

228.371 - Statistical Modelling for Engineers and Technologists

Week 1. Foundations of Inference

Dr. Daniel Walsh
IIMS 3.07 x 43603
d.c.walsh@massey.ac.nz

Dr Nihal Jayamaha
n.p.jayamaha@massey.ac.nz

Semester One - 2015

Class Reps

- ▶ Sign up at www.asa.ac.nz/classreps and identify yourself to me.
- ▶ Training Wed 11 March 2015.

Timetable

► Lectures:

Tue	1pm- 2pm	QB2
Wed	9am-10am	SC AUD
Thu	11am-12pm	SNW200

► Computer Labs:

Stream 1:

Wed	10am-11am	CL-QB4
Thu	1pm- 2pm	CL-QB4

Stream 2:

Wed	11am-12pm	CL-QB4
Thu	12pm- 1pm	CL-QB4

Assessment

Blah

- ▶ Assignment 1, due 24 March (5PM), 30%
 - ▶ Available soon, work on it as we progress through the topics
- ▶ Assignment 2, due 12 May (5PM), 30%
- ▶ Invigilated computer assessment, in weeks 13-15, 40%

Course Outline

Week 1	Foundations of Inference (DW)
Week 2-4	Linear Regression (DW)
Week 5-7	Design of Experiments (NJ)
Week 8-9	Generalised linear modelling (DW)
Week 10-12	Optimisation techniques (NJ)

Section Outline

- ▶ Sampling
- ▶ Confidence Intervals
 - ▶ Proportions
 - ▶ Difference in proportions
 - ▶ Means
 - ▶ Paired data
 - ▶ Difference in means
- ▶ Hypothesis Testing
 - ▶ Difference in proportions
 - ▶ One sample mean
 - ▶ Two sample means
 - ▶ Paired data

Random Sampling

- ▶ Each unit in the population equally likely to end up in the sample.
- ▶ Key to understanding how one random sample will differ from the next.
- ▶ Key to extrapolating from sample to population.

Random vs Haphazard Sampling

- ▶ Testing rats with different treatments for run time through maze.
- ▶ Groups decided by “*reaching in and grabbing a rat without really looking.*”
- ▶ Result?

Random vs Haphazard Sampling

- ▶ Grade computations in excel; those with a missing (excused) assessment are not computed automatically.
- ▶ Treat the remaining individuals as a sample to see how the class is doing?

Moral: Formal assignment of random numbers to the experimental units is the way to go.

Review: Central Limit Theorem

- ▶ Sample mean (and by extension sample proportion) tend to have a normal distribution.
- ▶ Standard deviation of this sample mean is less than the population standard deviation by a factor of \sqrt{n} .
- ▶ Foundation of most of the inferences we will learn.
- ▶ Depends on having a **random sample**.

Confidence Intervals (for proportions)

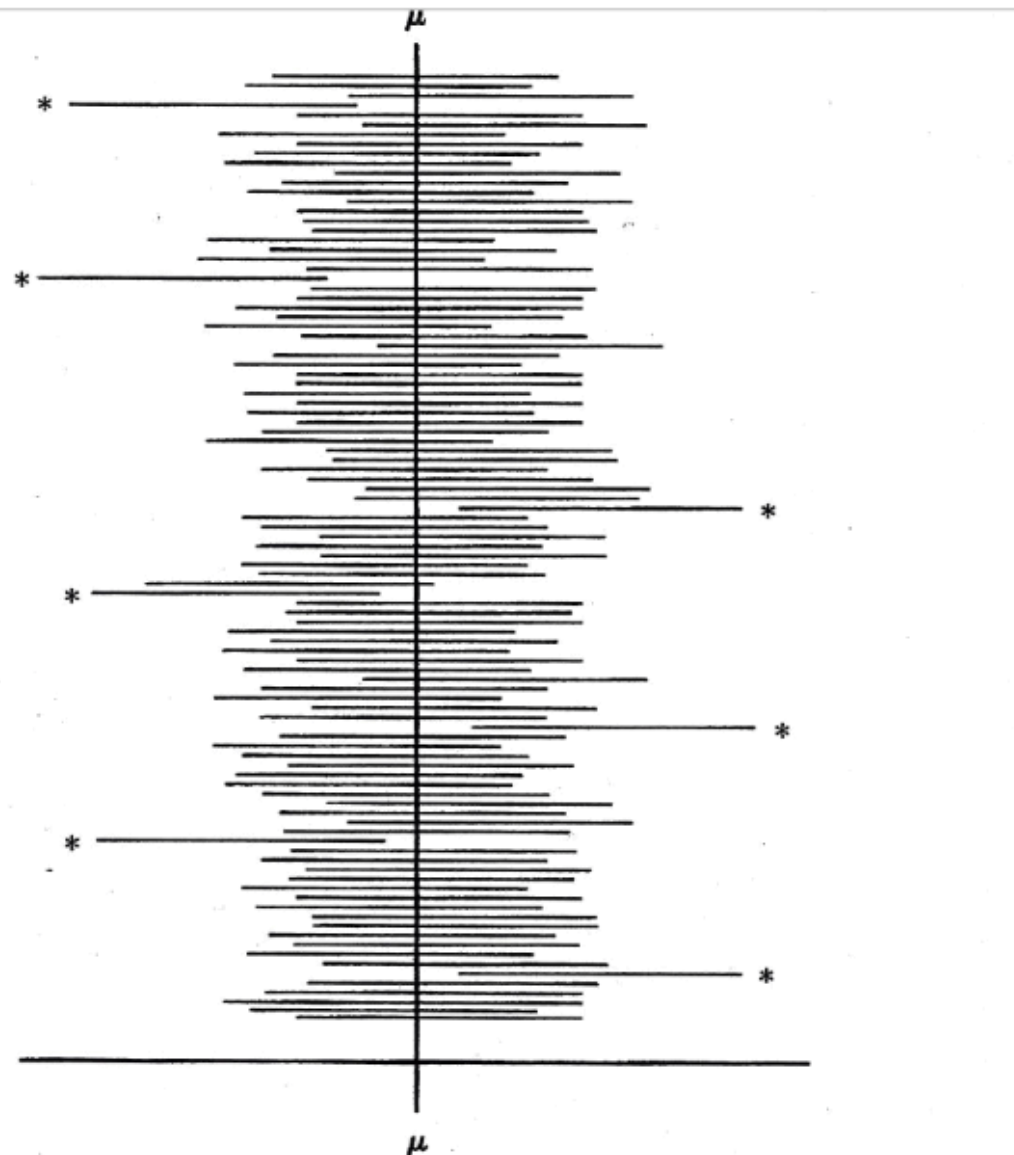
- ▶ Want to make statement “*With 95% confidence, the population proportion lies between x and y .*”
- ▶ By this we mean, if we repeated the experiment many times and used our method to construct the CI, 95% of those intervals would contain the true population proportion.

Rationale

- ▶ Most (roughly 95%) of sample proportions will be within 2 standard errors of the true proportion (property of the normal distribution).
- ▶ Therefore, a window of 2 (estimated) standard errors around the sample proportion should include the true proportion about 95% of the time.

Confidence Intervals

95% CIs for μ from 100 different samples (* indicates an interval that does not include μ)



General Format for Confidence Intervals

- ▶ For any confidence level, a confidence interval for either a population proportion or a population mean can be expressed as:

$$\text{Sample estimate} \pm \text{Multiplier} \times \text{Standard error}$$

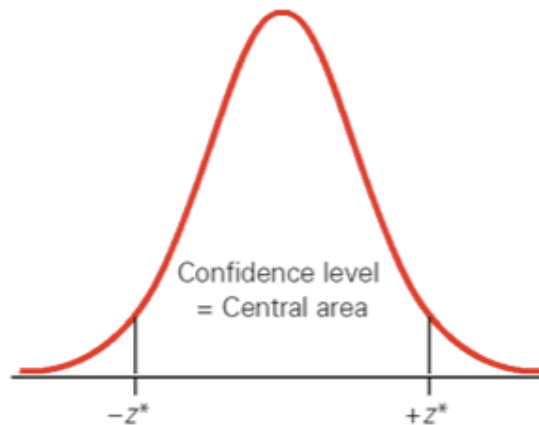
- ▶ The multiplier is affected by the distribution of the estimate and the choice of confidence level.
- ▶ Recall that the standard error is the standard deviation of our test statistic (e.g. sample proportion).

More about the Multiplier

Confidence Level	Multiplier	Confidence Interval
90	1.645 or 1.65	$\hat{p} \pm 1.65$ standard errors
95	1.96, often rounded to 2	$\hat{p} \pm 2$ standard errors
98	2.33	$\hat{p} \pm 2.33$ standard errors
99	2.58	$\hat{p} \pm 2.58$ standard errors

Note: Increase confidence level \Rightarrow larger multiplier.

- ▶ Multiplier, denoted as z^* , is the number such that the area between $-z^*$ and z^* under the standard normal curve corresponds to the desired confidence level.
- ▶ Use software.



Formula for a Confidence Interval for a Population Proportion p

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- ▶ \hat{p} is the sample proportion.
- ▶ z^* denotes the multiplier.
- ▶ $\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$ is the standard error of \hat{p} .

Case Study: Nicotine Patches vs Zyban

Study: New England Journal of Medicine 3/4/99

- ▶ 893 participants **randomly allocated** to **four treatment groups**: placebo, nicotine patch only, Zyban only, and Zyban plus nicotine patch.
- ▶ Participants **blinded**:
 - ▶ All used a patch (nicotine or placebo).
 - ▶ All took a pill (Zyban or placebo).
- ▶ Treatments used for nine weeks.

Computation of Zyban + Patch Interval

- ▶ 245 individuals in this group; 95 not smoking at the end of the 9 weeks (0.388).

- ▶ Estimated standard error:

$$\sqrt{0.388 \times (1 - 0.388)/245} = 0.031.$$

- ▶ Confidence interval

$$0.388 \pm 1.96 \times 0.031 (0.328, 0.449)$$

- ▶ Or `prop.test(95,245)` (more exact calculation; differences in third decimal place)

```
prop.test(95,245)
```

```
1-sample proportions test with continuity correction
```

```
data: 95 out of 245, null probability 0.5
```

```
X-squared = 11.902, df = 1, p-value = 0.0005607
```

```
alternative hypothesis: true p is not equal to 0.5
```

```
95 percent confidence interval:
```

```
0.3269738 0.4521169
```

```
sample estimates:
```

```
      p  
0.3877551
```

Case Study: Nicotine Patches vs Zyban

Treatment	Subjects	Proportion Not Smoking	Approx. 95% CI
Placebo only	160	.188	.13 to .25
Nicotine patch	244	.213	.16 to .26
Zyban	244	.348	.29 to .41
Zyban and nicotine patch	245	.388	.33 to .44

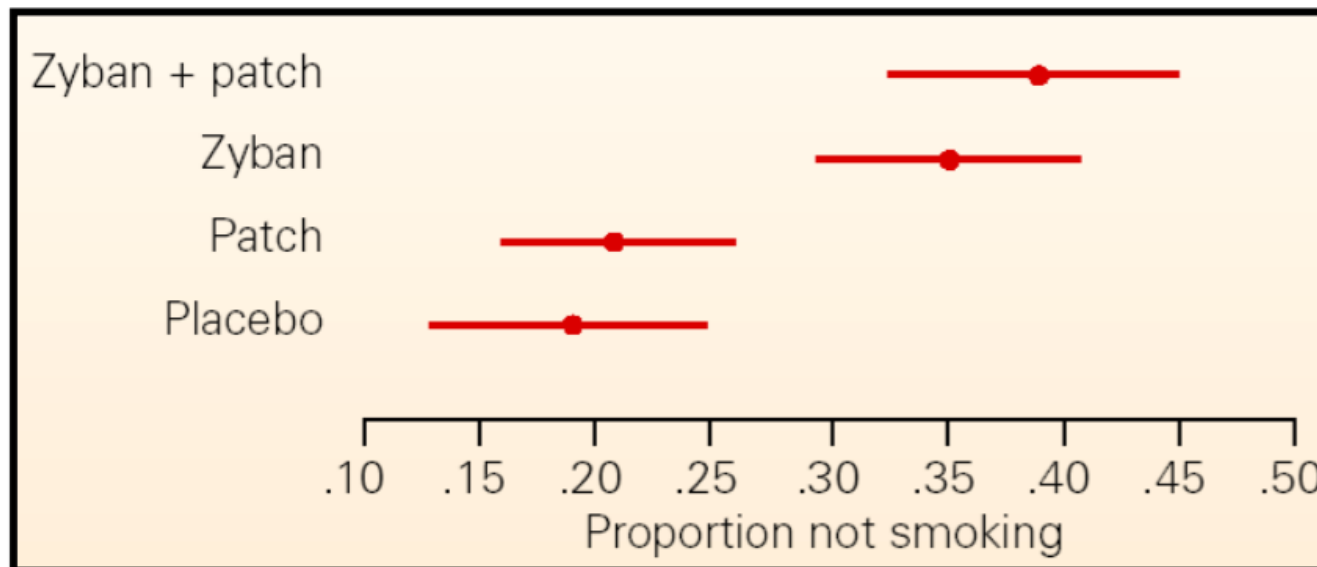
Conclusions:

Zyban is effective

(no overlap of Zyban and no Zyban CIs)

Nicotine patch is not particularly effective

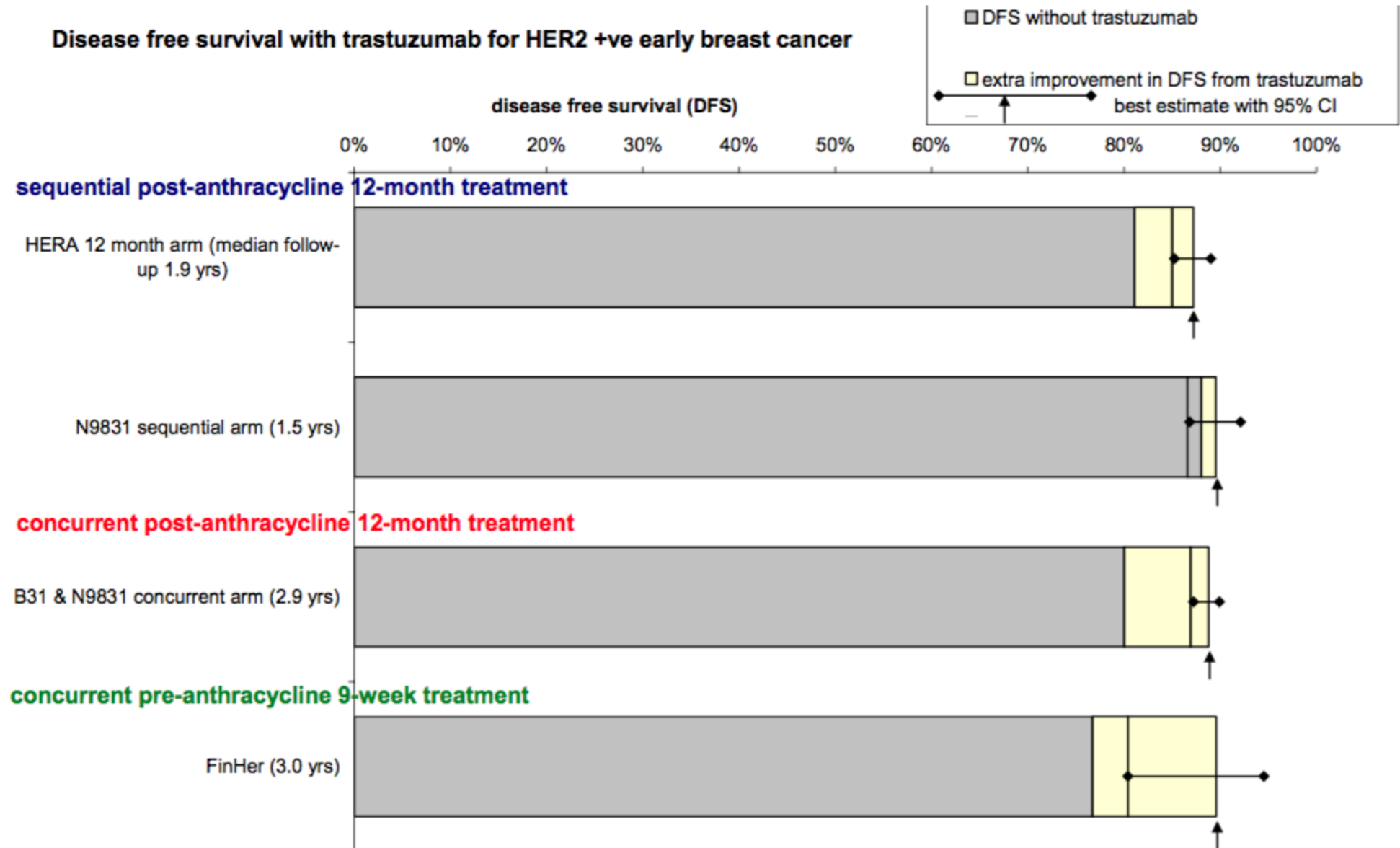
(overlap of patch and no patch CIs)



Application

- ▶ We will learn later in the paper about logistic regression, which would allow us to separate the effect of Zyban and nicotine patch, and look for interactions.
- ▶ Confidence intervals are still a useful way of displaying the results, especially to non specialists. See the following from Pharmac:

Disease free survival with trastuzumab for HER2 +ve early breast cancer



Notes on sample size

- ▶ Notice how the width of intervals depends on sample size (through the standard error).
- ▶ Intervals based on normal distribution, so sample size must be large enough for this to be reasonable:

$n\hat{p}, n(1 - \hat{p})$ at least 5

The Difference Between Two Proportions (Independent Samples)

A CI for the Difference Between Two Proportions (Independent Samples):

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

where z^* is the value of the standard normal variable with area between $-z^*$ and z^* equal to the desired confidence level.

Necessary Conditions

- ▶ **Condition 1:** Sample proportions are available based on independent, randomly selected samples from the two populations.
- ▶ **Condition 2:** All of the quantities

$$n_1\hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2, \text{ and } n_2(1 - \hat{p}_2)$$

are at least 5

Example: Patch vs Antidepressant (cont)

Study: $n_1 = n_2 = 244$ randomly assigned to each group.

Zyban: 85 of the 244 Zyban users quit smoking $\hat{p}_1 = 0.348$.

Patch: 52 of the 244 patch users quit smoking $\hat{p}_2 = 0.213$.

So, $\hat{p}_1 - \hat{p}_2 = 0.348 - 0.213 = 0.135$ and $s.e.(\hat{p}_1 - \hat{p}_2) = 0.040$.

$$0.135 \pm 1.96 \times 0.040 \Rightarrow 0.135 \pm 0.080 \Rightarrow 0.57 \text{ to } 0.213$$

Interval does not include the value 0, so it supports a difference between the success rates of the two methods. Note not exactly the same as querying whether the 1-sample intervals overlap.

R computation

```
prop.test (c(85,52), c(244,244))
```

2-sample test for equality of proportions with continuity correction

data: c(85, 52) out of c(244, 244)

X-squared = 10.3918, df = 1, p-value = 0.001266

alternative hypothesis: two.sided

95 percent confidence interval:

0.05231817 0.21817363

sample estimates:

prop 1	prop 2
0.3483607	0.2131148

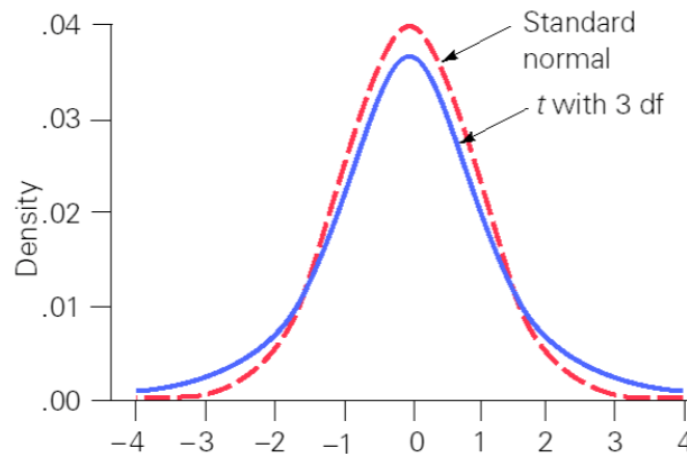
Confidence intervals for means

- ▶ The same thing . . . almost.
- ▶ When we estimate the standard deviation, σ , our intervals come out a little too small.
- ▶ We can fix this by using a multiplier from the t -distribution with $n - 1$ degrees of freedom.
- ▶ Approaches normal distribution as n approaches infinity.

Student's t -distribution: Replacing σ with s

Dilemma: We generally don't know σ . Using s we have:

$$t = \frac{\bar{x} - \mu}{s.e.(\bar{x})} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$$



If the sample size n is small, this standardised statistic will not have a $N(0, 1)$ distribution but rather a t -distribution with $n - 1$ degrees of freedom (df).

General CI for One Mean or Paired Data

A Confidence Interval for a Population Mean:

$$\bar{x} \pm t^* \times s.e.(\bar{x}) \Rightarrow \bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

where the multiplier t^* is the value in a t -distribution with degrees of freedom = $df = n - 1$ such that the area between $-t^*$ and t^* equals the desired confidence level.

Conditions:

- ▶ Population of measurements is bell-shaped and a random sample of any size is measured; or
- ▶ Population of measurements is not bell-shaped, but a large random sample is measured, $n \geq 30$.

Example of CI

- ▶ To detect insects, bats emit a sound and then listen for the echo. The distance between bat and insect when the insect is first detected varies according to some unknown distribution. Suppose a sample of $n = 100$ distances are recorded, and the following statistics are computed:

$$\bar{x} = 432\text{cm}$$

$$s = 18.1\text{cm}$$

- ▶ What is the 95% CI for the unknown mean μ ?

Computations

- ▶ Multiplier from t distribution is 1.98 (cf 1.96 for normal distribution).

```
qt(0.975,df=99)
```

```
[1] 1.984217
```

- ▶ $532 \pm 1.98 \times 18.1/\sqrt{100}$
- ▶ 532 ± 3.58 .
- ▶ (528.42, 535.58)
- ▶ We are still assuming the sample mean has a normal distribution; reasonable since $n \geq 30$.

A frequent situation call for a CI for the mean

- ▶ Often data consist of pairs of observations made on same unit (e.g. before and after some experimental treatment) or on matched pairs of alike subjects (twins, brothers etc).
- ▶ Difference in responses between them can be attributed to treatment.
- ▶ Interval is for the mean difference, based on differences for each pair.

One-sample t confidence interval for paired data: An example

Does aspirin influence blood-clotting ability? In a 1971 study, the amount of time for blood to clot in 12 adult males was measured both before and three hours after taking two aspirin tablets.

Total number of subjects in the experiment was 12. Sample mean for the before/after difference was 0.1084, with sample standard deviation of the difference was 0.5071.

Confidence Intervals for Differences

CI for difference in two means

- ▶ 2 cases: equal variance, unequal variance.
- ▶ R commands:
 - ▶ `t.test (x,y, var.equal=TRUE)`
 - ▶ `t.test(x,y) #default has var.equal=FALSE`
- ▶ Both assume independent random samples; either large samples (≥ 30) or normal population distribution for each group.

Pooled Confidence Interval

Pooled CI for the Difference Between Two Means (Independent Samples):

$$\bar{x}_1 - \bar{x}_2 \pm t^* \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where t^* is found using a t -distribution with $df = (n_1 + n_2 - 2)$ and s_p is the pooled standard deviation.

Equal Variance Assumption

- ▶ Often reasonable to assume the two populations have **equal population standard deviations**, or equivalently, equal population variances.
- ▶ Estimate of this variance based on the combined or “pooled” data is called the **pooled variance**. The square root of the pooled variance is called the **pooled standard deviation**:

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

General CI for Difference Between Two Means

A CI for the Difference Between Two Means (Independent Samples):

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the value in a t -distribution with area between $-t^*$ and t^* equal to the desired confidence level.

The df used depends on the observed variances.



Degrees of Freedom

The t -distribution is only approximately correct and df formula is complicated (Welch's approx):

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2} \right)^2}$$

Statistical software can use the above approximation.

Example: Small-sample CI for $\mu_1 - \mu_2$

An educator believes that certain new reading activities will help school children improve in their reading abilities.

One class of 21 students takes part in the activities, and one class of 23 students does not take part in the activities.

At the end of 8 weeks, the students are given a Degree of Reading Power (DRP) test.

Example continued : Test scores

Activity group:

24 61 59 46 43 44 52 43 58 67 62
57 71 49 54 43 53 57 49 56 33

No Activity group:

42 33 46 37 43 41 10 42 55 19 17 55
26 54 60 28 62 20 53 48 37 85 42

- Find 95% CI for the mean difference.

Assuming equal variance

- ▶ Center: $51.48 - 41.52 = 9.96$
- ▶ $df = 23 + 21 - 2 = 42$, multiplier $qt(.975, 42) = 2.018$
- ▶ Individual sd's: 11.01, 17.15
- ▶ $s_p = \sqrt{(11.01^2 * 20 + 17.15^2 * 22) / 42} = 14.55$
- ▶ $9.96 \pm 2.018 * 14.55 \sqrt{1/21 + 1/23} = (1.10, 18.82)$
- ▶ 95% confident the difference in mean scores is in this interval. Positive, could be small or large.

In R

```
activity <- c(24, 61, 59, 46, 43, 44, 52, 43, 58, 67, 62,
             57, 71, 49, 54, 43, 53, 57, 49, 56, 33)
noactivity <- c(42, 33, 46, 37, 43, 41, 10, 42, 55, 19, 17,
               55, 26, 54, 60, 28, 62, 20, 53, 48, 37, 85, 42)
t.test (noactivity, activity, var.equal=TRUE)
```

Two Sample t-test

```
data:  noactivity and activity
t = -2.2666, df = 42, p-value = 0.02863
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.817650  -1.091253
sample estimates:
mean of x mean of y
 41.52174  51.47619
```

Unequal Variances

```
t.test (noactivity, activity, var.equal=FALSE)
```

Welch Two Sample t-test

data: noactivity and activity

t = -2.3109, df = 37.855, p-value = 0.02638

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-18.67588 -1.23302

sample estimates:

mean of x mean of y

41.52174 51.47619

Rule of Thumb

- ▶ If the standard deviation differs by less than a factor of 2, the choice of method is not a big deal.
- ▶ If larger than two, use unequal variances.

Dictionary definition of a hypothesis

- ▶ *“A hypothesis is an idea which is suggested as a possible explanation for a particular situation or a condition, but which has not been proven to be correct.”* - Dictionary.
- ▶ A hypothesis test establishes if the hypothesis in question can be accepted or rejected based on evidence presented.

Preliminary concept about hypothesis testing

Suppose a person is accused of a crime. According to the NZ law, an accused is presumed innocent until proven guilty. There are 2 scenarios:

Ho: The defendant is **not guilty** vs

Ha: The defendant is **guilty**

Suppose you are a member of the jury; you need to present a verdict based on the evidence presented to the court, i.e. you need to decide if the defendant is guilty (i.e. reject H_0 in favour of H_a) or the defendant is not guilty (i.e. do not reject H_0).

Once you make the final verdict, you ask yourself if you have committed an error of judgement: (i) Has the criminal been erroneously freed? (ii) Has an innocent person been falsely convicted and penalised?

Null versus alternative

- ▶ **Null hypothesis:** H_0 is the hypothesis that is set up primarily to see whether it can be rejected. It can be viewed as a strawman set up to be knocked down.
- ▶ It is precisely what we do in criminal proceedings, where an accused is presumed to be innocent until his/her guilt has been established beyond a reasonable doubt. The presumption of innocence is a null hypothesis.

Null versus alternative

- ▶ **Alternative hypothesis:** H_a is the hypothesis which we accept when the null hypothesis is rejected.
- ▶ When our goal is to establish an assertion with substantive support obtained from the sample, the negation of the assertion is taken to be the null hypothesis H_0 , and the assertion itself is taken to be the alternative hypothesis H_a .

Referring to the criminal proceedings again, the intent of the hearing is to establish the assertion that the accused is guilty rather than to prove that he/she is innocent.

- ▶ A useful rule of thumb is:
 - ▶ If it is of interest to show that a hypothesis is **true**, make it the **alternative** hypothesis.
 - ▶ It is of interest to show that a hypothesis is **false**, make it the **null** hypothesis.

Decision table

Decision	H_0 False	H_0 true
Accept H_0	Type II Error :(Correct decision :)
Reject H_0	Correct decision :)	Type I Error :(

A medical analogy: false negative versus false positive

Consider a medical scenario in which you are tested for a disease. Most tests for diseases are not 100% accurate. In reading your results, the lab technician or physician must make the choice between 2 hypotheses:

Null hypothesis: You do not have the disease

Alternative hypothesis: You have the disease.

Choice 1: Based on the 'negative' result, the doctor declares you are healthy.

Potential error: You are actually diseased but have been told you are not, so your test is actually false negative.

Choice 2: Based on your 'positive' result, your doctor says that you are diseased.

Potential error: You are actually healthy but have been told you are diseased so the test is false positive.

Which error is more serious? It depends on the disease and on the consequence of a wrong diagnosis!

Type I and Type II errors

- ▶ Type I error is denoted by α .
- ▶ Type II error is denoted by β .

Naturally we would like to keep both α and β small. However, $\alpha + \beta$ is approximately constant so as α decreases, β increases.

A usual practice is to select α that is tolerable for the situation.

Significance level

Significance level or level of significance of the test is denoted by α :

This is the probability of making type I error. Thus a test with $\alpha = 0.01$ is said to have a significance level 0.01.

When the p -value (see slide after the next) is less than α , we say that the results are *statistically significant* at the α level.

Authors often report 'the result is significant' when the null is rejected in favour of the alternative. In most cases, the non-significant results are not reported; consequently other investigators would study the same problem again without knowing these hypotheses have been tested before - what a waste of time!

Test statistic

Test Statistics

- ▶ The function of the data that is used to decide between H_0 and H_a .
- ▶ A test statistic is a quantity computed from the sample data which is used to provide evidence about the plausibility of the null hypothesis.
- ▶ The sample proportion \hat{p} is normally distributed. We use z as the test statistic if testing hypotheses concerning p .

Interpretation of p -value

P -value: (observed significance level)

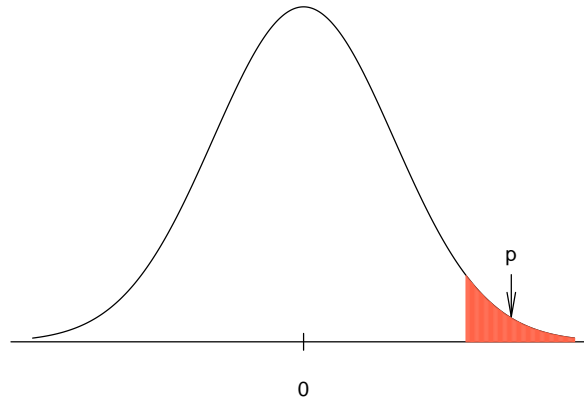
- ▶ The probability, calculated assuming H_0 is true, that the test statistic would take a value at least as extreme or more extreme than that actually observed is called the p -value of the test.
- ▶ It is the **tail probability** for which the statistic will take a value **more extreme** than that actually observed.

Meaning of a small p -value

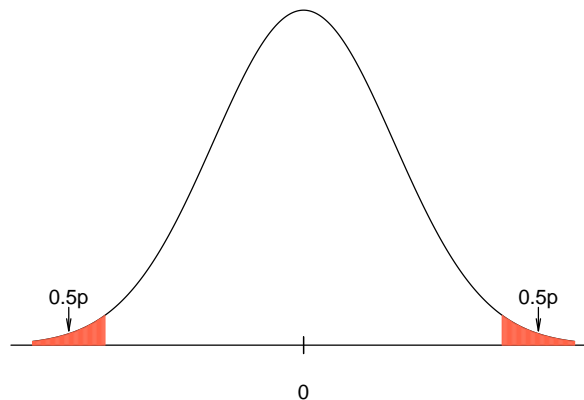
- ▶ The smaller the p -value, the more contradictory is the data to the null hypothesis H_0 ; or the smaller the p -value, the stronger evidence against H_0 provided by the data.
- ▶ “ P -value very small” means that very unlikely thing had happened if the null hypothesis had been true. So in a way it indicates that the null is unlikely to be true, so supports our rejection of the null.

Curves

- ▶ p -value of one tailed test:



- ▶ p -value of two tailed test:



Revision of jargon

Null and alternative hypotheses:

- ▶ H_0 : Claim about populations that is initially assumed to be true.
- ▶ H_a : Claim that is contradictory to the null hypothesis.

Type I and type II errors

- ▶ Reject the null hypothesis H_0 given it is true - type I error.
- ▶ Do not reject H_0 when it is false - type II error.

Test statistic

- ▶ Number based on the sample *which is used to provide evidence about the plausibility of the null hypothesis*.

p -value

- ▶ It is the tail probability for which the statistic (under H_0) is **more extreme** than the observed value of the statistic. Small p -value implies sufficient evidence against the null.

Significance level α

- ▶ This is the “predetermined” probability of making type I error. Reject H_0 if the p -value of the test is smaller than α .

Two-sample z test for a difference in two proportions

- ▶ Assumption: Both sample sizes are large.
- ▶ Hypotheses $H_0 : p_1 = p_2$ vs $H_a : p_1 \neq p_2$.
- ▶ Under the null hypothesis which implies equal variance, we use the pooled estimate of proportion which is:

$$\hat{p} = \frac{\text{Sum of all successes}}{\text{Total of two samples}} = \frac{x_1 + x_2}{n_1 + n_2}$$

- ▶ The test statistic is:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{p(1 - p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Test of a difference in proportions

- ▶ A product can be made with natural rubber or a synthetic substitute; is the proportion with marked deterioration under friction equal for the two materials? The result of an experiment is given below:

	Sample Size	Number deteriorating
Rubber	40	27
Synthetic	15	8

Test of a difference in proportions

- ▶ $np, n(1 - p) \geq 5$.
- ▶ $p_r - p_s = 27/40 - 8/15 = 0.142$.
- ▶ $p = (27 + 8)/(40 + 15) = 0.636$.

▶

$$\frac{0.142 - 0}{\sqrt{0.636(1 - 0.636)(1/40 + 1/15)}} = 0.9727$$

- ▶ Compare with standard normal distribution:

```
2*pnorm(0.9732, lower=FALSE)
[1] 0.3304539
```

- ▶ Alternatively:

```
prop.test (c(27,8), c(40,15))
2-sample test for equality of proportions with continuity correction

data:  c(27, 8) out of c(40, 15)
X-squared = 0.433, df = 1, p-value = 0.5105
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1953844  0.4787177
sample estimates:
   prop 1    prop 2 
0.6750000 0.5333333
```

Continuity correction

The difference in p -values is due to the continuity correction:

```
prop.test (c(27,8), c(40,15), correct=FALSE)
```

```
2-sample test for equality of proportions without continuity  
correction
```

```
data:  c(27, 8) out of c(40, 15)
```

```
X-squared = 0.9461, df = 1, p-value = 0.3307
```

```
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
```

```
-0.1495511  0.4328844
```

```
sample estimates:
```

```
prop 1    prop 2
```

```
0.6750000 0.5333333
```

Hypothesis Tests for Means

Tests concerning hypotheses about means

The **steps** common to all hypothesis tests are as follows:

- ▶ State or formulate the null hypothesis H_0 and the alternative.
- ▶ Find an appropriate statistic such as x .
- ▶ Computation of test statistic value under H_0 (i.e. when the parameter value specified by H_0 is true).
- ▶ Decide if it is one or two tailed test based on the alternative hypothesis.
- ▶ Set the significance level at $\alpha = 0.05$ or $\alpha = 0.01$.
- ▶ Find the p -value of the test.
- ▶ Decide whether to reject or not reject the null hypothesis based on the size of calculated p -value.
- ▶ If the calculated p -value is smaller than α we reject the null hypothesis in favour of the alternative at significance level α . Otherwise we say that we do not have enough evidence to reject the null hypothesis.

Testing Hypotheses About One Mean or Paired Data

Step 1: Determine null and alternative hypotheses

1. $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ (two-sided)
2. $H_0 : \mu = \mu_0$ versus $H_a : \mu < \mu_0$ (one-sided) or
 $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$
3. $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$ (one-sided) or
 $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$

Remember a p -value is computed assuming H_0 is true, and μ_0 is the value used for that computation.

Testing Hypotheses About One Mean or Paired Data

Step 2: Verify Necessary Data Conditions

- ▶ **Situation 1:** Population of measurements of interest is **approximately normal**, and a random sample of any size is measured. In practice, use method if shape is not notably skewed or no extreme outliers.
- ▶ **Situation 2:** Population of measurements of interest is **not approximately normal**, but a **large random sample** ($n \geq 30$) is measured. If extreme outliers or extreme skewness, better to have a larger sample.

Testing Hypotheses About One Mean or Paired Data

Continuing Step 2: The Test Statistic

- ▶ The *t*-**statistic** is a standardised score for measuring the difference between the sample mean and the null hypothesis value of the population mean:

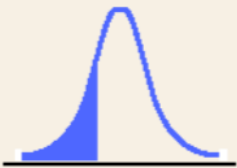
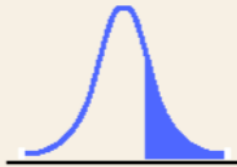

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

- ▶ This *t*-statistic has (approximately) a *t*-distribution with $df = n - 1$.

Testing Hypotheses About One Mean or Paired Data

Step 3: Assuming H_0 true, Find the p -value

- ▶ For H_a **less than**, the p -value is the area below t , even if t is positive.
- ▶ For H_a **greater than**, the p -value is the area above t , even if t is negative.
- ▶ For H_a **two-sided**, p -value is $2 \times$ area above $|t|$.

Statement of H_a		p -Value Area	t -Curve Region
$\mu < \mu_0$	(less than)	Area to the left of t (even if $t > 0$)	
$\mu > \mu_0$	(greater than)	Area to the right of t (even if $t < 0$)	
$\mu \neq \mu_0$	(not equal)	$2 \times$ area to the right of $ t $	

Use of R

```
x <- c(1.4,1.8,1.1,1.9,2.2,1.2)
stem (x)
```

The decimal point is at the |

```
1 | 124
1 | 89
2 | 2
```

```
t.test(x, mu=1.5, alternative="two.sided")
```

One Sample t-test

```
data: x
t = 0.5649, df = 5, p-value = 0.5965
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 1.144976 2.055024
sample estimates:
mean of x
 1.6
```

Use of R for performing one-sample t test

```
data <- read.table("Data/example.txt", header=TRUE)
```

```
data
```

```
  x
```

```
1 25.8
```

```
2 36.6
```

```
3 26.3
```

```
4 21.8
```

```
5 27.2
```

```
t.test(data,mu=25, alternative="greater")
```

```
One Sample t-test
```

```
data: data
```

```
t = 1.0382, df = 4, p-value = 0.1789
```

```
alternative hypothesis: true mean is greater than 25
```

```
95 percent confidence interval:
```

```
22.32433      Inf
```

```
sample estimates:
```

```
mean of x
```

```
27.54
```

Two-Sample t-Test

No assumption that the variances are equal

$$s.e(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Note: Degrees of freedom given by Welch's approximation (last lecture).

Test without assuming equal variance

```
x <- c(24, 61, 59, 46, 43, 44, 52, 43, 58, 67, 62,  
       57, 71, 49, 54, 43, 53, 57, 49, 56, 33)  
y <- c(42, 33, 46, 37, 43, 41, 10, 42, 55, 19, 17, 55,  
       26, 54, 60, 28, 62, 20, 53, 48, 37, 85, 42)  
t.test(x,y,var.equal=F)
```

Welch Two Sample t-test

data: x and y

t = 2.3109, df = 37.855, p-value = 0.02638

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

1.23302 18.67588

sample estimates:

mean of x mean of y

51.47619 41.52174

Tensile strength of liners

$n_1 = 10$ (no fusion)	$\bar{x}_1 = 2902.8$	$s_1 = 277.3$
$n_2 = 8$ (fused)	$\bar{x}_2 = 3108.1$	$s_2 = 205.9$

- ▶ We wish to test if the fusion process increases the average tensile strength.
- ▶ Normality test may be required to check the validity of the assumption (since the two sample sizes are quite small).
- ▶ The quantile-quantile plot, or Shapiro-Wilk test can be used for checking normality. (`qqnorm()`, `shapiro.test`).
- ▶ We are testing $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 < \mu_2$.

Tensile strength of liners: Examine data

```
x1 <- c(2748, 2700, 2655, 2822, 2511, 3149, 3257, 3213, 3220, 2753)
x2 <- c(3027, 3356, 3359, 3297, 3125, 2910, 2889, 2902)
c (mean(x1), mean(x2), sd(x1), sd(x2))
```

```
[1] 2902.8000 3108.1250 277.2651 205.8699
```

```
stem (x1)
```

The decimal point is 2 digit(s) to the right of the |

```
24 | 1
26 | 6055
28 | 2
30 | 5
32 | 126
```

```
stem (x2)
```

The decimal point is 2 digit(s) to the right of the |

```
28 | 9
29 | 01
30 | 3
31 | 3
32 |
33 | 066
```

Tensile strength of liners: Check normality

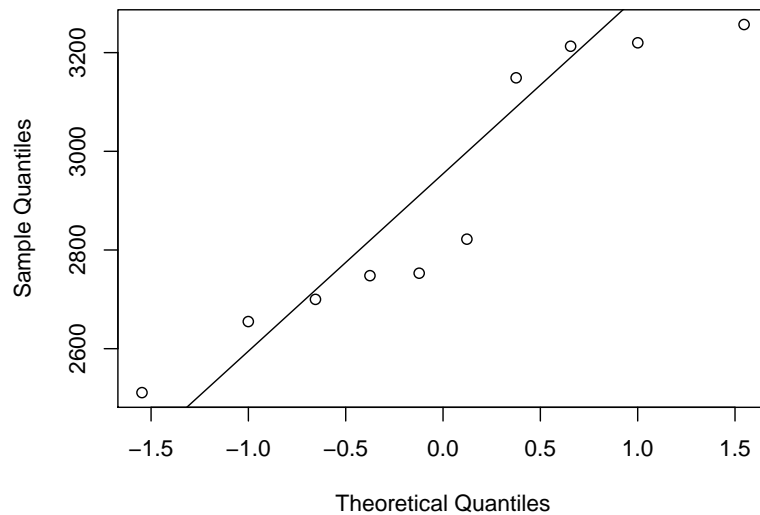
```
shapiro.test (x1)
```

Shapiro-Wilk normality test

data: x1

W = 0.87, p-value = 0.09984

Normal Q-Q Plot



```
qqnorm (x1)
```

```
qqline (x1)
```

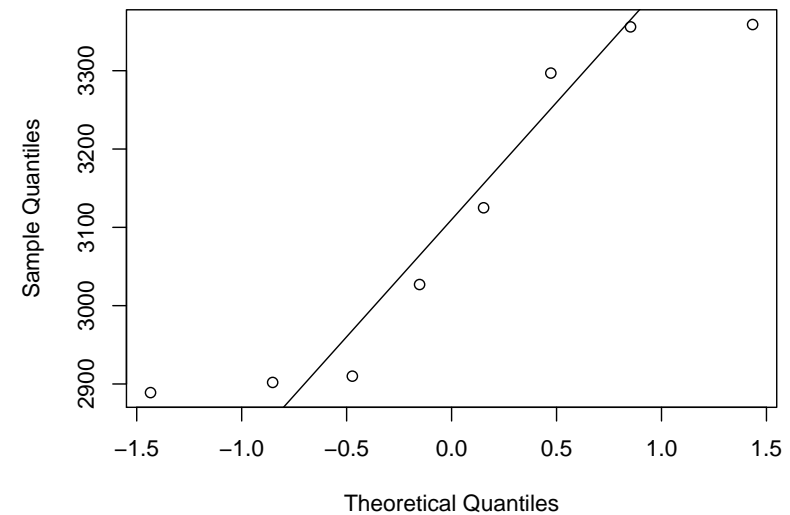
```
shapiro.test (x2)
```

Shapiro-Wilk normality test

data: x2

W = 0.847, p-value = 0.08881

Normal Q-Q Plot



```
qqnorm (x2)
```

```
qqline (x2)
```

Tensile strength of liners: Perform test

```
t.test(x1,x2, mu=0, alternative="less")
```

Welch Two Sample t-test

data: x1 and x2

t = -1.8018, df = 15.944, p-value = 0.04526

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -6.332257

sample estimates:

mean of x mean of y

2902.800 3108.125

Pooled Two-Sample t -test

Based on assumption that the two populations have **equal population standard deviations**: $\sigma_1 = \sigma_2 = \sigma$.

$$\text{Pooled standard deviation } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$\text{Pooled } s.e(\bar{x}_1 - \bar{x}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$t = \frac{\text{sample mean} - \text{null value}}{\text{pooled standard error}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Note: Pooled $df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

Tensile strength of liners: Equal variance assumption

```
t.test(x1,x2, mu=0, alternative="less", var.equal=TRUE)
```

Two Sample t-test

data: x1 and x2

t = -1.7414, df = 16, p-value = 0.0504

alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:

-Inf 0.5237307

sample estimates:

mean of x	mean of y
2902.800	3108.125

Note: Confidence interval is narrower due to assumption. More power.

Paired data

A paired t -test for testing a difference between 2 population means or treatment means will be used when 2 populations are matched in pairs. The experimental units within each pair are homogeneous.

(Make sure you know the difference between independent samples and paired data sample)

For example, we wish to test the effect “before and after” a drug is administered. The sample sizes n_1 and n_2 must be equal and they are given by n so we have n pairs of data.

Assumption: Distribution of differences is normal.

Paired comparisons are an example of **block designs** so that the experimental units are similar to avoid or minimise extraneous variations.

Paired data Example

- ▶ Two types of rubber are being considered for use in shoe soles. To compare durabilities, 25 volunteers agree to wear shoes having one type of rubber on one sole and the other type on the other sole. After 2 months the amount of wear is measured on each shoe.
- ▶ Observations on individuals “before and after” a certain test to see the effect of an “action” such as taking a prescribed drug.

Paired data example

```
dat <- read.table(file="Data/shoe.txt",  
                  header=TRUE)
```

```
dat
```

```
  A  B  
1 25 24  
2 24 22  
3 22 21  
4 19 16  
5 21 19  
6 23 21  
7 23 23  
8 22 21
```

```
shapiro.test (dat$A - dat$B)
```

Shapiro-Wilk normality test

```
data:  dat$A - dat$B
```

```
W = 0.9307, p-value = 0.5224
```

```
stem (dat$A - dat$B)
```

The decimal point is at the |

```
0 | 0  
1 | 000  
2 | 000  
3 | 0
```

```
t.test (dat$A - dat$B)
```

One Sample t-test

```
data:  dat$A - dat$B
```

```
t = 4.5826, df = 7, p-value = 0.002536
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
0.725995 2.274005
```

```
sample estimates:
```

```
mean of x
```

```
1.5
```

Note: We may use `t.test(datA, datB, paired=TRUE)`.

Paired t -test versus 2-sample t -test

- ▶ If a two-sample t -test were used in the paired data (by mistake), we would often increase the standard error so the calculated t -statistic will be smaller in absolute value than the correct one.
- ▶ Also we would end up using a t -curve with larger df so the t -curve will be less spread.
- ▶ The net result is we often would inflate the p -value of the test.
- ▶ **Getting this right is important!**

Paired t -test versus 2-sample t -test

```
attach (dat)  
t.test(A,B,mu=0)
```

Welch Two Sample t-test

data: A and B

$t = 1.3739$, $df = 12.951$, $p\text{-value} = 0.1928$

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

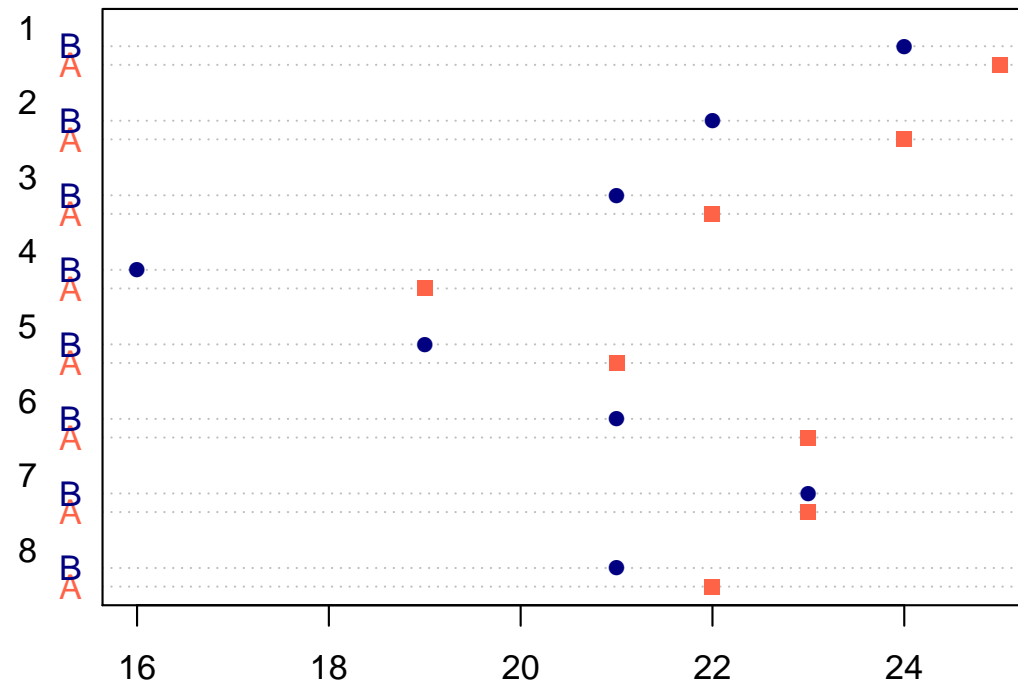
-0.8595412 3.8595412

sample estimates:

mean of x mean of y

22.375 20.875

Paired t -test versus 2-sample t -test



```
dotchart(rbind(A,B), pch=c(15,16), col=c("tomato", "navy"))
```

The two groups overlap considerably, but pairwise there is a difference.