

NST 1B Experimental Psychology

Statistics practical 2

Difference tests (1): parametric

Rudolf Cardinal & Mike Aitken

2 / 3 December 2003; Department of Experimental Psychology

University of Cambridge

Handouts:

- Answers to Examples 2 (from last time)
- Handout 3 (diff. tests 1)
- Examples 3 (diff. tests 1)
- Homophone practical data

pobox.com/~rudolf/psychology



*These slides are on the web.
No need to scribble frantically.*

pobox.com/~rudolf/psychology

Reminder: basic principles



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; the population mean of joggers is not the same as the population mean of non-joggers’.

Usually very hard to calculate the probability of a research hypothesis (sometimes because they’re poorly specified — for example, how *big* a difference?).

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.**’

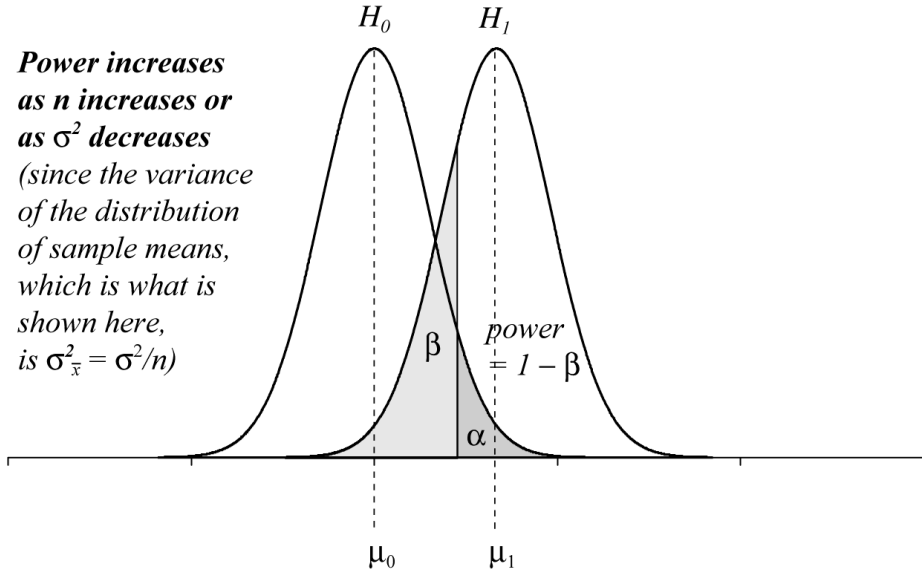
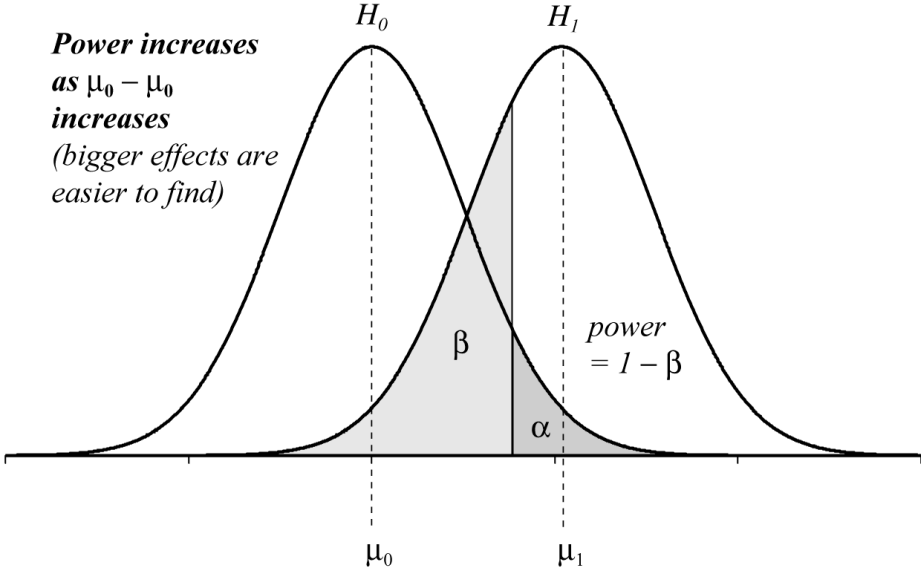
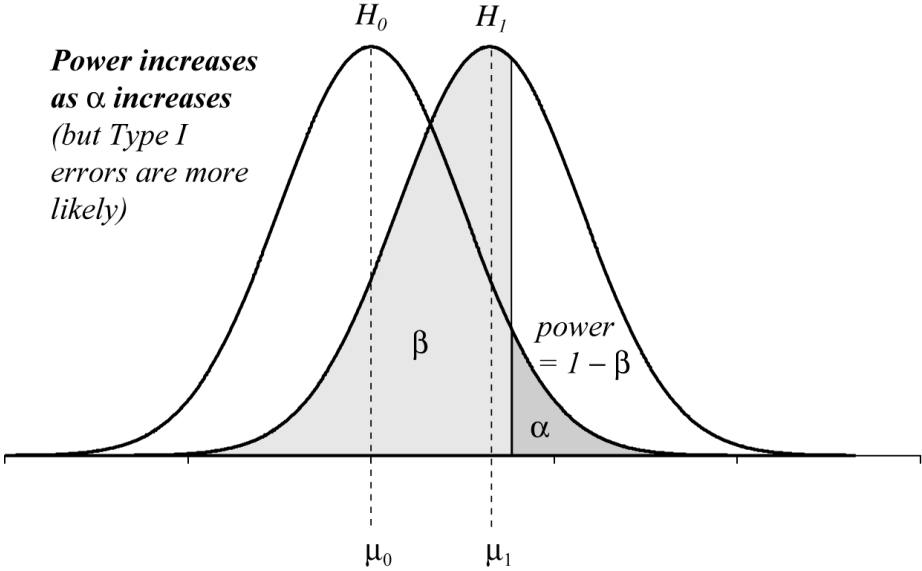
Calculate probability of finding the observed data (e.g. difference) if the null hypothesis is true. This is the *p* value.

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 , and errors we can make

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

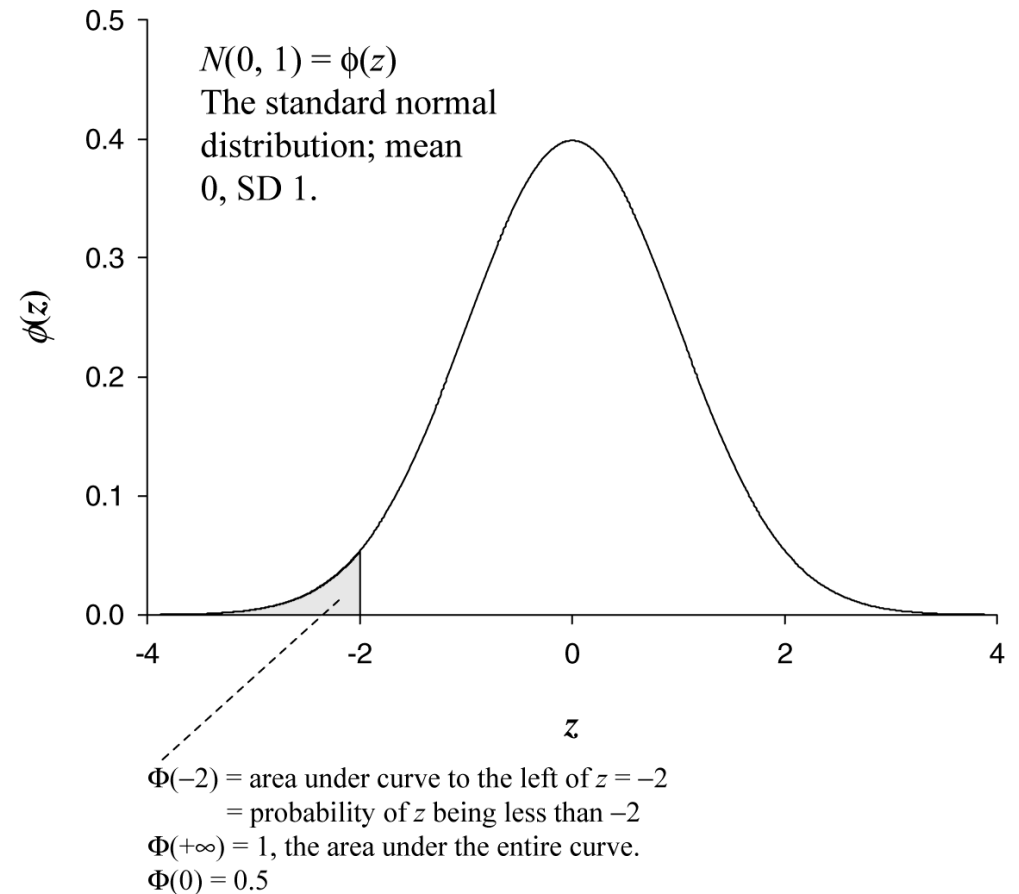
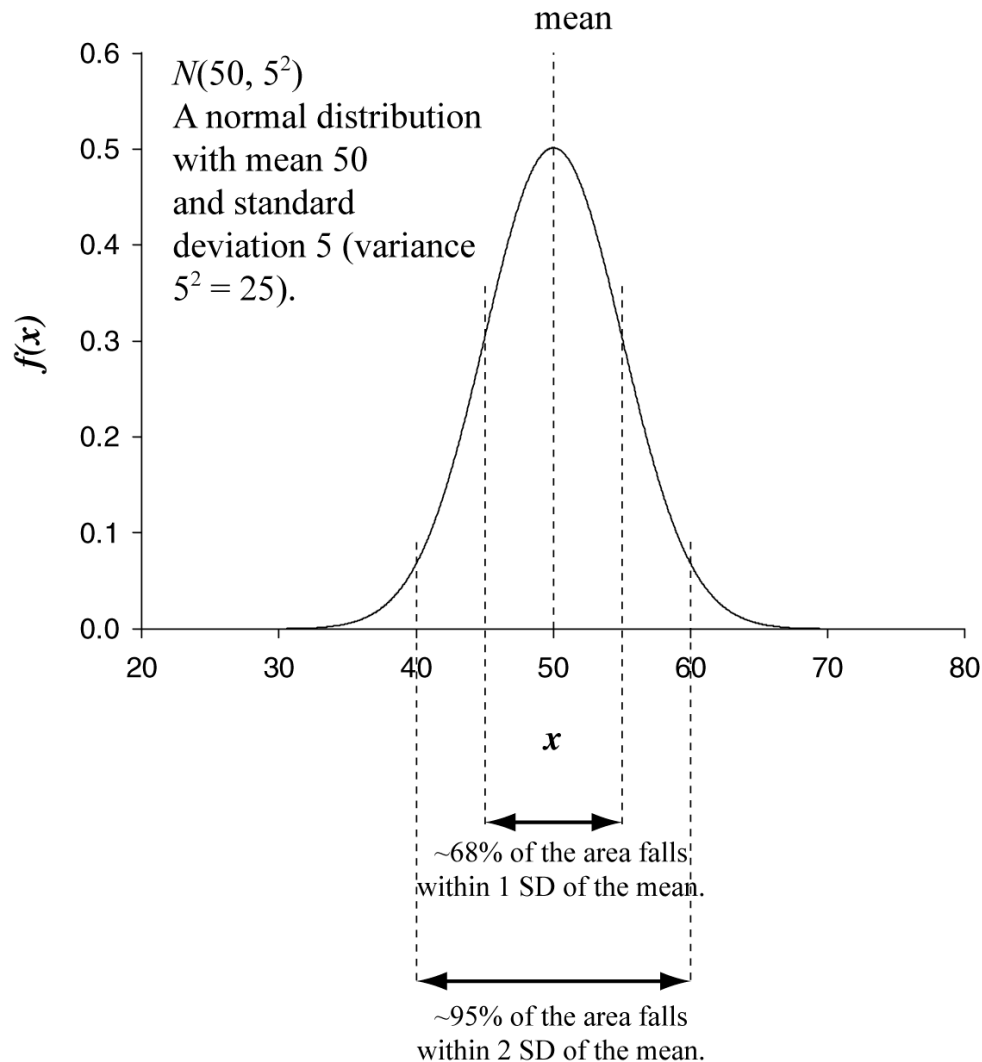
Power: the probability of FINDING a GENUINE effect



Reminder: Z



Reminder: the distribution of Z (the standard normal distribution)



The letter Z is used for the standard normal distribution (thus 'Z scores'). As before, ~68% of the area falls within 1 SD of the mean, and so on.

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

Reminder: If we know μ and σ , we can test hypotheses about **single** observations with a Z test

Example: we know IQs are distributed with a mean (μ) of 100 and a standard deviation (σ) of 15 in the healthy population. If we select a single person from our population, what is the probability that he/she has an IQ of **60 or less**?

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 100}{15} = -2.667$$

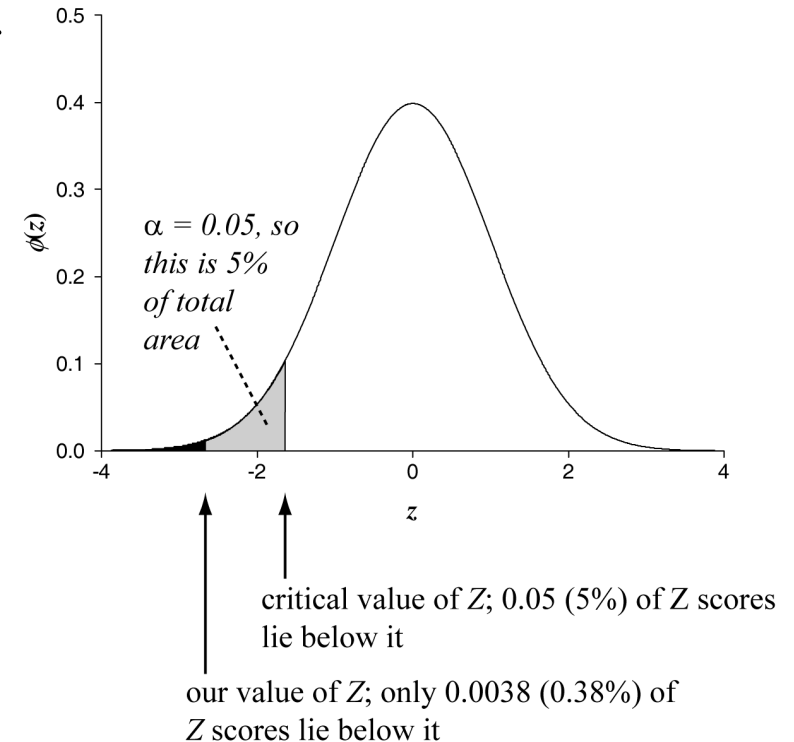
Our tables will tell us that the probability of finding a Z score less than +2.667 is 0.9962...

So the probability of finding a Z score less than -2.667 is $1 - 0.9962 = 0.0038$ (since the Z curve is symmetrical about zero).

If our **null hypothesis** is that the person *does* come from the healthy population, we might reject the null hypothesis if $p < 0.05$ (as in this example). This will happen whenever $Z < -1.64$. This is an example of a **critical value** of Z.

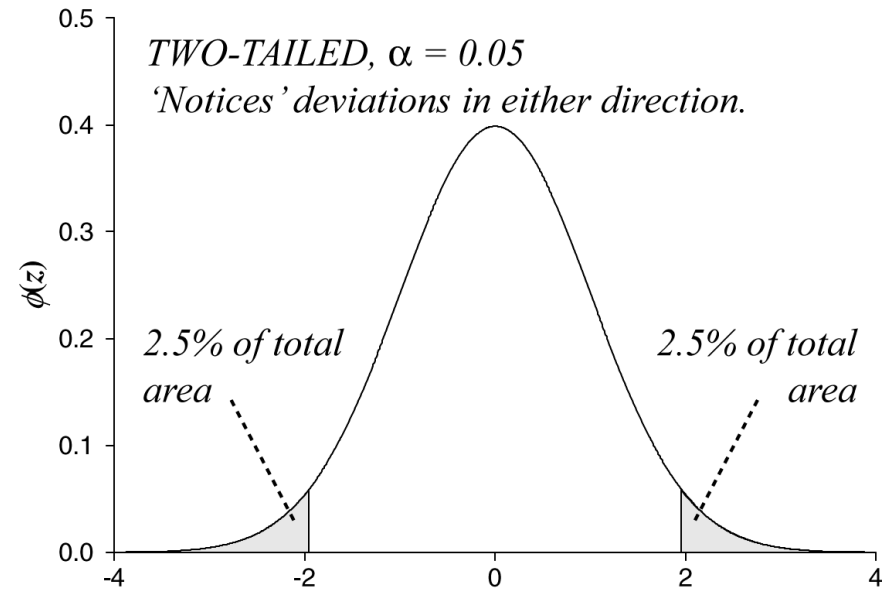
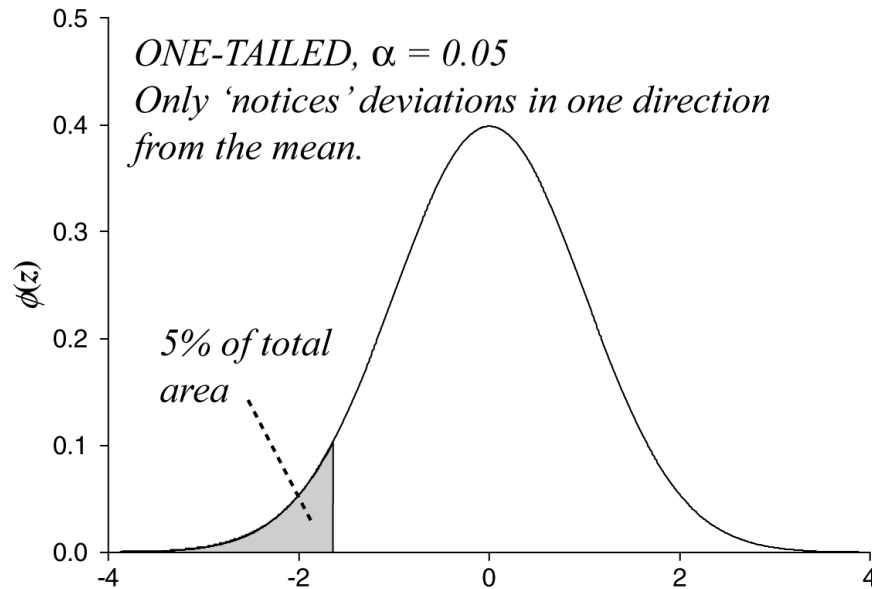
Therefore, this IQ is 2.667 standard deviations below the mean.

How likely is that?



Reminder: one- and two-tailed tests

We asked for the probability of finding an individual with an IQ of 60 **or less** in the normal population. This tests the null hypothesis H_0 : ‘the individual comes from the normal population with mean 100 and SD 15’. We calculate p , and would reject H_0 if $p < \alpha$ (where α is typically 0.05 by arbitrary convention).

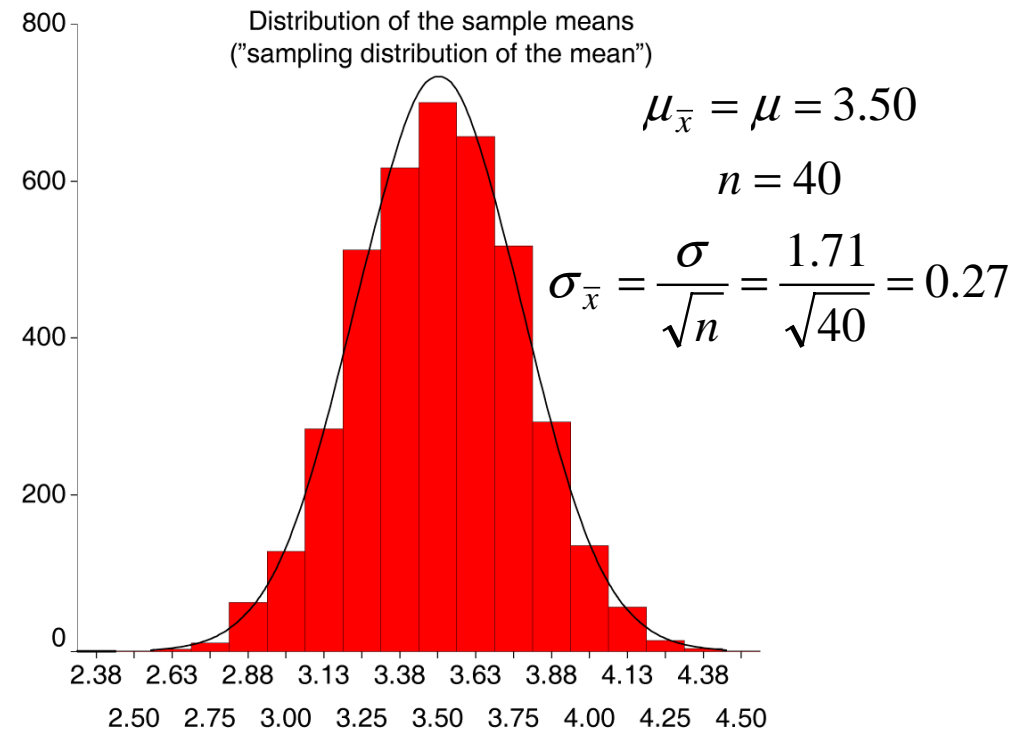
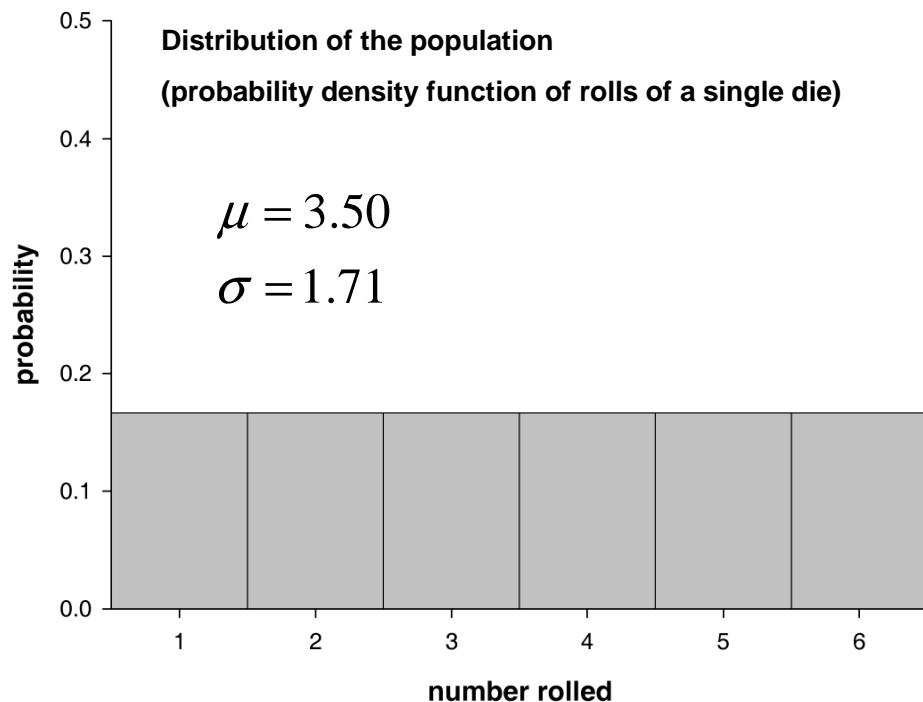


This is a **one-tailed** test. If we do **not** reject the null hypothesis, it means that the IQ is **not significantly less** than 100; it might be (a) not different from 100, or (b) *bigger* than 100. If we want to reject the null hypothesis if the IQ is **bigger or smaller**, we use a **two-tailed** test, and ‘allocate’ $\alpha/2$ for testing each tail to keep the overall Type I error rate at α . The critical values of Z will then be slightly larger (-1.96 and $+1.96$, as it happens).

The t test



The 'sampling distribution of the mean'



4000 samples, each with $n = 40$

SD (of sample means) = 0.27

mean (of all sample means) = 3.50

The Central Limit Theorem

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 lets us test hypotheses about **groups** of observations (**samples**). For a given n , we can find out the probability of obtaining a particular sample mean.

If we know the population SD, σ , we can test hypotheses about samples with a Z test

Example: we know IQs are distributed with a mean (μ) of 100 and a standard deviation (σ) of 15 in the healthy population. Suppose we take a **single sample** of 5 people and find their IQs are {140, 121, 95, 105, 91}. What is the probability of obtaining data **with this sample mean or greater** from the healthy population?

Well, we can work out our sample mean:

$$\bar{x} = 110.4$$

We know n :

$$n = 5$$

We know the mean of all sample means from this population:

$$\mu_{\bar{x}} = \mu = 100$$

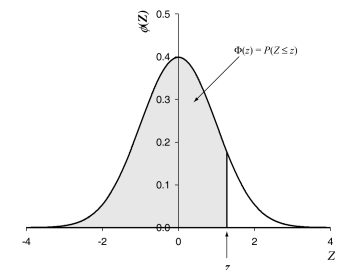
... and the standard deviation of all sample means:
(often called the **standard error of the mean**)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{5}} = 6.708$$


So we can work out a Z score:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{110.4 - 100}{6.708} = 1.55$$

Our tables will tell us that $P(Z < 1.55) = 0.9394$. So $P(Z > 1.55) = 1 - 0.9394 = 0.061$. We'd report $p = 0.061$ for our test.



But normally, we don't. So we have to use a **t test**.

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


If we don't know the **population** SD, σ , and very often we don't, we can't use this test.

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

Instead, we can calculate a number using the **sample SD** (which we can easily calculate) as an *estimator* of the population SD (which we don't know). **But this number, which we call t , does NOT have the same distribution as Z .**

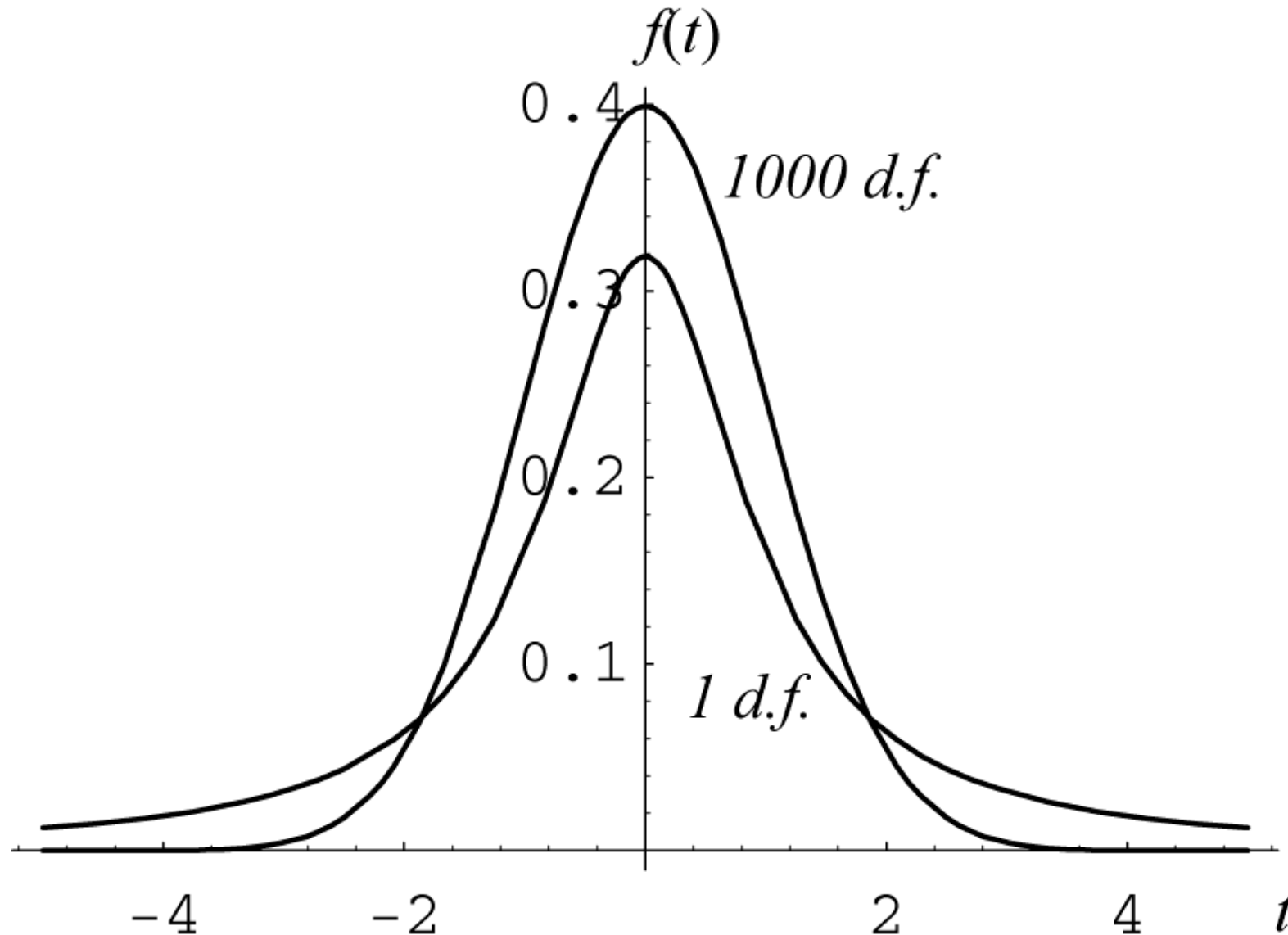
The distribution of t : “Student’s” (Gossett’s) t distribution

As is so often the case, beer made a statistical problem go away.



Student (Gossett, W.S.) (1908). The probable error of a mean. *Biometrika* **6**: 1–25.

The distribution of t when H_0 is true depends on the sample size (which determines the 'degrees of freedom', or d.f.)



When $\text{d.f.} = \infty$, the t distribution (under H_0) is the same as the normal distribution.

Degrees of freedom (*df*). (Few understand this well!)

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 (d.f. or *df*)**.

Or, the number of observations **free to vary**. (Example: 3 numbers and a mean.)

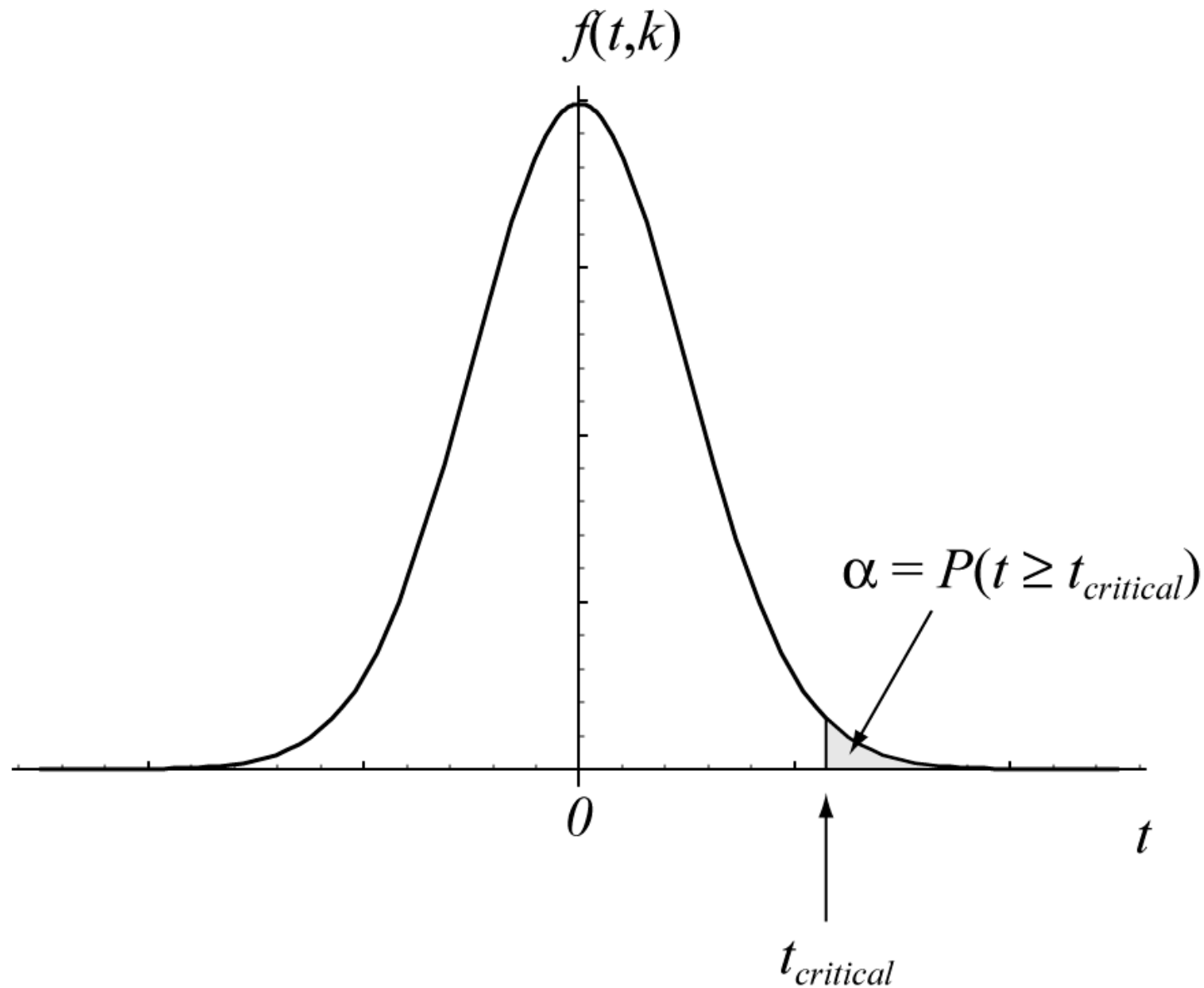
Or, the *df* is the number of measurements exceeding the amount absolutely necessary to measure the ‘object’ (or parameter) in question. To measure the length of a rod requires 1 measurement. If 10 measurements are taken, then the set of 10 measurements has 9 *df*.

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. For example, if the variance σ^2 is estimated (by s^2) from a random sample of n independent scores, then the number of degrees of freedom is equal to the number of independent scores (n) minus the number of parameters estimated as intermediate steps (one, as μ is estimated by \bar{x}) and is therefore $n-1$.

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

Two statistics are drinking in a bar. One turns to the other and asks ‘So how are you finding married life?’ The other replies ‘It’s okay, but you lose a degree of freedom.’ The first chuckles evilly. ‘You need a larger sample.’

Critical values of t (for a given number of d.f.)



When d.f. = ∞ , the t distribution (under H_0) is the same as the normal distribution.

The one-sample t test

We've just seen the logic behind this. We calculate t according to this formula:

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

df for this test (points to $n-1$)

sample mean (points to \bar{x})

test value (points to μ)

standard error of the mean (SEM) (standard deviation of the distribution of sample means) (points to $\frac{s_X}{\sqrt{n}}$)

sample SD (points to s_X)

*Degrees of freedom: we have n observations and have calculated one intermediate parameter (\bar{x} , which estimates μ in the calculation of s_X), so t has $n - 1$ *df*.*

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 one-sample t test: EXAMPLE (1)

It has been suggested that 15-year-olds should sleep 8 hours per night. We measure sleep duration in 8 such teenagers and find that they sleep {8.3, 5.4, 7.2, 8.1, 7.6, 6.2, 9.1, 7.3} hours per night. Does their group mean differ from 8 hours per night?

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

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

Enter descriptive statistics (SD) mode

Casio fx115s
 MODE 2

Clear the stats memory

SHIFT Scl
 AC

Enter values of x
 (e.g. 53)

5 3 M+ etc.
DATA DEL

Read out descriptive statistics:
 (see keypad and inside lid)

mean SHIFT \bar{x}
 1

sample SD ($n-1$ formula)

SHIFT $x\sigma_{n-1}$ n
 3

n RCL $x\sigma_{n-1}$ n
 3

Other Casio models

MODE →SD

SHIFT Scl
 AC =

SHIFT \bar{x}
 1 =

SHIFT $x\sigma_{n-1}$ n
 3 =

RCL C
 hyp

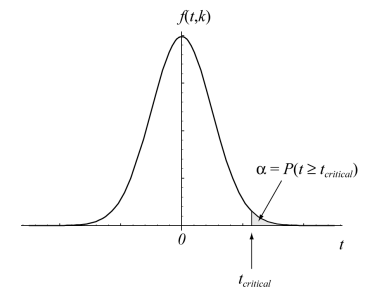
The one-sample t test: EXAMPLE (2)

It has been suggested that 15-year-olds should sleep 8 hours per night. We measure sleep duration in 8 such teenagers and find that they sleep **{8.3, 5.4, 7.2, 8.1, 7.6, 6.2, 9.1, 7.3}** hours per night.

Does their group mean differ from 8 hours per night?

sample mean	$\bar{x} = 7.4$
sample SD (s_X)	$s_X = 1.178$
population mean to test (μ)	$\mu = 8$
sample size (n)	$n = 8$
t	$t = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}} = \frac{7.4 - 8}{\frac{1.178}{\sqrt{8}}} = -1.44$
df	$df = n - 1 = 7$
critical value of t (use $\alpha = 0.05$ two-tailed)	for 7 df, and $\alpha = 0.05$ two tailed, critical $t = 2.365$

Since our $|t/|$ is not as large as the critical value, we do **not** reject the null hypothesis. **Not** ‘significant’; $p > 0.05$. We have **not** established that, as a group, they sleep less than 8h per night.



Paired and unpaired tests (related and unrelated data)

Now we'll look at t tests with **two samples**. In general, two samples can be **related** or **unrelated**.

- *Related*: e.g. measuring the same subject twice; measuring a large set of twins; ... any situation in which two measurements are *more likely to resemble each other than by chance alone* within the 'domain' of interest.
- *Unrelated*: where no two measurements are related.

Example: measuring digit span on land and underwater. Could use either

- **related (within-subjects) design**: measure ten people on land; measure same ten people underwater. 'Good' performers on land likely to be 'good' performers underwater; the two scores from the same subject are related.
- **unrelated (between-subjects) design**: measure ten people on land and another ten people underwater.

If there is 'relatedness' in your data, **your analysis must take account of it**.

- This may give you more power (e.g. if the data is paired, a paired test has more power than an unpaired test; unpaired test may give Type II error).
- beware **pseudoreplication**: e.g. measure one person ten times on land; measure another person ten times underwater; pretend that $n = 20$. In fact, $n = 2$, as repeated measurements of the same person do not add much more information — they're all likely to be similar. Get Type I errors.

The two-sample, paired t test

Very simple. 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')

The two-sample, paired t test: EXAMPLE (1)

Looking at **high-frequency words** only, does the **rate of errors** that you made while categorizing *homophones* differ from the error rate when categorizing *non-homophone (control)* words — i.e. is there a **non-zero homophone effect**? (Each subject categorizes both homophones and control words, so we will use a paired t test.)

Relevant difference scores are labelled % errors — *homophone effect* — *high f* on your summary sheet. **Mean = 3.1; standard deviation = 9.264; $n = 97$.**

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

test value for the differences (*zero* for the null hypothesis 'there is no difference', as in this case)

The two-sample, paired t test: EXAMPLE (2)

Looking at **high-frequency words** only, does the **rate of errors** that you made while categorizing *homophones* differ from the error rate when categorizing *non-homophone (control)* words — i.e. is there a **non-zero homophone effect**?

differences
 mean of differences
 sample SD (s_X) of differences
 mean diff. under null hypothesis (μ)
 sample size (n)

$$X = \{8.3, 8.3, 8.3, -8.3, 0, \dots\}$$

$$\bar{x} = 3.1$$

$$s_X = 9.264$$

$$\mu = 0$$

$$n = 97$$

t

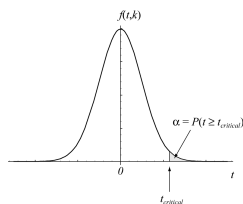
$$t = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}} = \frac{3.1 - 0}{\frac{9.264}{\sqrt{97}}} = 3.296$$

df

$$df = n - 1 = 96$$

critical value of t

for 96 df, and $\alpha = 0.05$ two tailed, critical $t \approx 2.00$



Since our t is larger than the critical value, we **reject** the null hypothesis. ‘Significant’; $p < 0.05$. In fact, $p < 0.01$, since critical t for $\alpha = 0.01$ and 96 df is less than 2.75. **You made more errors for homophones ($p < 0.01$ two-tailed).**

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...

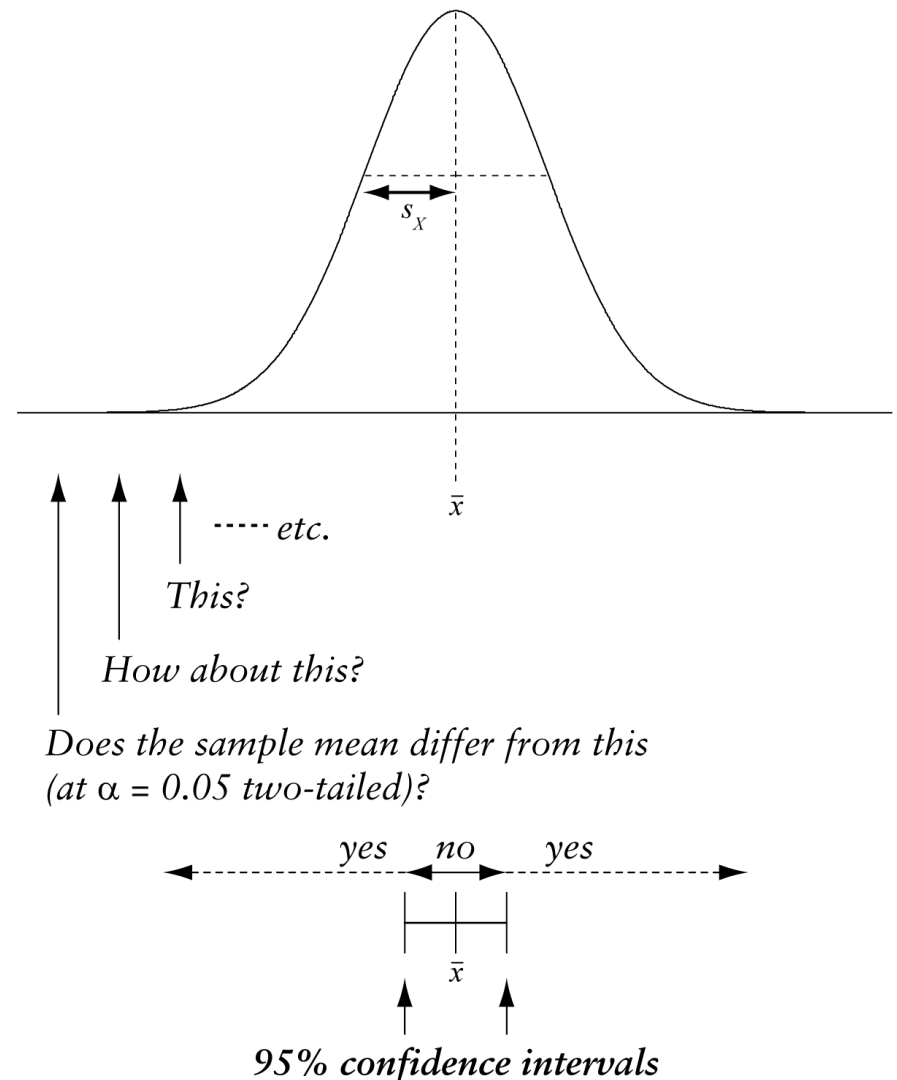
$$\text{Since } t_{n-1} = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}}$$

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

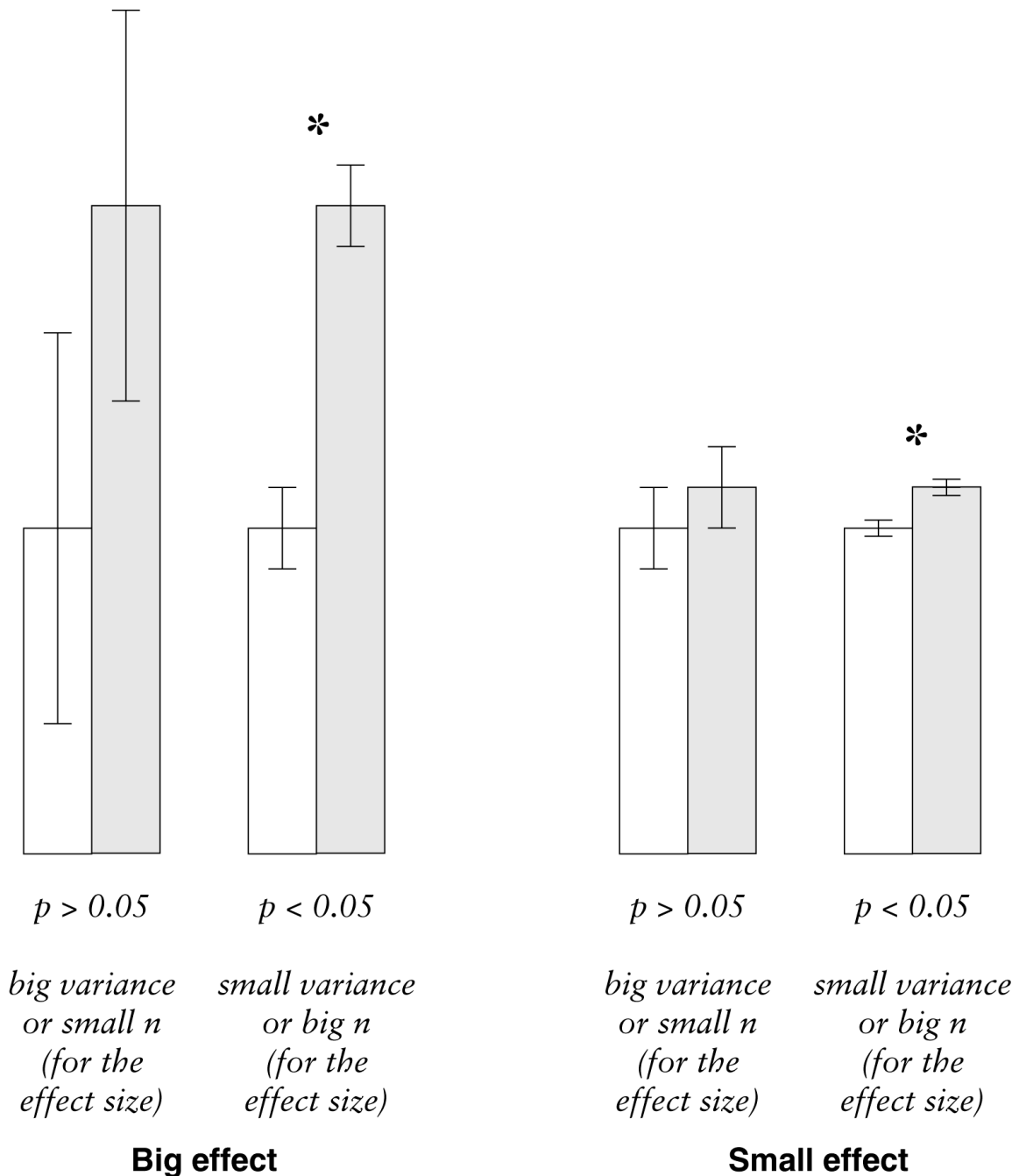
For our homophone example: sample mean = 3.1 (%), $s = 9.264$ (%). For $n = 97$ ($df = 96$), t_{critical} for $\alpha = 0.05$ two-tailed is approx. ± 1.96 . Therefore...

$$\pm 1.96 = \frac{3.1 - \mu}{\frac{9.264}{\sqrt{97}}} \text{ and therefore } \mu = 3.1 \pm 1.84$$

This means that there is a **95% chance that the true population mean homophone effect for high-frequency words is between 1.26% and 4.94%.**



Significance is not the same as effect size



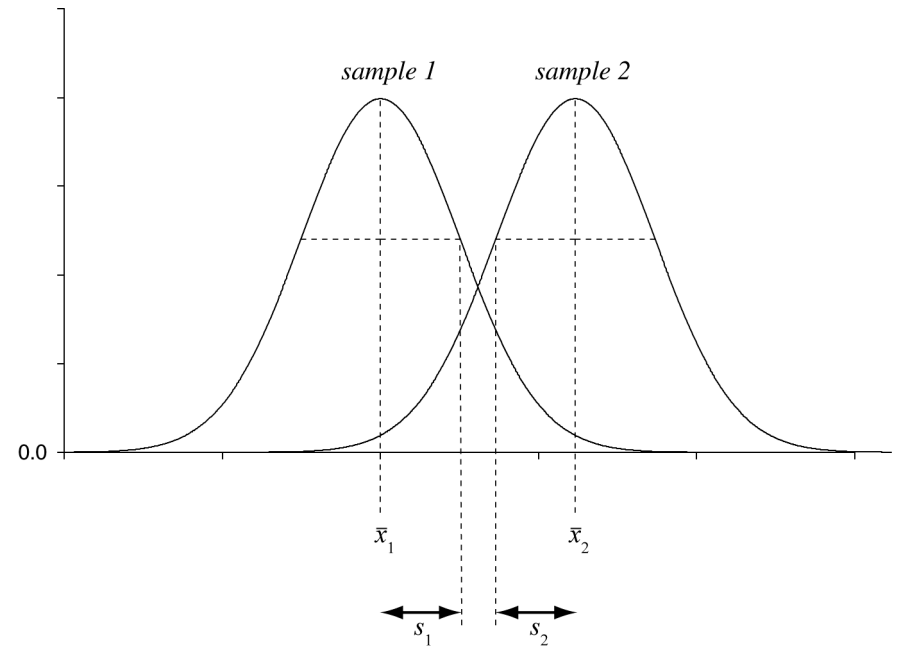
Reporting both may be useful (for example, giving the effect size with its 95% confidence interval; if the confidence interval includes 0, then the effect size is not significantly different from 0).

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

The two-sample, unpaired t test

How can we test the difference between two **independent** samples? In other words, do both samples come from underlying **populations** with the same mean? (= Null hypothesis.)

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.



As always,

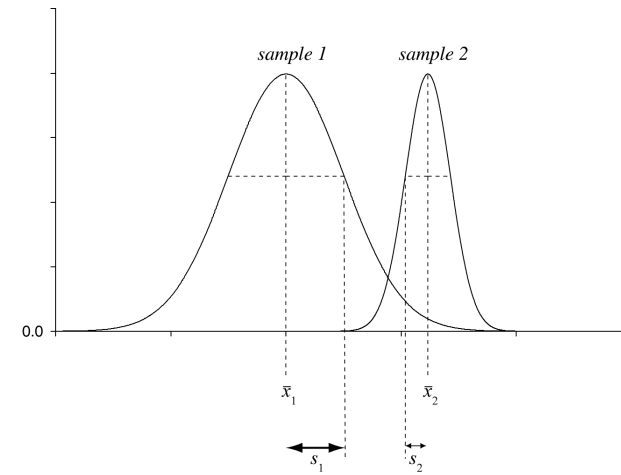
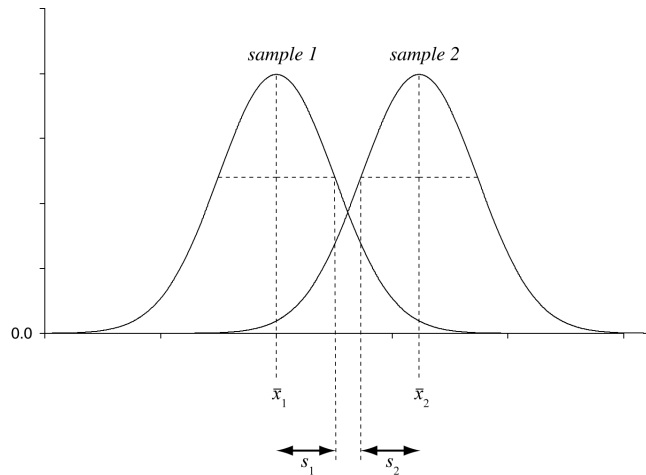
$$t = \frac{\text{something}}{\text{standard error of the something (= SD of an infinite set of samples of the something)}}$$

In this case,

$$t = \frac{\text{difference between the means}}{\text{standard error of the difference between the means (SED)}}$$

The two-sample, unpaired t test

Don't worry about how we calculate the SED (it's in the handout if you're bizarrely keen). The format of the t test depends (unfortunately) on whether the two samples have the same variance.



If the samples have the same variance:

- There's one formula for t if the samples are not the same size ($n_1 \neq n_2$), and a simpler formula if they are ($n_1 = n_2$).
- **Formulae are on the Formula Sheet.**
- We have $n_1 + n_2$ observations and estimated 2 parameters (the means, used to calculate the two SDs), so we have $n_1 + n_2 - 2$ *df*.

If they do not:

- the number we calculate does not have quite the same distribution as t .
- We calculate a number as before but call the result t' .
- We then test our t' as if it were a t score, but with a **different number of degrees of freedom**. Details on the Formula Sheet.

The two-sample, unpaired t test — EXAMPLE

Silly example... In high-frequency word categorization where those words are homophones, were there differences between males and females?

% errors —

Females: $n = 62$; mean = 8.5; SD = 7.314

Males: $n = 26$; mean = 12.5; SD = 10.607

Null hypothesis: no difference between males and females.

Let's use the unequal-variance form of the unpaired t test:

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

Look this up as if it were a t score, but **degrees of freedom = $(n_1 - 1)$ or $(n_2 - 1)$, whichever is smaller.**

The two-sample, unpaired t test — EXAMPLE (2)

Silly example... In low-frequency word categorization where those words are homophones (the hardest condition, judged by mean error rate), were there differences between males and females? **If we call females ‘group 1’ and males ‘group 2’...**

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{8.5 - 12.5}{\sqrt{\frac{7.314^2}{62} + \frac{10.607^2}{26}}} = -1.756$$

df : $(n_1 - 1) = 61$ or $(n_2 - 1) = 25$, whichever is smaller, i.e. **25**

Critical t for 25 df (for $\alpha = 0.05$ two-tailed) is **2.060**

Not a significant difference.

Caveat: some people were ignored because there wasn't enough of your name to judge your sex by it, or because I was incapable of predicting your sex from your name. So these data may not be wholly accurate!

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

So how can we tell if the variances are ‘the same’ or ‘different’ — either to choose the type of t test, or because we’re actually interested in differences in variability?

(a) We can look at them. It may be obvious.

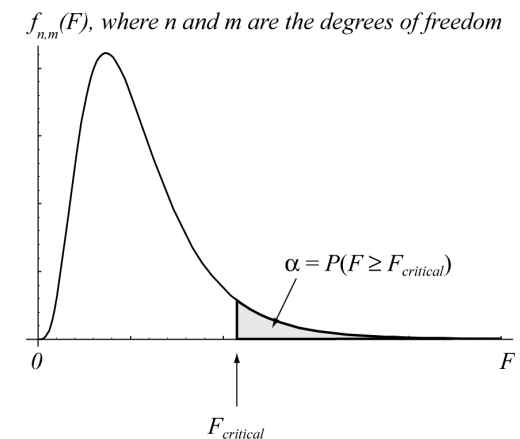
(b) We can perform a statistical test to compare the two variances.

A popular test — not the best one, but a reasonable and easy one — is the F test.

F is the ratio of two variances. Since our tables will give us critical values for $F > 1$ (but not $F < 1$), we make sure $F \geq 1$ by putting the **bigger variance on top**:

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

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*?’.**

F test: homophone example

Example: classifying homophones (high-frequency words).
Were males **more variable** than females when classifying homophones?

Males: $n = 26$, **SD** = 10.607

Females: $n = 62$, **SD** = 7.314

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

$$F_{25,61} = \frac{10.607^2}{7.314^2} = 2.103$$

From tables, critical F for 25,61 df (for $\alpha = 0.05$ two-tailed)... well, we haven't got it exactly, but the closest ($F_{25,60}$) is **1.87**.

Our F is bigger, so we reject the null hypothesis. **Males were significantly more variable than females** in this condition.

Put the biggest variance on top:

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

swap everything around if $s_1^2 < s_2^2$

Assumptions of the t test

- The mean is meaningful.

If you compare the football shirt numbers worn by England strikers who've scored more than 20 goals for their country with those worn by less successful strikers, you might find that the successful strikers have a mean shirt number that's 1.2 lower than the less successful strikers. So what?

- 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 not $n_1 = n_2$).

(There's a helpful clue to remember that one in the name of the test.) The t test is fairly **robust** to violations of this assumption (gives a good estimate of the p value) if $n_1 = n_2$, but not if $n_1 \neq n_2$.

Parametric and non-parametric tests

The t test is a **parametric** test: it makes assumptions about parameters of the underlying populations (such as the distribution — e.g. assuming that the data are normally distributed). If these assumptions are violated:

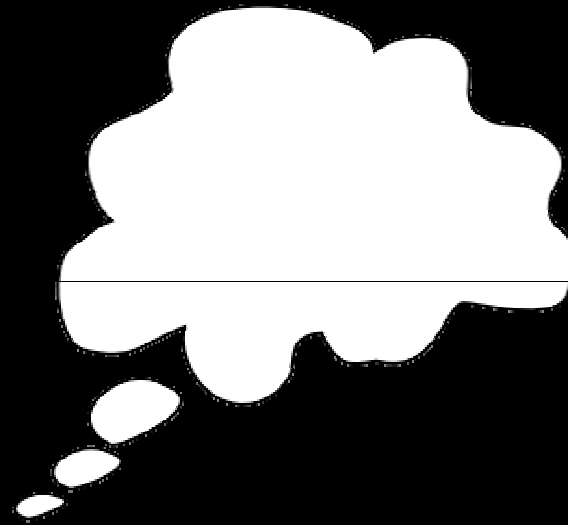
(a) we can *transform* the data to fit the assumptions better
(NOT covered at Part 1B level)

or (b) we can use a **nonparametric** ('distribution-free') test that doesn't make the same assumptions.

In general, if the assumptions of parametric tests are met, they are the most powerful. If not, we may need to use nonparametric tests. They may, for example, answer questions about medians rather than means. We'll cover some next time.



A final thought and a technique

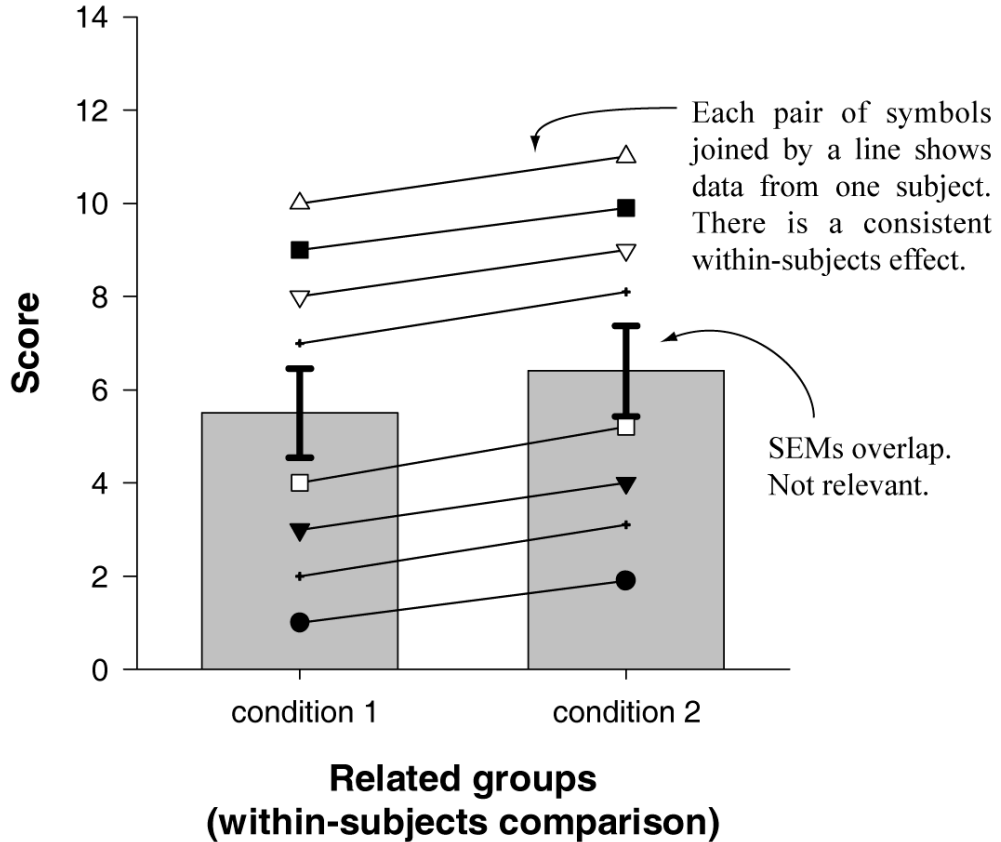
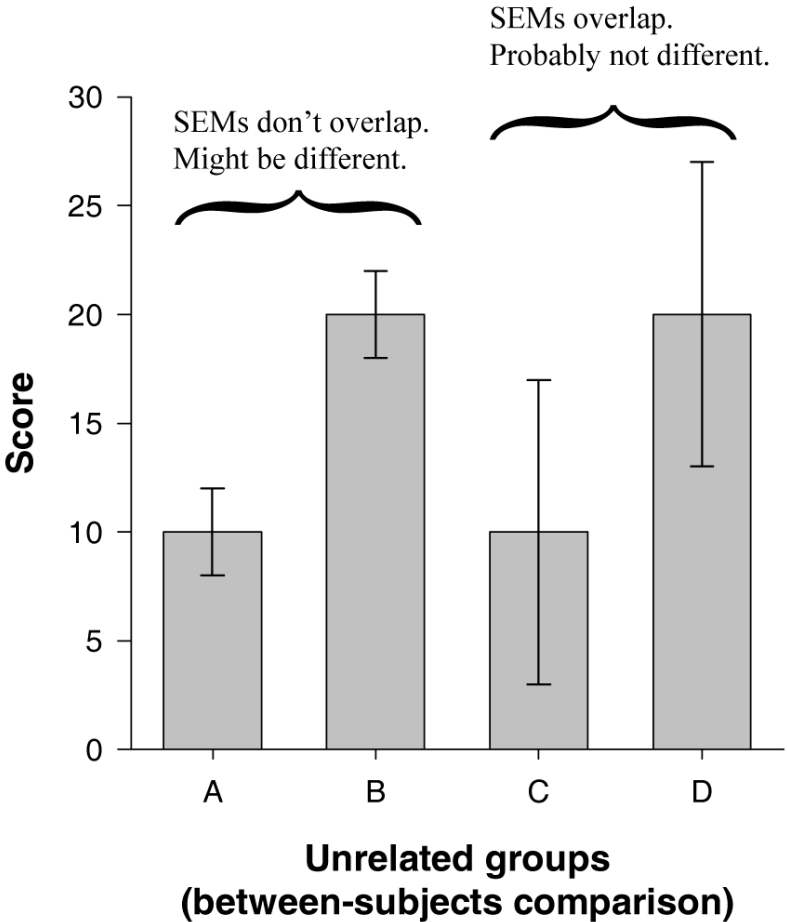


Reminder: multiple comparisons are potentially evil

Number of tests with $\alpha = 0.05$ per test	$P(\text{at least one Type I error if null hypothesis true})$ $= 1 - P(\text{no Type I errors if null hypothesis true})$
1	$1 - (1 - 0.05) = 0.05$
2	$1 - (1 - 0.05)^2 = 0.0975$
3	$1 - (1 - 0.05)^3 = 0.1426$
4	$1 - (1 - 0.05)^4 = 0.1855$
5	$1 - (1 - 0.05)^5 = 0.2262$
100	$1 - (1 - 0.05)^{100} = 0.9941$
n	$1 - (1 - 0.05)^n$

(But remember, you can't make a Type I error — saying something is significant when it isn't — at all *unless* the null hypothesis is *actually* true. So these are all 'maximum' Type I error rates.)

Drawing and interpreting between- and within-subject effects





© Maki Kawai