Lecture 3

Introduction to hypothesis testing: *t*-tests

假设检验的简介

用t检验比较各组均值

Comparing group means with *t*-tests

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

- *t*-tests extend the logic to comparisons between group means
  - it calculates *probabilities of differences* in group means

    计算各组均值差异的概率

Three scenarios

单样本T检验

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

- *Two sample t-tests*: do sample means differ? 独立样本T检验(两个样本)
  - Are !Kung men taller than !Kung women?

- *Paired t-test*: are two sets of measurements *from 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)

配对样本T检验

Why test for differences?

- If we want to know whether adult men are taller than adult women in the UK, we can:

1) Measure all ~20 million adult men and 20 million women, and compare their mean heights
   - Descriptive statistics; no test is required

or

2) You can take a sample of 100 men and a sample of 100 women, and compare their means
   - Now you need to test whether the difference is significant
   - What if the difference does not represent true differences, but was caused by sampling 200 non-representative people?

- A *t*-test is a way of testing whether a difference between two values is "significant" (to a conventional degree; for example, 95%)

检验两个数值之间的差异是否 " 显著 " 的方法(传统意义上的差异是95%)

t-test: test statistic *t*

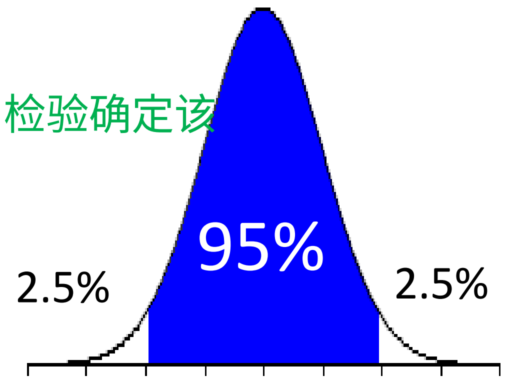
- *t*-test is based on the *t*-statistic, a '*t*-score' similar to a z-score:

t是两个值之间的sd，是差异化的z分数

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

- *t* is the standardised difference between two values; it is the z-score of difference
  - based on this probability and confidence intervals, t-test establishes whether this difference is 'significant' (i.e. 'too different')
    - = whether the are too different to be "the same" or from the same population
    - = whether the test value x and mean μ are significantly different from each other

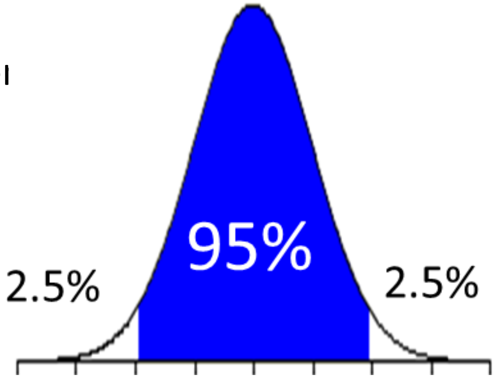
基于概率和置信区间，t检验确定该差异是否“ 显著 ”



他们是否差异太大而不能“ 相同 ”或来自同一群体
测试值x和平均值 μ 之间是否有显著差异

Standard error 标准误差

- An individual point (my height) has a standard deviation relative to a mean value (mean hight)
  - I deviate from mean value 样本值偏离平均值
- But a sample mean has an *error* relative to another value
  - a mean height of 180cm from a sample may be a wrong estimate of true mean of the UK populatio another population
  - the error may be caused by sample size (too small) etc. 误差可能由样本量(太小等)引起



- *sem* is the *standard error of the mean* sem是平均值的标准误差
  - measure of variation taking into account *sample size*
  - It tells you how much variation you expect from a sample mean if the trait has a given standard deviation and sample has a size n

sd/根号下(n-1)

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

标准化t值是指测试值偏离平均值的标准误差

- Standardised t-value is how many **standard errors** the test value deviates from the mean
  - Not how many **standard deviations** (as we calculated n Lecture 2

不是我们第二节课计算的多少个sd

零假设

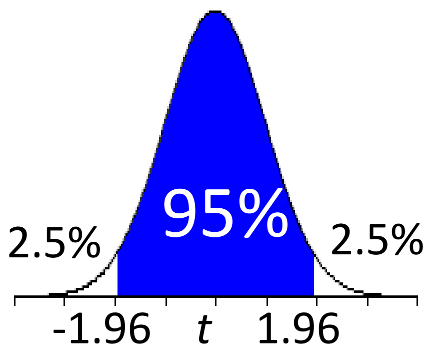
t-test: null hypothesis H_0

- t-test defines a *null hypothesis (H_o)*: 如果t分数在差异为95%的置信区间中
  - *If t*-score is within in the 95% confidence interval of differences:

= difference is not large enough (not ‘rare’ or unexpected enough)

= difference is not statistically significant

=difference is not ‘real’ but is just a random outcome from sampling



差异不够大
差异没有统计学意义
差异不真实，只是随机抽样结果

p值

p>0.05表示在95%CI内，接收零假设，无差异
p<0.05表示超出95%CI，拒绝零假设，接收备择假设

t-test: P values

t值和p值一一对应，可以查表得到

零假设是保守的，指组均值没有差异

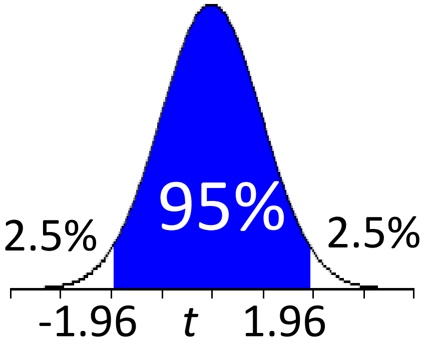
- Null hypothesis H_0 is *conservative*: there is NO difference in group means
  - **If P>0.05** (=probability of difference > 5%; inside 95% CI): null hypothesis is **accepted**: no difference
  - **If P<0.05** (=probability of difference < 5%; outside the 95% CI): null hypothesis is **rejected** and alternative hypothesis is accepted
    - there is a significant difference in means

- Since 95% CI is defined by t-scores between -1.96 and 1.96, for a significant difference, you need at least:
  - t < -1.96 至少需要
  - or t > 1.96

当样本较小时这些值会更高

Those values will be higher when samples are small

- (note: always include t-values when reporting test results, and not only P-values)



只有5%的差异被认为是“ 显著的 ”

显著性水平α

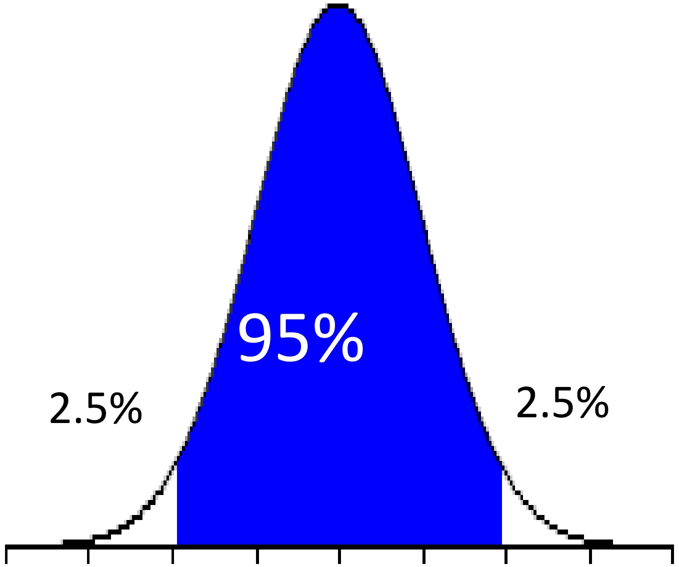
t-test: significance level α

- A 95% CI implies that only 5% of differences between mean and test values are considered 'significant'
  - *Significance level* of test, or α, is therefore 5%=0.05 显著性水平是0.05
  - A difference is significant if its probability, or *P value*, is lower than α 如果其概率或p值低于α，则差异显著

Significant difference: P value < α

如果进行t-test结果p=0.05，则我们“ 95%确定 ”差异是显著的

- If I run a t-test and result is P=0.05, we are "95% sure" difference is significant (because difference is (just) within a 95% CI)
  - If P=0.003, we are "99.7% sure" difference is significant
    - P < 0.05; 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% CI; it is still inside a 95% CI
    - 92% sure is not enough: we want to be at least 95% sure; difference is not significant or 'real' 92%不够，差异不显著或者不真实



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. mean height is low 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
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
```

结果是拒绝零假设，接收备择假设

如果对95%有信心，则平均值明显低于155

Syntax is very simple

- mu= test value

mu是测试值，给定的均值

- t = -14.39 低于平均值超过14个标准误差
  - very small t value! Over 14 standard errors below the mean 表明差异非常显著
  - suggests that difference is significant
- P=2.2e-16 is *R*’s way of saying ‘almost zero’ (P<0.05: significant difference) 几乎为0

有95%的把握平均身高在148到150之间

- 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

155在CI之外，显著差异，如果值在CI内则无差异

Outcome:

- reject null hypothesis, accept alternative hypothesis = true mean is not equal to 155 cm
- = mean height of !Kung women is significantly under 155cm, if 95% confidence in your result is enough

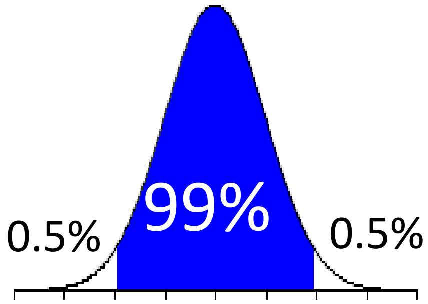
99% CI

改变显著性水平，加上conf.int

- To change significance level to α=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
```



难以证明显著差异

- 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?

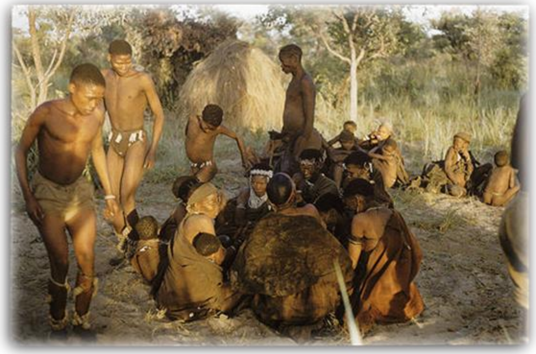
a) Is the null hypothesis accepted or rejected? Why?

b) Interpret the 95% CI

Re-run the test with a 99% CI

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

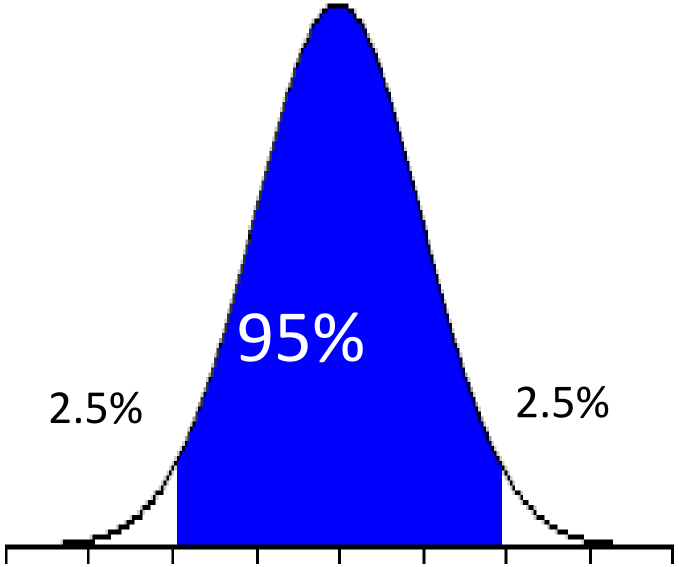
2) Two-sample *t*-test



- 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

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

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
```

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

R使用字母顺序或数字顺序，所以man在women前面

- **Notice that R uses alphabetical or numerical ordering for groups**
  - KungCensus$weight variable: 'man' before 'woman'
- Welch test is the t-test that calculates *sedm* as we did in previous slide

拒绝零假设(平均体重无差异)

- $t > 1.96$; $P<0.05$:
  - reject null hypothesis (=no difference in mean weights)
  - accept alternative hypothesis (=mean weights differ)

自由度看起来很奇怪，也是用均值

- Degrees of freedom look weird; they're calculated using means too

均值差，不包括0

- 95% CI is for ***difference of means***, and excludes zero
  - if it excludes zero, difference cannot be zero! 如果包括0，则差值不能为0
  - => difference is significant

Exercises:

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

b) Run the test differently by creating two separate files for women and men, and then comparing their weights with:

```
> t.test(variable 1, variable 2)
```

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

配对T检验

3) Paired *t*-test

两个测试值不独立

- 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' and '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
```

安装包

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*** 加上这个参数

```
> 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') 现在的顺序是先post再pre

- Result: significant difference between the groups 结果：两组之间有显著差异

- 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*

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

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

What do you conclude about the differences between two-sample and paired sample tests in this case?

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

单双侧的意思

- 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)

  因为所有的备择假设是“与均值不同”(太大或太小)

  - Only after you show they are different is that you can tell whether test value is smaller or larger than reference value

    只有在证明不同之后，才能知道测试值是小于还是大于参考值

- 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!

  在某些情况下，这是唯一选择

  - 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!)

  15岁可能不会长，但不会萎缩

- In this case, for a 95% CI, the 'rare' 5% are placed one side of the curve only!!! 在这种情况下，对于95%CI，“罕见”的5%仅位于曲线的一侧

- If you want to run one-tailed –tests, add arguments *alt='g'* for greater than, or *alt='l'* for less than 如果要单侧测试，添加参数

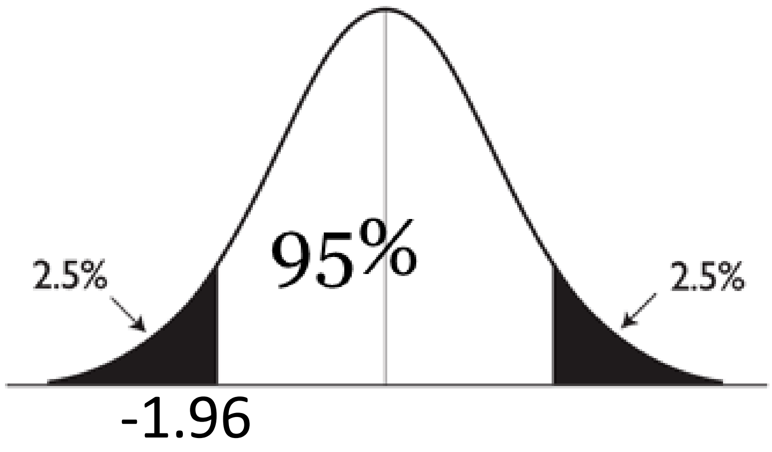
One- vs. two-tailed *t*-tests

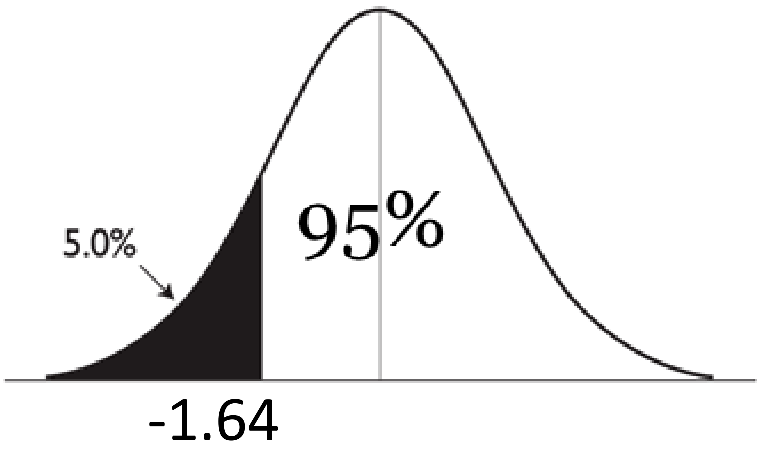
- There are important differences between the one- and two-tailed tests

- In a 95% CI, a two-tailed test splits the extreme 5% into two 2.5% parts

双侧将5%平分在两边，单侧放在一边

- 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





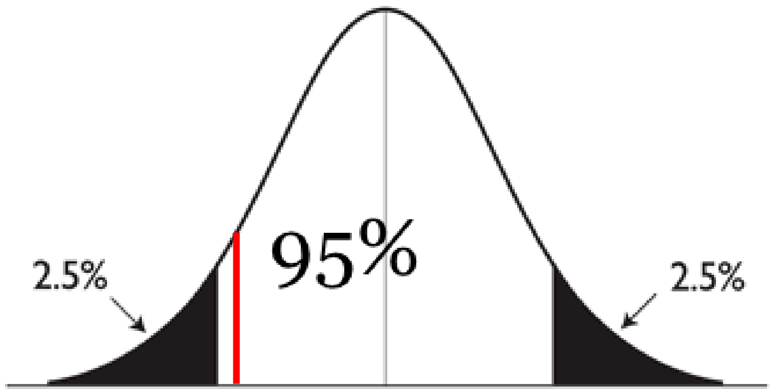
One- vs. two-tailed *t*-tests

- This means there is a temptation to *cheat* and switch from two-tailed (and a non-significant result)…

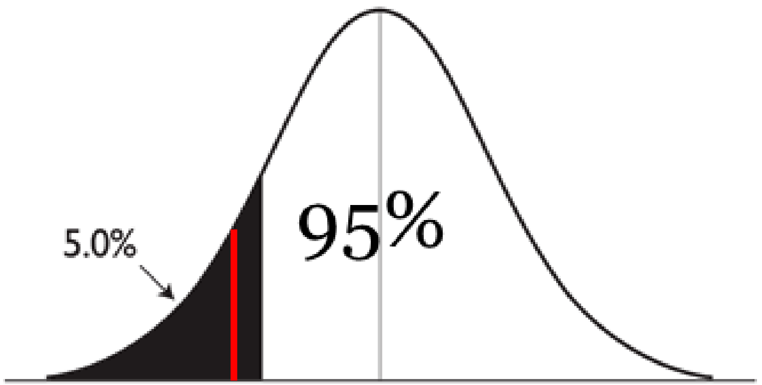
- …to a one-tailed test (and a significant difference between means)

t=-1.8时双侧没有不同，单侧有显著差异

- 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

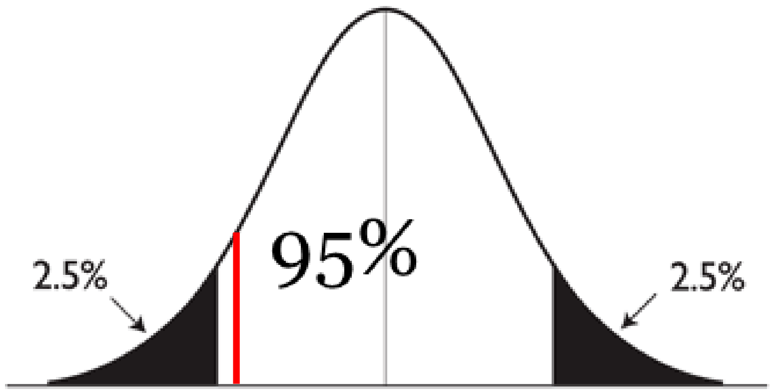
One- vs. two-tailed *t*-tests

- It is wrong to run a one-tailed test just because it is easier to prove that groups are different

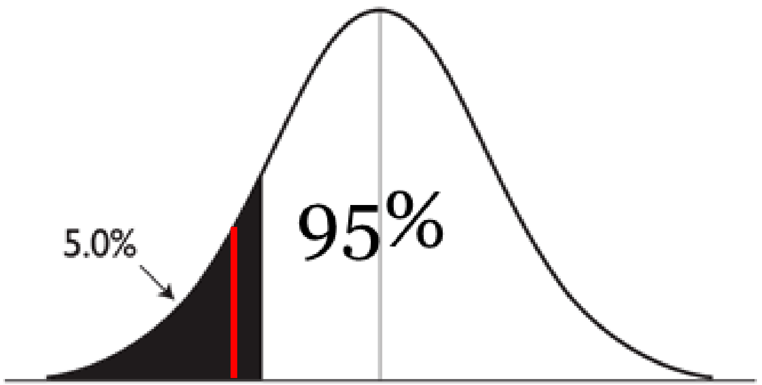
- Example: to test for differences between male and female height, you should *always* run a two-tailed test; you shouldn't argue that "males are always taller"

notes:

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



x=-1.8: inside CI



x=-1.8: outside CI

Conclusions

置信区间和所有t检验都假设正态分布

- 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: 目前的趋势是在报告的简介结果中提供p值，置信区间和t值

- **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)**

零假设总是假设两个被比较的均值没有不同

- 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*

  如果置信区间包括0，差异不显著

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

单侧检验不太常用(很难证明合理性)

Exercises:

```
library(ISwR)
```

File *kfm* (ISwR library)

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

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? 每一年的年份是相同的，所以用paired

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

```
t.test(kfm$weight ~ kfm$sex)
t.test(kfm$dl.milk ~ kfm$sex)
t.test(humanlongevity$longevity.women, humanlongevity$longevity.men, paired = T)
```