Lecture 3

Introduction to hypothesis testing: *t*-tests

Comparing group means with *t*-tests

- We've seen that when variables show a bell-shaped distribution, the normal curve can be used as a model to calculate cumulative probabilities of values and confidence intervals

- t-tests extend the logic so as to compare group means
  - done through calculation of *probabilities of differences* in group means

Three scenarios

- *One-sample t-test*: does a group differ from a reference value?
  - Is daily caloric intake of children from a village school in Ghana significantly below the WHO recommended value?

- *Two sample t-tests*: do groups differ?
  - Are !Kung men taller than !Kung women?

- *Paired t-test*: are two sets of measurements *of the same individuals* different?
  - Did blood pressure in patients differ before and after a new treatment was introduced? (the two samples are from the same patients)

t-test: test statistic

- *t*-test is based on the *t*-statistic, a '*t*-score' similar to a z-score:

$$t = \frac{x - \mu}{sem}$$

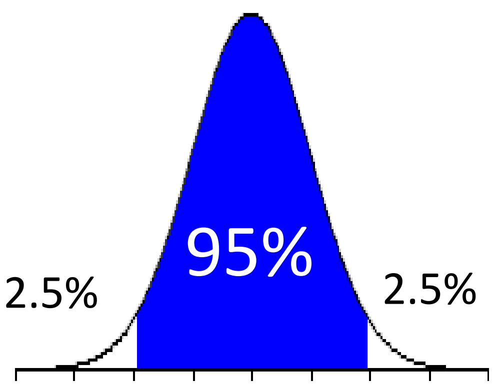
mean

Standard error of the mean

- t is the standardised difference between two values
  - t-test evaluates the probability of this difference
  - Based on this probability and confidence intervals, test establishes whether this difference is 'significant' (i.e. 'too different')
  - i.e. whether the test value x and mean μ are significantly different from each other
- *sem* is the *standard error of the mean*
  - measure of variation taking into account *sample size*

sd

$$sem = {}^{\sigma}/_{\sqrt{n-1}}$$



- As seen, 'significant difference' is a matter of convention
  - 'significantly different' implies that difference between values is *outside our 95% confidence interval*

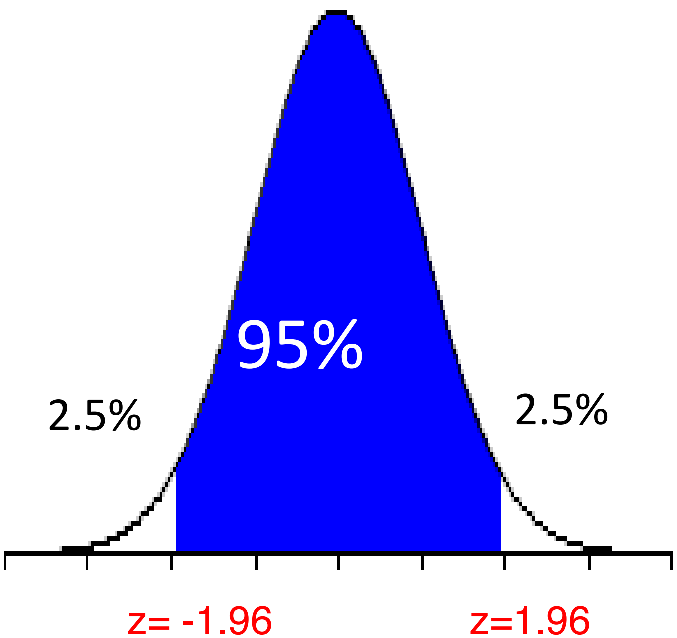
t-test: procedure

T-test >> P value

- A 95% CI implies that only 5% of differences between mean and test values are considered 'significant'
  - *Significance value* of test, or *P value*, is therefore P=5%=0.05

- If I run a t-test and result is P=0.04, we are "96% sure" difference is not significant (because difference is within a 96% CI)
  - If P=0.003, we are "99.7% sure" difference is significant
    - that's enough (we want to be at least 95% sure)
  - If P=0.08, we are only "92% sure" difference is significant
    - P=0.08 would only be outside a 92%
    - not enough; difference is not significant or 'real'

???



t-test: procedure

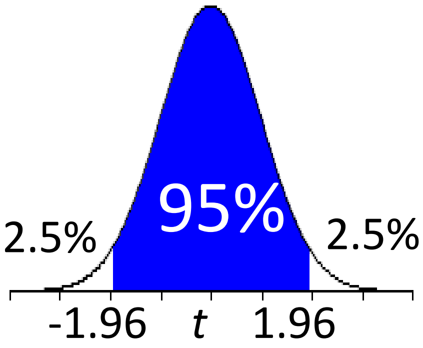
- t-test defines a *null hypothesis (H_o)*:
  - *t*-score (standardised difference between a sample mean and a test value) is contained in the 95% confidence interval,

  = difference is not large enough ('rare' enough)

  = there is no significant difference between values

  =difference is not 'real' but is just an outcome from sampling

  - if you sample the same population twice (two groups of British voters, selected with the same criteria) and ask about their voting intention, they may differ a little; but they are still samples from the same population



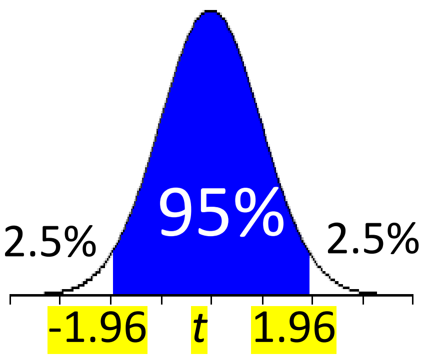
****

t-test: t values

- Null hypothesis is *conservative*: there is NO difference in means
  - If P>0.05, null hypothesis is accepted: no difference
  - If P<0.05, null hypothesis is rejected and alternative hypothesis is accepted: significant difference to the mean

***T test contributes to P value

more than 95% happen



- Since 95% CI is defined by z-scores between -1.96 and 1.96, for significant difference you need:
  - either $t < -1.96$
  - or $t > 1.96$

- Also, do include t-values when reporting test results (and not only P-values)

1) One-sample *t*-test in *R*

Example: Based on our census, can we say that height of !Kung women is significantly different from 155 cm?

- or is the difference just by chance, i.e. they seem to be small due to small sample etc?

- Sample size= 181 adult female heights from 264 cases excluding NAs
- mean=149.5cm, sd=5.12
- test value: 155 cm

One-sample *t*-test in *R*

\> t.test(kaf$height, mu=155)

One Sample t-test

data: kaf$height

Degree of freedom

t = -14.39, df = 180, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 155

95 percent confidence interval:

148.7721 150.2741

sample estimates:

mean of x

149.5231

The narrowness is not defined by sample size, affected by sd

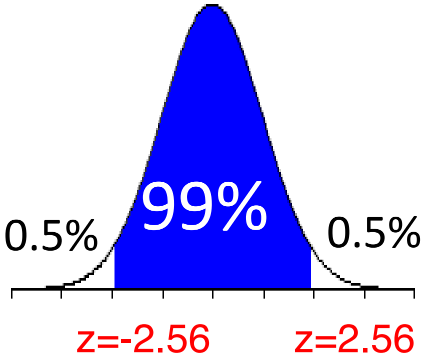
- Syntax is very simple
  - mu= test value
- t = -14.39
  - difference must be significant
- P=2.2e-16 is *R*'s way of saying 'zero'
  - P<0.05: significant difference
- 95% CI: we are 95% sure that mean height of !Kung adult females is between 148.77 and 150.27
  - 155cm is outside CI; significant difference
  - If test value is within CI, no difference
- Outcome:
  - Reject null hypothesis, accept alternative hypothesis = true mean is not equal to 155 cm

99% CI

- To change significance level to P=0.01, add *conf.int=0.99*

```
> t.test(kaf$height, mu=155, conf.level=0.99)
     One Sample t-test

data:  kaf$height
t = -14.39, df = 180, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 155
99 percent confidence interval:
 148.5322 150.5139
sample estimates:
mean of x
 149.5231
```



Can be calculated by qnorm(0.005/0.985)

- Basic stats are the same (t, P), but 99% CI is wider; harder to demonstrate significant difference
- Still: reject null hypothesis
- **Now you're '99% sure' that !Kung adult female height differs from 155cm**

Exercises:

Is the mean weight of !Kung adult females significantly different from 40kg?

rejected

a) Is the null hypothesis accepted or rejected? Why?

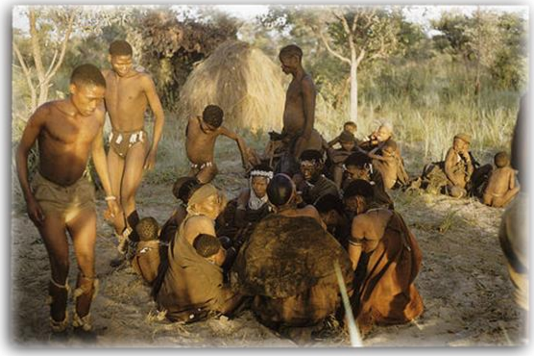
b) Interpret the 95% CI

Re-run the test with a 99% CI

c) is the null hypothesis accepted or rejected? Why?

For one-sample test's mean, if mean stays in between 95CI values, Ho is rejected/accepted?

2) Two-sample *t*-test



- Second, you may also want to test whether *two* samples are significantly different

- Westernised European men are typically heavier than women: is this also true for the !Kung?

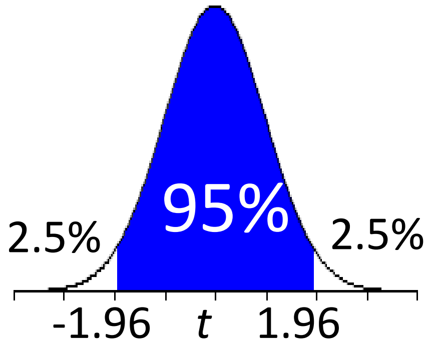
- Test procedure is similar: *t*-statistic is now the difference between the *means of the two compared groups*

- If male height is μ_1 and female height is μ_2,

$$t = \frac{\mu_1 - \mu_2}{sedm}$$

- Why sedm (=standard error of the difference of means)?
  - instead of one *sem*, now we have two (one from each group); we use *sedm* a combination of both

$$sedm = \sqrt{sem_1{}^2 + sem_2{}^2}$$



CI results go from minus to plus, including t=0, meaning Ho could be true and cannot be excluded
CI results go from plus to plus, excluding t=0, meaning Ho is rejected

Two-sample *t*-test in *R*

- Our file has one column for weight and one for sex; first possible syntax is:

```
> t.test(kc$weight ~ kc$sex)
    Welch Two Sample t-test
data:  kc$weight by kc$sex
t = 4.9926, df = 584.101, p-value = 7.874e-07
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 3.657924 8.402225
sample estimates:
 mean in group man mean in group woman
          38.91039            32.88031
```

T is greater than 1.96, significant difference

95/

- ps. samples include children too, hence the different mean height values

- **Notice that R uses alphabetical or numerical order for groups**
  - 'man' before 'woman'
- Welch test is the t-test that calculates *sedm* as we did
- *t* <<<-1.97; P<0.05:
  - reject null hypothesis (no difference in mean weights)
  - accept alternative hypothesis (weights differ)
- Degrees of freedom look weird; they're calculated using means too
- 95% CI is for *difference* of means, and excludes zero
  - if it excludes zero, difference cannot be zero!

Exercises:

a) Run the same two-sample test with a 99% CI; do weight in men and women differ? Why?

reject Ho

b) Run the test using the alternative syntax below:

> t-test(variable 1, variable 2)

(hint: what is variable 1? And variable 2?

3) Paired *t*-test

- A paired test is used when the two compared measurements are not independent
  - for example, two paired measurements from the same individual (typically, comparison between 'before vs. after')

Example

- The file *intake* has data on pre- and post-menstrual calorie consumption in 11 women; is there a difference?

- Select *Packages* tab (bottom right panel)
- Install and then run library *ISwR* (by ticking box)
- Enter *intake* to see *intake* file

> > intake #this is a file in the library *ISwR*

Paired *t*-test

- It is *incorrect* to run a two-sample test in this case, because the two samples are not independent; *pre* and *post* measurements taken from the same individual (i.e. paired)

- But you can define the difference *d* as a new variable

  d = post-pre

- i.e., we are no longer taking two measurements from each person: we are measuring only one variable:
  - **the variation (or 'delta') in calorie consumption for the same individual before and after**

Now we just test whether *d* is significantly different from zero, as in a one-sample test

- Paired *t*-test is thus a one-sample *t*-test with test value=0

- To run a paired *t*-test: just add ***paired=T***

```
> t.test(intake$post, intake$pre, paired=T)

        Paired t-test

data:  intake$post and intake$pre
t = -11.941, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1566.838 -1074.072
sample estimates:
mean of the differences
              -1320.455
```

- (now group order is determined as ‘post’, then ‘pre’)

- Result: significant difference between the groups

- This makes sense: information that measurements are paired is very relevant to the test
  - intake dataset: *every* women reduces calorie consumption from pre to post
  - this information is lost in a two-sample *t*-test, which first calculates means for *post* and pre, and then calculates their difference

Exercises:

Run the same test with a 99% CI

a) What happens to P value?
b) Is there a significant difference?

Now run a two-sample t-test on *pre* and *post*

(wrong, don't do this)

c) With 95% CI, is there a significant difference
d) With a 99% CI, is there a significant difference?

One- vs. two-tailed *t*-tests

One-tailed test only happens under the circumstances of one-directional/ sided event

- All *t*-tests we've run so far are *two-tailed* because the alternative hypothesis is that 'mean is *different* from *x*' (i.e. either too large or too small)
  - Only after you show they are different is that you can tell whether test value is smaller or larger than reference value

- But sometimes you may want to test only whether a mean is *smaller than* or *greater than* a value; in some cases, this is the only option!
  - suppose you measure the height of a sample of British girls aged 15, and another sample of girls aged 16.
  - the question was: are girls still growing between ages 15 and 16?

One- vs. two-tailed *t*-tests

- In this case, you can run a **one-tailed** *t*-test comparing data on heights at age 15 and 16
  - the alternative hypothesis is now more specific: mean height at age 16 is GREATER THAN (not just different from) mean height at age 15

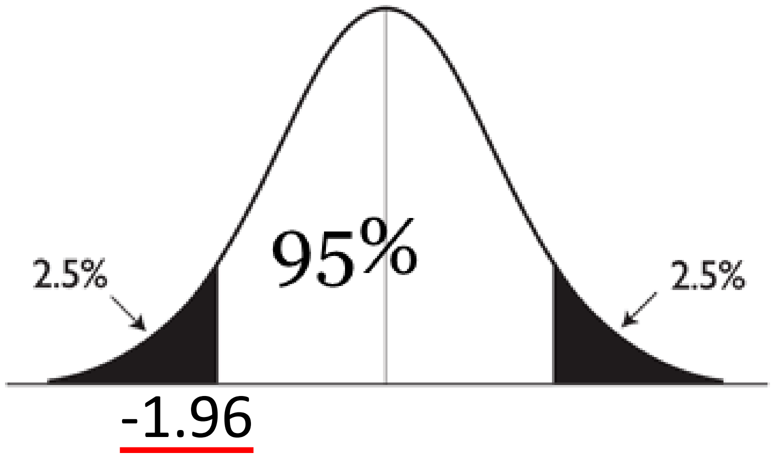
  (justification: 15-year-old girls may not grow, but they don't shrink!)

- In this case, for a 95% CI, the 'rare' 5% are placed one side of the curve only!!!

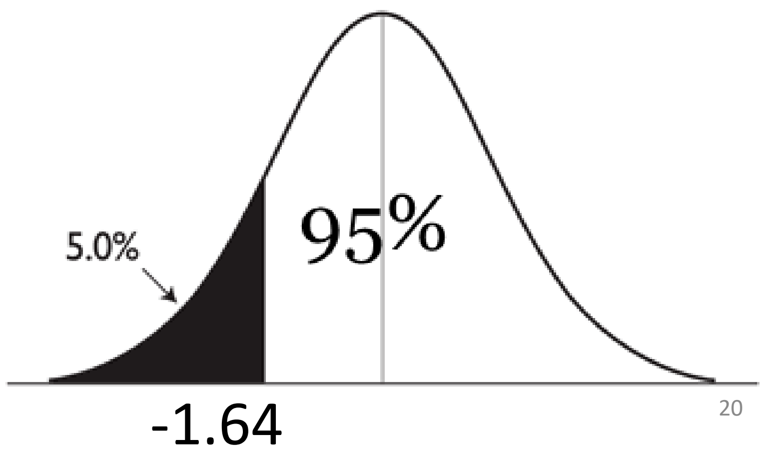
- If you want to run one-tailed –tests, add arguments *alt='g'* for greater than, or *alt='l'* for less than

One- vs. two-tailed *t*-tests

- There are important differences between the one- and two-tailed tests
- In a 95% CI, a two-tailed test splits the extreme 5% into two 2.5% parts

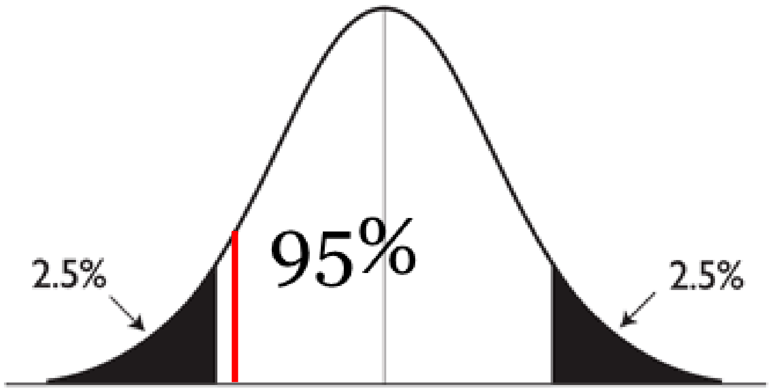


- But the one-tailed test places the whole 5% on one side only, and therefore creates a larger, more ‘inclusive’ single tail
  - The t-value corresponding to cumulative probability 0.05 is now t=-1.64

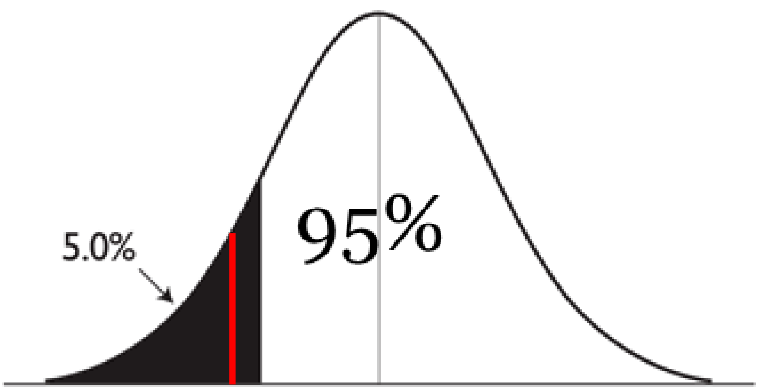


One- vs. two-tailed *t*-tests

- This means there is a temptation to *cheat* and switch from two-tailed (and a non-significant result)…
- …to a one-tailed test (and a significant difference between means)
- Example: imagine my t value is t=-1.8; this is inside a two-tailed 95% CI (not different) but outside a one-tailed 95% CI (significantly different)



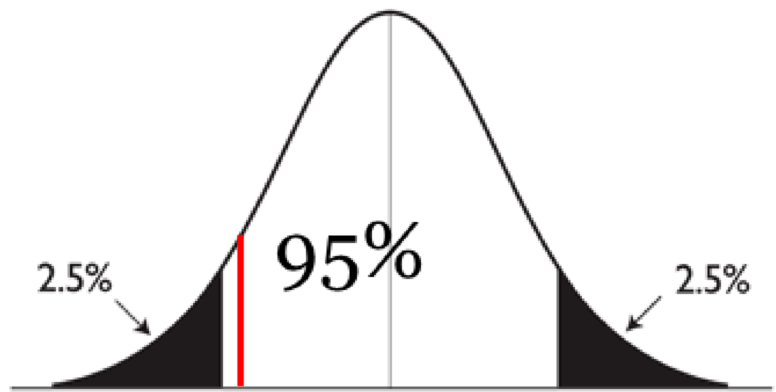
x=-1.8: inside CI



x=-1.8: outside CI

One- vs. two-tailed *t*-tests

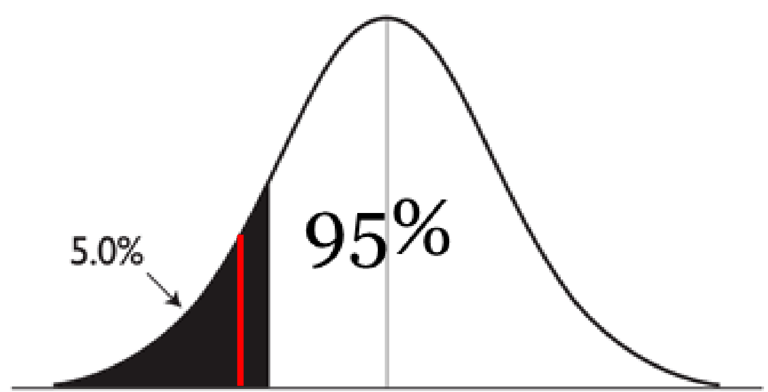
- It is wrong to run a one-tailed test just because it is easier to prove that groups are different
- Example: to test for differences between male and female height, you should *always* run a two-tailed test; you shouldn't argue that "males are always taller"



x=-1.8: inside CI

notes:

- it is hard to draw the line between 'young girl don't shrink' ('ok') and 'men are always taller' ('cheating'); use common- sense
- one-tailed tests are much more rarely used than two-tailed tests



x=-1.8: outside CI

Conclusions

- Confidence intervals and all *t*-tests assume a normal distribution
  - That's why you do not *prove* differences; you compare groups and give an estimate of the *probability* that they are different or similar

Important:

- **Current trend is to provide confidence intervals and t-values in addition to *P* values when reporting results of tests in general (not just *t*-tests)**

- Null hypothesis is always that the two compared means are *not* different (i.e. one value is a relatively frequent value around the other mean)

- It is easy to interpret t-tests: for a confidence level of 95%, *if $P<0.05$ then difference is statistically significant* (groups differ); if *$P>0.05$, there is no statistically significant difference*
  - Or: if *confidence interval includes 0, difference is not significant*

- One-tailed *t*-tests are less commonly used (they are harder to justify)

```
t.test(kfm$weight~kfm$sex)
OR
kfm_man<-subset(kfm, sex=="boy")
kfm_woman<-subset(kfm, sex=="girl")
t.test(kfm_man$weight, kfm_woman$weight)
```

Exercises:

two sample test

File *kfm* (ISwR library)

weight ~ sex

a) File has data on sex and weight of babies; is weight in boys and girls significantly different? Less Accept

b) Is breast milk intake (variable dl.milk) significantly different in boys and girls?

Longevity in men and women (file humanlongevity)

c) We want to compare longevity in women and men; look at the data in file human longevity. Which t-test do we need to run Paired test

d) Is there a significant difference between men and women in longevity?