question, we must decide what kind of techniques to use. Should we use parametric or nonparametric techniques? **There's no necessarily 'right' answer.** We could summarize the theoretical arguments like this:

- The purist might say that the Beck Depression Inventory (BDI) is an **ordinal** rating scale; higher scores indicate 'more depression' somehow, but the differences between two ratings are not quantitatively meaningful (i.e. the difference between 0 and 10 is not necessarily the same as between 20 and 30) — like our example on p. 7 of Army ranks (lieutenant → captain → major, etc.). This approach would mean that parametric techniques would not be applicable and we should use a nonparametric analysis.
- Alternatively, we might treat the BDI as an **interval** scale for the purposes of analysis, which would allow us to use statistics such as the mean and standard deviation, and appropriate parametric tests (assuming their other assumptions are met). There are several rationales for this, discussed for example by Velleman & Wilkinson (1993), who argue that the 'meaning' of a scale is largely what you make of it — so if you're happy to speak about a '3-point change in a BDI score' then you should be happy to use parametric techniques here. As Francis Bacon (1620) said, truth emerges more readily from error than from confusion. And while 'common' doesn't imply 'correct', it is certainly common to analyse rating scales with parametric techniques — the first paper on depression I looked at for this purpose analysed the Hamilton Depression Rating Scale using analysis of variance, which is a parametric technique (Mayberg *et al.*, 2000); the second I found did the same with BDI scores (Allen *et al.*, 1998).

Being pragmatic, are there other barriers to using one or other technique?

- Part (b), comparing men and women before treatment, could be approached parametrically with an unpaired *t* test or non-parametrically with a Mann–Whitney *U* test, so no problem there.
- Part (c) could be approached parametrically with Pearson's *r* or non-parametrically with Spearman's r_s correlation, so no problem there.
- But part (a) asks which treatment is more effective — so we have to look at some difference between pre-treatment and post-treatment scores for each subject. If we take the differences (e.g. Post minus Pre scores) and analyse these, in whatever way, we have already made the assumption of an interval scale of measurement simply by calculating those differences, so we might as well use parametric techniques *unless their other assumptions are violated*. If we don't do this, the only information available is whether a subject improved or not. Since 15 / 15 Cogth patients improved, and 14 / 15 Couns patients improved, we're never going to find a significant difference with some form of categorical test (and a χ^2 test won't be valid since there will be expected values <5 and highly uneven across rows / columns, which violates the assumption of normality).

I'll illustrate both techniques below. Which will the examiners prefer? I've never been a IB examiner, but if I were marking this, I'd accept either (particularly if some justification were given to show that you'd thought about the issue). You'll note that both actually give the same answers in terms of 'significant or not' judgements at conventional levels of α. And you may be influenced by the amount of work involved — ranking 30 scores is perhaps error-prone during an exam, whereas your calculator will do much of the work for the parametric tests.

The answer…

(P.T.O.)

Parametric version of Q1 (2000 Paper 1)

(a)
First, we calculate the difference between pre-therapy and post-therapy depression scores (Post – Pre), collapsing across (ignoring) sex. We find they are:

Cogth: {–10, –7, –11, –9, –13, –22, –18, –14, –12, –10, –11, –14, –9, –13, –11}
Couns: {–5, –4, +1, –6, –4, –8, –4, –7, –3, –9, –7, –8, –3, –8, –4}

The Beck Depression Inventory gives high scores to depressed people, and low scores to non-depressed people. So calculating the scores this way (Post – Pre), the better treatment will have the lower (more negative) difference score. The mean score in the Cogth group is –12.267 (n = 15, SD 3.770, variance 14.21); the mean score in the Couns group is –5.267 (n = 15, SD 2.658, variance 7.067). So the Cogth treatment appears to do a better job at reducing depression scores. Is this a significant difference? Let's run a two-sample unpaired t test. First, we run an F test to see if the difference between the variances is significant:

$$F_{n_1-1,n_2-1} = F_{14,14} = \frac{s_1^2}{s_2^2} = \frac{14.21}{7.067} = 2.01, NS$$

Since it wasn't, we can use the 'equal variances assumed' version of the two-sample unpaired t test, with the simpler formula since $n_1 = n_2$. So we can calculate t (with 28 df) as follows:

$$t_{n_1+n_2-2} = t_{28} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(-12.27) - (-5.267)}{\sqrt{\frac{14.21}{15} + \frac{7.067}{15}}} = \frac{-7.003}{1.191} = -5.88, p < .01$$

Therefore, the Cogth group improved significantly more than the Couns group.

(b)
Before treatment, the men's scores were {42, 35, 32, 28, 26, 38, 33, 34, 29, 23} — a mean of 32 (n = 10, SD = 5.696, variance = 32.444). The women's scores were {20, 18, 17, 19, 21, 15, 22, 21, 27, 19, 19, 17, 18, 20, 23, 13, 24, 28, 24, 17} — a mean of 20.1 (n = 20, SD = 3.782, variance = 14.31). Let's run an unpaired t test again. First our homogeneity-of-variance check:

$$F_{9,19} = \frac{32.444}{14.31} = 2.267, NS$$

So we can use the 'equal variances assumed' version of the two-sample unpaired t test, but this time since the group ns are different, we must use the full formula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{9 \times 32.444 + 19 \times 14.31}{28} = 20.14$$

$$t_{n_1+n_2-2} = t_{28} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{32 - 20.1}{\sqrt{\frac{20.14}{10} + \frac{20.14}{20}}} = \frac{11.9}{1.738} = 6.847, p < .01$$

So the men were significantly more depressed than the women before treatment.

(c)
To answer this question, we *correlate* the Pre (X) and Post (Y) scores in the Cogth group, and see if that correlation is significant. Our X–Y pairs are {20, 10}, {18, 11}, etc. If your calculator gives you r directly, fine. Otherwise, we'll use the long-winded formula for the covariance:

$$\text{cov}_{XY} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n - 1} = \frac{4779 - \frac{362 \times 178}{15}}{14} = 34.52$$

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{34.52}{7.52 \times 5.17} = 0.888$$

This is high, so likely to be significant; let's check that with the usual t test:

$$t_{n-2} = t_{13} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.888\sqrt{13}}{\sqrt{1-0.788}} = 6.954, p < .01$$

It is. There is a significant positive correlation (r = 0.888, p < .01) between depression before and after treatment in the cognitive therapy group.

(Beware: if you had found no correlation, you would not be able to say 'no' definitively to question (c) — remember, you would need to sketch a scatter plot, because r only measures *linear* relationships.)

Nonparametric version of Q1 (2000 Paper 1)

(a)

For reasons summarized above, treating the BDI as an ordinal scale is slightly half-baked here (in my opinion). The only reasonable way to approach it non-parametrically is to view the BDI as an interval scale for the purposes of calculating difference scores, but then to analyse these non-parametrically. First, we calculate the difference between pre-therapy and post-therapy depression scores (Post – Pre), collapsing across (ignoring) sex. We find they are:

Cogth: {–10, –7, –11, –9, –13, –22, –18, –14, –12, –10, –11, –14, –9, –13, –11}

Couns: {–5, –4, +1, –6, –4, –8, –4, –7, –3, –9, –7, –8, –3, –8, –4}

We then perform a Mann–Whitney *U* test. Both groups are the same size, so we'll arbitrarily call the Cogth group 'Group 1' (n = 15) and the Couns group 'Group 2' (n = 15). When we rank the scores from 1–30 we have the following ranks:

Group 1 (Cogth): 11.5, 20, 9, 14, 5.5, 1, 2, 3.5, 7, 11.5, 9, 3.5, 14, 5.5, 9

Group 2 (Couns): 23, 25.5, 30, 22, 25.5, 17, 25.5, 20, 28.5, 14, 20, 17, 28.5, 17, 25.5

So our sums of ranks are $R_1 = 126$, $R_2 = 339$.

We calculate

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 126 - \frac{15\times16}{2} = 6 \text{ and } U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 339 - \frac{15\times16}{2} = 219$$

so U = 6. (We also verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$ to check our arithmetic.) The critical value of $U_{15,15}$ is 65 for two-tailed α = 0.05, so our U is significant. The Cogth group did significantly better than the Couns group.

(b)

Before treatment, the men's scores were {42, 35, 32, 28, 26, 38, 33, 34, 29, 23} (n = 15). The women's scores were {20, 18, 17, 19, 21, 15, 22, 21, 27, 19, 19, 17, 18, 20, 23, 13, 24, 28, 24, 17} (n = 20). We can compare these with a Mann–Whitney *U* test. We call the men's scores Group 1, because it's the smaller group. We rank all the scores (1–30). By group, they are:

Group 1 (M) 30, 28, 25, 22.5, 20, 29, 26, 27, 24, 16.5 *rank sum = 248*

Group 2 (F) 11.5, 6.5, 4, 9, 13.5, 2, 15, 13.5, 21, 9, 9, 4, 6.5, 11.5, 16.5, 1, 18.5, 22.5, 18.5, 4 *rank sum = 217*

We calculate

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 248 - \frac{10\times11}{2} = 193 \text{ and } U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 217 - \frac{20\times21}{2} = 7$$

so U = 7. (We also verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$ to check our arithmetic.) The critical value of $U_{10,20}$ is 56 for two-tailed α = 0.05, so our U is significant. The men were significantly more depressed than the women before treatment.

(c)

To calculate Spearman's r_s, we **rank** the Pre and Post scores in the Cogth group. We have these raw scores:

Pre	20, 18, 17, 19, 21, 42, 35, 32, 15, 28, 22, 21, 26, 27, 19
Pre rank (X)	6, 3, 2, 4.5, 7.5, 15, 14, 13, 1, 12, 9, 7.5, 10, 11, 4.5
Post	10, 11, 6, 10, 8, 20, 17, 18, 3, 18, 11, 7, 17, 14, 8
Post rank (Y)	6.5, 8.5, 2, 6.5, 4.5, 15, 11.5, 13.5, 1, 13.5, 8.5, 3, 11.5, 10, 4.5

We correlate the ranks as X, Y pairs, e.g. {6, 6.5}, {3, 8.5}… If your calculator gives you r directly, fine. Otherwise, we'll use the formula for the covariance:

$$\text{cov}_{XY} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{1200.25 - \frac{120\times120}{15}}{14} = 17.16$$

$$r_s = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{17.16}{4.464\times4.452} = 0.863$$

For n = 15, the tables tell us that the critical value of $|r_s|$ is 0.689 at the two-tailed α = 0.01 level, and our value is bigger than this, so our nonparametric correlation is significant at $p < .01$. So there is a significant positive correlation between depression before and after treatment in the cognitive therapy group.

Q2 (2000 Paper 2) n / a — experimental design question.
See Section 9 for advice on these questions.

Q3 (2001 Paper 1) *(a) Recognition v recall*

There are some hidden ambiguities in this question.

First off, we want to test if subjects differed for recognition and recall scores. This calls for either a paired *t* test (parametric) or a Wilcoxon matched-pairs signed-rank test (non-parametric). If the assumptions of the *t* test are met (i.e. the differences between each pair of data come from a normally-distributed population), the *t* test has more power.

The difference scores (recall minus recognition) are 7, –11, –10, –7, –10, 0, 0, –18, –1, 3.
Placed in order, they are –18, –11, –10, –10, –7, –1, 0, 0, 3, 7.

Parametric or nonparametric? A quick glance (or histogram, or stem-and-leaf plot) suggests that this isn't the best normal distribution in the world, so you may favour the Wilcoxon test. But note this: later, the question wants you not only to correlate these variables (which could be accomplished either with Pearson's *r* or Spearman's r_s) but to perform a regression, for which you only know parametric techniques (which assume normality of both the recall and recognition scores and that the difference between two normal random variables should also be normally distributed).… so you could probably argue the case either way.

If you had run a one-sample *t* test, you'd have got this: the difference scores have a mean of –4.7 and an SD of 7.689; $n = 10$. So

$$t_{n-1} = t_9 = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}} = \frac{-4.7 - 0}{\frac{7.689}{\sqrt{10}}} = \frac{-4.7}{2.431} = -1.933$$

From tables, you'd have found that $p < .05$ for a one-tailed test but $.05 < p < .1$ for a two-tailed test.

If you'd used the Wilcoxon test, there are 8 non-zero difference scores and you'd have found $T_8 = 5.5$, for which $p < .05$ one-tailed but $.05 < p < .1$ two-tailed.

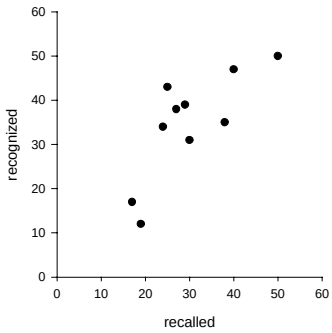
So both approaches give the same answer (which, if you actually chose to run them both, should reassure you that they're giving valid answers).

But what of the original question — 'determine whether recognition performance is better than recall performance'? This phrasing suggests a one-tailed test, and recognition performance did give the higher scores, so you may say 'significant, $p < .05$ one-tailed'. Alternatively, you may decide that a real researcher would not want to ignore a difference in the opposite direction, and say 'not significant, $p > .05$ two-tailed'. Since you wouldn't want to be accused either of misinterpreting the question or answering a slightly daft scientific question, you could just say 'one-tailed $p < .05$ but two-tailed $.05 < p < .1$', showing that you know what you're talking about, and then choose and defend a choice of a one- or a two-tailed test. It's your understanding of what's going on that counts, rather than any particular choice you justify sensibly.

(b) scatterplot, correlation, regression

'Construct a scatter plot of the number of pictures recognised against the number recalled.'

You could assign either to the *X* and *Y* axes. Looking ahead, we're going to be asked to predict recognition from recall, so let's call recall *X* and recognition *Y* so we're on familiar ground, predicting *Y* from *X*.



'Did those subjects who recalled more pictures also recognise more of them?'

That's asking 'was there a significant correlation between recall and recognition?' Either calculate *r* with your calculator, or work it out like this. Again, I'll call recall *X* and recognition *Y*:

$$\text{cov}_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{11{,}205 - \frac{299 \times 346}{10}}{9} = 95.51$$

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{95.51}{10.14 \times 12.14} = 0.776$$

Is this significant? We work out a *t* score:

$$t_{n-2} = t_8 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.776\sqrt{8}}{\sqrt{1-0.602}} = 3.479,\ p < .01$$

So the answer's yes; there is a significant positive correlation (subjects who recognized more also recalled more).

'Plot on your graph the line that best predicts recognition performance from recall performance.'

We need to calculate the regression equation

$$\hat{Y} = a + bX$$

We therefore want to find

$$b = \frac{\text{cov}_{XY}}{s_X^2} = r\frac{s_Y}{s_X}$$

and

$$a = \bar{y} - b\bar{x}$$

So we have

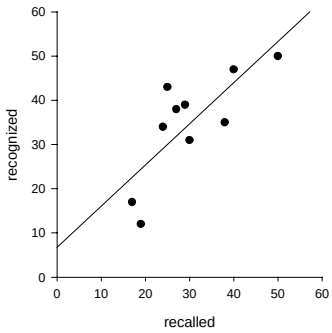
$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{95.51}{(10.14)^2} = r\frac{s_Y}{s_X} = 0.776\frac{12.14}{10.14} = 0.929$$

$$a = \bar{y} - b\bar{x} = 34.6 - 0.929 \times 29.9 = 6.823$$

So our regression equation is

$$\hat{Y} = 6.823 + 0.929X$$

We can plot this on our scatterplot, and we're done.



Q4 (2001 Paper 2) n / a — experimental design question.
See Section 9 for advice on these questions.

Q5 (2002 Paper 1) A correlation and regression question…

(a) … estimate the rate of 'mental rotation' from the data.

The trick in the question is that the raw data you've given don't represent the rotation angles, since 'the transformation is thought to be carried out over the most direct route'. So first, we must calculate the corrected data:

original (°)	*0*	*45*	*90*	*135*	*180*	*225*	*270*	*315*
rotation (°)	**0**	**45**	**90**	**135**	**180**	**135**	**90**	**45**
RT (ms)	518	563	638	781	896	738	625	552
error (%)	0.87	0.84	1.47	2.44	4.20	2.29	1.33	0.53

We want to predict RT (Y) from the (corrected) rotation angle (X). We have a series of paired values, $n = 8$, and we'd like to find the regression equation

$$\hat{Y} = a + bX$$

We therefore want to find

$$b = \frac{\text{cov}_{XY}}{s_X^2} = r\frac{s_Y}{s_X}$$

and

$$a = \bar{y} - b\bar{x}$$

We can easily find $\bar{x} = 90^\circ$, $\bar{y} = 663.9$ ms, $s_X = 58.92°$, $s_Y = 130.5$ ms. Your calculator should also give you r directly; if not, calculate the covariance

$$\text{cov}_{XY} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{530{,}190 - \frac{720\times 5311}{8}}{7} = 7457$$

We don't actually need to find r if we're doing it by hand, but a calculator will give us:

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{7457}{58.92\times 130.5} = 0.970$$

So our regression coefficients are

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{7457}{(58.92)^2} = r\frac{s_Y}{s_X} = 0.970\frac{130.5}{58.92} = 2.148 \text{ ms / degree}$$

$$a = \bar{y} - b\bar{x} = 663.9 - 2.148\times 90 = 470.6 \text{ ms}$$

and our regression equation is

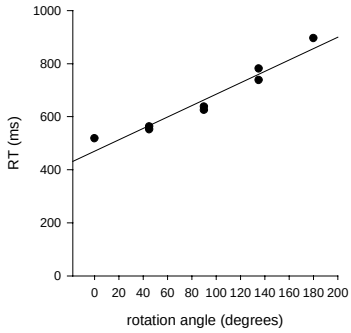
$$\hat{Y} = 470.6 + 2.148X$$

I've tacked on the units of a and b above — you can easily work out what they must be, since you start with a number in degrees (X), and multiplying it by b gives you a number in ms (Y), and a must have the same units as Y. The reason for doing this is because these numbers are what you're actually after. The answer to 'what is the rate of mental rotation?' is 2.148 ms / degree (b)…

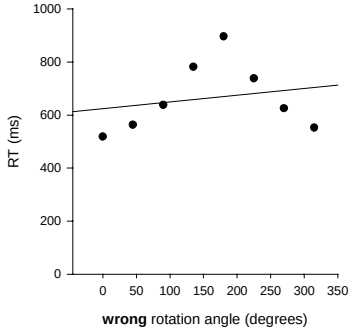
(b) What is your best estimate of the time occupied by the remaining components of the reaction time?

… and the remaining reaction time is 470.6 ms (a).

Although the question doesn't require it, you would be well advised to do a quick sketch of a scatterplot — for one thing, to see if there really is a linear relationship between X and Y:



A nice straight line. One benefit to doing a scatterplot in this question is that it might save you the embarrassment of failing to adjust the rotation angle. If you were in an exam-induced daze and didn't fix the angles, your scatterplot would look like this:



… which certainly isn't a linear relationship, and might make you notice that something was amiss.

(c) Is there a significant relationship between reaction time and error rate?

Here we go again — but this time, the question only asks 'is there' a relationship, not 'what is' the relationship. So we only need to calculate a correlation and test its significance. This time let's call RT X and error rate Y (or as you see fit). Either get r from your calculator or calculate

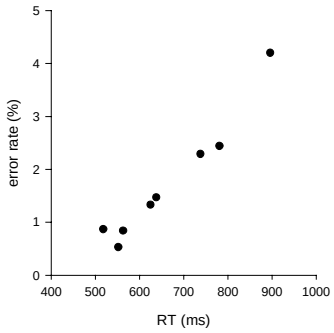
$$\text{cov}_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{10{,}344.11 - \frac{5311 \times 13.97}{8}}{7} = 152.83$$

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{152.83}{130.5 \times 1.202} = 0.974$$

Such a high correlation is certain to be significant with 8 observations. To test this formally, we calculate a t score:

$$t_{n-2} = t_6 = \frac{0.974\sqrt{6}}{\sqrt{1-0.9487}} = 10.53,\ p < .01$$

(since the critical value for t with 6 df is 3.707 for a two-tailed $\alpha = 0.01$). So there is a significant relationship between RT and error rate. The scatterplot **(not required but a sketch may be helpful)** is shown below.



But please note: if there had been no significant correlation, you would not have been able to say there was no relationship, just no linear relationship. This is where scatterplots help (is there a nonlinear relationship?), so you're always advised to sketch one, however rough it is.

*This final part is **not** part of the question… but you might be wondering whether there's just a relationship between RT and error rate because RT is related to the rotation angle, and errors are related to the rotation angle. We'd work out the correlation between RT (call it X) and error rate (call it Y), 'partialling out' the effects of rotation angle (call it Z). First we need to work out the one correlation we haven't worked out yet — between rotation angle (corrected so all the angles are ≤180°, of course) and error rate. It's 0.909. Then we can calculate*

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} = \frac{0.974 - 0.970 \times 0.909}{\sqrt{(1-0.970^2)(1-0.909^2)}} = \frac{0.09227}{0.1013} = 0.911$$

So even accounting for the relationships between RT and rotation angle, and between error rate and rotation angle, there's a pretty strong relationship between RT and error rate (though it's less than the correlation we first worked out of 0.974).

Q6 (2002 Paper 2) n / a — experimental design question.
See Section 9 for advice on these questions.

Q7 (2003 Paper 1) *(a)*
We begin with a confidence interval question. Our sample has a mean of 614.8 and a standard deviation of 47.98, and n = 10. Since we've been told to assume a normal distribution and find the 95% CI, we can use our formula:

$$\text{Confidence intervals} = \bar{x} \pm \frac{s_X}{\sqrt{n}} t_{critical(n-1)df}$$

For 95% confidence intervals, we want critical values of t for α = 0.05 two-tailed, with n–1 = 9 *df*. From our tables, this critical value is 2.262. So we can plug that into our formula:

$$95\%\text{ confidence intervals} = 614.8 \pm \left(\frac{47.98}{\sqrt{10}} \times 2.262 \right) = 614.8 \pm 34.32$$

So there is a 95% probability that the true population mean lies within the range 580 to 649 (to 3 *sf*).

(b)
The second part of this question asks about a difference between groups. Regardless of all the irrelevant guff about the experimental design, the key point is that we have two independent groups of scores, n = 6 per group. So we'd like to run an two-sample unpaired t test. We can justify this since we've already been told that reaction times on this task are from a normally-distributed population (and we can see that they're not grossly non-normal). The first group has mean = 555.2, SD = 57.26, variance = 3279. The second group has mean 601.8, SD = 37.02, variance = 1370. First, we should check the variances are not significantly different with an F test:

$$F_{n_1-1,n_2-1} = F_{5,5} = \frac{s_1^2}{s_2^2} = \frac{3279}{1370} = 2.39, NS$$

Since the variances are not significantly different and $n_1 = n_2$ (= n), we can use our simple formula for the two-sample unpaired t test assuming equal variances:

$$t_{2n-2} = t_{10} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} = \frac{555.2 - 601.8}{\sqrt{\frac{3279 + 1370}{6}}} = \frac{-46.6}{27.84} = -1.674, NS$$

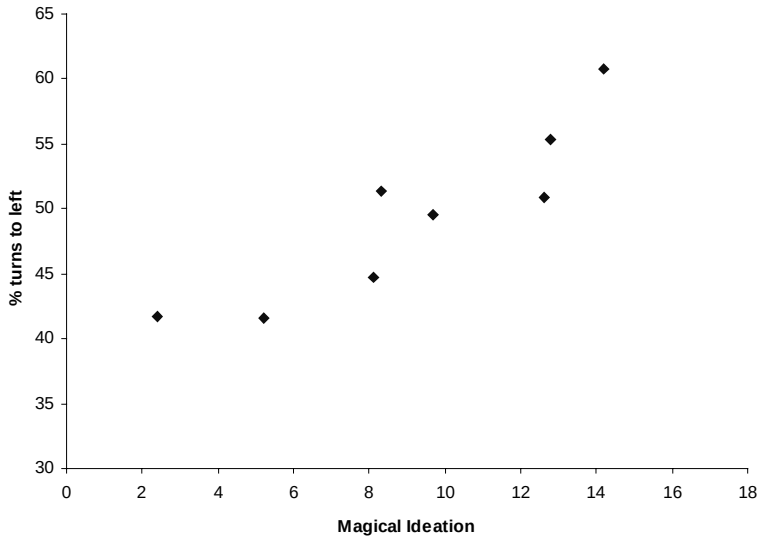
So the difference between groups was not significant ($p > .1$ two-tailed).

Q8 (2003 Paper 2) n / a — experimental design question.
See Section 9 for advice on these questions.

Q9 (2004 Paper 1) *(a)*
We're considering only the subjects who received placebo, so these are the data that should be used for the scatterplot and correlation to examine the relationship between magical ideation and left-turning behaviour:

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
MI	8.1	5.2	9.7	12.8	12.6	14.2	2.4	8.3
Left %	44.7	41.6	49.6	55.3	50.9	60.8	41.7	51.4

The scatterplot looks something like this:



Calculation of Pearson's correlation coefficient gives:

$$r = 0.902$$
$$n = 8$$
$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 5.109 \text{ (degrees of freedom: } n - 2 = 6)$$

From our tables of critical values of t, we can therefore establish that $p < 0.01$ two-tailed; there is a significant linear correlation.

The proportion of variation accounted for is 0.813, or 81.3% (= r^2).

(It would be less optimal to use a nonparametric approach: the question informs us that the data are from a normally-distributed population, validating the use of Pearson's r. Furthermore, r^2 isn't particularly meaningful for the nonparametric approach with Spearman's correlation, so it's difficult to answer the question about the proportion of variability accounted for without

using the parametric approach.)

(b)
This part of the question calls for a test to compare two unrelated groups. The optimal test (in terms of power) is the unpaired *t* test. The data are may be summarized like this:

	placebo	drug
	44.7	61.9
	41.6	57.2
	49.6	63.2
	55.3	53.1
	50.9	52.3
	60.8	47.3
	41.7	54.2
	51.4	48.3
mean	49.5	54.69
n	8	8
stdev	6.689	5.790
variance	44.74	33.52
SEM	2.365	2.05

A preliminary *F* test establishes that the variances are not significantly different:

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2} = F_{7,7} = \frac{44.74}{33.52} = 1.33, \text{NS}$$

Therefore, since the group sizes are equal, we perform the *t* test as follows:

$$t_{2n-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} = \frac{49.5 - 54.69}{\sqrt{\frac{44.74 + 33.52}{8}}} = t_{14} = -1.659, \textit{NS}$$

So there is not a significant difference between the dopamine agonist and placebo.

Alternative, but less powerful approach:
The question has explicitly given us the assumptions of the *t* test, so it is less optimal (less powerful) to use a nonparametric approach. However, if you chose to use a nonparametric approach regardless, the correct test to use would be the Mann–Whitney *U* test:

placebo	ranks	drug	ranks
44.7	3	61.9	15
41.6	1	57.2	13
49.6	6	63.2	16
55.3	12	53.1	10
50.9	7	52.3	9
60.8	14	47.3	4
41.7	2	54.2	11
51.4	8	48.3	5
	n = 8		n = 8
	rank sum 53 = R_1		rank sum 83 = R_2

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 53 - \frac{8\times 9}{2} = 17 \text{ and } U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 83 - \frac{8\times 9}{2} = 47$$

So Mann–Whitney $U = 17$. Critical U for $n_1 = n_2 = 8$ is 14 (from tables), so NS.

Optimal, but not taught at NST 1B level:
The optimal technique would be actually be an analysis of covariance or ANCOVA (using MI scores as covariate); this takes account of variations in MI scores, and having done so, it turns

out that there is an effect of the drug ($p < 0.001$). But that's beyond this course!

Q10 (2004 Paper 2) n / a — experimental design question.
See Section 9 for advice on these questions.

8.7. Answers to Examples 7: mixed

Q1 0.62% ($Z = -2.5$, corresponding to $p = 0.0062$)

Q2 No (preliminary F test of variances, $F_{14,14} = 1.366$, NS; unpaired t test assuming equal variances, $t_{28} = 1.1397$, NS, exact $p = .264$)

Q3 (a) 0, –2.5, –5, 5, 1.4

(b) 30, 70, 69.5, 24.2, 66.5, 46.7

(c) 137

(d) 0

(e) 57

Q4 No; $\chi^2_2 = 2.134, NS, p = .344$. (Note that the expected values for left and right turns are not equal — there might some bias that promotes left or right turning independent of the rats' breeding. That's why there are are several rat groups. Calculate the expected values as usual for a 2×2 contingency table.)

Q5 Correlating the rank order of the amounts with the rank order of their arrival time (Spearman's correlation coefficient) gives $r_s = -0.732$ ($n = 15$), $p < .01$ two-tailed.

Q6 (a) mean 45, SD 15

(b) 81.8%

(c) between 34.9 and 55.1 (to 1 *dp*)

Q7 Yes (paired t test, $t_{19} = 2.27$, exact $p = .035$).

Q8 No ($r = -0.19$, $t_{10} = -0.613$, NS).

Q9 No (preliminary F test to check homogeneity of variances $F_{7,14} = 1.36$, NS; unpaired two-sample t test assuming equal variances $t_{21} = 1.71$, NS, exact $p = .59$).

Q10 (a) [only 100,000 people above IQ of] 143

(b) [num. people with IQ between 95 and 105] 13,055,860 (should really round this to 3 *sf*!)

(c) [… between 60 and 70] 945,982 (should really round this to 3 *sf*!)

Q11 Yes. A preliminary F test gives $F_{9,11} = 2.982$, two-tailed $p = .09$. Whether you decide this is OK and proceed to an equal-variances-assumed unpaired t test ($t_{20} = 2.7999$, $p = .011$ two-tailed), or prefer a t test not based on the equal-variances assumption ($t'_9 = 2.6696$, $p = .026$ two-tailed) — which is probably better (most people use two-tailed $\alpha = 0.1$ for preliminary F tests) — you still find a significant difference even with two-tailed t tests.

Q12 Yes! $\chi^2_{12} = 78.72, p = 7.2 \times 10^{-12}$.

This test assumes equal independence of observations (that no student got more than one degree, for example), normality (no expected values very small and no dramatic skew of row / column values — this looks OK), and inclusion of non-occurrences (that we've included all the students and universities that we studied). The first of these is the one that we'd have to be most careful about checking!

Q13 No significant difference; Mann–Whitney $U_{7,13} = 44$, NS.

Q14 Yes; $\chi^2_4 = 15.52, p = .0037$.

Q15 Rearrange the data into a 2×2 contingency table: {scratching / not scratching} versus {tooth-baring / not tooth-baring}. Then we find $\chi^2_1 = 6.88, p = .0087$, so yes, the two are 'connected'. The way they're connected seems to be that they're unlikely to occur together!

Q16 Three examples of (potentially) relevant tests:

(a) Are husbands' preferences correlated with those of their wives? No ($r = 0.310$, $t_{24} = 1.60$, NS).

(b) Do the husbands' preferences vary more or less than the wives'? There is a significant difference in variance ($F_{25,25} = 6.29$, $p < .001$) — husbands' preferences vary more.

(c) Are husbands' preferences (mean 2.61) significantly different from their wives' (mean 2.42)? No (paired t test, $t_{25} = 0.385$, NS).

(d) Is there a relationship between sex and number of children? Perhaps I should rephrase that… is there a relationship between gender and desired number of children? Yes. Categorize people by gender and by number of children desired (e.g. '0', '1', '2', '3', '4', 'more than 4') and perform a contingency χ^2 test. If you use these categories, you find $\chi^2_5 = 15.81, p = .0074$. However, this approach assumes independence of husbands' and wives' preferences, which is perhaps questionable (the answer to (a) just tells you there's no overall linear correlation, which isn't the same thing).

Thanks to MRFA for (d), which I didn't think of. You could also apply nonparametric versions of some of the

above (simply a different strategy — these additional tests don't provide new information, so running a parametric and a nonparametric version of the same test wouldn't 'count' as two tests). The only rather artificial extra question I can come up with is to use the order information, e.g. 'is there a relationship between the mean number of children preferred by a couple and the order in which the couples were asked?' (no; correlation between the couple order and the rank order of the couple's mean preference, $r_s = -0.326$, $p > .1$ two-tailed).

Q17 Yes; $t_9 = 3.21$, $p = .011$ by paired t test.

Q18 Not quite: correlation $r = -0.507$, $t_{13} = -2.122$, $p = .054$.

Q19 No, unless you were in a particularly suspicious mood: $\chi^2_5 = 10.36$, $p = .066$.

Q20 Yes; Wilcoxon signed-rank $T_{15} = 21$, $p < .05$ two-tailed.

Q21 Yes. You have to calculate 'hits' from what you know (misses and totals). Then $\chi^2_3 = 17.27$, $p = .00062$.

Q22 A bit nasty, this, and **way** beyond exam standard. You'd need to calculate *which* normal distribution *best* fits the data, and then see whether there's a significant difference between the actual values and the values predicted by this normal distribution. You could do it like this:

(a) Take the dioptre measurement as your variable. So you have 11 observations with value –5, 29 observations with value –4… up to 5 observations with value +5. There are a total of $n = 1000$ observations. The sum of the observations is $(11 \times -5) + (29 \times -4) + \ldots + (5 \times +5)$: we have $\sum x = -174$. So the mean is $\bar{x} = \frac{\sum x}{n} = \frac{-174}{1000} = -0.174$. Now we need the SD:

$$\sum x^2 = \left(11\times(-5)^2\right) + \left(29\times(-4)^2\right) + \ldots + \left(5\times5^2\right) = 2322$$

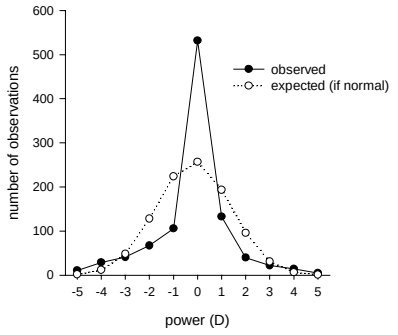
$$s_X = \sqrt{s_X^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{2322 - \frac{(-174)^2}{1000}}{999}} = 1.515$$

(b) The second phase is to determine the expected values for a normally-distributed variable with a mean of –0.174 and an SD of 1.515 if we measured them in a categorical way. It would be reasonable to suppose that the '0D' category includes people measuring –0.5 to +0.5D; the '1D' category includes people from +0.5 to +1.5D, and so on. So we can calculate Z scores for each boundary between dioptre categories, work out the proportion we expect to find in each category, and multiple by 1000 to get the number of observations we expect. That gives us expected values like this:

–5D	–4D	–3D	–2D	–1D	0D	+1D	+2D	+3D	+4D	+5D
1.93	11.90	48.27	128.35	224.13	257.05	193.63	95.79	31.10	6.62	0.92

(c) Finally, we run a χ^2 test with these as the expected values. Since we have made the observed and expected values agree as to n, $\bar{x}$, and s_X, we have lost *three* degrees of freedom (rather than a 'conventional' χ^2 test, in which we make them agree only as to n, and lose only one df). So $df = k - 3$ (where k is the number of categories) = 8. We find that $\chi^2_8 = 534.4$, $p = 2.87\times10^{-110}$, which we could call 'very significant indeed'! The lens powers *do* deviate significantly from a normal distribution.

Out of interest, here are the histograms of the observed values and corresponding expected values (for the same n) based on a normal distribution having the same mean and standard deviation. (Trivia: the posh name for this type of deviation from normality — too peaked in the middle — is *leptokurtic*.)



9. Experimental design tips and glossary

9.1. About the experimental design questions

There's no 'right' answer to an experimental design question. Good experimental design requires that you understand the question well — if you don't know what retroactive and proactive interference are, for example, you'll have difficulty with Q3(c). But *in general*, what things should you think about when designing experiments?

One vital thing is to establish **what you'll measure.** *What numbers will you actually write down?* For example, for exam question Q3(a), will you measure baby looking times? Baby approach distances? Proportion of occasions on which the baby approaches? Another example: exam Q8 (part 2) asks about investigating whether aspirin actively affects the operation of the active mechanism in the cochlea. Will you measure detection thresholds at different frequencies? Will you measure frequency selectivity (and how)? Will you measure otoacoustic emissions? To choose, you need to know what aspects of hearing actually depend on the active mechanism of the cochlea — obviously, not all aspects of hearing do.

You also need to determine your **subjects.** Will you use people? Owls? Rats? Dissected cochleas? If you use humans, who? Psychology undergraduates? People recruited from newspaper adverts? Will you impose restrictions on the age or sex of your subjects? Will you exclude them if they have a history of mental illness, head injury, or ototoxic antibiotic use? If you think the specifics are not important, you might simply say that you'd recruit subjects with normal hearing. Sometimes your experimental technique influences who you recruit: you can't give PET scans to young women (potential egg damage from radiation), you can't put people with implanted magnetic metal into an MRI scanner (it accelerates hard and tends to kill), and you might be careful before giving a drug that makes people unhappy to those with a history of severe depression.

Will you use a **correlative** or a **causal** technique? In our aspirin example, will you seek out people who use lots of aspirin and compare them to people who don't? If so, do you need to match these groups somehow? What might confound your interpretation of any differences? Or will you use a causal technique, in which you give aspirin to subjects in some fashion and compare them to people who didn't receive aspirin?

Many psychological experiments are based on simple intervention studies, where treatments are controlled by the experimenter. If done properly, these allow inferences to be made about whether the treatment *caused* an effect. They may use between-subjects (between-groups) or within-subjects designs

- A simple *between-groups* design: assign subjects (at random) to groups. Treat each group differently, before giving them all the same test. If the test results for the groups are different, this is evidence that the treatment influences performance.
- A simple *within-subjects* design: test the same subjects repeatedly after (or during) different treatments, counterbalancing for the order in which you give the treatments and so on. If the results differ across treatment conditions, this is evidence that the treatment influences performance.
- Sometimes we have to use more complex experimental designs. For example, when more than one treatment is given, they can all be given in a between-subjects fashion, or all in a within-subjects fashion, or some between- and some within-subjects.
- Sometimes we have to consider differences between groups that are not assigned by the experimenter — such as gender, age, IQ, prior illness, prior drug use. The exact interpretation of differences between groups in such an experiment may be more complicated, as the effect of one variable may be **confounded** by another. For example, if you give all the male subjects treatment A, and all the female subjects treatment B, you won't be able to tell whether a difference between the groups is due to the treatment difference or the sex difference — these two variables are confounded. In general, **random assignment** of subjects to groups is a good way to get around this problem.

So will you use a **between-subjects** or a **within-subjects** design? In our aspirin example, will you test subjects on aspirin and the same subjects off aspirin? Or will you give some subjects aspirin and some not?

- If you use a within-subjects design, will you test 'off' aspirin and then 'on'? Or 'on' then 'off'? Will there be effects of practice on the task that affect the interpretation in this case? Will the drug have permanent or long-lasting effects? Should you randomize or counterbalance the order (so some subjects get aspirin first and some get placebo first)?
- When you're not giving the drug, will you give nothing, or a placebo? If you use a placebo, will the experimenter be 'blind' as to the condition the subject is in? The importance of the **experimenter's awareness,** or lack of it, applies more generally whenever there is the possibility that the experimenter's expectations may influence the subject, or influence the recording of the data, consciously or unconsciously.
- If you use a between-subjects design, how will you assign subjects to groups?

Sometimes the question asks about how you will **analyse** the data you collect. The design of your experiment partly determines the analytical technique. Do you have **quantitative** or **categorical** data? The use of a between-subjects or a within-subjects design influences the 'relatedness' of your data, and may determine whether you will use **related** (e.g. paired) or **unrelated** (unpaired) statistical tests. What will your **null hypothesis** be? There may be some things you can't specify in advance (e.g. you may prefer to use parametric tests like the *t* test, but you don't always know whether their assumptions will be met until you collect the data; if their assumptions are violated, you may need to use nonparametric tests instead — and you can simply say that).

Good designs are **simple,** and answer the question clearly. If you find an effect in your experiment, will the **interpretation** be simple?

Sometimes you may need to design a **series of experiments** rather than a single experiment. Think about what each experiment should aim to establish; keep each experiment as simple as possible. Sometimes the most sensible choice of the next experiment depends on the results of previous experiments; you can do no more than anticipate likely outcomes or lines of investigation if you are outlining a proposed series of experiments.

Good designs are also **practical.** If your design calls for the use of a zero-gravity cell culture environment, it may be impractical or expensive — but if the question is important (will it save lives? improve the lot of millions?), maybe it's worth it. On the other hand, if you could do the same experiment with a set of headphones and a signal generator, that's probably preferable. If your design involves inducing permanent hearing damage in volunteer humans, it's highly questionable **ethically.** When using animals, experimenters always seek to *refine* experiments to minimize suffering and distress, *reduce* the number of animals used, and *replace* live animals with alternatives when possible.

Finally, your design may be excellent, or it may just be the best thing you can think of during the exam. If you spot **problems** or flaws in your design, discuss them. You're not expected to design something perfect, but if you know there are problems, talk about them.

Many of these issues were also discussed in Section 1.

9.2. Glossary of jargon

You may encounter the following terms in relation to experimental design. This glossary is here to help you to understand descriptions of experiments; you don't have to use the jargon for your own answers (though you can if you want).

- **Between-groups design.** Same as *between-subjects design.*
- **Between-subjects design.** A design in which individual subjects are each given only one treatment; different treatments are given to different groups of subjects. For example, we might assign subjects to group A or group B; each individual has their performance on a task measured after being given either sugar (for group A), or amphetamine (for group B). See also *within-subjects design.*
- **Blind.** Unaware. 'Blind to' means 'unaware of'. For example, in a *single-blind* drug study, the subject is unaware of whether he has taken an active drug or a *placebo*. If the subject is not blind, his expectations of the treatment's effects may influence the results. For example, if the subject is a depressed patient in a study examining the effects of a new antide-

pressant, he may expect good things of the drug, and therefore feel good about the fact that he's receiving it, and feel less depressed. See also *placebo, double-blind*.

- **Clever Hans effect.** The fact that subtle and unintentional cueing by an experimenter, which reflects the experimenter's own expectations, may influence subjects. Wilhelm von Osten was a retired German schoolmaster who attempted to teach his Russian stallion Hans arithmetic. Von Osten would show Hans numbers on cards; the horse would tap out the number, or the answer to simple mathematical problems, with its hoof. A group of observers including a zoologist, a vet, and a politician were convinced; they could detect no fraud or cueing on the part of von Osten. Finally, Oskar Pfungst conducted a very thorough series of experiments in 1907 to investigate Hans's performance (Pfungst, 1907). First, he established that if von Osten did not know the number or answer, Hans did not succeed. Second, he showed that if the horse could not see von Osten, it also failed to get the right answer. Pfungst then turned his attention to von Osten, and noticed that he made almost imperceptible alterations in posture when interrogating the horse. Von Osten inclined his head as the horse began to tap the ground, and straightened slightly, lifting his eyebrows and flaring his nostrils slightly, when the horse approached the correct answer. The horse had learned to respond on the basis of these cues. Pfungst then stood in front of the horse himself, remained silent, showed no cards, and yet made Hans tap his hoof and cease using slight head movements. Pfungst went on to perform experiments in which he played Hans's role; he showed that over 90% of human subjects provided subtle bodily cues as the correct answer was approached (attributed to 'tension release'), just as von Osten had. See also *double-blind*.
- **Confound.** Two variables are confounded when their effects are impossible to distinguish. Suppose we want to establish whether a drug (call it treatment A) influences performance on a particular task, compared with placebo (call it treatment B). If you give all the male subjects treatment A, and all the female subjects treatment B, you can't tell whether a difference between the groups is due to the treatment difference or the sex difference — these two variables are confounded. In general, random assignment of subjects to groups is a good way to get around this problem (see also *randomization*). Common confounding factors worth thinking about are time (see *order effects*) and who collects the data.
- **Control.** OED: 'A standard of comparison for checking inferences drawn from an experiment; specifically a patient, specimen, etc., similar to the one(s) being investigated but not subjected to the same treatment.' If two groups are compared and one receives some critical treatment but the other does not, we refer to the latter as the *control group*. (See also *placebo, sham*.)
- **Counterbalancing.** A method of avoiding *confounding* among variables. Suppose subjects are tested on both an auditory reaction time task and a visual reaction time task. If all the subjects are tested on the auditory task first, the task order (first versus second) is confounded with the task type (auditory versus visual), and *order effects* such as a *practice effect* may account for any observed differences. The experiment should have been designed better: task order and task type should have been counterbalanced, such that half the subjects were given the auditory task first and half were given the visual task first. When there are several conditions, a *Latin square* is often use to design the counterbalancing.
- **Dependent variable.** A variable that you measure, but do not control. Compare *independent variable*.
- **Double-blind.** Where both the experimenter and the subject are unaware of the treatment that the subject receives. Otherwise, either the experimenter's expectations, or the subject's, or both, may influence the results. For example, if Alice is interviewing patients about their mood as part of a study looking at antidepressant effects, and is aware that the patient has been on a new drug of which Alice has a high opinion, she may expect the patient to be happier, therefore be and appear happier herself, and therefore influence the answers. See *clever Hans effect*. The 'gold standard' for a medical drug trial is a *double-blind placebo-controlled* study (see also *placebo, control*).
- **External validity** (also known as **generality** or **applicability**). The degree to which your experimental results can be applied to other populations and settings. If you examine reaction times in Cambridge IB psychology undergraduates, to what extent can your results be generalized to Cambridge undergraduates? To UK undergraduates? To undergraduates in general? To people in general? Compare *internal validity*.
- **Factorial design.** When an experimenter is interested in the effects of two or more treatments, it is common to analyse them in a factorial design. Suppose we are interested in the effects of nicotine on psychomotor performance. We might be interested in the effects of both (1) nicotine dose, and (2) the difficulty of the task. So one variable ('factor') is drug dose; the other is task difficulty. Suppose there are three doses (none, low, high) and two

levels of task difficulty (easy, hard). In a factorial design, we test every combination of the factors (i.e. none / easy, none / hard, low / easy, low / hard, high / easy, high / hard). Factorial designs can have more than two factors. We have *not* covered the statistical techniques required to analyse factorial designs, but they are widely used in research.

- **Independent variable.** A variable that you control or manipulate (e.g. drug versus placebo). Compare *dependent variable*.
- **Internal validity.** To what degree are you justified in drawing conclusions from your data? Basically, was your experiment any good? If you have overlooked a *confound*, you may be unable to interpret your data in the way you had hoped, and you do not have good internal validity. Compare *external validity*.
- **Interpretation bias.** When the interpretation of evidence is influenced, sometimes inappropriately, by prior beliefs. *Confirmation bias* — people's tendency to notice and remember evidence that confirms their beliefs or decisions, and to ignore, dismiss, or forget evidence that is discrepant. *Rescue bias* — discounting data by finding selective faults in the experiment. *Auxiliary hypothesis bias* — introducing *ad hoc* modifications to imply that an unanticipated finding would have been otherwise had the experimental conditions been different. *Mechanism bias* — being less sceptical when underlying science furnishes credibility for the data. *"Time will tell" bias* — the phenomenon that different scientists need different amounts of confirmatory evidence. *Orientation bias* — the possibility that the hypothesis itself introduces prejudices and errors and becomes a determinate of experimental outcomes. (See Kaptchuk, 2003.)
- **Latin square.** (You don't need to know this!) A way to *counterbalance* the order in which subjects are tested in different conditions. A Latin square is a *n* by *n* grid in which each of *n* symbols appears exactly once in each row and once in each column. This can be used to allocate subjects to different treatment orders. If there are four treatments (A, B, C, D) and each subject must be tested in each condition once, a good Latin square is ABCD, CADB, BDAC, DCBA. If you randomly assign subjects to these four treatment orders, you will have successfully counterbalanced for treatment order. Each treatment immediately precedes and follows the other conditions once, known as a digram-balanced Latin square. (A much less good Latin square is ABCD, BCDA, CDAB, DABC, known as cyclic.)
- **Matching.** One way to avoid *confounds* between variables. Suppose we want to compare the performance of two groups of people on a reaction-time task; one group will receive drug and the other will receive placebo. We want our groups *not* to differ in any variable except drug v. placebo — otherwise our treatment and that variable will be confounded. Matching would involve deliberately measuring all sorts of things that we think might be relevant (e.g. IQ, age, sex, reaction time on a different task…) and assigning subjects to groups so that both groups have a similar distribution of sex, age, IQ, and so on. See also *randomization*, which is the other very important method to use in this situation; randomization deals with all the other variables you haven't thought about.
- **Order effect.** A concern of *within-subjects* designs, in which subjects are each tested several times: the order in which you test subjects may be an important factor. This may be due to *practice* (performance getting better with time), or other factors such as fatigue or boredom (getting worse with time), lingering drug effects (e.g. drug not yet fully gone from the body from a previous occasion; tolerance develops to drug with time), etc.
- **Orientation bias.** When prior expectations influence the collection of data. For example, psychology graduate students, when informed that rats were specially bred for maze brightness, found that these rats outperformed those bred for maze dullness, despite both groups really being standard laboratory rats assigned at random. (See Kaptchuk, 2003.)
- **Placebo.** Literally, 'I shall please'; a pill or procedure prescribed by a doctor for the psychological benefit of obtaining a prescription, rather than any physiological effect. Often a sugar pill. A *placebo effect* is an effect caused by a placebo. In research design, to establish the effect of a drug, one must compare it with something similar in all respects except for the drug itself. Therefore it's not wise to give one group of people the drug and the other (control) group nothing; the control group should be given a placebo. If the drug is in pill form, the placebo should be an inactive pill, labelled identically; if the drug is an injection, the placebo should be some inactive substance that is also injected (e.g. saline solution, or whatever liquid 'vehicle' the drug is dissolved in).
- **Practice effect.** If subjects are tested repeatedly, they may get better as a result of practice. An example of an *order effect.* Suppose you want to measure the effect of amphetamine on performance of the computer game Tetris, using a *within-subjects design*. You could give your subjects a *placebo* and test them; you could then give them amphetamine and retest them. Suppose they're better the second time: is this due to the drug or to the effects of practice? You can't tell: the two are *confounded.* You should have *counterbalanced* the or-

der assignment.

- **Publication bias.** The tendency of referees and journal editors (and sometimes the scientists concerned) to publish studies (or not) based on the direction or strength of the study's findings. For example, journals like to report 'significant' effects, because they seem more interesting, meaning that studies failing to find that a treatment has an effect may be less likely to be published. As a consequence, too high a proportion of what you read in journals would suggest the treatment *does* have an effect.
- **Randomization.** Random assignment of subjects to groups and / or treatment conditions is an important way to avoid inadvertent *confounds* (q.v.). Suppose we want to compare the performance of two groups of people on a reaction-time task. One group will receive a drug and the other will receive a placebo. We want our groups *not* to differ in any factor except that which we're manipulation — otherwise our treatment and that variable will be confounded. We might attempt to *match* groups (see *matching*) for relevant variables. But we probably can't explicitly match groups on every variable that might potentially be a confound; eventually we need a mechanism to decide which group a subject goes in, and that method should be random assignment. So in our example, if we have plenty of subjects, we could just randomly assign them to the drug group or the placebo group. Or we could match them a bit better by ranking them in order of reaction time performance and, working along from the best to the worst, take pairs of subjects (from the best pair to the worst pair), and from each pair assign one to the drug group and one to the placebo group at random. Random assignment takes care of all the factors you haven't thought of — for example, if your subjects are all going to do an IQ test in your suite of testing rooms, you should seat them randomly, in case one room's hotter than the others, or nearer the builders' radio outside, or whatever.
- **Sham.** Similar to *placebo*; the term is often used to refer to practical (e.g. surgical) procedures. If one group of rats receives amygdala lesions, the appropriate control group is probably not a set of unoperated rats, but a set of rats who have received 'sham' surgery — surgery identical except for the omission of the toxin that destroys the amygdala, say.
- **Within-subjects design.** A design in which individual subjects are each given more than one treatment, at different times. For example, each individual has their performance on a task measured after being given sugar, and on a separate occasion after being given amphetamine. These designs are often statistically powerful (they need fewer subjects to detect effects of the treatment), since differences between subjects' ability to perform the task don't contribute to the measurement error. Problems with these designs include *practice* and *order* effects; attention must be paid to proper *counterbalancing* of the order in which subjects are tested in different conditions.

10. Tables and formulae

Notation used

X	random variable that can take many values
x	single observation from the random variable X
Σx	The sum of all values of x
μ	Population mean
$\bar{x}$	Sample mean
σ	Population standard deviation
s	Sample standard deviation
σ^2	Population variance
s^2	Sample variance
H_0	null hypothesis
H_1	alternative hypothesis (research hypothesis)
p	probability of obtaining the observed data if H_0 is true
α	significance level = probability of making a Type I error (rejecting H_0 when it is true)
β	probability of making a Type II error (accepting H_0 when it is false)

Descriptive statistics

mean

$$\bar{x} = \frac{\sum x}{n}$$

population variance

$$\sigma_X^2 = \frac{\sum (x-\mu)^2}{n}$$

sample variance

$$s_X^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

population standard deviation

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

sample standard deviation

$$s_X = \sqrt{s_X^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

The normal distribution

Converting any normal distribution $N(\mu,\sigma^2)$ to the standard normal distribution $Z = N(0,1)$

$$z = \frac{x-\mu}{\sigma}$$

Correlation and regression

Sample covariance of two variables X and Y (the left-hand expression is the 'conceptual' formula; the right-hand one is mathematically identical but quicker to compute)

$$\text{cov}_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1}$$

Pearson product–moment correlation coefficient (varies from –1 to +1)

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y}$$

Adjusted r (always positive)

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}}$$

Is r significantly different from zero? A t test with n – 2 degrees of freedom

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

This test assumes (1) the variance of Y is roughly the same for all values of X, i.e. homogeneity of variance; (2) for all values of X, the corresponding values of Y should be normally distributed; (3) X and Y are both normally distributed. Look up the value of t in the tables of the t distribution to see if it is significant.

Regression, predicting Y from X

linear regression equation

$$\hat{Y} = a + bX$$

coefficients

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\text{cov}_{XY}}{s_X^2} = r\frac{s_Y}{s_X}$$

Difference tests — parametric

Standard error of the mean (SEM)

$$s_{\bar{x}} = \frac{s_X}{\sqrt{n}}$$

One-sample t test and two-related-sample (paired) t test

$$t_{n-1} = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

where the null hypothesis is that $\bar{x} = \mu$. For a one-sample *t* test, $\bar{x}$ and s_X refer to the mean and standard deviation of the observations from the single sample; for a two-sample *t* test, they refer to the mean and standard deviation of the *differences* between the two samples in each pair. The *t* test has $n - 1$ degrees of freedom.

Two-sample t test for unrelated samples — where the variances of the two groups are equal

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where s_p^2 is the pooled variance and the null hypothesis is that $\bar{x}_1 = \bar{x}_2$. The denominator is the standard error of the differences between means (SED). The *t* test has $n_1 + n_2 - 2$ degrees of freedom. **If the two samples are of equal size** ($n_1 = n_2 = n$), a simpler formula can be used:

$$t_{2n-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

Two-sample t test for unrelated samples — where the variances of the two groups are unequal

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(Note that this is written *t*′, not *t*.) Test this just as if it were a *t* score, but with fewer degrees of freedom: **degrees of freedom = ($n_1 - 1$) or ($n_2 - 1$),** whichever is smaller.

Assumptions of t tests

(1) The *t* test assumes that the underlying populations of the scores (or difference scores, for the paired *t* test) are normally distributed.
(2) For a two-sample test, in order to use the equal-variance *t* test, we assume the two samples come from populations with equal variances ($\sigma_1^2 = \sigma_2^2$). If this is not the case, especially if $n_1 \neq n_2$, we should use the unequal-variance version of the *t* test.

The F test for differences between two variances (used to choose the form of the t test)

Put the larger variance on top of the ratio:

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2} \text{ if } s_1^2 > s_2^2 \qquad F_{n_2-1,n_1-1} = \frac{s_2^2}{s_1^2} \text{ if } s_2^2 > s_1^2$$

The subscripts on the *F* are the numbers of degrees of freedom in the numerator and denominator, respectively (either $n_1 - 1$ and $n_2 - 1$, or $n_2 - 1$ and $n_1 - 1$). If the calculated value of *F* exceeds the critical value for the relevant α and degrees of freedom, reject the null hypothesis that the two samples come from populations with equal variances, and use the *unequal variances* form of the *t* test to test for differences between the means of the two samples. If the calculated value of *F* is not significant, assume that the populations have equal variances, and use the *equal variances* form of the *t* test to test for differences between the two means.

The *F* test assumes that the underlying populations are normally distributed.

Difference tests — nonparametric

How to rank data

Place the data in ascending numerical order. Assign them ranks, starting with rank 1 for the smallest datum. If two or more data are tied for two or more ranks, assign the *mean* of those ranks to be each datum's rank.

The Mann–Whitney U test for two independent samples

1. Call the smaller group 'group 1', and the larger group 'group 2', so $n_1 < n_2$. (If $n_1 = n_2$, choose at random.)
2. Calculate the sum of the ranks of group 1 (= R_1) and group 2 (= R_2).
3. $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$
4. $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$
5. The Mann–Whitney statistic U is the smaller of U_1 and U_2.

As a check, verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$. The null hypothesis is that the two samples come from identical populations.

The Wilcoxon matched-pairs signed-rank test for two related samples

1. Calculate the difference score for each pair of samples.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or –).
4. Add up all the ranks for difference scores that were positive; call this T^+.
5. Add up all the ranks for difference scores that were negative; call this T^-.
6. The Wilcoxon matched-pairs statistic T is the smaller of T^+ and T^-.

As a check, verify that $T^+ + T^- = \frac{n(n+1)}{2}$. The null hypothesis is that the difference scores are symmetrically distributed about zero.

The Wilcoxon signed-rank test for one sample

Calculate a difference score ($x - M$) for each score x, and proceed as above. The null hypothesis is that the scores are symmetrically distributed with a median of M.

Chi-square (χ^2) test

Regardless of the type of test,

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$ where O = observed value, E = expected value.

For a goodness-of-fit test (one categorical variable; the expected proportions in each category are known beforehand) with c categories, there are $c - 1$ degrees of freedom.

For a contingency test (two categorical variables) with R rows and C columns, there are $(R - 1)(C - 1)$ degrees of freedom, and the expected values are given by

$$E(row_i, column_j) = \frac{R_i C_j}{n}$$

where R_i is the row total for row i, C_j is the column total for row j, and n is the total number of observations.

The χ^2 test assumes equal independence of observations, normality (no values of E less than 5), and inclusion of all observations (including non-occurrences).

Confidence intervals

Normal distribution; population mean (μ) and SD (σ) known

Confidence intervals = $\mu \pm \sigma Z_{\text{critical}}$. (For 95% confidence intervals, $Z_{\text{critical}} = 1.96$.)

Normal distribution; sample mean and SD known

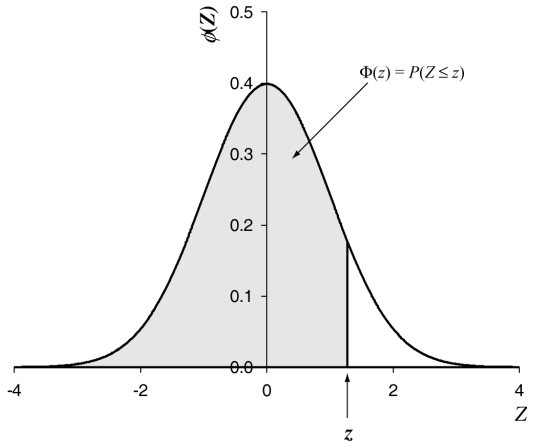
Confidence intervals = $\bar{x} \pm \frac{s_X}{\sqrt{n}} t_{critical(n-1)df}$. (For 95% confidence intervals, use t for $\alpha = 0.05$ two-tailed.)

The standard normal distribution, $Z = N(0,1)$

Mean = 0. Standard deviation = 1 (i.e. one Z point = one SD).

Cumulative distribution function $\Phi(z)$ is the area under the probability density function to the left of z (see figure).

This table gives the cumulative distribution function. **"If I know a Z score, what is the probability that a number $\leq z$ comes from a standard normal distribution?"**



If you have a Z score of 1.14, read down the left-hand side until you get to the row labelled '1.1', then read across until you get to the column labelled '0.04'. The number you reach is $\Phi(1.14)$. If you have a **negative Z score, $-z$,** calculate $1 - \Phi(z)$. For example, the probability associated with a Z score of –1.91 is (1 – 0.9719) = 0.0281. **If you want to know the probability that a number > z comes from a standard normal distribution, it's 1 minus the probability that a number $\leq z$ comes from the distribution. The 'significance level' of a Z score is the probability that a number *equal to or more extreme* than Z ($\geq z$ if z is positive, $\leq z$ if z is negative) comes from this distribution.**

Z	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	**0.9750**	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	**0.9951**	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	**0.9995**	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Source: Microsoft Excel 97 NORMSDIST function

Probabilities corresponding to a two-tailed α of 0.05, 0.01, and 0.001 are shown in bold. (These correspond to an α for each tail of 0.025, 0.005, and 0.0005.)

Spearman's correlation coefficient for ranked data, r_s

Here are the **critical values of $|r_s|$** (the absolute magnitude of r_s, ignoring any + or – sign) for different values of n and α. If your value of $|r_s|$ is **bigger than** the critical value, you would reject the null hypothesis. (If the entry in the table is blank, it is not possible to reject the null hypothesis.)

One-tailed α	**0.05**	**0.025**	**0.01**	**0.005**
Two-tailed α	**0.10**	**0.05**	**0.02**	**0.01**
n				
1				
2				
3				
4				
5	0.900			
6	0.829	0.886	0.943	
7	0.714	0.786	0.893	
8	0.643	0.738	0.833	0.881
9	0.600	0.683	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.523	0.623	0.736	0.818
12	0.497	0.591	0.703	0.780
13	0.475	0.566	0.673	0.745
14	0.457	0.545	0.646	0.716
15	0.441	0.525	0.623	0.689
16	0.425	0.507	0.601	0.666
17	0.412	0.490	0.582	0.645
18	0.399	0.476	0.564	0.625
19	0.388	0.462	0.549	0.608
20	0.377	0.450	0.534	0.591
21	0.368	0.438	0.521	0.576
22	0.359	0.428	0.508	0.562
23	0.351	0.418	0.496	0.549
24	0.343	0.409	0.485	0.537
25	0.336	0.400	0.475	0.526
26	0.329	0.392	0.465	0.515
27	0.323	0.385	0.456	0.505
28	0.317	0.377	0.448	0.496
29	0.311	0.370	0.440	0.487
30	0.305	0.364	0.432	0.478

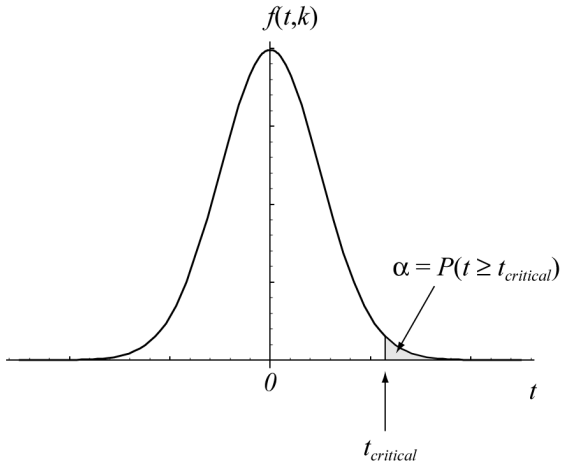
Source: Olds EG (1938), Annals of Mathematical Statistics 9. Note: there is considerable variation in published tables of critical values of $|r_s|$, because computing them is very difficult and there are many techniques for computing approximate values.

If $n > 30$, calculate a value of t instead:

$$t_{n-2} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^{\,2}}}$$

and test this using the tables of the t distribution with $n - 2$ degrees of freedom.

The *t* distribution



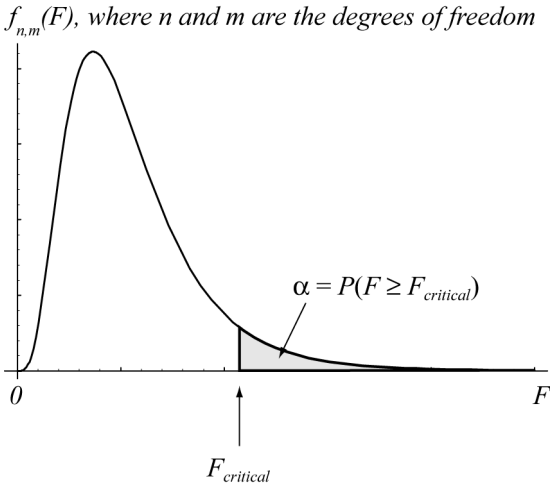
There would not be space here to give a *p* value for every possible combination of a *t* score and a certain number of degrees of freedom. So here are the **critical values** of *t* for different values of degrees of freedom and α. If your value of *t* is **bigger than** the critical value, you would reject the null hypothesis. If you have a **negative** value of *t*, just drop the minus sign (the *t* distribution is symmetrical about $t = 0$).

One-tailed α:	**0.05**	**(0.025)**	**0.01**	**(0.005)**
Two-tailed α:	**0.1**	**0.05**	**(0.02)**	**0.01**
df				
1	6.314	12.706	31.821	63.656
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
⋮	⋮	⋮	⋮	⋮
∞	1.645	1.960	2.326	2.576

Source: Microsoft Excel 97 TINV function, except for ∞ row (NORMSDIST function)

(Explanation of '∞' entry: for ∞ *df*, critical values of *t* are the same as critical values of *z*, because the *t* distribution approaches a normal distribution as $df \to \infty$.)

The *F* distribution



There would not be space here to give a *p* value for every possible combination of a *F* score, a certain number of degrees of freedom (numerator and denominator), and α. So here are the **critical values** of *F* for different values of degrees of freedom and α. There are three tables, for three different levels of α. If your value of *F* is **bigger than** the critical value, you would reject the null hypothesis.

Critical values of *F*, α = 0.05 (one-tailed, e.g. for ANOVA), equivalent to α = 0.1 if used for a two-tailed test ('having put the bigger variance on top, are they different?')

	Numerator *df*																
Denominator *df*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**15**	**20**	**25**	**30**	**40**	**50**
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	…	245.95	248.02	249.26	250.10	251.14	251.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	…	19.43	19.45	19.46	19.46	19.47	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	…	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	…	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	…	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	…	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	…	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	…	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	…	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	…	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	…	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	…	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	…	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	…	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	…	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	…	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	…	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	…	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	…	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	…	2.20	2.12	2.07	2.04	1.99	1.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	…	2.15	2.07	2.02	1.98	1.94	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	…	2.11	2.03	1.97	1.94	1.89	1.86
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	…	2.07	1.99	1.94	1.90	1.85	1.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	…	2.04	1.96	1.91	1.87	1.82	1.79
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	…	2.01	1.93	1.88	1.84	1.79	1.76
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	…	1.92	1.84	1.78	1.74	1.69	1.66
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	…	1.87	1.78	1.73	1.69	1.63	1.60
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	…	1.84	1.75	1.69	1.65	1.59	1.56
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	…	1.75	1.66	1.60	1.55	1.50	1.46
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	…	1.72	1.62	1.56	1.52	1.46	1.41
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	…	1.69	1.59	1.53	1.48	1.42	1.38
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	…	1.68	1.58	1.52	1.47	1.41	1.36

Source: Microsoft Excel 97 FINV function

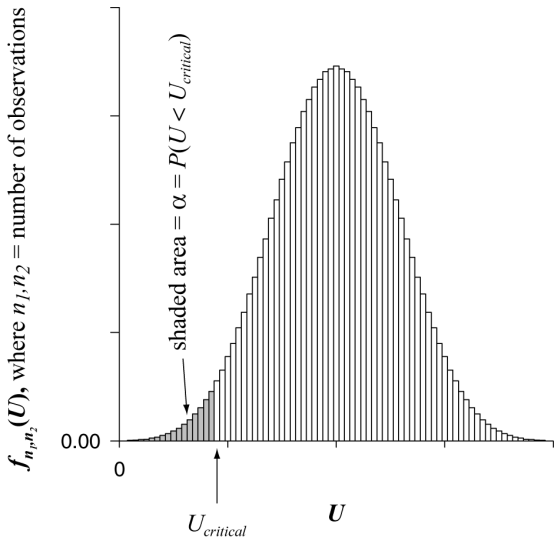
Critical values of F, α = 0.025 (one-tailed, e.g. for ANOVA), equivalent to α = 0.05 if used for a two-tailed test ('having put the bigger variance on top, are they different?')

	Numerator *df*																
Denominator *df*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**15**	**20**	**25**	**30**	**40**	**50**
1	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	…	984.87	993.08	998.09	1001.4	1005.6	1008.1
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	…	39.43	39.45	39.46	39.46	39.47	39.48
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	…	14.25	14.17	14.12	14.08	14.04	14.01
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	…	8.66	8.56	8.50	8.46	8.41	8.38
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	…	6.43	6.33	6.27	6.23	6.18	6.14
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	…	5.27	5.17	5.11	5.07	5.01	4.98
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	…	4.57	4.47	4.40	4.36	4.31	4.28
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	…	4.10	4.00	3.94	3.89	3.84	3.81
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	…	3.77	3.67	3.60	3.56	3.51	3.47
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	…	3.52	3.42	3.35	3.31	3.26	3.22
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	…	3.33	3.23	3.16	3.12	3.06	3.03
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	…	3.18	3.07	3.01	2.96	2.91	2.87
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	…	3.05	2.95	2.88	2.84	2.78	2.74
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	…	2.95	2.84	2.78	2.73	2.67	2.64
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	…	2.86	2.76	2.69	2.64	2.59	2.55
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	…	2.79	2.68	2.61	2.57	2.51	2.47
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	…	2.72	2.62	2.55	2.50	2.44	2.41
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	…	2.67	2.56	2.49	2.44	2.38	2.35
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	…	2.62	2.51	2.44	2.39	2.33	2.30
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	…	2.57	2.46	2.40	2.35	2.29	2.25
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	…	2.50	2.39	2.32	2.27	2.21	2.17
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	…	2.44	2.33	2.26	2.21	2.15	2.11
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	…	2.39	2.28	2.21	2.16	2.09	2.05
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	…	2.34	2.23	2.16	2.11	2.05	2.01
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	…	2.31	2.20	2.12	2.07	2.01	1.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	…	2.18	2.07	1.99	1.94	1.88	1.83
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	…	2.11	1.99	1.92	1.87	1.80	1.75
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	…	2.06	1.94	1.87	1.82	1.74	1.70
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	…	1.94	1.82	1.75	1.69	1.61	1.56
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	…	1.90	1.78	1.70	1.64	1.56	1.51
500	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	…	1.86	1.74	1.65	1.60	1.52	1.46
1000	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	…	1.85	1.72	1.64	1.58	1.50	1.45

Critical values of F, α = 0.01 (one-tailed, e.g. for ANOVA), equivalent to α = 0.02 if used for a two-tailed test ('having put the bigger variance on top, are they different?')

	Numerator *df*																
Denominator *df*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**15**	**20**	**25**	**30**	**40**	**50**
1	4052.1	4999.3	5403.5	5624.2	5763.9	5858.9	5928.3	5980.9	6022.4	6055.9	…	6156.9	6208.6	6239.8	6260.3	6286.4	6302.2
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	…	99.43	99.45	99.46	99.47	99.48	99.48
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	…	26.87	26.69	26.58	26.50	26.41	26.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	…	14.20	14.02	13.91	13.84	13.75	13.69
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	…	9.72	9.55	9.45	9.38	9.29	9.24
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	…	7.56	7.40	7.30	7.23	7.14	7.09
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	…	6.31	6.16	6.06	5.99	5.91	5.86
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	…	5.52	5.36	5.26	5.20	5.12	5.07
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	…	4.96	4.81	4.71	4.65	4.57	4.52
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	…	4.56	4.41	4.31	4.25	4.17	4.12
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	…	4.25	4.10	4.01	3.94	3.86	3.81
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	…	4.01	3.86	3.76	3.70	3.62	3.57
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	…	3.82	3.66	3.57	3.51	3.43	3.38
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	…	3.66	3.51	3.41	3.35	3.27	3.22
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	…	3.52	3.37	3.28	3.21	3.13	3.08
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	…	3.41	3.26	3.16	3.10	3.02	2.97
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	…	3.31	3.16	3.07	3.00	2.92	2.87
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	…	3.23	3.08	2.98	2.92	2.84	2.78
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	…	3.15	3.00	2.91	2.84	2.76	2.71
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	…	3.09	2.94	2.84	2.78	2.69	2.64
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	…	2.98	2.83	2.73	2.67	2.58	2.53
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	…	2.89	2.74	2.64	2.58	2.49	2.44
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	…	2.81	2.66	2.57	2.50	2.42	2.36
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	…	2.75	2.60	2.51	2.44	2.35	2.30
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	…	2.70	2.55	2.45	2.39	2.30	2.25
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	…	2.52	2.37	2.27	2.20	2.11	2.06
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	…	2.42	2.27	2.17	2.10	2.01	1.95
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	…	2.35	2.20	2.10	2.03	1.94	1.88
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	…	2.19	2.03	1.93	1.86	1.76	1.70
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	…	2.13	1.97	1.87	1.79	1.69	1.63
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	…	2.07	1.92	1.81	1.74	1.63	1.57
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	…	2.06	1.90	1.79	1.72	1.61	1.54

The Mann–Whitney *U* statistic



Here are the **critical values** of U for different values of n_1 and n_2. Only critical values for $\alpha = 0.05$ (two-tailed) are given. If your value of U is **smaller than** the critical value, you would reject the null hypothesis. (If the value shown in the table is zero, it is not possible to reject the null hypothesis, since U cannot be smaller than zero.)

Critical values of U, $\alpha = 0.05$ (two-tailed) or $\alpha = 0.025$ (one-tailed)

n_2 / n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2		0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	3	3	3	3
3			0	0	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
4				1	2	3	4	5	5	6	7	8	9	10	11	12	12	13	14	15
5					3	4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
6						6	7	9	11	12	14	15	17	18	20	22	23	25	26	28
7							9	11	13	15	17	19	21	23	25	27	29	31	33	35
8								14	16	18	20	23	25	27	30	32	35	37	39	42
9									18	21	24	27	29	32	35	38	40	43	46	49
10										24	27	30	34	37	40	43	46	49	53	56
11											31	34	38	41	45	48	52	56	59	63
12												38	42	46	50	54	58	62	66	70
13													46	51	55	60	64	68	73	77
14														56	60	65	70	75	79	84
15															65	71	76	81	86	91
16																76	82	87	93	99
17																	88	94	100	106
18																		100	107	113
19																			114	120
20																				128

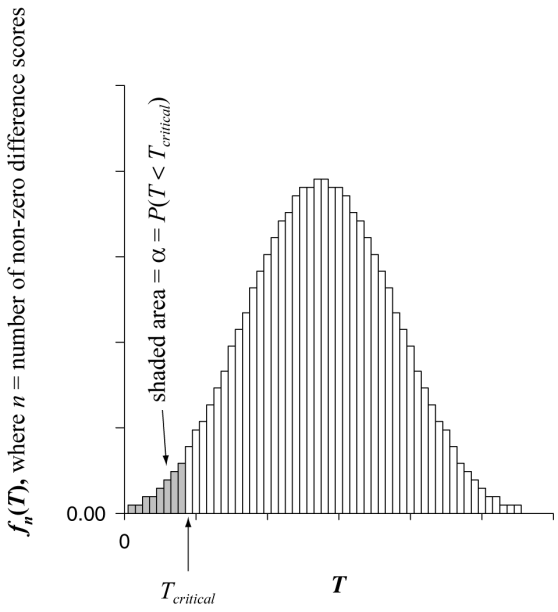
Source: R (http://www.r-project.org/), qwilcox(one-tailed α, n_1, n_2) gives q such that $P(U < q) \leq \alpha$.

If $n_2 > 20$, use the normal approximation instead. Calculate a Z score

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

and test this using the tables of the standard normal distribution Z.

The Wilcoxon signed-rank *T* statistic



Here are the **critical values** of *T* for different values of *n* (where *n* is the number of non-zero difference scores) and *α*. If your value of *T* is **smaller than** the critical value, you would reject the null hypothesis. (If the value shown in the table is zero, it is not possible to reject the null hypothesis, since *T* cannot be smaller than zero.)

One-tailed *α*	**0.05**	**0.025**	**0.01**	**0.005**
Two-tailed *α*	**0.10**	**0.05**	**0.02**	**0.01**
n				
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	1	0	0	0
6	3	1	0	0
7	4	3	1	0
8	6	4	2	1
9	9	6	4	2
10	11	9	6	4
11	14	11	8	6
12	18	14	10	8
13	22	18	13	10
14	26	22	16	13
15	31	26	20	16
16	36	30	24	20
17	42	35	28	24
18	48	41	33	28
19	54	47	38	33
20	61	53	44	38
21	68	59	50	43
22	76	66	56	49
23	84	74	63	55
24	92	82	70	62
25	101	90	77	69

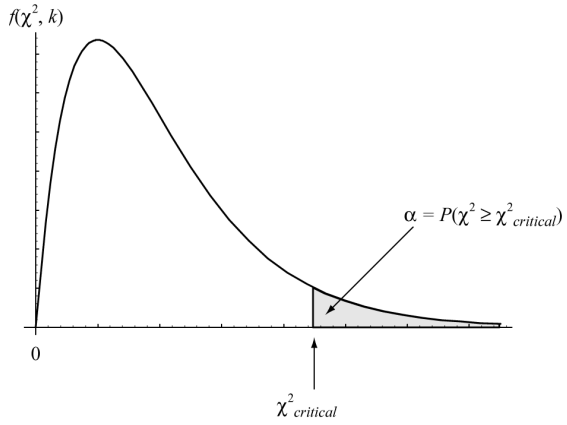
Source: R (http://www.r-project.org/), qsignrank(one-tailed α, n) gives q such that P(T < q) ≤ α.

If *n* > 25, use the normal approximation instead. Calculate a *Z* score

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

and test this using the tables of the standard normal distribution *Z*.

The χ^2 distribution



There would not be space here to give a *p* value for every possible combination of a χ^2 score and a certain number of degrees of freedom. So here are the **critical values** of χ^2 for different values of degrees of freedom (k) and α. If your value of χ^2 is **bigger than** the critical value, you would reject the null hypothesis.

	α		
d.f.	**0.05**	**0.01**	**0.001**
1	3.84	6.63	10.83
2	5.99	9.21	13.82
3	7.81	11.34	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.51
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.12
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.68	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.31
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
26	38.89	45.64	54.05
27	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70
⋮	⋮	⋮	⋮
40	55.76	63.69	73.40
50	67.50	76.15	86.66
60	79.08	88.38	99.61
70	90.53	100.43	112.32
80	101.88	112.33	124.84

Source: Microsoft Excel 97 CHIINV function

(end of tables and formulae booklet)

References

If you want to follow something up in a textbook, I recommend starting with Howell (1997).

Abelson, R. P. (1995). *Statistics As Principled Argument*, Lawrence Erlbaum, Hillsdale, New Jersey.

Allen, J. J. B., Schnyer, R. N. & Hitt, S. K. (1998). The efficacy of acupuncture in the treatment of major depression in women. *Psychological Science* **9:** 397-401.

Bacon, F. (1620). *Novum Organon*.

Bland, J. M. & Altman, D. G. (1994). Some examples of regression towards the mean. *British Medical Journal* **309:** 780.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin* **57:** 49-64.

Edwards, A. W. F. (1986). More on the too-good-to-be-true paradox and Gregor Mendel. *The Journal of Heredity* **77:** 138.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? (Reprinted in Stern, C. and Sherwood, E.R. The Origin of Genetics: A Mendel Source Book. W.H. Freeman and Co., San Francisco and London). *Annals of Science* **1:** 115-137.

Frank, H. & Althoen, S. C. (1994). *Statistics: Concepts and Applications*, Cambridge, Cambridge University Press.

Howell, D. C. (1997). *Statistical Methods for Psychology*. Fourth edition, Wadsworth, Belmont, California.

Kaptchuk, T. J. (2003). Effect of interpretive bias on research evidence. *British Medical Journal* **326:** 1453-1455.

Keppel, G. (1991). *Design and analysis: a researcher's handbook*. Third edition, Prentice-Hall, London.

Mayberg, H. S., Brannan, S. K., Tekell, J. L., Silva, J. A., Mahurin, R. K., McGinnis, S. & Jerabek, P. A. (2000). Regional metabolic effects of fluoxetine in major depression: Serial changes and relationship to clinical response. *Biological Psychiatry* **48:** 830-843.

Mendel, J. G. (1866). Versuche über Pflanzenhybriden. *Verhandlungen des Naturforschenden Vereins Brünn* **4:** 3-47.

Myers, J. L. & Well, A. D. (1995). *Research Design and Statistical Analysis*, Lawrence Erlbaum, Hillsdale, New Jersey.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*, John Wiley & Sons, Chichester.

Pfungst, O. (1907). *Das Pferd des Herrn von Osten (Der Kluge Hans). Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie*, J.A. Barth, Leipzig.

Pilgrim, I. (1984). The too-good-to-be-true paradox and Gregor Mendel. *The Journal of Heredity* **75:** 501-502.

Pilgrim, I. (1986a). Rebuttal to A.W.F. Edwards' Communication. *The Journal of Heredity* **77:** 138.

Pilgrim, I. (1986b). A Solution to the Too-good-to-be-true Paradox and Gregor Mendel. *The Journal of Heredity* **77:** 218-220.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2:** 110–114.

Velleman, P. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading for classifying statistical methodology. *The American Statistician* **47**.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29:** 350-362.

Winer, B. J. (1971). *Statistical principles in experimental design*. Second edition, McGraw-Hill, New York.