Lecture 10

T - tests

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Normal theory

Previously: permutation tests and bootstrapping to obtain estimated p-values, confidence intervals respectively.

Now the normal theory stand-point.

Simulation for hypothesis testing and confidence intervals is a relatively new technique.

Normal theory is traditional and has upsides.

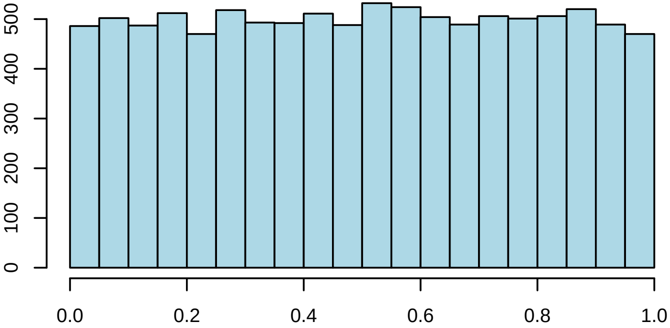
Today: central limit theorem and t-tests.

Central Limit theorem

Normal distribution approximates the binomial distribution but has a much more central fundamental role

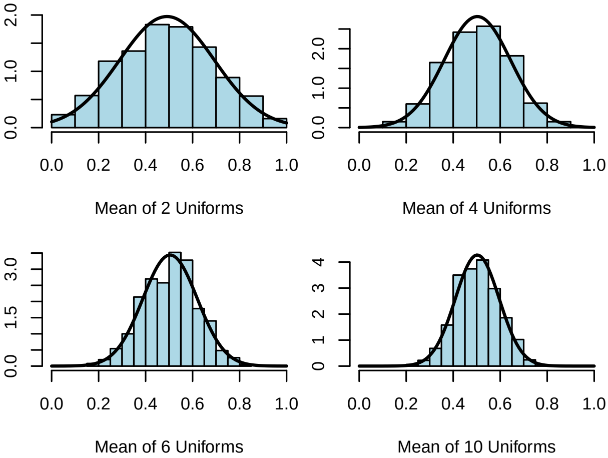
Uniform distribution

$$X_1, X_2, \ldots, X_n \sim \mathcal{U}(0, 1)$$



Average of uniform

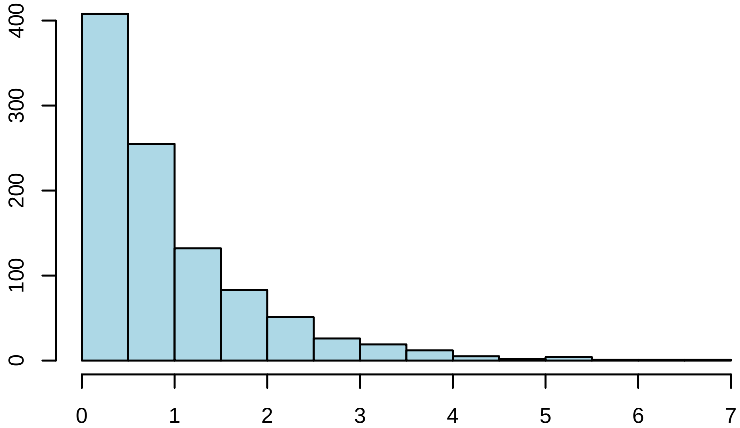
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$



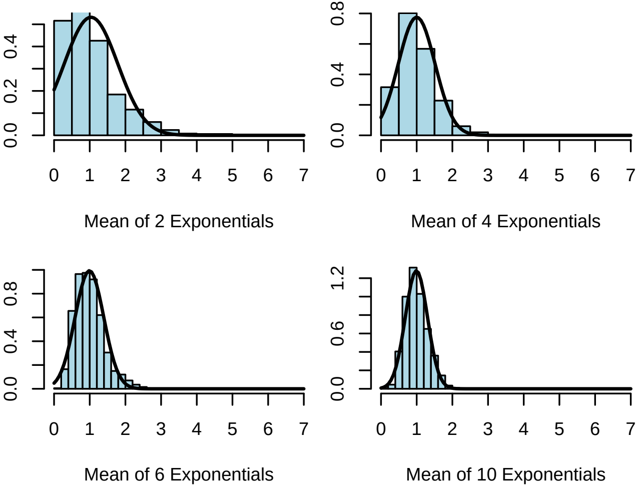
Exponential distribution

Let's look at another: the exponential distribution $\mathcal{E}(\lambda)$ with pdf:

$$f(x) = \lambda e^{-\lambda x}.$$



Average of exponentials



The mean gets closer to resembling a normal curve.

Sampling distribution

Sampling distribution

Let S be a statistic considered a random variable. Consider now n instances $S_1, S_2, \ldots, S_n$. Then its *sampling distribution* is how those S_i are distributed.

A function of the n times we sample.

As n gets large it gets closer to the population distribution of the statistic.

Central Limit Theorem

This tendency towards a normal distribution holds in general.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be iid[a] with mean μ and variance σ^2, and let

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i .$$

Then $\sqrt{n}(\bar{X}_n - \mu)$ converges (in distribution) to a Normal distribution as $n \to \infty$.

[a] independent and identically distributed

Central Limit Theorem

For large enough number of samples n from any distribution, the mean will approximately follow a normal distribution.

Rate of convergence?

"large enough n" depends on the distribution of original X_i.

Symmetric smooth distributions, good approximation for quite small n (10)

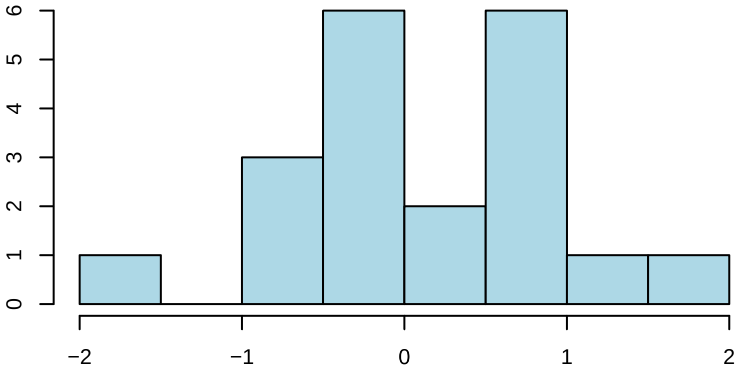
Skewed, bimodal or discrete distributions → larger n needed

$n > 30$ is a rule of thumb.

Testing for normality

Formal testing procedures – the Kolmogorov-Smirnov or Shapiro-Wilk test – we stick to a graphical check.

Just look at histogram?



quantile quantile plot or QQ plot

Idea: sort data and plot against what you should see if the data were exactly normally distributed.

Obtaining a QQ plot

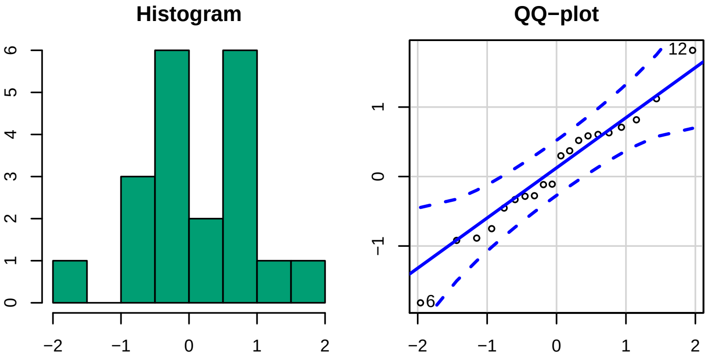
1. Let the data $y_1, \ldots, y_n$ be pre-sorted in increasing order. These divide the line into $n+1$ intervals: $n-1$ between each pair of ys and 2 at the ends.
2. Calculate a set of points $z_1, \ldots, z_n$ that divide the line into $n+1$ intervals based on equal normal probabilities such that by

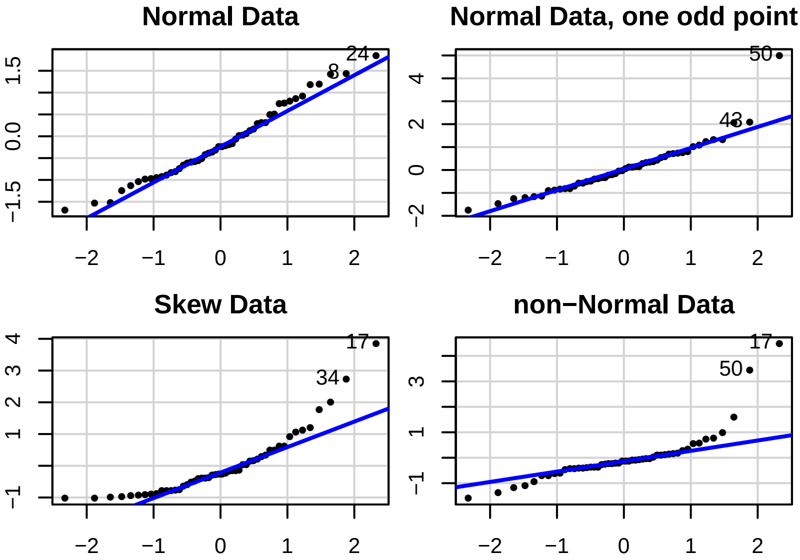
$$P(Z < z_i) = \frac{i}{n+1}$$

where Z is standard normal.

3. Plot each y_i against the corresponding z_i.

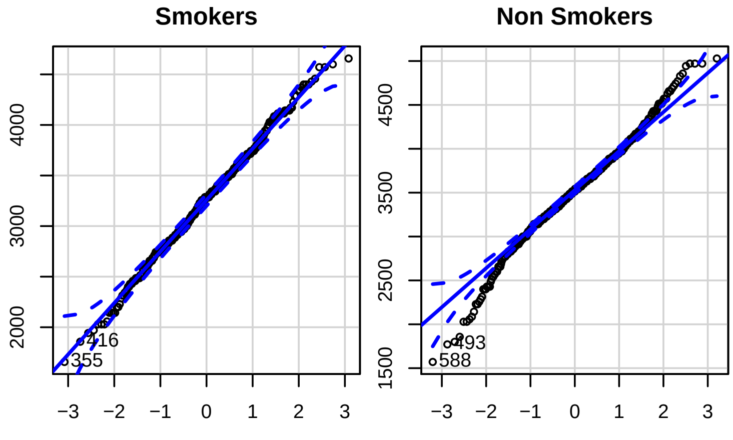
If the data are normal it should be close to a straight line.


Histogram
0
1
2
3
4
5
6
−2
−1
0
1
2
QQ−plot
12
6
−1
0
1
−2
−1
0
1
2


Normal Data
Normal Data, one odd point
Skew Data
non−Normal Data
24
8
50
43
17
34
17
50
1.5
0.0
−1.5
4
2
0
−2
4
3
2
1
−1
3
1
−1
−2
−1
0
1
2

Is the birth weight data normal?

Combined data is two distributions together – so seperate makes sense



Non-smokers look a bit non-normal: Central Limit Theorem tells us the mean distribution will be approximately normal.

The t test

Developed by William Gosset in 1908 – published under pen name Student, hence full name Student's *t*-test.

The *t*-test is one of the most misunderstood/feared/abused methods in statistics.

A t-test is any statistical hypothesis test "whose test statistic follows a Student's t-distribution under the null hypothesis".

Standardisation

Let $(x_1, \ldots, x_n)$ be data with mean $\bar{x}$ and standard deviation s.

1. Shift data by $\bar{x}$ so that its mean is 0.
2. Then scale by s to get a standard deviation of 1.

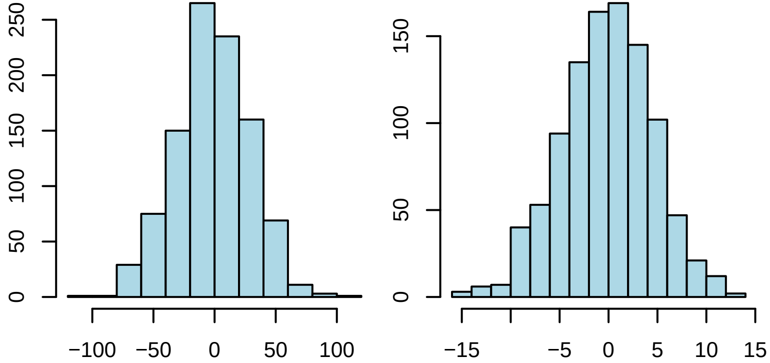
This gives standardised values:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

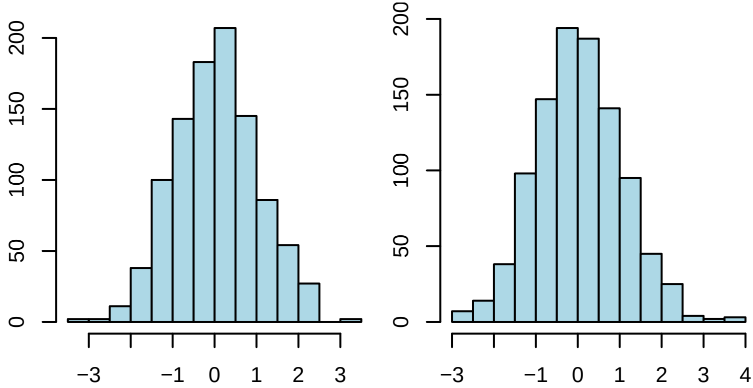
Exercise

Prove that the set of all z_i have zero mean and standard deviation of 1.

Differences in means for birth weight and store location data



birth weight and sales histograms after standardisation



Standard Error

1. The standard error (SE) of a statistic is the standard deviation of its sampling distribution.
2. When the statistic is the sample mean, it is called the standard error of the sample mean (SEM).

SEM

Can sometimes be estimated using a simulation – such as calculating 1000 means and computing their standard deviation.

In practise, the population variance is unknown, so we use the approximation.

SEM approximation

$$\text{SEM} = \frac{s}{\sqrt{n}},$$

where s is the data standard deviation and n the sample size.

One sample t - test

Suppose have some data whose sample mean is $\bar{x}$, and our hypothesis is that $\mu = \bar{x}$. We wish to evaluate if it's true.

So we use the t-statistic:

$$t = \frac{\bar{x} - \mu}{\text{SEM}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where s is the sample standard deviation and n is the sample size.

The hypotheses that arise are:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0 \text{ etc.}$$

The t-statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

can be shown to follow what is known as a t-distribution with $n - 1$ degrees of freedom.

Two sample t test

In maternal smoking problem, we are interested in seeing if there is a statistically significant difference in means.

Assume the underlying variances in the populations are the same

Let n_1 and n_2 be the number of elements in each group, and s_1 and s_2 be the (sample) standard deviations of each group.

$$\text{t-statistic} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

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

s_p is called the pooled standard deviation.

Simulate or use the t distribution

We could use this t-statistic, could simulate using permutations, obtain a p-value, and decide if there is enough evidence to accept the alternative hypothesis.

If the null hypothesis is true, the t-statistic will follow a t distribution, for large enough sample size (at least 30).

This brings us to the **t-test**.

Birth weight data

Summary	Smoke=No	Smoke=Yes
n	742	484
Mean	3515.64	3260.29
Standard Deviation	497.1	517.11

$$s_p^2 = \frac{(742-1)\times 497.1^2 + (484-1)\times 517.11^2}{742+484-2} = 255114.55$$

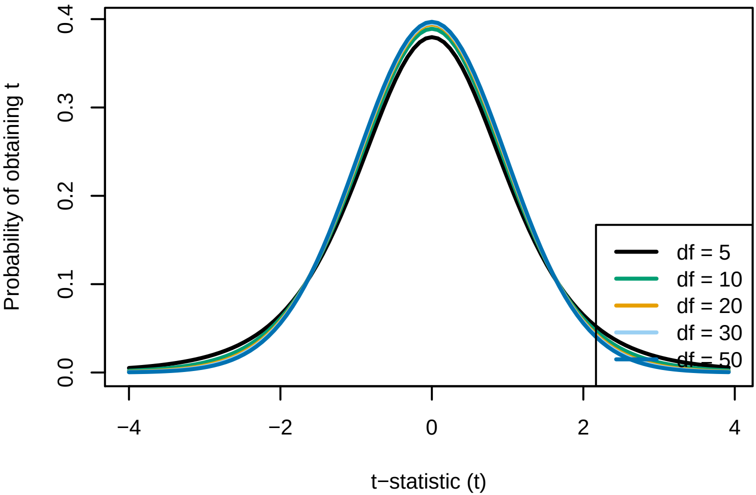
and

$$s_p = 505.09.$$

Thus the t-statistic is:

$$t = \frac{3515.64 - 3260.29}{505.09\sqrt{\frac{1}{742}+\frac{1}{484}}} = 8.653$$

t distribution



One parameter called degrees of freedom (df). In the two-sample test this is $n_1 + n_2 - 2$.

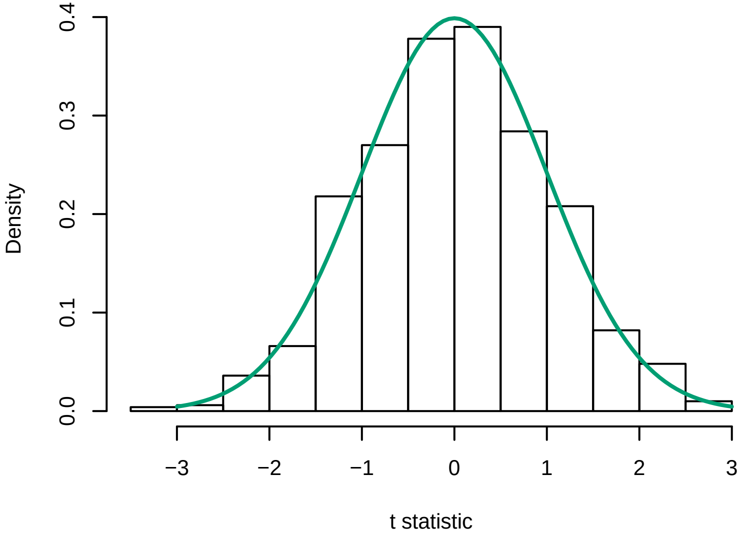
t distribution

It represents the sample distribution the statistic will follow provided:

the null hypothesis holds and population distribution of the statistic follows a normal distribution.

The central limit theorem helps establish this as a reasonable assumption.

So to estimate a p-value for a given hypothesis we can use this distribution rather than simulating as before.


Density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
t statistic

Birth weight R test

```
Two Sample t-test
data: bwt by smoke
t = 8.6527, df = 1224, p-value < 2.2e-16
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval: 197.4554 313.2520
sample estimates:
mean in group no mean in group yes
3515.639 3260.285
```

Office sales data

```
Two Sample t-test
data: sales by office
t = 1.9314, df = 98, p-value = 0.05632
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval: -0.2379317 17.5511689
sample estimates:
mean in group east mean in group west
162.6991 154.0425
```

Confidence intervals from a t distribution

Previously we estimated confidence intervals using bootstrapping.

Just as with hypothesis testing, we can also approximate confidence intervals using the *t*-distribution.

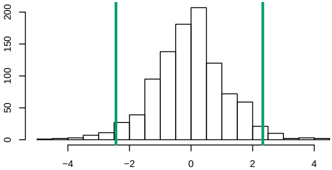
From the t-test above, provided the true difference in means is zero, then

$$\frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, provided the sample sizes are large enough (thanks to the guarantees of the central limit theorem).

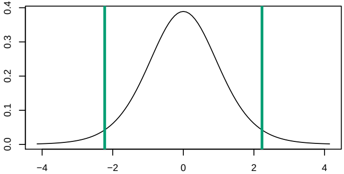
So instead of bootstrapping, we could simulate from the t-distribution, and find those points that have 2.5% less and 2.5% greater.

For example, simulating from a *t*-distribution for 1000 simulations:



In fact, this is overkill. We know already the formula for the t-distribution so we can use it directly.

Here is the 95% interval of *t* distribution with df $= 10$:



Confidence intervals

The middle interval of a t distribution will always be plus and minus the same number because of symmetry.

The dark lines haves 2.5% of the t-distribution to left and 2.5% to right. For 10 degrees of freedom, these values are -2.228, 2.228. We often write this number as $t_{0.025,10}$, which is in this case 2.228.

It means that 95% of the time a t-statistic with 10 degrees of freedom, is between $-t_{0.025,10}$ and $+t_{0.025,10}$, or more generally 95% of the time:

$$(\bar{x}_1 - \bar{x}_2) - t_{0.025,(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$< \mu_1 - \mu_2 <$$

$$(\bar{x}_1 - \bar{x}_2) + t_{0.025,(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Birth weight confidence intervals

For the birth weight data we get 95 percent confidence interval:
197.4554 313.2520

Two sample t - test summarised

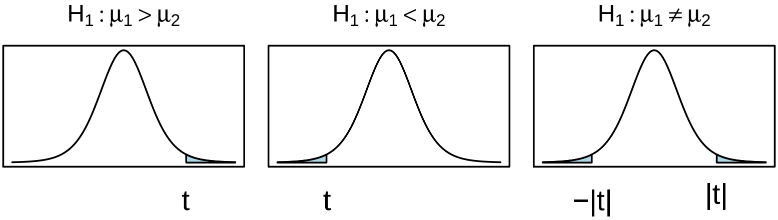
- $H_1 : \mu_1 > \mu_2$ $1 - pt(t, df)$
- $H_1 : \mu_1 < \mu_2$ $pt(t, df)$
- $H_1 : \mu_1 \neq \mu_2$ $2 \times (1 - pt(abs(t), df))$

```
t.test(x~grp, var.equal=TRUE)
```

or

```
t.test(x1,x2, var.equal=TRUE)
```

The area representing the p value for each hypothesis test is then:



A 95% confidence interval for the actual difference in means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{1/n_1 + 1/n_2}$$

where $t_{\alpha/2}$ is derived from a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

For 95% $\alpha = 0.05$, `t = qt(1-0.05/2, df)`

Paired data

From http://www.statsci.org/data/oz/nzhelmet.html

After purchasing a batch of flight helmets that did not fit the heads of many pilots, the NZ Airforce decided to measure the head sizes of all recruits.

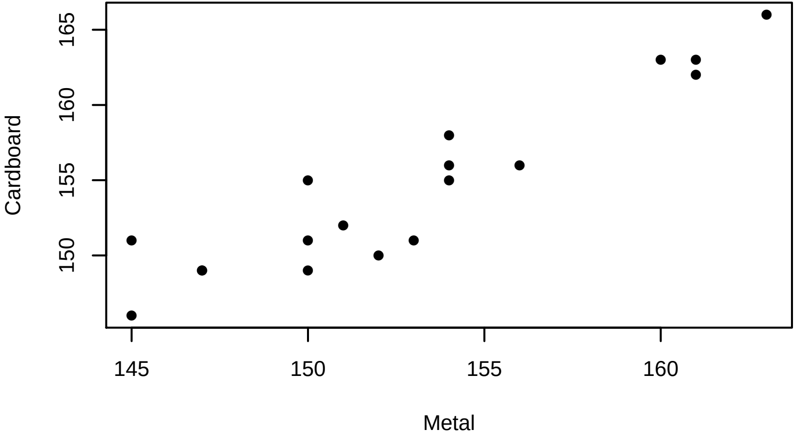
Information was collected to determine the feasibility of using cheap cardboard callipers to make the measurements, instead of metal ones which were expensive and uncomfortable.

The data lists the head diameters of 18 recruits measured once using cardboard callipers and again using metal callipers. *The question was whether there is any systematic difference between the two sets of callipers.*

Every head is measured twice: once with cardboard and once with metal callipers.

Here's an extract:

```
  Cardboard Metal
1       146   145
2       151   153
3       163   161
4       152   151
5       151   145
6       151   150
```


Cardboard
165
160
155
150
145
150
155
160
Metal

Paired t statistic

differences between the pairs.

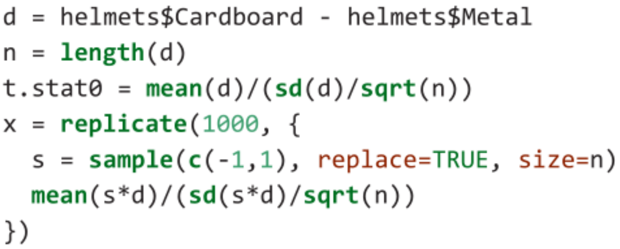
$$d_i = x_i - y_i$$

If there is no difference the d_i would have mean zero. So a t-statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $\bar{d}$ is the mean and s_d is the standard deviation of the d_i.

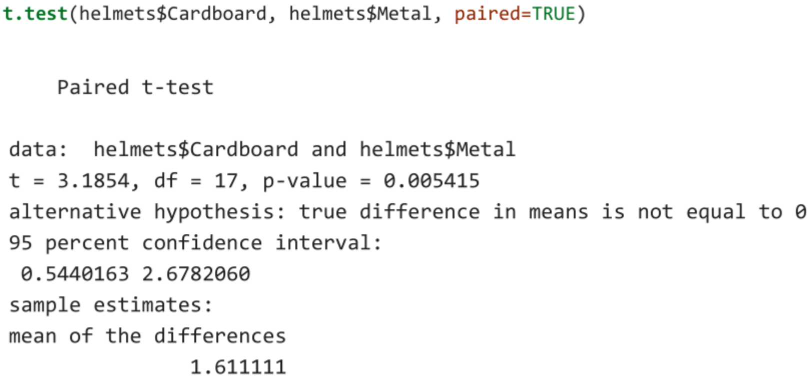
p-value by simulation



```
d = helmets$Cardboard - helmets$Metal
n = length(d)
t.stat0 = mean(d)/(sd(d)/sqrt(n))
x = replicate(1000, {
  s = sample(c(-1,1), replace=TRUE, size=n)
  mean(s*d)/(sd(s*d)/sqrt(n))
})
```

One output $p = 0.009$

Use t-distribution directly



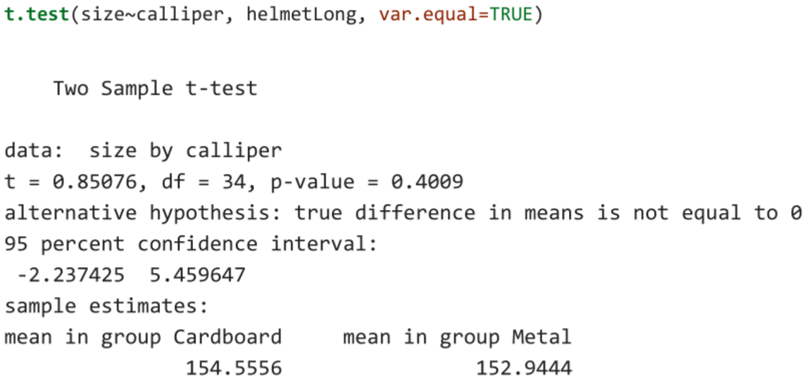
```
t.test(helmets$Cardboard, helmets$Metal, paired=TRUE)

    Paired t-test

data:  helmets$Cardboard and helmets$Metal
t = 3.1854, df = 17, p-value = 0.005415
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5440163 2.6782060
sample estimates:
mean of the differences
               1.611111
```

Paired vs. unpaired data

Suppose we simply compared cardboard against metal?

size	calliper
153	Metal
163	Metal
155	Cardboard
154	Metal
160	Metal
151	Cardboard
147	Metal
163	Cardboard
150	Metal
154	Metal

Paired vs. unpaired data



```
t.test(size~calliper, helmetLong, var.equal=TRUE)


    Two Sample t-test

data:  size by calliper
t = 0.85076, df = 34, p-value = 0.4009
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.237425  5.459647
sample estimates:
mean in group Cardboard     mean in group Metal
               154.5556                152.9444
```

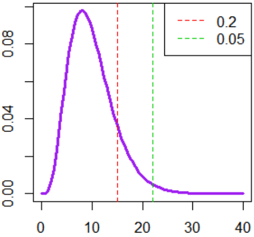
Hypothesis testing outcomes

There are four possibilities:

1. Not rejecting H_0 when you shouldn't — Good
2. Rejecting H_0 when you should — Good
3. Not rejecting H_0 when you should — Bad
4. Rejecting H_0 when you shouldn't — Bad

Two good outcomes and two bad outcomes.

The "Critical Value" controls the occurrence of options 1 & 2.



Type I & II errors

False positives and false negatives

- Type I error — finding evidence **against** the null although it is actually **true**. Rejecting H_0 when it is correct!
- Type II error — failing to find evidence **against** the null when it is actually **false**. Not rejecting H_0 when it is wrong!

Type I errors are easy to control.

Type II errors are harder to control.

Controlling Type I errors

Type I error controlled by statistical significance $p < \alpha$

Suppose null is true.

For $\alpha = 0.05$, probability of type I error is 5%.

Type II errors and power

Power

$$\text{power} = 1 - \text{Prob}(\text{type II error})$$

It is the probability of finding evidence **against** the null when it is actually **false**.

Testing method and/or more samples gives more power.

Pairing gave more power.

More (correct) assumptions / narrower hypothesis yield more power.