NST 1B Experimental Psychology
Statistics practical 5

Statistics: revision

Rudolf Cardinal & Mike Aitken
29 / 30 April 2004
Department of Experimental Psychology
University of Cambridge

Handouts:

- Answers to Examples 5 (χ^2)
- Answers to Examples 6 (past papers + exp. design)
- Answers to Examples 7 (mixed)

pobox.com / ~rudolf / psychology

Background: null hypothesis testing

Reminder: the logic of null hypothesis testing

Research hypothesis (H_1): e.g. measure weights of 50 joggers and 50 non-joggers; research hypothesis might be 'there is **a difference** between the weights of joggers and non-joggers'.

Null hypothesis (H_0): e.g. 'there is **no difference** between the population means of joggers and non-joggers; any observed differences are **due to chance.'**

Calculate probability of finding the observed data (e.g. difference) **if the null hypothesis is true.** This is the ***p* value.**

If *p* very small, reject null hypothesis ('chance alone is not a good enough explanation'). Otherwise, retain null hypothesis (Occam's razor: chance is the simplest explanation). **Criterion level of *p* is called α.**

Reminder: α, errors we can make, and power

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

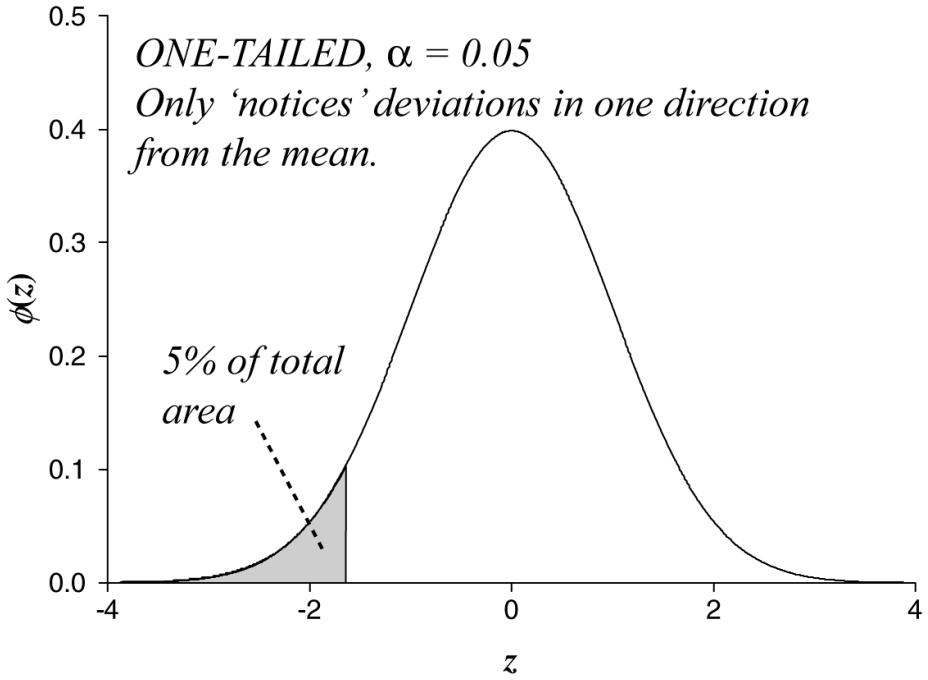
α is the probability of declaring something 'significant' when there's no genuine effect (of making a Type I error).

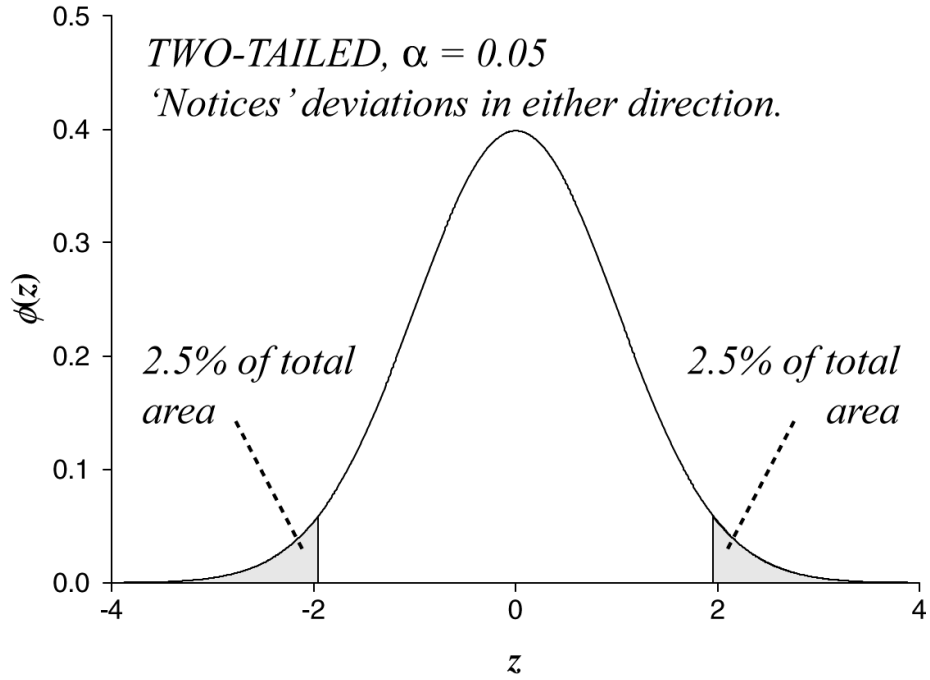
Power is the probability of finding a genuine effect (of *not* making a Type II error).

Power is higher with

- a big effect!
- large samples (big *n*)
- small variability (small σ)

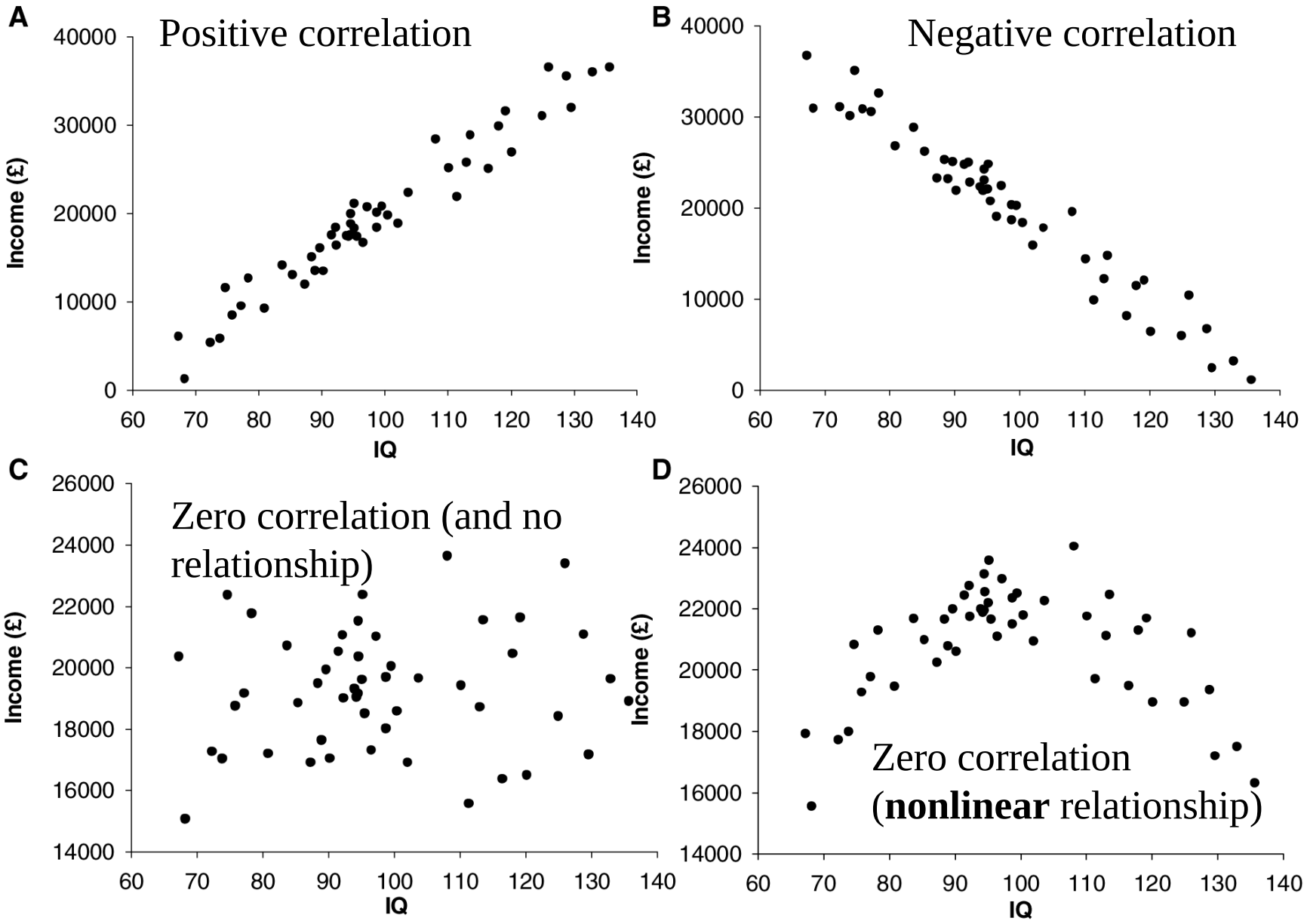
Reminder: one- and two-tailed tests





Correlation and regression

Scatter plots show the relationship between two variables



r, the Pearson product–moment correlation coefficient

$$r_{XY}$$

r varies from –1 to +1.

***r* does not depend on which way round *X* and *Y* are.**

Your calculator calculates *r*.

You should know how to use your calculator for statistical functions, including *r* and SDs.

For **descriptive statistics** (mean, SD, etc.):

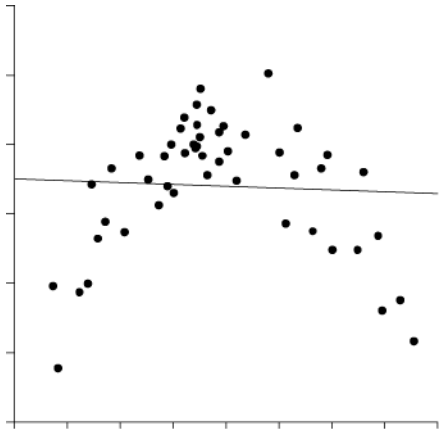
	Casio fx115s	Other Casio models
Enter descriptive statistics (SD) mode	MODE 2	MODE →SD
Clear the stats memory	SHIFT Scl AC	SHIFT Scl AC =
Enter values of x (e.g. 53)	5 3 M+ (DATA DEL) etc.	
Read out descriptive statistics: mean *(see keypad and inside lid)*	SHIFT $\bar{x}$ 1	SHIFT $\bar{x}$ 1 =
sample SD (n–1 formula)	SHIFT $x\sigma_{n-1}$ n 3	SHIFT $x\sigma_{n-1}$ n 3 =
n	$x\sigma_{n-1}$ n RCL 3	C RCL hyp

For correlation and **linear regression** (r, a, b):

	Casio fx115s	Other Casio models
Enter linear regression (LR) mode	MODE 3	MODE →REG→Lin
Clear the stats memory	SHIFT Scl C	SHIFT Scl AC =
Enter values of x, y pairs (e.g. x = 53, y = 17)	5 3 [(--- (x_D, y_D) 1 7 M+ (DATA DEL) etc.	5 3 , 1 7 M+ etc.
Read out desired coefficients *(see keypad and inside lid)*		
r	SHIFT r 9	SHIFT r (=
a (A)	SHIFT A 7	SHIFT A 7 =
b (B)	SHIFT B 8	SHIFT B 8 =

Never give up! Never surrender! Never forget...

‘Zero correlation’ doesn’t imply ‘no relationship’.



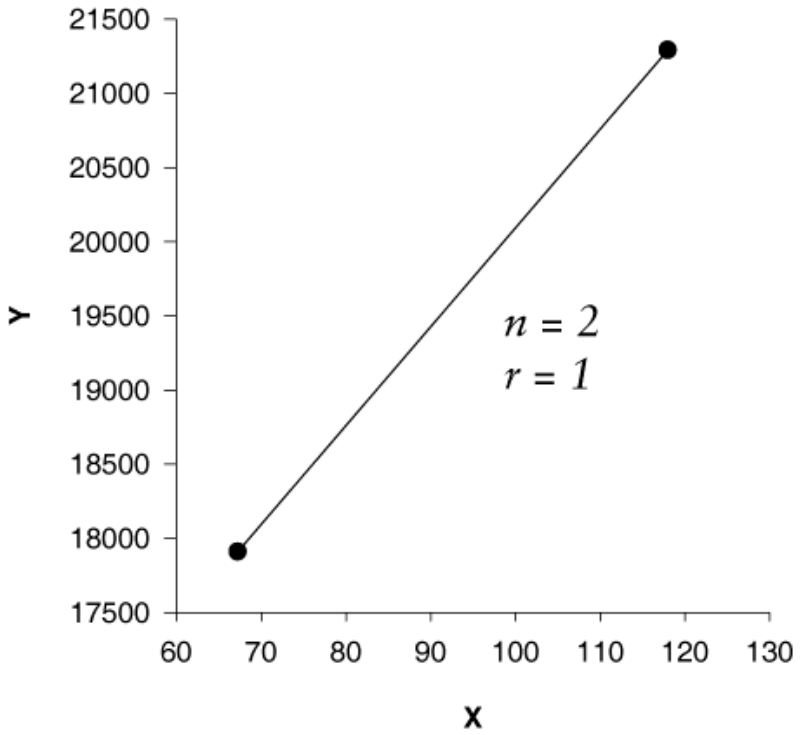
So always draw a scatter plot.

Correlation does not imply causation.

Adjusted *r*

You've calculated *r*, the correlation in your sample.

"What is your best estimate of the correlation in the population?"



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

‘Is my correlation significant?’ Our first *t* test.

Null hypothesis: the correlation in the population is zero ($\rho = 0$).
Calculate *t*:

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

Assumptions you make when you test hypotheses about ρ

Basically, the data shouldn't look too weird. We must assume

- that the variance of Y is roughly the same for all values of X.

- that X and Y are both normally distributed

- that for all values of X, the corresponding values of Y are normally distributed, and vice versa

r_s: Spearman's correlation coefficient for **ranked** data

This is a nonparametric version of correlation. You can use it when you obtain **ranked** data, or when you want to do significance tests on *r* but your data are *not* normally distributed.

- Rank the *X* values.
- Rank the *Y* values.
- Correlate the *X* **ranks** with the *Y* **ranks.** (You do this in the normal way for calculating *r*, but you call the result r_s.)

- To ask whether the correlation is 'significant', use the table of critical values of Spearman's r_s in the *Tables and Formulae* booklet.

How to rank data

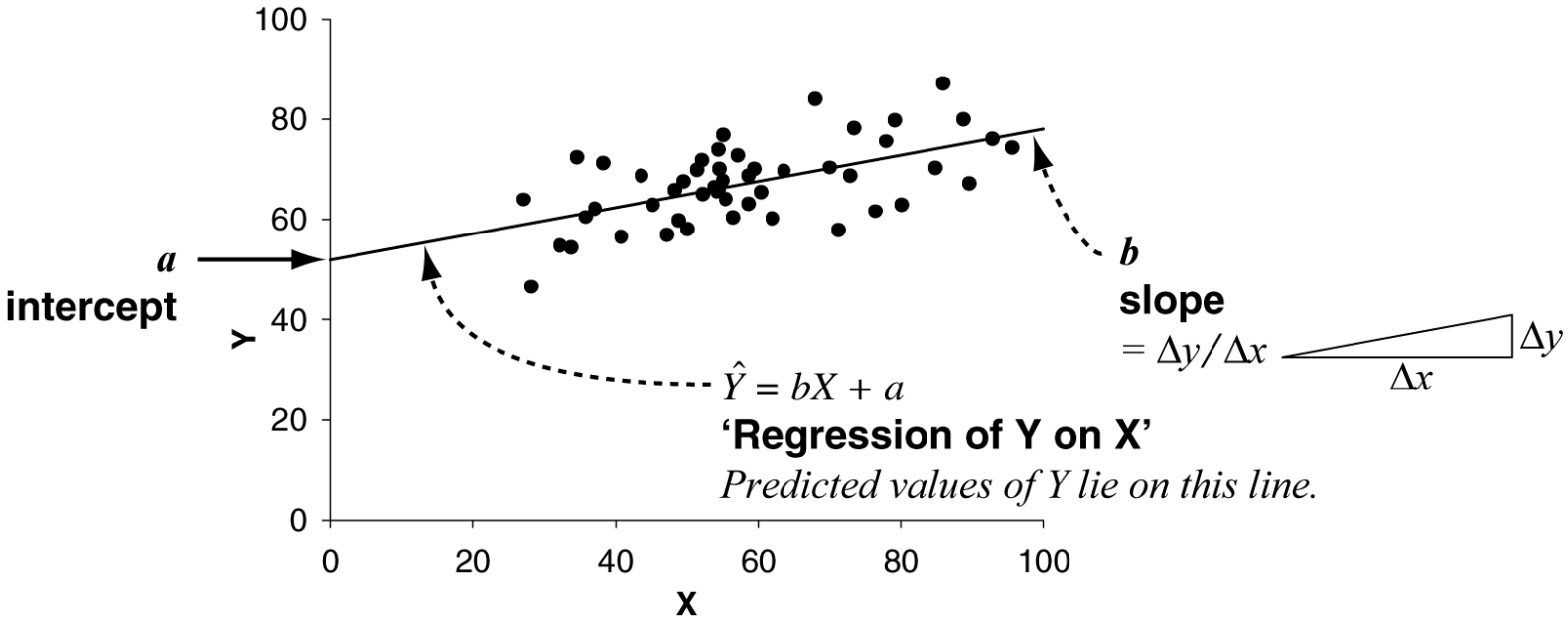
Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

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:

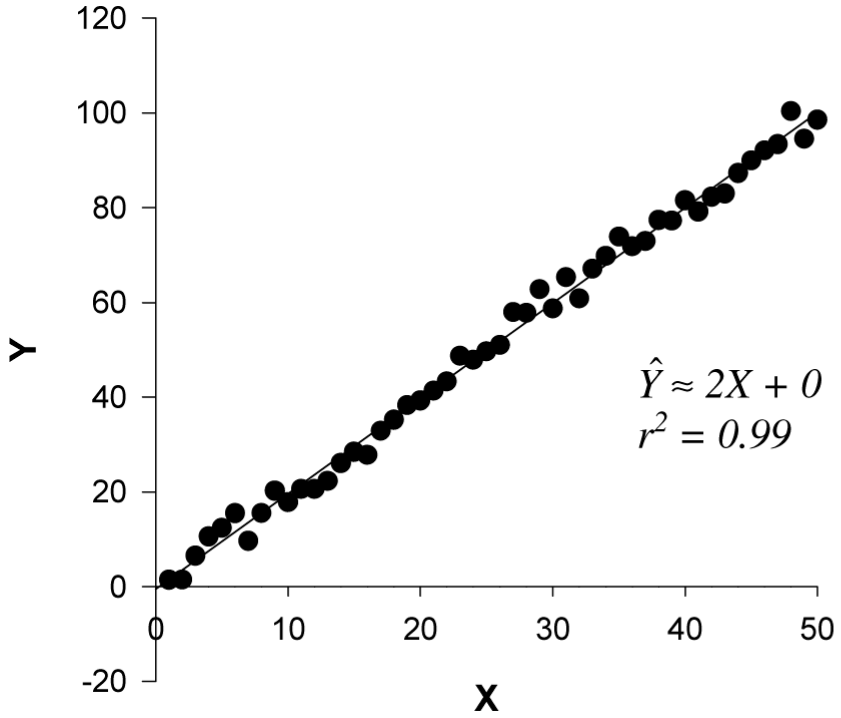
X:	5	8	9	12	12	15	16	16	16	17
rank:	1	2	3	4.5	4.5	6	8	8	8	10

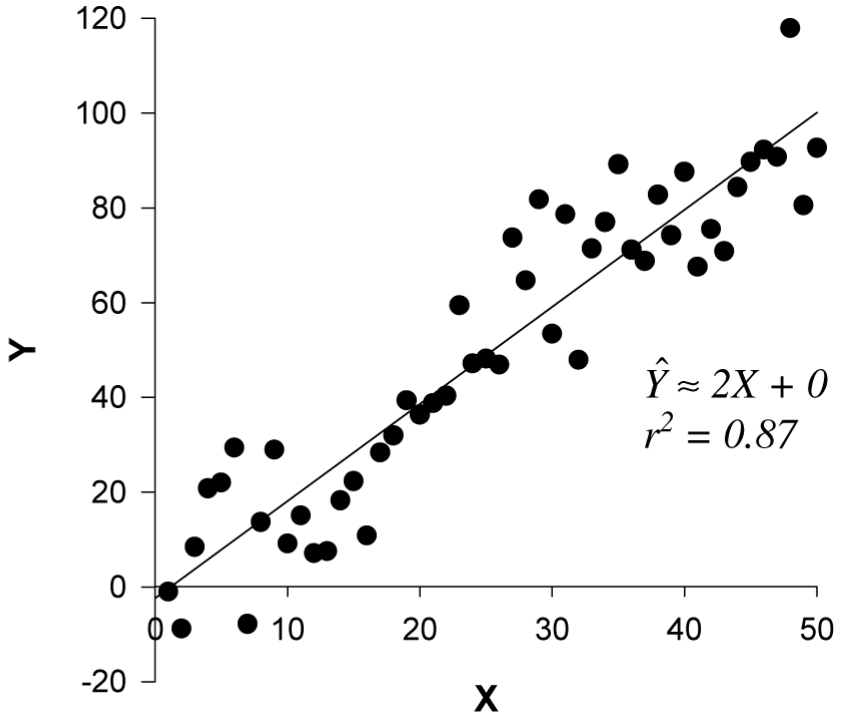
Regression: e.g. predicting *Y* from *X* (≠ predicting *X* from *Y*)



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

r^2 means something important — the proportion of the variability in Y predictable from the variability in X





Parametric difference tests (t tests)

The one-sample *t* test

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

df for this test

sample mean

test value

standard error of the mean (SEM) (standard deviation of the distribution of sample means)

sample SD

The null hypothesis is that the sample comes from a population with mean μ.

Look up the critical value of t (for a given α) using your tables of t for the correct number of degrees of freedom ($n - 1$). If your $|t|$ is bigger, it's significant.

The two-sample, paired *t* test

Calculate the **differences** between each pair of observations. Then perform a **one-sample** *t* test on the differences, comparing them to zero. (Null hypothesis: the mean difference is zero.)

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

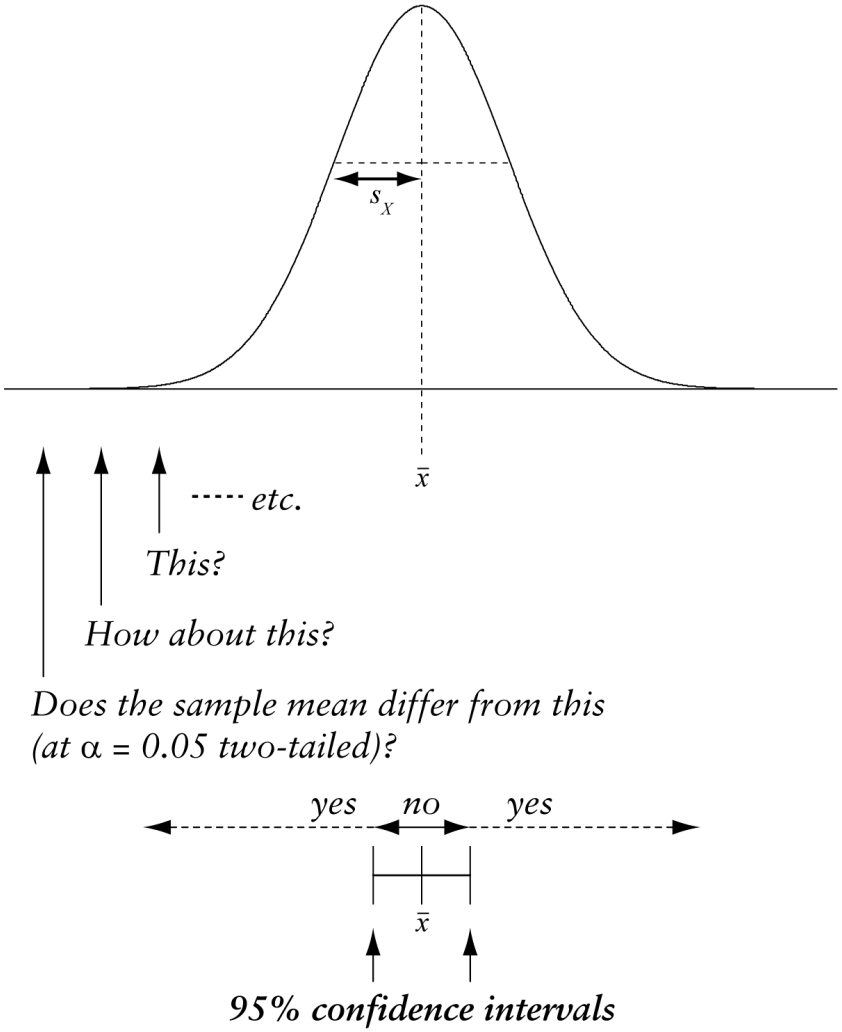
test value for the differences (**zero** for the null hypothesis 'there is no difference')

Confidence intervals using *t*

If we know the mean and SD of a sample, we could perform a *t* test to see if it differed from a given number. We could repeat that for every possible number...

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

therefore $\mu = \bar{x} \pm \left(t_{\text{critical for } n-1\, df} \times \dfrac{s_X}{\sqrt{n}} \right)$

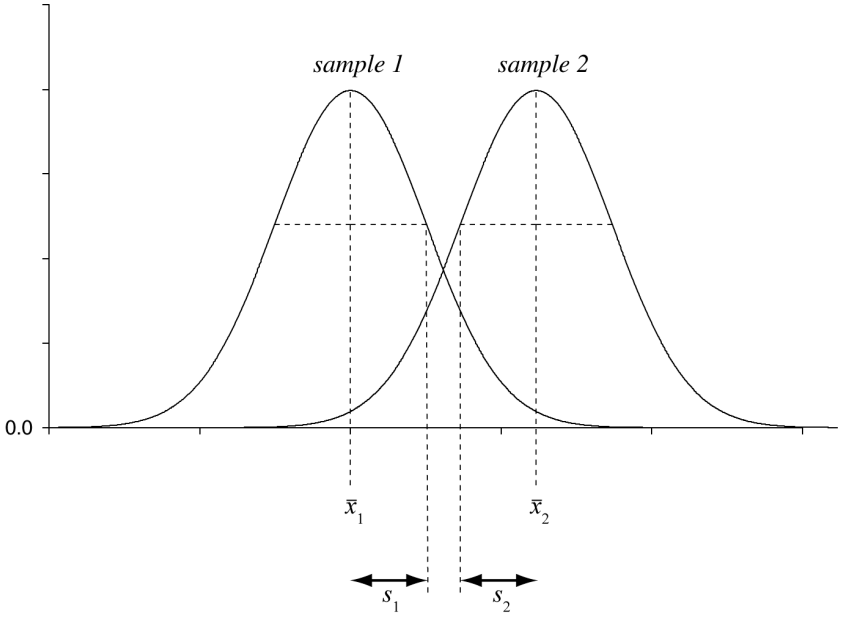


This means that there is a **95% chance that the true population mean** is within this confidence interval.

The two-sample, unpaired *t* test

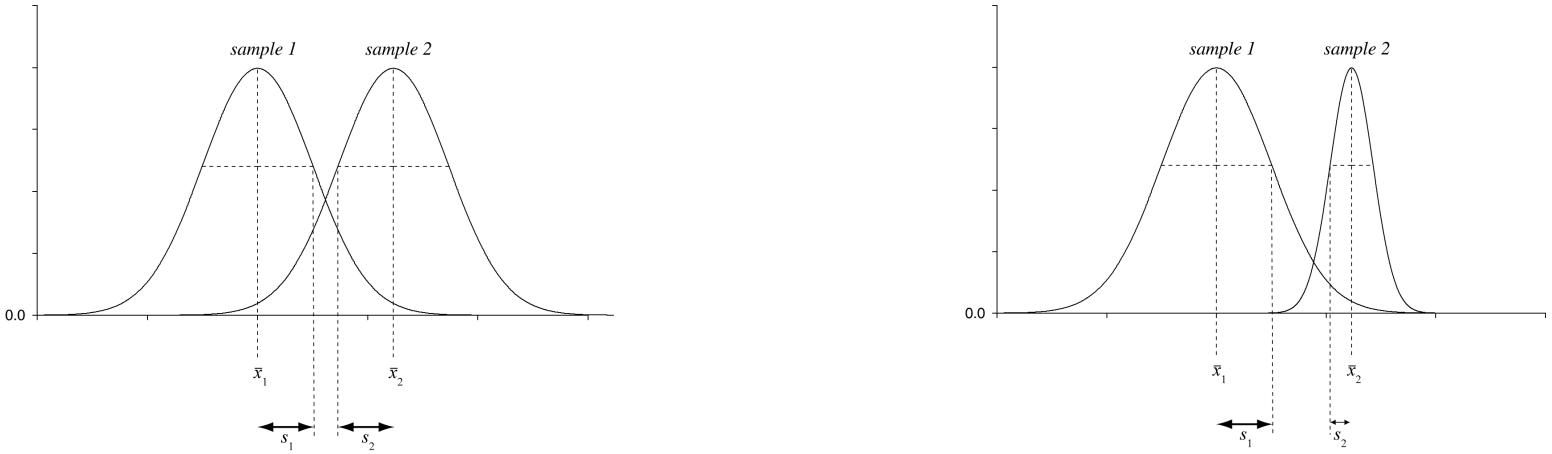
Two **independent** samples.
Are they different?
Null hypothesis: both samples come from underlying **populations** with the same mean.

Basically, if the sample means are very far apart, as measured by something that depends (somehow) on the variability of the samples, then we will reject the null hypothesis.



The two-sample, unpaired *t* test

The format of the *t* test depends (unfortunately) on whether the two samples have the same variance.



Formulae are on the Formula Sheet.

‘Are the variances equal or not?’ The *F* test

To perform an *F* test, put the **bigger variance on top:**

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

$$\text{swap everything around if } s_1^2 < s_2^2$$

Null hypothesis is that the variances are the same ($F = 1$). If our *F* exceeds the critical *F* for the relevant number of *df* (note that there are separate *df* for the numerator and the denominator), we reject the null hypothesis. Since we have ensured that $F \geq 1$, we run a one-tailed test on *F* — so double the stated one-tailed α to get the two-tailed α for the question ‘are the variances *different?*’.

Assumptions of the *t* test

- The mean is meaningful.

- The underlying scores (for one-sample and unpaired *t* tests) or difference scores (for paired *t* tests) are **normally distributed.**

Rule of thumb: if $n > 30$, you're fine to assume this. If $n > 15$ and the data don't look too weird, it's probably OK. Otherwise, bear this in mind.

- To use the equal-variance version of the unpaired two-sample *t* test, the two samples must come from populations with equal variances (whether or not $n_1 = n_2$).

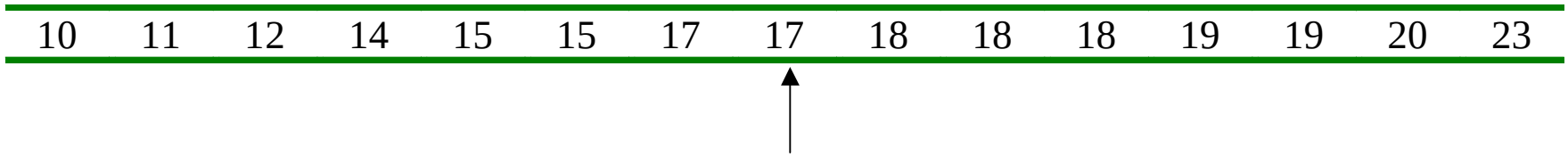
Nonparametric difference tests

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

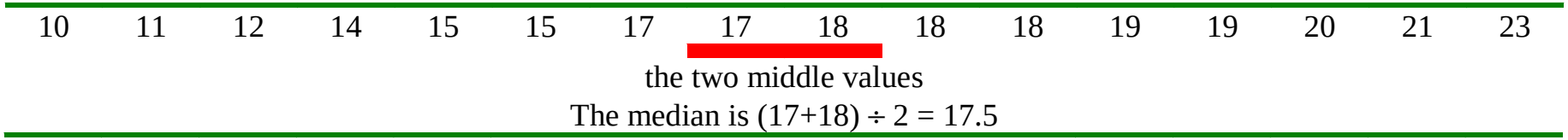
The median

The **median** is the value at or below which 50% of the scores fall when the data are arranged in numerical order.

If *n* is odd, it's the middle value (here, 17):



If *n* is even, it's the mean of the two middle values (here, 17.5):



Medians are less affected by outliers than means

Ranking removes ‘distribution’ information

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.

Two unrelated samples: the Mann–Whitney *U* test

Calculating *U*: see formula sheet.

Determining a significance level from *U*: see formula sheet.

- If $n_2 \leq 20$, look up the **critical value** for *U* in your tables. (The critical value depends on n_1 and n_2.) If your *U* is **smaller** than the critical value, it's significant (you reject the null hypothesis).
- If $n_2 > 20$, use the **normal** approximation (see formula sheet).

Null hypothesis: the two samples were drawn from identical populations. If we assume the distributions are similar, a significant Mann–Whitney test suggests that the **medians** of the two populations are different.

Time-saving tip…

If the ranks **do not overlap at all,** $U = 0$.

Example:

Group A	55	65	75	80	82.5
Ranks	**6**	**7**	**8**	**9**	**10**
Group B	10	15	39	40	48
Ranks	**1**	**2**	**3**	**4**	**5**

$$U = 0$$

If you find a significant difference…

If you conduct a Mann–Whitney test and find a significant difference, **which group had the larger median and which group had the smaller median?**

You have to calculate the medians; you can't tell from the rank sums.

Two related samples: Wilcoxon matched-pairs signed-rank test

Calculating *T*: see formula sheet.

Determining a significance level from *T*: see formula sheet.

- If $n \leq 25$, look up the **critical value** for T in your tables. If your T is **smaller** than the critical value, it's significant (you reject the null hypothesis).
- If $n > 25$, use the **normal** approximation (see formula sheet).

Null hypothesis: the distribution of differences between the pairs of scores is symmetric about zero. Since the median and mean of a symmetric population are the same, this can be restated as 'the differences between the pairs of scores are symmetric with a mean and median of zero'.

One sample: Wilcoxon signed-rank test with only one sample

Very easy.

Null hypothesis: the median is equal to M.

For each score x, calculate a difference score $(x - M)$. Then proceed as for the two-sample Wilcoxon test using these difference scores.

The χ^2 test: for categorical data

Goodness of fit test: ONE categorical variable

100 people choose between chocolate and garibaldi biscuits.
Expected (*E*): 50 chocolate, 50 garibaldi.
Observed (*O*): 65 chocolate, 35 garibaldi.

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

If χ^2 is big enough, we will reject the null hypothesis.

***df* = categories – 1**

Contingency test: TWO categorical variables

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

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*

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

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

***df* = (rows – 1) × (columns – 1)**

Assumptions of the χ^2 test

- **Independence of observations.**
  - **Mustn't** analyse data from several subjects when there are multiple observations per subject. Need one observation per subject. **Most common cock-up?**
  - Can analyse data from *only* one subject — then all observations are equally independent — but conclusions only apply to that subject.

- **Normality.** Rule of thumb: **no *E* value less than 5.**

- **Inclusion of non-occurrences.** Petrol station example:

Obtained values	Men	Women
Support booze	17	11

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

Experimental design questions

Experimental design questions: tips (see also Answers 6)

- No ‘right’ answer. Need to understand the science behind the question.
- What will you measure? What numbers will you actually write down?
- Subjects?
- Correlative (measurement) or causal (interventional, ‘true experimental’) study?

- For interventional studies, will you use a between- or within-groups design? Within-subjects designs are often more powerful but order effects may be a problem: need appropriate counterbalancing.
- Consider confounds (confounding variables). What is the appropriate control condition? Remember blinding and placebo / sham techniques.

- Keep it simple. Is your design the simplest way to answer the question? If you find an effect, will it be simple to interpret? If you don’t, what will that tell you?
- How will you analyse your data? What will your null hypothesis be? But remember, this is *not* the main focus of the questions!
- Will you need a series of experiments? Will you alter your plans based on the result of the first few experiments? Do you need to outline a plan?
- Consider ethics and practicality.
- If you think of problems with your design, discuss them.

Numerical past paper questions

Approach to numerical exam questions

- **Read the question carefully.**

- **Where appropriate, consider steps in choice of test.**

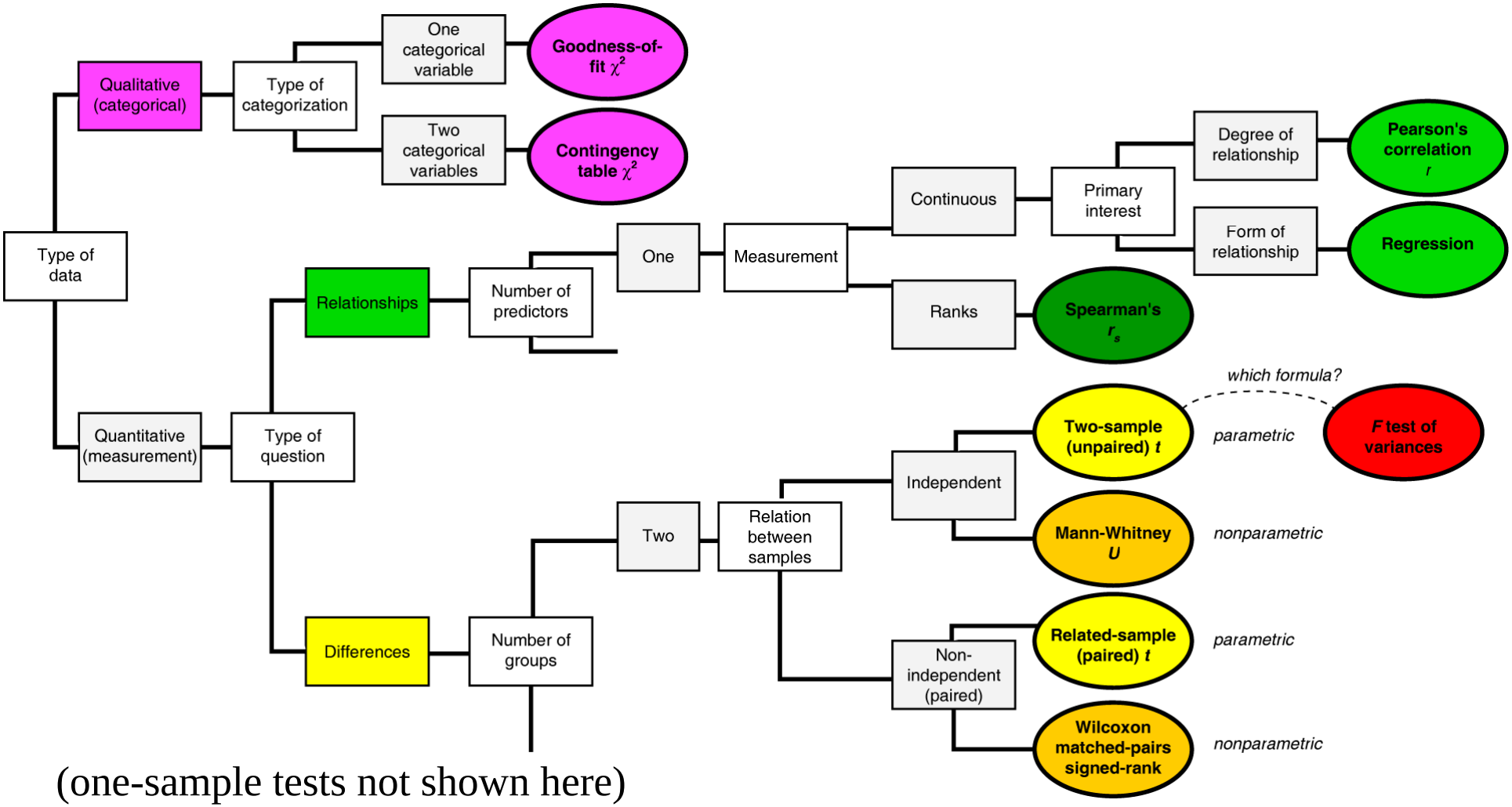
  e.g. What is question / hypothesis?
  Type of data?
  Number of groups / conditions?
  Correlation or differences between groups?
  Related or unrelated samples?
  Parametric / non-parametric test?
  Direction of test?

- **Choice of formula? Any other considerations?**

  e.g. perform *F* test to choose equal- / unequal-variance form of *t* test?

- **Which data to put in formulae?**

  if χ^2: draw table and note how to work out expected frequencies
  not all data has to go into each test (depends on question)
  sometimes you need to calculate difference scores
  sometimes you need to convert scores (e.g. percentages $\rightarrow$ raw scores)



(one-sample tests not shown here)

- **categorical** (= nominal, classificatory) data: use χ^2.
- **parametric tests** preferred (more sensitive) when assumptions are met. Need data on an interval / ratio scale; scores with roughly a normal distribution; etc.
- **nonparametric tests** have fewer assumptions (so use if the assumptions of a parametric test are violated) and can deal with ranked data.

Examples 6, 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

Examples 6, Q2 (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?

Examples 6, Q4 (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?

Examples 6, Q6 (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.

Good luck!