Interval estimation: Part 1
(Module 3)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

School of Mathematics and Statistics
University of Melbourne

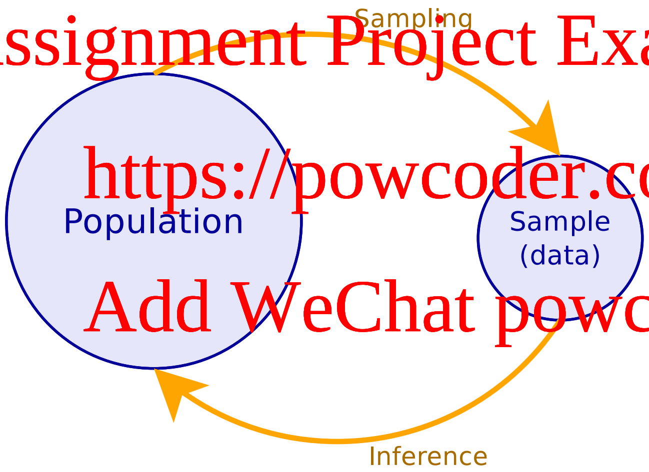
Semester 2, 2022

Aims of this module

- Introduce the idea of **quantifying uncertainty** and describe some methods for doing so
- Explain interval estimation, particularly **confidence intervals**, which are the most common type of interval estimate
- Describe some important probability distributions that appear in many statistical procedures
- Work through some common, simple inference scenarios

Outline

Statistics: the big picture



How useful are point estimates?

Example: surveying Melbourne residents as part of a disability study. The results will be used to set a budget for disability support.

Estimate from survey: 5% of residents are disabled

What can we conclude?

Estimate from a second survey: 2% of residents are disabled

What can we now conclude?

What other information would be useful to know?

Going beyond point estimates

- Point estimates are usually only a starting point
- Insufficient to conclusively answer real questions of interest
- Perpetual lurking questions:
  - How confident are you in the estimate?
  - How accurate is it?
- We need ways to **quantify** and **communicate** the **uncertainty** in our estimates.

Outline

Report $\mathrm{sd}(\hat{\Theta})$?

Previously, we calculated the variance of our estimators.

Reminder: $\mathrm{sd}(\hat{\Theta}) = \sqrt{\mathrm{var}(\hat{\Theta})}$

This tells us a typical amount by which the estimate will vary from one sample to another, and thus (for an unbiased estimator) how close to the true parameter value it is likely to be.

Can we just report that? (Alongside our estimate, $\hat{\theta}$)

Problem: this is usually an expression that depends on the parameter values, which we don't know and are trying to estimate.

Estimate $\mathrm{sd}(\hat{\Theta})$!

We know how to deal with parameter values. . . we estimate them!

Let's estimate the standard deviation of our estimator.

A common approach: substitute point estimates into the expression for the variance.

Example:

Consider the sample proportion, $\hat{p} = X/n$.

We know that $\mathrm{var}(\hat{p}) = \frac{p(1-p)}{n}$.

Therefore, an estimate is $\widehat{\mathrm{var}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$.

If we take a sample of size $n = 100$ and observe $x = 30$, we get

$$\hat{p} = 30/100 = 0.3,$$

$$\widehat{\mathrm{se}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.3 \times 0.7}{100}} = 0.046.$$

We refer to this estimate as the standard error and write:

$$\mathrm{se}(\hat{p}) = 0.046$$

Standard error

The standard error of an estimate is the estimated standard deviation of the estimator.

Notation:

- Parameter: θ
- Estimator: $\hat{\Theta}$
- Estimate: $\hat{\theta}$
- Standard deviation of the estimator: $\mathrm{sd}(\hat{\Theta})$
- Standard error of the estimate: $\mathrm{se}(\hat{\theta})$

Note: some people also refer to the standard deviation of the estimator as the standard error. This is potentially confusing, best to avoid doing this.

Reporting the standard error

There are many ways that people do this.

Suppose that $\hat{p} = 0.3$ and $\mathrm{se}(\hat{p}) = 0.046$.

Here are some examples:

- $0.3\ (0.046)$
- 0.3 ± 0.046
- $0.3 \pm 0.092 \quad [= 2 \times \mathrm{se}(\hat{p})]$

This now gives us some useful information about the (estimated) accuracy of our estimate.

Back to the disability example

More info:

- First survey: $5\% \pm 4\%$
- Second survey: $2\% \pm 0.1\%$

What would we now conclude?

What result should we use for setting the disability support budget?

Outline

Interval estimates

Let's go one step further. . .

The form est ± error can be expressed as an interval, $(\text{est} - \text{error}, \ \text{est} + \text{error})$.

This is an example of an interval estimate.

More general and more useful than just reporting a standard error. For example, it can cope with skewed (asymmetric) sampling distributions.

How can we calculate interval estimates?

Example

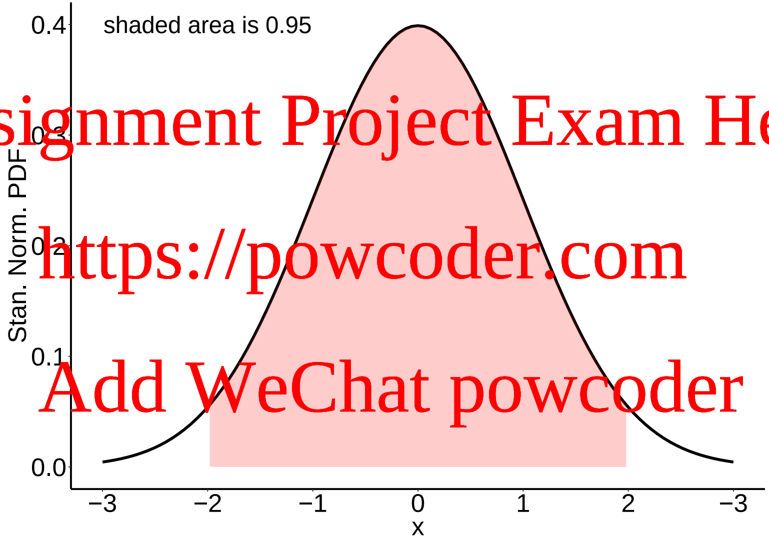
Random sample (iid): $X_1, \ldots, X_n \sim \mathrm{N}(\mu, 1)$

The sampling distribution of the sample mean is $\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n})$.
Since we know that $\Phi^{-1}(0.025) = -1.96$, we can write:

$$\Pr\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 1 - 2 \times 0.025 = 0.95$$

or, equivalently,

$$\Pr\left(\mu - 1.96\frac{1}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{1}{\sqrt{n}}\right) = 0.95$$


shaded area is 0.95
Stan. Norm. PDF
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
−3
x

Rearranging gives:

$$\Pr\left(\bar{X} - 1.96\frac{1}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{1}{\sqrt{n}}\right) = 0.95$$

This says that the interval $(\bar{X} - 1.96/\sqrt{n}, \, \bar{X} + 1.96/\sqrt{n})$ has probability 0.95 of containing the parameter μ.

We use this as an interval estimator.

The resulting interval estimate, $(\bar{x} - 1.96/\sqrt{n}, \, \bar{x} + 1.96/\sqrt{n})$ is called a 95% confidence interval for μ.

Sampling distribution of the interval estimator

- Is this an estimator?
- Does it have a sampling distribution?
- What does it look like?
- There are **two** statistics here, the endpoints of the interval:

$$\Pr(\mathbf{L} < \mu < \mathbf{U}) = 0.95$$

- They will have a joint (bivariate) sampling distribution

Example

For the previous example:

- Realisations of the interval will have a fixed width but a random location
- The randomness is due to $\bar{X}$
- Sampling distribution:

$$L \sim \mathrm{N}\left(\mu - 1.96\frac{1}{\sqrt{n}}, \frac{1}{n}\right)$$

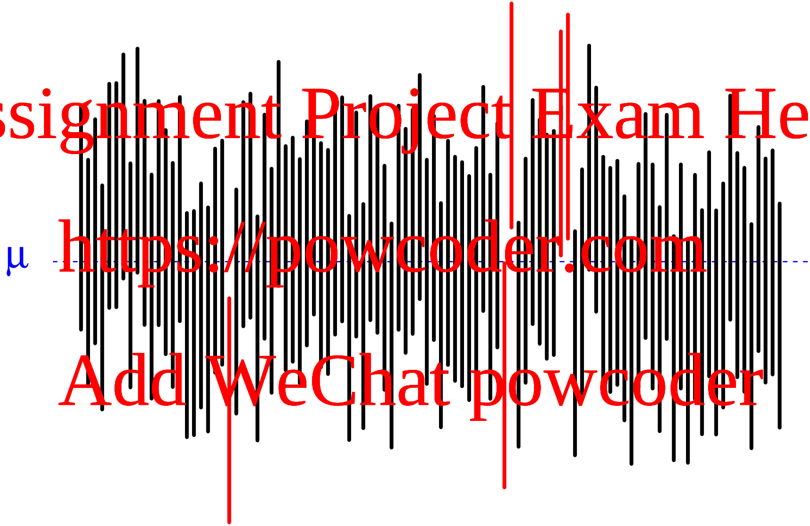
$$U \sim \mathrm{N}\left(\mu + 1.96\frac{1}{\sqrt{n}}, \frac{1}{n}\right)$$

$$U - L = 2 \times 1.96\frac{1}{\sqrt{n}}$$

- Can write it more formally as a bivariate normal distribution:

$$\begin{bmatrix} L \\ U \end{bmatrix} \sim \mathrm{N}_2\left(\begin{bmatrix} \mu - 1.96\frac{1}{\sqrt{n}} \\ \mu + 1.96\frac{1}{\sqrt{n}} \end{bmatrix}, \frac{1}{n}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$$

- Note that here we have perfect correlation, $\mathrm{cor}(L, U) = 1$
- Usually, easier and more useful to think about realisations of the actual interval...


μ

Interpretation

- This interval estimator is a **random interval** and is calculable from our sample. The parameter is fixed and unknown.
- Before the sample is taken, the probability the random interval contains μ is 95%.
- After the sample is taken, we have a realised interval. It no longer has a probabilistic interpretation; it either contains μ or it doesn't.
- This makes the interpretation somewhat tricky. We argue simply that it would be unlucky if our interval did *not* contain μ.
- In this example, the interval happens to be of the form, $\mathrm{est} \pm \mathrm{error}$. This will be the case for many of the confidence intervals we derive.

Example (more general)

Random sample (iid): $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, and assume that we know the value of σ^2.

The sampling distribution of the sample mean is $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$.
Let $\Phi^{-1}(1-\alpha/2) = c$, so we can write:

$$\Pr\left(-c < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < c\right) = 1-\alpha$$

or, equivalently,

$$\Pr\left(\mu - c\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + c\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Rearranging gives:

$$\Pr\left(\bar{X} - c\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + c\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

The following random interval contains μ with probability $1 - \alpha$:

$$\left(\bar{X} - c\frac{\sigma}{\sqrt{n}}, \bar{X} + c\frac{\sigma}{\sqrt{n}}\right)$$

Observe $\bar{x}$ and construct the interval. This gives a $100 \times (1 - \alpha)\%$ confidence interval for the population mean μ.

Worked example

Suppose $X \sim \mathrm{N}(\mu, 36^2)$ represents the lifetime of a light bulb, in hours. Test 27 bulbs, observe $\bar{x} = 1478$.

Let $c = \Phi^{-1}(0.975)$. A 95% confidence interval for μ is:

$$\bar{x} \pm c\left(\frac{\sigma}{\sqrt{n}}\right) = 1478 \pm 1.96\left(\frac{36}{\sqrt{27}}\right) = [1464, 1492]$$

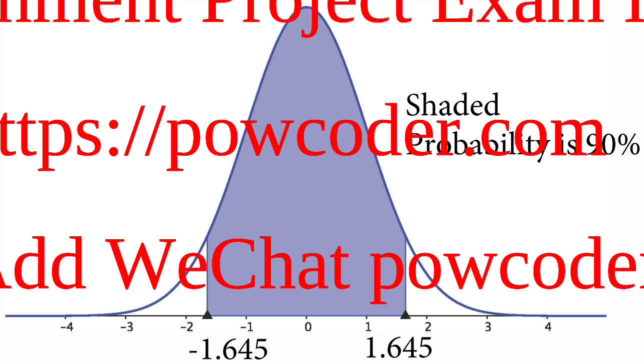
In other words, we have good evidence that the mean lifetime for a light bulb is approximately 1,460–1,490 hours.

Example (CLT approximation)

- If the distribution is not normal, we can use the Central Limit Theorem if n is large enough, $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \approx N(0,1)$
- Example: X is the amount of orange juice consumed (g/day) by an Australian. Know $\sigma = 96$. Sampled 576 Australians and found $\bar{x} = 133$ g/day.
- An approximate 90% CI for the mean amount of orange juice consumed by an Australian, regardless of the underlying distribution for individual orange juice consumption, is:

$$133 \pm 1.645 \left(\frac{96}{\sqrt{576}} \right) = [126, 140]$$

- In some studies, n is small because observations are expensive.


Shaded
Probability is 90%
-4
-3
-2
-1
0
1
2
3
4
-1.645
1.645

Definitions

- An interval estimate is a pair of statistics defining an interval that aims to convey an estimate (of a parameter) with uncertainty.
- A confidence interval is an interval estimate constructed such that the corresponding interval estimator has a specified probability, known as the confidence level, of containing the true value of the parameter being estimated.
- We often use the abbreviation CI for 'confidence interval'.

General technique for deriving a CI

- Start with an estimator, T, whose sampling distribution is known
- Write the central probability interval based on its sampling distribution,

$$\Pr\left(\pi_{0.025} < T < \pi_{0.975}\right) = 0.95$$

- The endpoints will depend on the parameter, θ, so can write it as,

$$\Pr\left(a(\theta) < T < b(\theta)\right) = 0.95$$

- Invert it to get a random interval for the parameter,

$$\Pr\left(b^{-1}(T) < \theta < a^{-1}(T)\right) = 0.95$$

- Substitute observed value, t, to get an interval estimate,

$$\left(b^{-1}(t),\, a^{-1}(t)\right)$$

Challenge problem (exponential distribution)

Take a random sample of size n from an exponential distribution with rate parameter λ.

1. Derive an exact 95% confidence interval for λ.
2. Suppose your sample is of size 9 and has sample mean 3.93.
   2.1 What is your 95% confidence interval for λ?
   2.2 What is your 95% confidence interval for the population mean?

3. Repeat the above using the CLT approximation (rather than an exact interval).

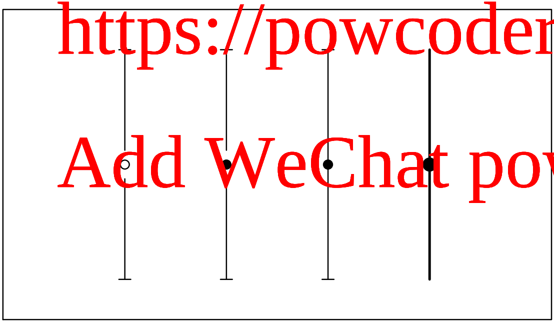
Recap

- A **point estimate** is a single number that is our 'best guess' at the true parameter value. In other words, it is meant to be the 'most plausible' value for the parameter, given the data.
- However, this doesn't allow us to adequately express our uncertainty of this estimate.
- An **interval estimate** aims to provide a **range** of values that are plausible based on the observed data. This allows us to more adequately express our uncertainty of the estimate, by giving an indication of the various plausible alternative true values.
- The most common type of interval estimate is a **confidence interval**.

Graphical presentation of CIs

Draw CIs as 'error bars'

There are various graphical styles that people use



Width of CIs

The width of a CI is controlled by various factors:

- inherent variation in the data
- choice of estimator
- confidence level
- sample size

For example, the width for the normal distribution example was:

$$2c\frac{\sigma}{\sqrt{n}}$$

where $c = \Phi^{-1}\left(1 - \alpha/2\right)$.

Interpreting CIs

- Narrower width usually indicates stronger/greater evidence about the plausible true values for the parameter being estimated
- **Very wide CI** $\Rightarrow$ usually cannot conclude much other than that we have insufficient data
- **Moderately wide CI** $\Rightarrow$ conclusions often depend on the location of the interval
- **Narrow CI** $\Rightarrow$ more confident about the possible true values, often can be more conclusive
- What constitutes 'wide' or 'narrow', and how conclusive/useful the CI actually is, will depend on the context of the study question

Three important distributions

- χ^2-distribution
- t-distribution
- F-distribution

Chi-squared distribution

- Also written in text as χ^2-distribution
- Single parameter: $k > 0$, known as the degrees of freedom
- Notation: $T \sim \chi^2_k$ or $T \sim \chi^2(k)$
- The pdf is:

$$f(t) = \frac{t^{\frac{k}{2}-1} e^{-\frac{t}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, \quad t \geq 0$$

- Mean and variance:

$$\mathbb{E}(T) = k$$
$$\mathrm{var}(T) = 2k$$

- The distribution is bounded below by zero and is right-skewed

- Arises as the sum of iid standard normal rvs:

$$Z_i \sim \mathrm{N}(0,1) \quad \Rightarrow \quad T = Z_1^2 + \cdots + Z_k^2 \sim \chi_k^2$$

- When sampling from a normal distribution, the sample variance follows a χ^2-distribution:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Student's t-distribution

- Also known as simply the t-distribution
- Single parameter: $k > 0$, the degrees of freedom (same as for χ^2)
- Notation: $T \sim t_k$ or $T \sim t(k)$
- The pdf is:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < t < \infty$$

- Mean and variance:

$$\mathbb{E}(T) = 0, \quad \text{if } k > 1$$

$$\text{var}(T) = \frac{k}{k-2}, \quad \text{if } k > 2$$

- The t-distribution is similar to a standard normal but with 'wide' tails
- As $k \to \infty$, then $t_k \to \mathrm{N}(0, 1)$
- If $Z \sim \mathrm{N}(0, 1)$ and $U \sim \chi^2(r)$, and they are independent, then

$$T = \frac{Z}{\sqrt{U/r}} \sim t_r$$

- This arises when considering the sampling distributions of statistics from a normal distribution, in particular:

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

F-distribution

- Also known as the Fisher-Snedecor distribution
- Parameters: $m, n > 0$, the degrees of freedom (same as before)
- Notation: $W \sim \mathrm{F}_{m,n}$ or $W \sim \mathrm{F}(m, n)$
- If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent then

$$F = \frac{U/m}{V/n} \sim \mathrm{F}_{m,n}$$

- This arises when comparing sample variances (see later)

Pivots

Recall our general technique that starts with a probability interval using a statistic with a known sampling distribution:

$$\Pr\left(a(\theta) < T < b(\theta)\right) = 0.95$$

The easiest way to make this technique work is by finding a function of the data and the parameters, $Q(X_1, \ldots, X_n; \theta)$, whose **distribution does not depend** on the parameters. In other words, it is a random variable that has the same distribution regardless of the value of θ.

The quantity $Q(X_1, \ldots, X_n; \theta)$ is called a pivot or a pivotal quantity.

Remarks about pivots

- The **value** of the pivot can depend on the parameters, but its **distribution** cannot.
- Since pivots are a function of the parameteres as well as the data, they are usually **not** statistics.
- If a pivot is also a statistic, then it is called an ancillary statistic.

Examples of pivots

- We have already seen the following result for sampling from a normal distribution with known variance:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

  Therefore, Z is a pivot in this case.

- If we know the distribution of the pivot, we can use it to write a probability interval, and start deriving a confidence interval.
- For example, in the normal case with known variance,

$$\Pr\left(a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < b\right)$$

  where a and b are fixed values that do not depend on μ.

Common scenarios: overview

Normal distribution:

- Inference for a single mean
  - Known σ
  - Unknown σ
- Comparison of two means
  - Known σ
  - Unknown σ
  - Paired samples
- Inference for a single variance
- Comparison of two variances

Proportions:

- Inference for a single proportion
- Comparison of two proportions

Normal, single mean, known σ

Random sample (iid): $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, and assume that we know the value of σ.

We've seen this scenario already in previous examples.

Use the pivot:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0, 1).$$

Normal, single mean, unknown σ

Random sample (iid): $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, and σ is unknown.

A pivot for μ in this case is:

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} is the t-distribution with $n-1$ degrees of freedom.

Now proceed as before.

Given α, let c be the $(1-\alpha/2)$ quantile of t_{n-1}.
We then write:

$$\Pr\left(-c < \frac{\bar{X}-\mu}{S/\sqrt{n}} < c\right) = 1-\alpha.$$

Rearranging gives:

$$\Pr\left(\bar{X} - c\frac{S}{\sqrt{n}} < \mu < \bar{X} + c\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

and for observed $\bar{x}$ and s, a $100 \cdot (1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{x} - c\frac{s}{\sqrt{n}}, \, \bar{x} + c\frac{s}{\sqrt{n}}\right).$$

Example (normal, single mean, unknown σ)

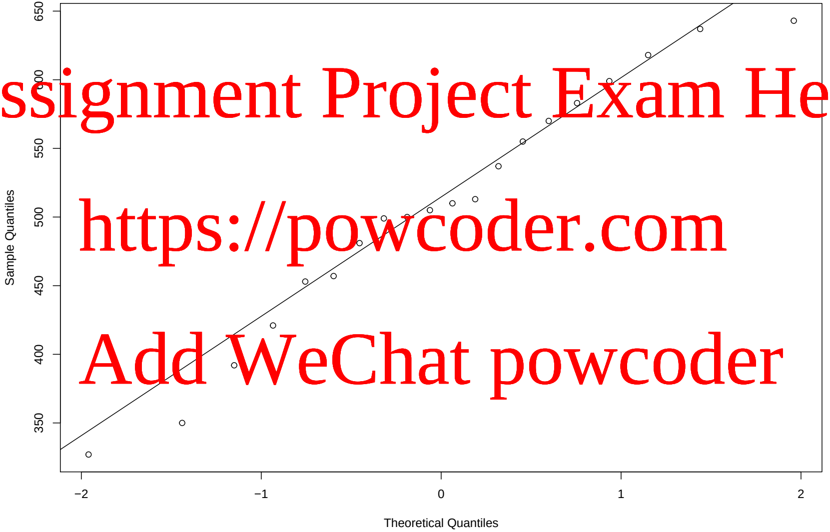
$X \sim \mathrm{N}(\mu, \sigma^2)$ is the amount of butterfat produced by a cow. Examining $n = 20$ cows results in $\bar{x} = 507.5$ and $s = 89.75$. Let c be the 0.95 quantile of t_{19}, we have $c = 1.729$. Therefore, a 90% confidence interval for μ is,

$$507.50 \pm 1.729 \left(\frac{89.75}{\sqrt{20}} \right) = [472.80, 542.20]$$

```
> butterfat
 [1] 481 537 513 583 453 510 570 500 457 555 618 327
[13] 350 643 499 421 505 637 599 392

> t.test(butterfat, conf.level = 0.9)

        One Sample t-test

data:  butterfat
t = 25.2879, df = 19, p-value = 4.311e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 472.7982 542.2018
sample estimates:
mean of x
    507.5

> sd(butterfat)
[1] 89.75082
```

```
> qqnorm(butterfat, main = "")
> qqline(butterfat, probs = c(0.25, 0.75))
```

This gives us the following QQ plot...


Sample Quantiles
Theoretical Quantiles
350
400
450
500
550
600
650
−2
−1
0
1
2

Remarks

- CIs based on a t-distribution (or a normal distribution) are of the form:

$$\text{estimate} \pm c \times \text{standard error}$$

  for an appropriate quantile, c, which depends on the sample size (n) and the confidence level ($1-\alpha$).

- The t-distribution is appropriate if the sample is from a normally distributed population.
- Can check using a QQ plot (in this example, looks adequate).

- If not normal but n is large, can construct approximate CIs using the normal distribution (as we did in a previous example). This is usually okay if the distribution is continuous, symmetric and unimodal (i.e. has a single 'mode', or maximum value).
- If not normal and n small, distribution-free methods can be used. We will cover these later in the semester.

Normal, two means, known σ

Suppose we have two populations, with means μ_X and μ_Y, and want to know how much they differ.

Random samples (iid) from each population:
$X_1, \dots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \dots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$

The two samples must be independent of each other.

Assume σ_X^2 and σ_Y^2 are known.
Then we have the following pivot (why?):

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathrm{N}(0, 1)$$

Defining c as in previous examples, we then write,

$$\Pr\left(-c < \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} < c\right) = 1 - \alpha$$

Rearranging as usual gives the $100 \cdot (1 - \alpha)\%$ confidence interval for $\mu_X - \mu_Y$ as

$$\bar{x} - \bar{y} \pm c\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

...but it is rare to know the population variances!

Normal, two means, unknown σ, many samples

What if we don't know σ_X^2 and σ_Y^2?

If n and m are large, we can just replace σ_X and σ_Y by estimates, e.g. the sample standard deviations S_X and S_Y.
Rationale: these will be good estimates when the sample size is large.

The (approximate) pivot is then:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \approx \mathrm{N}(0, 1)$$

This gives the following (approximate) CI for $\mu_X - \mu_Y$:

$$\bar{x} - \bar{y} \pm c\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

Normal, two means, unknown σ, common variance

But what if the sample sizes are small?

If we assume a common variance, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, we can find a pivot, as follows.

Firstly,

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathrm{N}(0, 1)$$

Also, since the samples are independent,

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

because U is the sum of independent χ^2 random variables.

Moreover, U and Z are independent. So we can write,

$$T = \frac{Z}{\sqrt{U/(n+m-2)}} \sim t_{n+m-2}$$

Substituting and rearranging gives,

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

$$S_P = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$$

is the pooled estimate of the common variance.

Note that the unknown σ has disappeared (cancelled out), therefore making T a pivot (why?).

We can now find the quantile c so that

$$\Pr(-c < T < c) = 1 - \alpha$$

and rearranging as usual gives a $100 \cdot (1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$:

$$\bar{x} - \bar{y} \pm c \cdot s_P \sqrt{\frac{1}{n} + \frac{1}{m}}$$

where

$$s_P = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}$$

Example (normal, two means, unknown common variance)

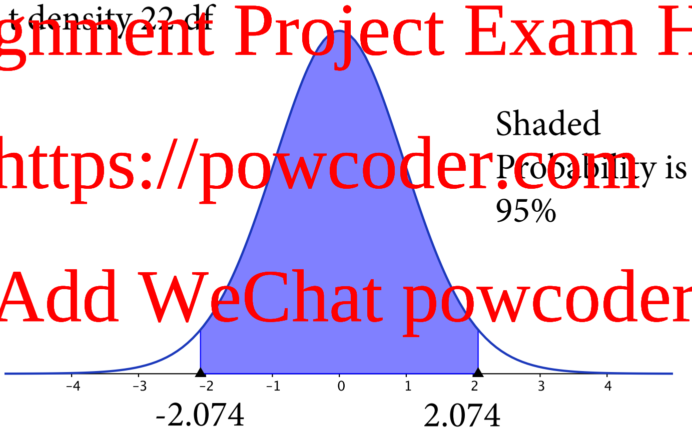
Two independent groups of students take the same test. Assume the scores are normally distributed and have a common unknown population variance.

We have sample sizes $n = 9$ and $m = 15$, and get the following summary statistics: $\bar{x} = 81.31$, $\bar{y} = 78.61$, $s_x^2 = 60.76$, $s_y^2 = 48.24$.

The pivot has df $9 + 15 - 2 = 22$ degrees of freedom. Using the 0.975 quantile of t_{22}, which is 2.074, the 95% confidence interval is:

$$81.31 - 78.61 \pm 2.074 \sqrt{\frac{8 \times 60.76 + 14 \times 48.24}{22}} \sqrt{\frac{1}{9} + \frac{1}{15}}$$

$$= [-3.65, 9.05]$$


t density 22 df
Shaded Probability is 95%
-4
-3
-2
-1
0
1
2
3
4
-2.074
2.074

Normal, two means, unknown σ, different variances

What if the sample sizes are small and pretty sure that $\sigma_X^2 \neq \sigma_Y^2$?

Then we can use Welch's approximation:

$$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

which approximately follows a t_r-distribution with degrees of freedom given by:

$$r = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

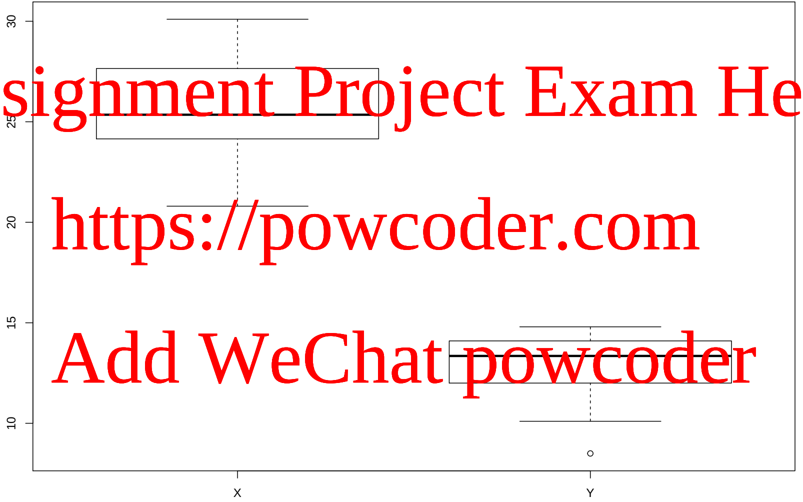
This is often the default for constructing confidence intervals.

Example (normal, two means, unknown different variances)

We measure the force required to pull wires apart for two types of wire, X and Y. We take 20 measurements for each wire.

	1	2	3	4	5	6	7	8	9	10
X	28.8	24.4	30.1	25.6	26.4	23.9	22.1	22.5	27.6	28.1
Y	14.1	12.2	14.0	14.6	8.5	12.6	13.7	14.8	14.1	13.2

	11	12	13	14	15	16	17	18	19	20
X	20.8	27.7	24.4	25.1	24.6	26.3	28.2	22.2	26.3	24.4
Y	12.1	11.4	10.1	14.2	13.6	13.1	11.9	14.8	11.1	13.5


30
25
20
15
10
X
Y

Some heavily edited R output...

Different variances:

```
> t.test(X, Y,
+        conf.level = 0.95)

t = 18.8003
df = 33.086
95% CI: 11.23214 13.95786
```

Pooled variance:

```
> t.test(X, Y,
+        conf.level = 0.95,
+        var.equal  = TRUE)

t = 18.8003
df = 38
95% CI: 11.23879 13.95121
```

Remarks

- From box plots: look like very different population means and possibly different spreads
- The Welch approximate t-distribution is appropriate so a 95% confidence interval is 11.23–13.96
- If we assumed equal variances, the confidence interval becomes slightly narrower, 11.24–13.95
- Not a big difference!

Normal, paired samples

- As before, we are interested in the difference between the means of two sets of observations, $\mu_D = \mu_X - \mu_Y$
- This time, we observe the measurements in pairs, $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Each pair is observed independently of each other pair, but the members of each pair could be related
- We can exploit this extra information (the relationship within pairs) to both simplify and improve our estimate
- Let $D_i = X_i - Y_i$ be the differences of each pair
- Often reasonable to assume $D_i \sim \mathrm{N}(\mu_D, \sigma_D^2)$

- We can now use our method of inference for a single mean!
- A $100 \cdot (1 - \alpha)\%$ confidence interval for μ_D is:

$$\bar{d} \pm c \frac{s_d}{\sqrt{n}}$$

where c is the $1 - \alpha/2$ quantile of t_{n-1}.

Example (normal, paired samples)

The reaction times (in seconds) to a red or green light for 8 people are given in the following table. Find a 95% CI for the mean difference in reaction time.

	Red (X)	Green (Y)	$D = X - Y$
1	0.30	0.24	0.06
2	0.43	0.27	0.16
3	0.23	0.36	-0.13
4	0.32	0.41	-0.09
5	0.41	0.38	0.03
6	0.58	0.38	0.20
7	0.53	0.51	0.02
8	0.46	0.61	-0.15

Summary statistics:
$n = 8$,
$\bar{d} = 0.0125$,
$s_d = 0.129$

95% CI:

$$0.0125 \pm 2.365\frac{0.129}{\sqrt{8}}$$
$$= [-0.095, 0.12]$$

(2.365 is the 0.975 quantile of t_7)

Normal, single variance

Random sample (iid): $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$

This time we wish to infer σ, rather than μ

A pivot for σ is:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Now we need the $\alpha/2$ and $1-\alpha/2$ quantiles of χ^2_{n-1}.
Call these a and b. In other words,

$$\Pr\left(a < \frac{(n-1)S^2}{\sigma^2} < b\right) = 1 - \alpha$$

Rearranging gives

$$\begin{aligned} 1-\alpha &= \Pr\left(\frac{a}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{b}{(n-1)S^2}\right) \\ &= \Pr\left(\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}\right) \end{aligned}$$

So a $100 \cdot (1-\alpha)\%$ confidence interval is

$$\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right]$$

Example (normal, single variance)

Sample $n = 13$ seeds from a $\mathrm{N}(\mu, \sigma^2)$ population.

Observe a mean sprouting time of $\bar{x} = 18.97$ days, and sample variance $s^2 = 128.41/12$.

A 90% confidence interval for σ^2 is:

$$\left[\frac{128.41}{21.03}, \frac{128.41}{5.226}\right] = [6.11, 24.6]$$

with the 0.05 and 0.95 quantiles from a χ^2_{12} distribution being 5.226 and 21.03.

Normal, two variances

Now we wish to compare the variances of two normally distributed populations. Random samples (iid) from each population:
$X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$

We will compute a confidence interval for σ_X^2/σ_Y^2. Start by defining:

$$\frac{\frac{S_Y^2}{\sigma_Y^2}}{\frac{S_X^2}{\sigma_X^2}} = \frac{\left[\frac{(m-1)S_Y^2}{\sigma_Y^2}\right]/(m-1)}{\left[\frac{(n-1)S_X^2}{\sigma_X^2}\right]/(n-1)}$$

This is the ratio of independent χ^2 random variables divided by their degrees of freedom and hence has an $\mathrm{F}_{m-1,n-1}$ distribution. This doesn't depend on the parameters and is thus a pivot.

We now need the $\alpha/2$ and $1-\alpha/2$ quantiles of $\mathrm{F}_{m-1,n-1}$. Call these c and d. In other words,

$$1-\alpha = \Pr\left(c \le \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \le d\right) = \Pr\left(c\frac{S_X^2}{S_Y^2} \le \frac{\sigma_X^2}{\sigma_Y^2} \le d\frac{S_X^2}{S_Y^2}\right)$$

Rearranging gives the $100\cdot(1-\alpha)\%$ confidence interval for σ_X^2/σ_Y^2 as

$$\left[c\frac{s_x^2}{s_y^2}, d\frac{s_x^2}{s_y^2}\right]$$

Example (normal, two variances)

Continuing from the previous example, $n = 13$ and $12s_x^2 = 128.41$.
A sample of $m = 9$ seeds from a second strain gave $8s_y^2 = 36.72$.

The 0.01 and 0.99 quantiles of $\mathrm{F}_{8,12}$ are 0.176 and 4.50.

Then a 98% confidence interval for σ_X^2/σ_Y^2 is

$$\left[0.176\frac{128.41/12}{36.72/8}, 4.50\frac{128.41/12}{36.72/8}\right] = [0.41, 10.49]$$

Not very useful! Too wide.

Single proportion

- Observe n Bernoulli trials with unknown probability p of success,

$$X_1, X_2, \ldots, X_n \sim \mathrm{Be}(p)$$

- We want a confidence interval for p

- Recall that the sample proportion of successes $\hat{p} = \bar{X}$ is the maximum likelihood estimator for p and is unbiased for p

- The central limit theorem shows for large n,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0, 1)$$

- Rearranging the corresponding probability statement as usual and estimating p by $\hat{p}$ gives the approximate $100 \cdot (1 - \alpha)\%$ confidence interval as

$$\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example (single proportion)

- In the Newspoll of 3rd April 2017, 36% of 1,708 voters sampled said they would vote for the Government first if an election were held on that day. What is a 95% confidence interval for the population proportion of voters who would vote for the Government first?
- The sample proportion has an approximate normal distribution since the sample size is large so the required confidence interval is:

$$0.36 \pm 1.96\sqrt{\frac{0.36 \times 0.64}{1708}} = [0.337, 0.383]$$

- It might be nice to round to the nearest percentage for this example. This gives us the final interval: 34%–38%

Example 2 (single proportion)

- In a survey, $y = 185$ out of $n = 351$ voters favour a particular candidate. Note that $185/351 = 0.527$. An approximate 95% confidence interval for the proportion of the population supporting the candidate is

$$0.527 \pm 1.96\sqrt{\frac{0.527 \times 0.473}{351}} = [0.475, 0.579]$$

- The candidate is not guaranteed to win despite $\hat{p} > 0.5$!

Two proportions

- We now wish to compare proportions between two different samples: $Y_1 \sim \mathrm{Bi}(n_1, p_1)$, $Y_2 \sim \mathrm{Bi}(n_2, p_2)$
- Use the approximate pivot

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx \mathrm{N}(0, 1)$$

- This gives the approximate CI

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

Example (two proportions)

Following on from the previous Newspoll example...

- At the previous poll, with 1,824 voters sampled, there were 37% of voters who reported that they would vote for the Government first. Has the vote dropped? What is a 90% confidence interval for the difference in proportions in the population on the two occasions?

- The CI is

$$0.36 - 0.37 \pm 1.6449 \sqrt{\frac{0.36 \times 0.64}{1708} + \frac{0.37 \times 0.63}{1824}} = [-0.037, 0.017]$$

- This interval comfortably surrounds 0, meaning there is no evidence of a change in public opinion.
- This analysis allows for sampling variability in both polls, so is the preferred way to infer whether the vote has dropped.

Example 2 (two proportions)

Two detergents. First successful in 63 out of 91 trials, the second in 42 out of 79.

Summary statistics: $\hat{p}_1 = 0.692$, $\hat{p}_2 = 0.532$

90% confidence interval for the difference in proportions is:

$$0.692 - 0.532 \pm 1.645\sqrt{\frac{0.692 \times 0.308}{91} + \frac{0.532 \times 0.468}{79}}$$
$$= [0.038, 0.282]$$

Very wide! Need greater sample size to get more certainty.