Interval estimation: Part 1

(Module 3)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019

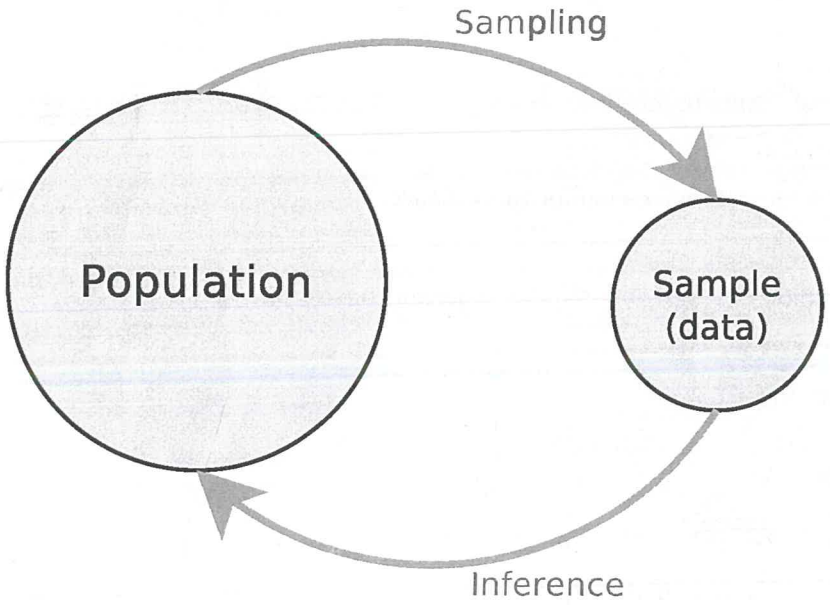
Contents

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 distribution that appear in many statistical procedures
- Work through some common, simple inference scenarios

1 The need to quantify uncertainty

Statistics: the big picture



We have learnt how to do basic inference, using point estimates. What's next?

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.

2 Standard error

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

Previously, we calculated the variance of our estimators.

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

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 $\text{sd}(\hat{\Theta})$**!**

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

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

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

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

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

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

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: $\text{sd}(\hat{\Theta})$
- Standard error of the estimate: $\text{se}(\hat{\theta})$

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 $\text{se}(\hat{p}) = 0.046$.

Here are some examples:

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

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?

3 Confidence intervals

3.1 Introduction

Interval estimates

Let's go one step further...

The form est $\pm$ error can be expressed as an interval, (est $-$ error, est $+$ 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

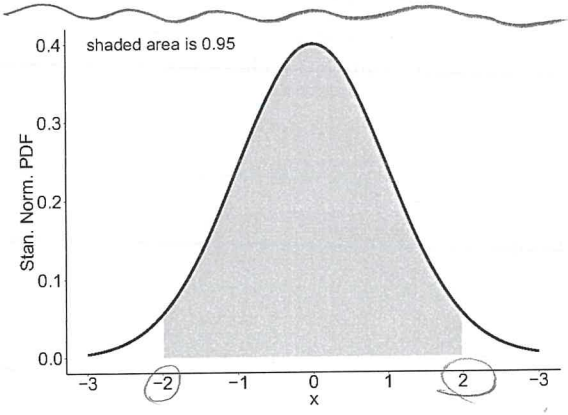
Random sample (iid): $X_1, \ldots, X_n \sim \mathrm{N}(\mu, 1)$

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,

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



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...

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, est $\pm$ 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 $1-\alpha$:

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

Observe $\bar{x}$ and construct the interval. This gives a $100 \cdot (1-\alpha)\%$ *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:

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

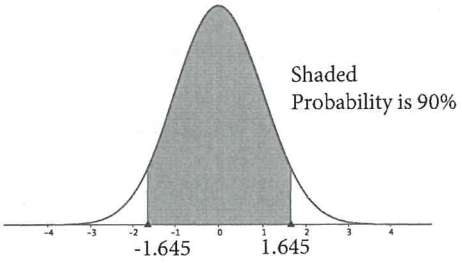
In other words, we have good evidence that the mean lifetime for a light bulb is approximately 1,460 1,490 hours.

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.



3.2 Definition

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.
   (a) What is your 95% confidence interval for λ?
   (b) What is your 95% confidence interval for the population mean?
3. Repeat the above using the CLT approximation (rather than an exact interval).

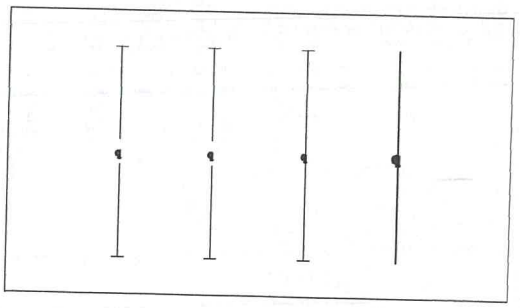
Recap

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

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

3.3 Important distributions

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:

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

- Mean and variance:

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

- The distribution is bounded below by zero and is *right-skewed*
- Arises as the sum of iid standard normal rvs:

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

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

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

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 $t_k \to \text{N}(0,1)$
- If $Z \sim \text{N}(0,1)$ 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: $W \sim \text{F}_{m,n}$ or $W \sim \text{F}(m,n)$
- If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent then

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

- This arises when comparing sample variances (see later)

3.4 Pivots

Pivots 枢轴量

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 parameters 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.

3.5 Common scenarios

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

$$\bar{X} \pm c\cdot\frac{\sigma}{\sqrt{n}}$$

Normal, single mean, known σ

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

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

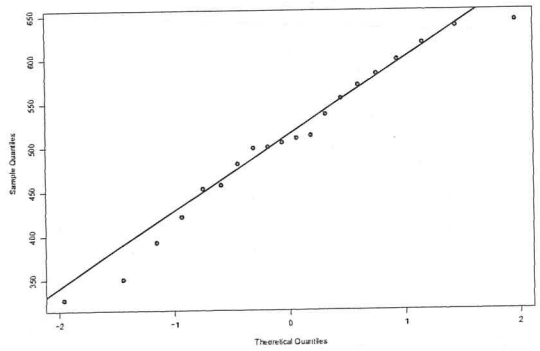
Example (normal, single mean, unknown σ)

$X \sim \mathrm{N}(\mu, \sigma^2)$ 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...



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 independent

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: $X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$

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

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

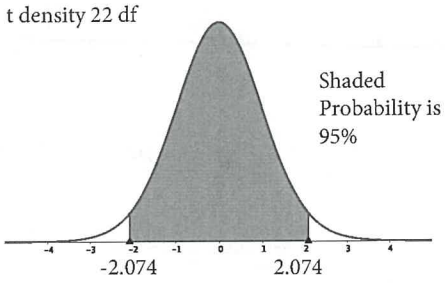
Example (normal, two means, unknown common variance)

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:

$$\begin{aligned} &81.31 - 78.61 \pm 2.074\sqrt{\frac{8 \times 60.76 + 14 \times 48.24}{22}}\sqrt{\frac{1}{9} + \frac{1}{15}} \\ &= [-3.65, 9.05] \end{aligned}$$



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:

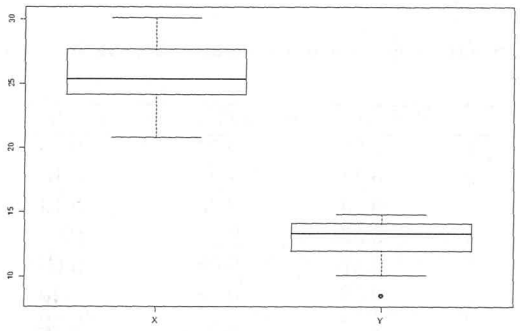
$$r \approx \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

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
```



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): $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$

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,

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

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

$$0.527 \pm 1.96\sqrt{\frac{0.527 \times 0.573}{351}} = [0.475, 0.579]$$

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

Two proportions

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

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

- 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.

Interval estimation: Part 2

(Module 4)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019

Contents

Aims of this module

- Explain some less common scenarios where confidence intervals are used
- Describe some general aspects of confidence intervals
- Introduce **prediction intervals**, an interval estimator in the context of predicting a future value
- Explain how to calculate the sample size required for a study

1 Confidence intervals

1.1 Less common scenarios

Less common scenarios: overview

- One-sided CIs
- CIs based on discrete statistics

One-sided confidence intervals

We can construct one-sided confidence intervals, e.g. just an upper or lower bound.

For example, if we sample from $N(\mu, \sigma^2)$ with known σ:

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

where $c = \Phi^{-1}(1-\alpha)$. Rearranging gives

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

and therefore a *one-sided* $100 \cdot (1-\alpha)\%$ confidence interval for μ is

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

Remarks

- The main thing to remember is to start with a one-sided probability statement about the pivot.
- In this example, we obtained a lower bound.
- To get an upper bound, start with an inequality in the other direction.
- Other scenarios are analogous. For example, if σ is unknown then replace σ with s and let c be a quantile from t_{n-1} rather than $N(0,1)$.
- Since we only need one tail probability, we don't need to separate α into two parts. That's why we use the $1-\alpha$ quantile here rather than the $1-\alpha/2$ quantile.

Example (one-sided interval)

A winemaker requires a minimum concentration of 10 g/L of sugar in the grapes used to make a certain wine. In a sample of 30 units she finds an average concentration of 11.9 g/L and a standard deviation of 0.96. Is that high enough?

She calculates a 95% lower bound (one-sided CI) as follows:

$$\bar{x} - c\frac{s}{\sqrt{n}} = 11.9 - 1.699 \times \frac{0.96}{\sqrt{30}} = 11.60$$

where $c = 1.699$ is the 0.95 quantile from t_{29}.

On that basis, she is confident that the average sugar content is adequately high.

Example using R

Recall the butterfat example from the previous module. Now re-doing using one-sided CIs...

```
> t.test(butterfat,
+        conf.level  = 0.90,
+        alternative = "less")

...

90 percent confidence interval:
    -Inf 534.146
```

```
> t.test(butterfat,
+        conf.level  = 0.90,
+        alternative = "greater")

...

90 percent confidence interval:
 480.854     Inf
```

Confidence intervals based on discrete statistics*

Our starting point has been probability intervals like:

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

What if T is discrete? For example, $T \sim \text{Bi}(n, \theta)$

Limitation: $a()$ and $b()$ can only take specific (discrete) values.

$\Rightarrow$ Cannot guarantee an exact probability (confidence level).

$\Rightarrow$ Inversion is messy.

Usually aim for something close, with 'at least' probability. For example,

$$\Pr\left(a(\theta) \leqslant T \leqslant b(\theta)\right) \geqslant 0.95$$

where:

- $a(\theta)$ is the largest value of x such that $\Pr(x \leqslant T \mid \theta) \geqslant 0.975$
- $b(\theta)$ is the smallest value of x such that $\Pr(T \leqslant x \mid \theta) \geqslant 0.975$

How do we invert these?

For an observed value t_{obs} (of T), we have:

- c is such that $\Pr(t_{\text{obs}} \leqslant T \mid \theta = c) = 0.025$
- d is such that $\Pr(T \leqslant t_{\text{obs}} \mid \theta = d) = 0.025$

Then, the 'at least' 95% confidence interval is (c, d).

1.2 General techniques

CIs from MLEs

Maximum likelihood estimators have many convenient properties. We will cover some of the theory later in the semester. For now, it is useful to know the following...

Let,

$$V(\theta) = -\frac{\partial^2 \ln L}{\partial \theta^2}$$

This is known as the *observed information function*. It can be used to estimate the standard deviation of the MLE:

$$\text{se}(\hat{\theta}) = \frac{1}{\sqrt{V(\hat{\theta})}}$$

Moreover, the MLE is **asymptotically unbiased** and **asymptotically normally distributed**.

Therefore, for large sample sizes, we can construct approximate CIs using:

$$\hat{\theta} \pm \frac{c}{\sqrt{V(\hat{\theta})}}$$

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

Example (approximate CI from MLE)

Sampling (iid) from: $X \sim \text{Exp}(\theta)$. Previously we found that $\hat{\theta} = \bar{X}$ and

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2}.$$

Differentiate once more,

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{2\sum x_i}{\theta^3}$$

and so we have,

$$\text{se}(\hat{\theta}) = \left(-\frac{n}{\hat{\theta}^2} + \frac{2\sum x_i}{\hat{\theta}^3}\right)^{-\frac{1}{2}}$$

and an approximate 95% confidence interval is given by $\hat{\theta} \pm 1.96\,\text{se}(\hat{\theta})$.

Review of general methods for constructing CIs

Methods:

- Invert a probability interval based on a known sampling distribution (use a pivot)
- Use the asymptotic MLE result

Common approximations:

- Normality (based on the CLT or the asymptotic MLE)
- Substitute parameter estimates into the expression for the standard deviation of the estimator

1.3 Properties

CIs are random intervals

Recall: the CI estimator is a **random interval**.

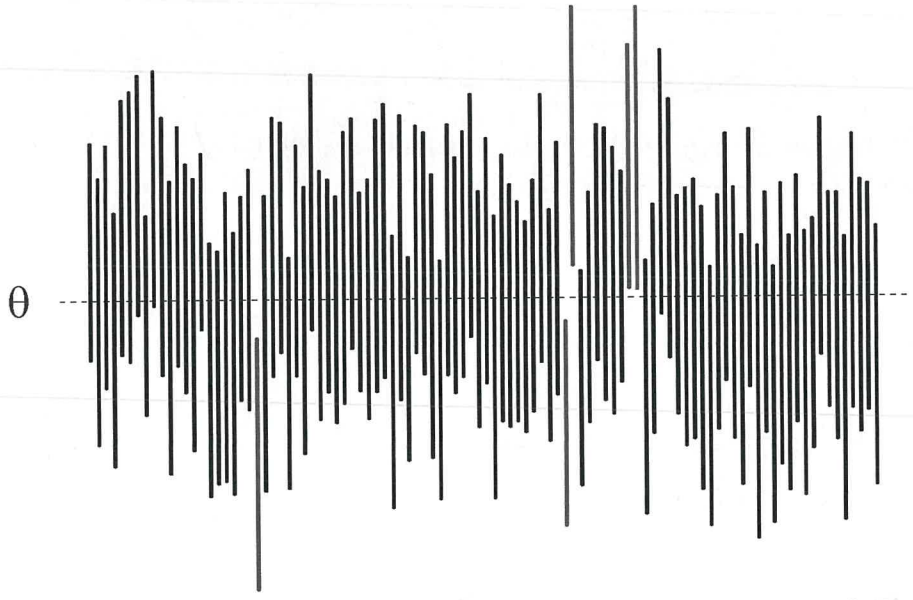
A CI consists of two statistics: the lower bound and the upper bound of the interval. They both have sampling distributions.

The random elements are therefore the endpoints, not the parameter:

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

Contrast this with a probability statement for a statistic:

$$\Pr(l < \mathbf{T} < u) = 0.95$$



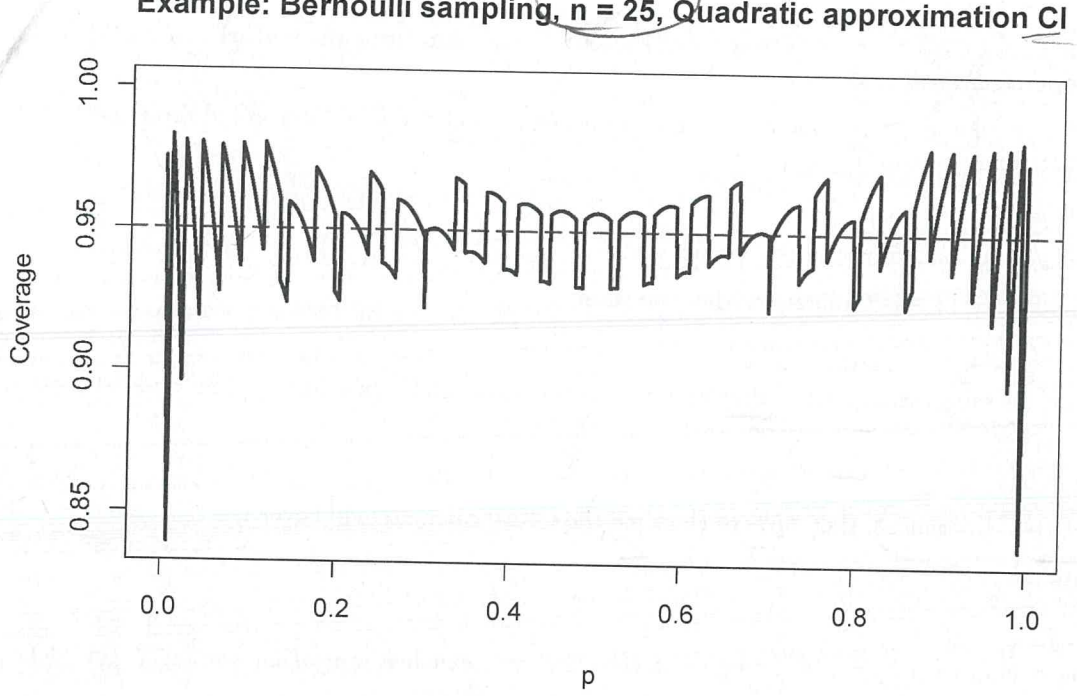
Coverage

The *coverage* or *coverage probability* of a confidence interval (estimator) is the probability it contains the true value of the parameter,

$$C = \Pr(L < \theta < U)$$

Usually this is equal to the confidence level, which is also known as the *nominal coverage probability.*

However, due to various approximations we use, the actual coverage achieved may vary from the confidence level.



More detail about the quadratic approximation will be shown in the tutorials and lab classes.

1.4 Choice of confidence level

Choice of confidence level

This is somewhat arbitrary.

If very high:

- More likely to capture the true value.
- Impractically wide; won't act as a useful guide for showing plausible values based on the data.
- It will place too much emphasis on tails of the sampling distribution, which aren't actually all that likely.

If very low:

- More 'useful' in the sense of being more selective about the possible values of the parameter.
- This comes at the expense of the loss of 'confidence', i.e. not as certain about whether the true value is captured inside the interval.

Choice of confidence level: some guidelines

- **95%** is a very common convention. If you follow this, it will rarely be questioned. Others may be expecting this, so always be clear if you deviate from it.
- **90%** can also be a reasonable choice.
- **50%** is sometimes useful, due to easy interpretation. A good use case: plotting a large number of overlapping intervals, to reduce visual clutter.
- The choice can **vary by application**, and you may even use different choices for the same problem (e.g. 50% for a particular plot, but 95% when reporting a headline result in text).
- Whatever you choose, remember that the true value is **never guaranteed** to be inside the interval. There is **always** a chance it will be outside.

1.5 Interpretation

Explaining CIs

The probability associated with a CI (i.e. the confidence level) relates to the **sampling procedure**. In particular, it refers to **hypothetical repeated samples**.

Once a specific sample is observed and a CI is calculated, the confidence level **cannot** be interpreted probabilistically in the context of the specific data at hand.

It is incorrect to say things like:

- This CI has a 95% chance of including the true value
- We can be '95% confident' that this CI includes the true value

Don't do it!

The probability only has a meaning when considering potential replications of the whole sampling and estimation procedure.

We can only say something like:

- If we were to repeat this experiment, then 95% of the time the CI we calculate will cover the true value.

(This is a bit of a mouthful...)

In practice:

- If you are reporting results to people who know what they are, you can just state that the "95% confidence interval is..."
- If people want to know what this means, use an intuitive notion like, "it is the set of plausible values of the parameter that are consistent with the data". (Note: this is not actually true in general, but will be accurate enough for all of the examples we cover this semester.)
- If you need to actually explain what a CI is precisely, you need to explain it in terms of repeated sampling. (No shortcuts!)

Communicating results: general tips

- Describe the extent of your uncertainty
- Emphasise a range a plausible values
- Phrase results in terms of the **degree of evidence** (e.g. 'strong/modest/weak evidence of...')

1.6 Summary

Confidence intervals: summary

- Interval estimates are the most common way to quantify uncertainty.
- Confidence intervals are the most common type of interval estimate.
- Confidence intervals are straightforward to construct if we know or can approximate the sampling distribution of the statistic and can construct a pivot.
- We have looked at some well known (and widely used) examples for means, variances and proportions.
- We can derive CIs, whether exact or approximate, for a variety of scenarios, and have techniques for constructing them in general.
- 95% CIs are the most common convention.

2 Prediction intervals

Prediction intervals

Suppose we want to estimate the value of a *future observation*, rather than a parameter of the distribution. We usually call this *prediction* rather than 'estimation'.

We have available data that arose from the same probability distribution. Can we use this to come up with an interval estimate?

Yes. Easiest to see with an example...

Example (prediction interval)

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

Let X^* be a future observation on X, independent of those currently observed.

By independence, we have:

$$\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n})$$
$$X^* \sim \mathrm{N}(\mu, 1)$$
$$\bar{X} - X^* \sim \mathrm{N}(0, 1 + \frac{1}{n})$$

Therefore we can write,

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

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

From this we get a *95% prediction interval* for X,

$$\bar{x} \pm 1.96\sqrt{1+\frac{\sigma^2}{n}}$$

Compare this with the 95% confidence interval for μ,

$$\bar{x} \pm 1.96\sqrt{\frac{\sigma^2}{n}}$$

Remarks

- The prediction interval for X is much wider than the confidence interval for μ.
- As $n \to \infty$, the width of the confidence interval shrinks to zero, but the width of the prediction interval tends to the width of the corresponding population probability interval ($\mu \pm 1.96$).
- This makes sense: we get complete certainty about μ, but each observation on X has inherent variability (in this case, a variance of 1).
- In the prediction interval estimator, all quantities are random variables:

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

3 Sample size determination

Sample size determination: overview

We are planning a study. How much data do we need?

It depends on how much precision is required, often measured by the desired width of a confidence interval.

We will go through two estimation scenarios:

- Means
- Proportions

Example: sample size for means

Random sample (iid): $X_1, \ldots, X_n \sim \mathrm{N}(\mu, 15^2)$

Want a 95% confidence interval of width 2 (i.e. $\bar{x} \pm 1$).

The confidence interval will be given by $\bar{x} \pm 1.96\frac{15}{\sqrt{n}}$.

So we need,

$$1.96\frac{15}{\sqrt{n}} = 1$$

which gives

$$\sqrt{n} = 29.4, \quad \text{or} \quad n \approx 864.36$$

and so for our study we need sample size of at least **865**.

Sample size for means

Confidence interval of the form:

$$\bar{x} \pm c\frac{\sigma}{\sqrt{n}} = \bar{x} \pm \epsilon$$

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

For a prespecified ϵ, we have:

$$\epsilon = c\frac{\sigma}{\sqrt{n}}, \quad \text{or} \quad n = \left(\frac{c\sigma}{\epsilon}\right)^2$$

Example 2: sample size for means

A researcher plans to select a sample of first-grade girls in order to estimate their mean height μ. The sample is required to be large enough to get an estimate to within 0.5 cm. From previous studies we know $\sigma \approx 2.8$ cm.

$$n = \left(\frac{c\sigma}{\epsilon}\right)^2 = \left(\frac{1.96 \times 2.8}{0.5}\right)^2 = 120.47$$

The researcher selects 121 girls.

Sample size for proportions

Confidence interval is of the form:

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

where $c = \Phi^{-1}(1 - \alpha/2)$. In order to have $\hat{p} \pm \epsilon$ for a given ϵ, we need:

$$\epsilon = c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \text{or} \quad n = \frac{c^2\,\hat{p}(1-\hat{p})}{\epsilon^2}$$

Can use a preliminary estimate of $\hat{p}$ if this is available. Otherwise, note that $\hat{p}(1-\hat{p}) \leqslant \frac{1}{4}$, which means we can use $n = c^2/(4\epsilon^2)$ as a conservative choice.

$$n = \frac{c^2}{4\epsilon^2} \cdot 8$$

Example: Sample size for proportions

The unemployment rate has been 8% for a while. A researcher wishes to take new sample to estimate it and wants to be 'very certain', by using a 99% CI, that the new estimate is within 0.001 of true proportion.

$$n = \frac{c^2\,\hat{p}(1-\hat{p})}{\epsilon^2} = \frac{2.576^2 \times 0.08 \times (1-0.08)}{0.001^2} \approx 488394$$

At this stage the researcher panics and says they don't really need to be that sure!

Try again... a 98% CI and a difference of 0.01 gives $n = 3982$, which is more practical, although possibly still a bit large.

R markdown.

Regression

(Module 5)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019

Contents

Aims of this module

- Introduce the concept of **regression**
- Show a simple model for studying the relationship between two variables
- Discuss correlation and how it relates to regression

1 Introduction

Relationships between two variables

We have studied how to do estimation for some simple scenarios:

- iid samples from a single distribution (X_i)
- comparing iid samples from two different distributions (X_i & Y_j)
- differences between paired measurements ($X_i - Y_i$)

We now consider how to analyse bivariate data more generally, i.e. two variables, X and Y, measured at the same time, i.e. as a pair.

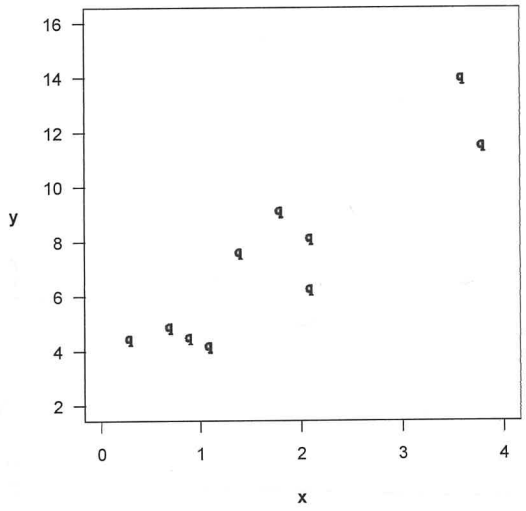
The data consist of pairs of data points, (x_i, y_i).

These can be visualised using a *scatter plot.*

Example data

x_i	y_i
1.80	9.18
1.40	7.66
2.10	6.33
0.30	4.51
3.60	14.04
0.70	4.94
1.10	4.24
2.10	8.19
0.90	4.55
3.80	11.57

$n = 10$



Scatter plot.

2 Regression

Regression

Often interested in how Y depends on X. For example, we might want to use X to predict Y.

In such a setting, we will assume that the X values are known and fixed (henceforth, x instead of X), and look at how Y varies given x.

Example: Y is a student's final mark for Statistics, and x is their mark for the prerequisite subject Probability. Does x help to predict Y?

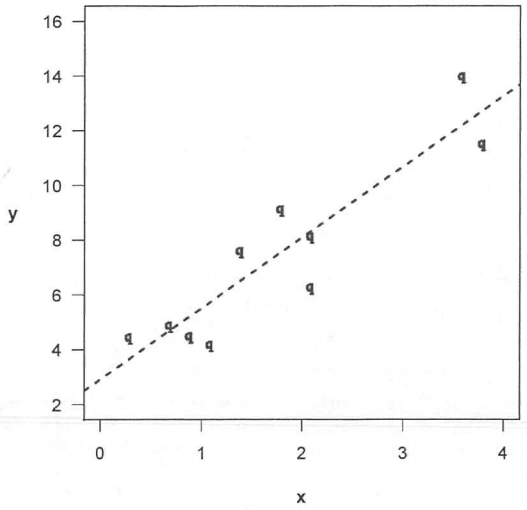
The *regression* of Y on x is the conditional mean, $\mathbb{E}(Y \mid x) = \mu(x)$.

The regression can take any form. We consider *simple linear regression*, which has the form of a **straight line**:

$$\mathbb{E}(Y \mid x) = \alpha + \beta x \quad \text{and} \quad \text{var}(Y \mid x) = \sigma^2.$$

Example: simple linear regression model

$$\mathbb{E}(Y \mid x) = \alpha + \beta x$$
$$\text{var}(Y \mid x) = \sigma^2$$



Terminology

- Y is called a *response* variable. Can also be called an *outcome* or *target* variable. Please do **not** call it the 'dependent' variable.
- x is called a *predictor* variable. Can also be called an *explanatory* variable. Please do **not** call it an 'independent' variable.
- $\mu(x)$ is called the *(linear) predictor function* or sometimes the *regression curve* or the *model equation*.
- The parameters in the predictor function are called *regression coefficients*.

Why 'regression'?

It is strange terminology, but it has stuck.

Refers to the idea of *'regression to the mean'*: if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement, and vice versa.

First described by Sir Francis Galton when studying the inheritance of height between fathers and sons. In doing so, he invented the technique of simple linear regression.

Linearity

A regression model is called *linear* if it is linear in the coefficients.

It doesn't have to define a straight line!

Complex and non-linear functions of x are allowed, as long as the resulting predictor function is a linear combination (i.e. an additive function) of them, with the coefficients 'out the front'.

For example, the following are linear models:

$$\mu(x) = \alpha + \beta x + \gamma x^2$$

$$\mu(x) = \frac{\alpha}{x} + \frac{\beta}{x^2}$$

$$\mu(x) = \alpha \sin x + \beta \log x$$

The following are NOT linear models:

$$\mu(x) = \alpha \sin(\beta x)$$
$$\mu(x) = \frac{\alpha}{1+\beta x}$$
$$\mu(x) = \alpha x^{\beta}$$

...but the last one can be re-expressed as a linear model on a log scale (by taking logs of both sides),

$$\mu^*(x) = \alpha^* + \beta \log x$$

3 Simple linear regression

Estimation goals

Back to our simple linear regression model:

$$\mathbb{E}(Y \mid x) = \alpha + \beta x \quad \text{and} \quad \text{var}(Y \mid x) = \sigma^2.$$

- We wish to estimate the slope (β), the intercept (α), the variance of the errors (σ^2), their standard errors and construct confidence intervals for these quantities.
- Often want to use the fitted model to make predictions about future observations (i.e. predict Y for a new x).
- Note: the Y_i are not iid. They are independent but have different means, since they depend on x_i.
- We have not (yet) assumed any specific distribution for Y, only a conditional mean and variance.

Reparameterisation

Changing our model slightly...

Let $\alpha_0 = \alpha + \beta\bar{x}$, which gives:

$$\begin{aligned}\mathbb{E}(Y \mid x) &= \alpha + \beta x \\ &= \alpha_0 + \beta(x - \bar{x})\end{aligned}$$

Now our model is in terms of α_0 and β.

This will make calculations and proofs simpler.

3.1 Point estimation of the mean

Least squares estimation

Choose α_0 and β to minimize the sum of squared deviations:

$$H(\alpha_0, \beta) = \sum_{i=1}^{n} (y_i - \alpha_0 - \beta\,(x_i - \bar{x}))^2$$

Solve this by finding the partial derivatives and setting to zero:

$$0 = \frac{\partial H(\alpha_0, \beta)}{\partial \alpha_0} = 2\sum_{i=1}^{n} [y_i - \alpha_0 - \beta(x_i - \bar{x})](-1)$$
$$0 = \frac{\partial H(\alpha_0, \beta)}{\partial \beta} = 2\sum_{i=1}^{n} [y_i - \alpha_0 - \beta(x_i - \bar{x})](-(x_i - \bar{x}))$$

These are called the *normal equations*.

Least squares estimators

Some algebra yields the *least square estimators*,

$$\hat{\alpha}_0 = \bar{Y}, \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Another expression for $\hat{\beta}$ is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

These are equivalent, due to the following result:

$$\sum(x_i - \bar{x})(Y_i - \bar{Y}) = \sum(x_i - \bar{x})Y_i.$$

Can also then get an estimator for α:

$$\begin{aligned}\hat{\alpha} &= \hat{\alpha}_0 - \hat{\beta}\bar{x} \\ &= \bar{Y} - \hat{\beta}\bar{x}.\end{aligned}$$

And also an estimator for the predictor function,

$$\begin{aligned}\hat{\mu}(x) &= \hat{\alpha} + \hat{\beta}x \\ &= \hat{\alpha}_0 + \hat{\beta}(x - \bar{x}) \\ &= \bar{Y} + \hat{\beta}(x - \bar{x}).\end{aligned}$$

Ordinary least squares

This method is sometimes called *ordinary least squares* or *OLS*.

Other variants of least squares estimation exist, with different names. For example, 'weighted least squares'.

Example: least squares estimates

For our data:

$$\begin{aligned}\bar{x} &= 1.78 \\ \bar{y} &= 7.52 = \hat{\alpha}_0 \\ \hat{\alpha} &= 2.91 \\ \hat{\beta} &= 2.59\end{aligned}$$

The fitted model equation is then:

$$\hat{\mu}(x) = 2.91 + 2.59x$$

```
> rbind(y, x)
  [,1] [,2] [,3] [,4]  [,5] [,6] [,7] [,8] [,9] [,10]
y 9.18 7.66 6.33 4.51 14.04 4.94 4.24 8.19 4.55 11.57
x 1.80 1.40 2.10 0.30  3.60 0.70 1.10 2.10 0.90  3.80

> model1 <- lm(y ~ x)
> model1

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x
      2.911        2.590
```

Properties of these estimators

What do we know about these estimators?

They are all linear combinations of the Y_i,

$$\hat{\alpha}_0 = \sum_{i=1}^{n} \left(\frac{1}{n}\right) Y_i$$

$$\hat{\beta} = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{K}\right) Y_i$$

where $K = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

This allows us to easily calculate means and variances.

Means?

$$\mathbb{E}(\hat{\alpha}_0) = \mathbb{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n}[\alpha_0 + \beta(x_i - \bar{x})] = \alpha_0$$

$$\mathbb{E}(\hat{\beta}) = \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{K}\mathbb{E}(Y_i) = \frac{1}{K}\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + (x_i - \bar{x})\beta)$$

$$= \frac{1}{K}\sum_{i=1}^{n}(x_i - \bar{x})\alpha + \frac{K}{K}\beta = \beta$$

This also implies, $\mathbb{E}(\hat{\alpha}) = \alpha$ and $\mathbb{E}(\hat{\mu}(x)) = \mu(x)$, and so we have that all of the estimators are unbiased.

Variances?

$$\text{var}(\hat{\alpha}_0) = \text{var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(Y_i) = \frac{\sigma^2}{n}$$

$$\text{var}(\hat{\beta}) = \text{var}\left(\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{K}Y_i\right) = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{K}\right)^2\text{var}(Y_i)$$

$$= \frac{1}{K^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\,\text{var}(Y_i) = \frac{1}{K^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\,\sigma^2$$

$$= \frac{1}{K^2}\sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{K^2}\sigma^2 K$$

$$= \frac{\sigma^2}{K}$$

Similarly,

$$\text{var}(\hat{\alpha}) = \left(\frac{1}{n} + \frac{\bar{x}^2}{K}\right)\sigma^2$$

$$\text{cov}(\hat{\alpha}_0, \hat{\beta}) = 0$$

$$\text{var}(\hat{\mu}(x)) = \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{K}\right)\sigma^2$$

Can we get their standard errors?

We need an estimate of σ^2.

3.2 Interlude: Analysis of variance

Analysis of variance: iid model

For $X_i \sim \mathrm{N}(\mu, \sigma^2)$ iid,

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

Analysis of variance: regression model

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&\quad = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}) + \hat{\alpha}_0 + \hat{\beta}(x_i - \bar{x}) - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&\quad = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}) + (\hat{\alpha}_0 - \alpha_0) + (\hat{\beta} - \beta)(x_i - \bar{x}))^2 \\
&\quad = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))^2 + n(\hat{\alpha}_0 - \alpha_0)^2 + K(\hat{\beta} - \beta)^2
\end{aligned}$$

Note that the cross-terms disappear. Let's see...

The cross-terms...

$$\begin{aligned}
t_1 &= 2\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))(\hat{\alpha}_0 - \alpha_0) \\
t_2 &= 2\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))(\hat{\beta} - \beta)(x_i - \bar{x}) \\
t_3 &= 2\sum_{i=1}^{n}(x_i - \bar{x})(\hat{\beta} - \beta)(\hat{\alpha}_0 - \alpha_0)
\end{aligned}$$

Since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0) = \sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$, the first and third cross-terms are easily shown to be zero.

For the second term,

$$\begin{aligned}
\frac{t_2}{2(\hat{\beta} - \beta)} &= \sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta}K \\
&= \sum_{i=1}^{n}Y_i(x_i - \bar{x}) - \sum_{i=1}^{n}Y_i(x_i - \bar{x}) \\
&= 0
\end{aligned}$$

Therefore, all the cross-terms are zero.

Back to the analysis of variance formula...

$$\sum_{i=1}^{n}(Y_i - \alpha_0 - \beta(x_i - \bar{x}))^2$$
$$= \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))^2 + n(\hat{\alpha}_0 - \alpha_0)^2 + K(\hat{\beta} - \beta)^2$$

Taking expectations gives,

$$n\sigma^2 = \mathbb{E}(D^2) + \sigma^2 + \sigma^2$$
$$\Rightarrow \mathbb{E}(D^2) = (n-2)\sigma^2$$

where

$$D^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))^2.$$

3.3 Point estimation of the variance

Variance estimator

Based on these results, we have an unbiased estimator of the variance,

$$\hat{\sigma}^2 = \frac{1}{n-2}D^2.$$

The inferred mean for each observation is called its *fitted value*, $\hat{Y}_i = \hat{\alpha}_0 + \hat{\beta}(x_i - \bar{x})$.

The deviation from each fitted value is called a *residual*, $R_i = Y_i - \hat{Y}_i$.

The variance estimator is based on the sum of squared residuals, $D^2 = \sum_{i=1}^{n} R_i^2$.

Example: variance estimate

For our data:

$$d^2 = 16.12$$
$$\hat{\sigma}^2 = 2.015$$
$$\hat{\sigma} = 1.42$$

3.4 Standard errors of the estimates

Standard errors

We can subsitute $\hat{\sigma}^2$ into the formulae for the standard deviation of the estimators in order to calculate standard errors.

For example,

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{K}$$
$$\Rightarrow \text{se}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{K}}$$

Example: standard errors

For our data:

$$\mathrm{se}(\hat{\alpha}_0) = \frac{\hat{\sigma}}{\sqrt{n}} = 0.142$$

$$\mathrm{se}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{K}} = 0.404$$

$$\mathrm{se}(\hat{\mu}(x)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{K}} = 1.42 \times \sqrt{\frac{1}{10} + \frac{(x-1.78)^2}{12.34}}$$

3.5 Confidence intervals

Maximum likelihood estimation

Want to also construct confidence intervals. This requires further assumptions about the population distribution.

Let's assume a normal distribution:

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2).$$

Alternative notation (commonly used for regression/linear models):

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2).$$

Let's maximise the likelihood...

Since the Y_i's are independent, the likelihood is:

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(y_i - \alpha_0 - \beta(x_i - \bar{x}))^2}{2\sigma^2}\right\} \\ -\ln L(\alpha, \beta, \sigma^2) &= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \beta(x_i - \bar{x}))^2 \\ &= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} H(\alpha_0, \beta) \end{aligned}$$

The α_0 and β that maximise the likelihood (minimise the log-likelihood) are the same as those that minimise the sum of squares, H.

The OLS estimates are the same as the MLEs!

What about σ^2?

Differentiate by σ, set to zero, solve...

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n} D^2$$

This is biased. Prefer to use the previous, unbiased estimator,

$$\hat{\sigma}^2 = \frac{1}{n-2} D^2$$

Sampling distributions

The $Y_1, \cdots, Y_n$ are independent normally distributed random variables.

Except for $\hat{\sigma}^2$, our estimators are linear combinations of the Y_i so will also have normal distributions, with mean and variance as previously derived.

For example,

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{K}\right).$$

Moreover, we know $\hat{\alpha}_0$ and $\hat{\beta}$ are independent, because they are bivariate normal rvs with zero covariance.

Using the analysis of variance decomposition (from earlier), we can show that,

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Therefore, we can define pivots for the various mean parameters. For example,

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{K}} \sim t_{n-2}$$

and

$$\frac{\hat{\mu}(x) - \mu(x)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{K}}} \sim t_{n-2}$$

This allows us to construct confidence intervals.

Example: confidence itervals

For our data, a 95% CI for β is:

$$\hat{\beta} \pm c\frac{\hat{\sigma}}{\sqrt{K}} = 2.59 \pm 2.31 \times 0.404 = (1.66, 3.52)$$

where c is the 0.975 quantile of t_{n-2}.

A 95% CI for $\mu(3)$ is:

$$\hat{\mu}(3) \pm c \times \mathrm{se}(\hat{\mu}(3)) = 10.68 \pm 2.31 \times 0.667 = (9.14, 12.22)$$

3.6 Prediction intervals

Deriving prediction intervals

Use the same trick as we used for the simple model,

$$\begin{aligned} Y^* &\sim \mathrm{N}\left(\mu(x^*), \sigma^2\right) \\ \hat{\mu}(x^*) &\sim \mathrm{N}\left(\mu(x^*), \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{K}\right)\sigma^2\right) \\ Y^* - \hat{\mu}(x^*) &\sim \mathrm{N}\left(0, \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{K}\right)\sigma^2\right) \end{aligned}$$

A 95% PI for Y^* is given by:

$$\hat{\mu}(x^*) \pm c\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{K}}$$

Example: prediction interval

A 95% PI for Y^* corresponding to $x^* = 3$ is:

$$10.68 \pm 2.31 \times 1.42 \times \sqrt{1 + \frac{1}{10} + \frac{(3 - 1.78)^2}{12.34}} = (7.06, 14.30)$$

Much wider than the corresponding CI, as we've seen previously.

3.7 R examples

```
> model1 <- lm(y ~ x)
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01970 -1.05963  0.02808  1.04774  1.80580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9114     0.8479   3.434 0.008908 **
x             2.5897     0.4041   6.408 0.000207 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.419 on 8 degrees of freedom
Multiple R-squared:  0.8369,Adjusted R-squared:  0.8166
F-statistic: 41.06 on 1 and 8 DF,  p-value: 0.0002074

> # Confidence intervals for mean parameters
> confint(model1)
                2.5 %   97.5 %
(Intercept) 0.9560629 4.866703
x           1.6577220 3.521623

> # Data to use for prediction.
> data2 <- data.frame(x = 3)

> # Confidence interval for mu(3)
> predict(model1, newdata = data2, interval = "confidence")
       fit      lwr      upr
1 10.6804 9.142823 12.21798

> # Prediction interval for y when x = 3.
> predict(model1, newdata = data2, interval = "prediction")
       fit   lwr     upr
1 10.6804 7.064 14.2968
```

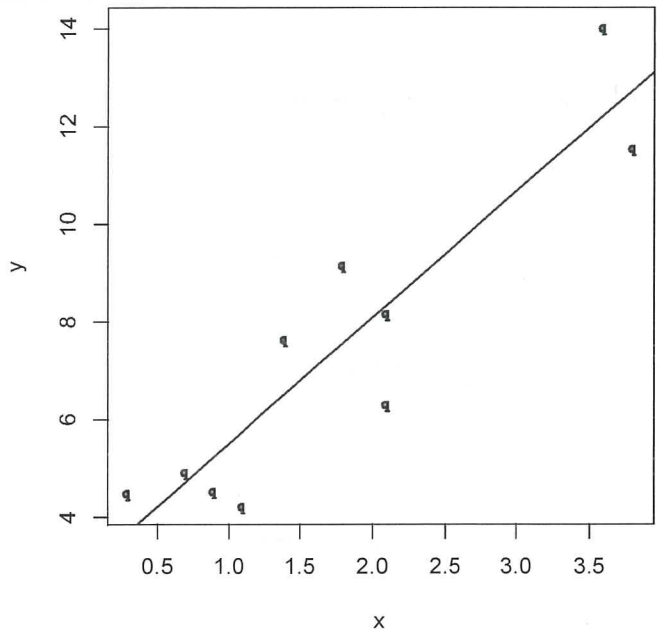
R example explained

- The `lm` (linear model) command fits the model.
- `model1` is an object that contains all the results of the regression needed for later calculations.
- `summary(model1)` acts on `model1` and summarizes the regression.
- `predict` can calculate CIs and PIs.
- R provides more detail than we need at the moment. Much of the output relates to hypothesis testing that we will get to later.

Plot data and fitted model

```
> plot(x, y, col = "blue")
> abline(model1, col = "blue")
```

The command `abline(model1)` adds the fitted line to a plot.



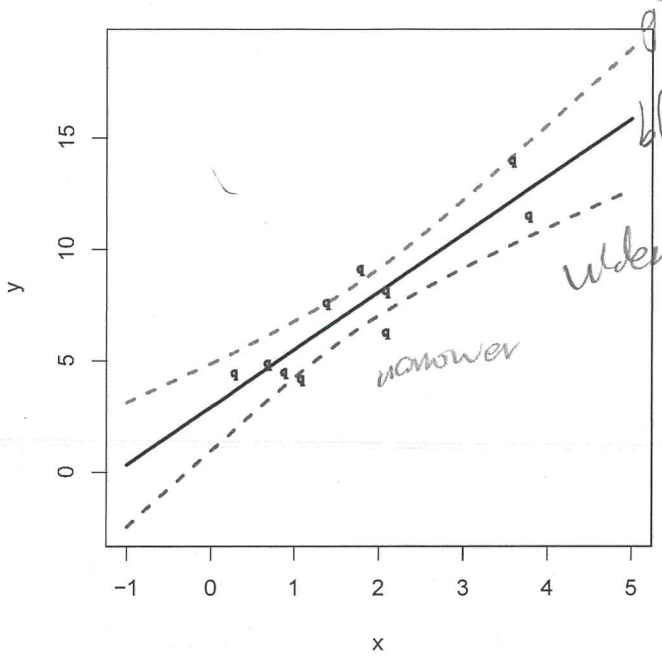
Fitted values and CIs for their means

```
> predict(model1, interval = "confidence")
         fit       lwr       upr
1   7.572793  6.537531  8.608056
2   6.536924  5.442924  7.630925
3   8.349695  7.272496  9.426895
4   3.688285  1.963799  5.412771
5  12.234204 10.247160 14.221248
6   4.724154  3.280382  6.167925
7   5.760023  4.546338  6.973707
8   8.349695  7.272496  9.426895
9   5.242088  3.921478  6.562699
10 12.752138 10.603796 14.900481
```

Confidence band for the mean

```
> data3 <- data.frame(x = seq(-1, 5, 0.05))
> y.conf <- predict(model1, data3, interval = "confidence")
> head(cbind(data3, y.conf))
     x       fit       lwr      upr
1 -1.00 0.3217104 -2.468232 3.111653
2 -0.95 0.4511941 -2.295531 3.197919
3 -0.90 0.5806777 -2.122943 3.284298
4 -0.85 0.7101613 -1.950472 3.370794
5 -0.80 0.8396449 -1.778124 3.457414
6 -0.75 0.9691286 -1.605906 3.544164
```

```
> matplot(data3$x, y.conf, type = "l", lty = c(1, 2, 2),
+         lwd = 2, xlab = "x", ylab = "y")
> points(x, y, col = "blue")
```

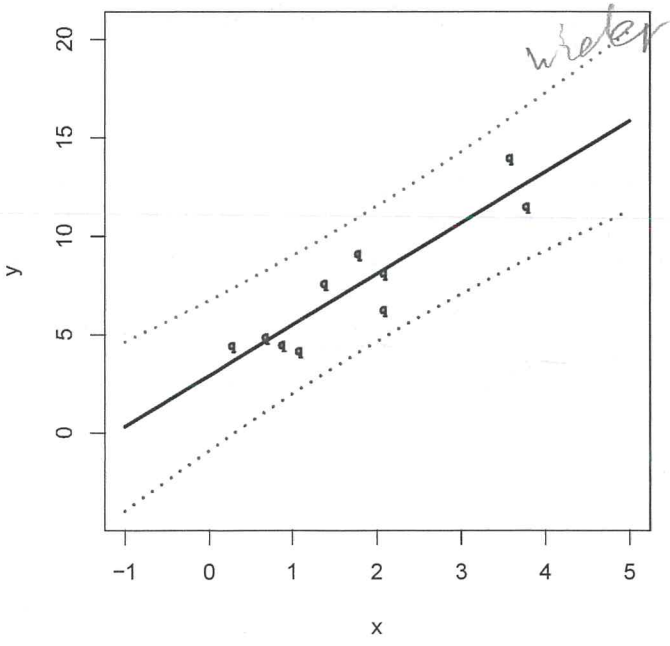


$$\frac{\hat{\mu}(x) - \mu(x)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{lc}}} \sim t_{n-2}$$

~~conf~~ CI

Prediction bands for new observations

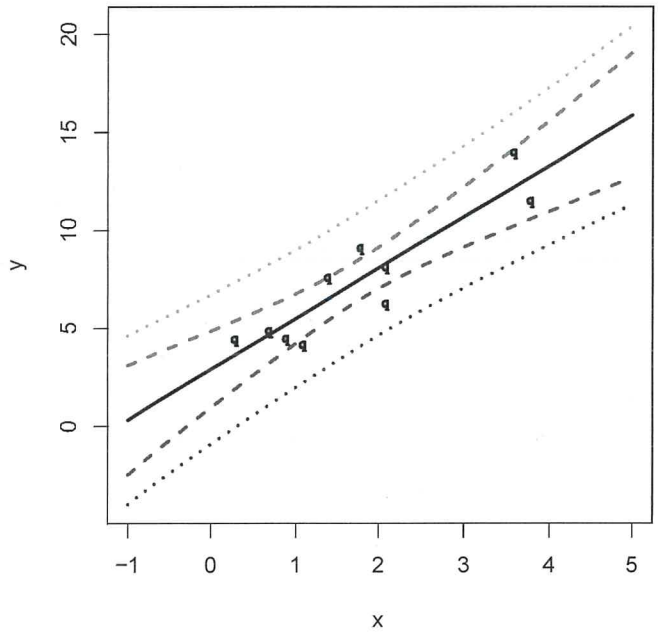
```
> y.pred <- predict(model1, data3, interval = "prediction")
> head(cbind(data3, y.pred))
      x       fit       lwr      upr
1 -1.00 0.3217104 -3.979218 4.622639
2 -0.95 0.4511941 -3.821827 4.724215
3 -0.90 0.5806777 -3.664763 4.826119
4 -0.85 0.7101613 -3.508034 4.928357
5 -0.80 0.8396449 -3.351646 5.030936
6 -0.75 0.9691286 -3.195606 5.133863

> matplot(data3$x, y.pred, type = "l", lty = c(1, 3, 3),
+         lwd = 2, xlab = "x", ylab = "y")
> points(x, y, col = "blue")
```



PI

Both bands plotted together

```
> matplot(data3$x, y.pred, type = "l", lty = c(1, 2, 2, 3, 3),
+         lwd = 2, xlab = "x", ylab = "y")
> points(x, y, col = "blue")
```



3.8 Model checking

Checking our assumptions

What modelling assumptions have we made?

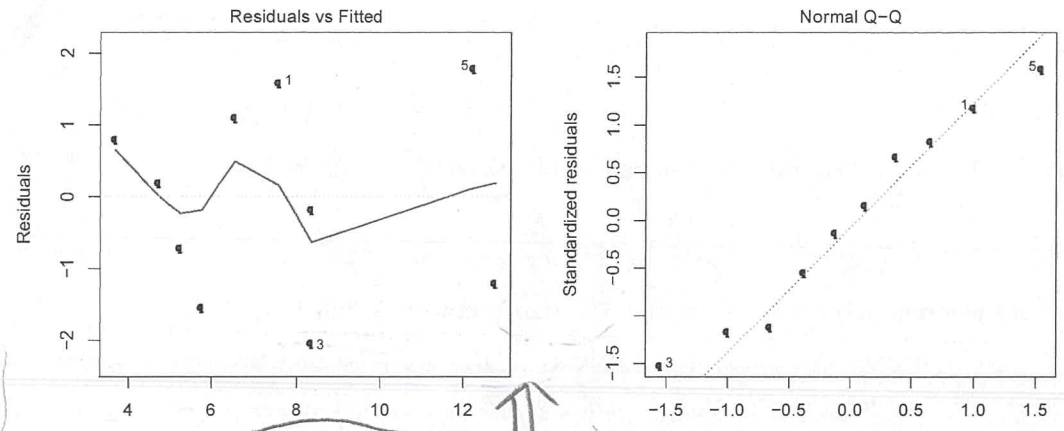
- Linear model for the mean
- Equal variances for all observations (*homoscedasticity*)
- Normally distributed residuals

Ways to check these:

- Plot the data and fitted model together (done!)
- Plot residuals vs fitted values
- QQ plot of the residuals

In R, the last two of these are very easy to do:

```
> plot(model1, 1:2)
```



4 Further regression models

Multiple regression

- What if we have more than one predictor?
- Observe $x_{i1}, \ldots, x_{ik}$ as well as y_i (for each i)
- Can fit a *multiple regression model*:

$$\mathbb{E}(Y \mid x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- This is linear in the coefficients, so is still a linear model
- Fit by method of least squares by minimising:

$$H = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2$$

- Take partial derivatives, etc., and solve for $\beta_0, \ldots, \beta_k$.
- The subject Linear Statistical Models (MAST30025) looks into these types of models in much more detail.

Two-sample problem

- The two-sample problem can be expressed as a linear model!
- Sample $Y_1, \ldots, Y_n \sim \mathrm{N}(\mu_1, \sigma^2)$ and $Y_{n+1}, \ldots, Y_{n+m} \sim \mathrm{N}(\mu_2, \sigma^2)$.
- Define *indicator variables* (x_{i1}, x_{i2}) where $(x_{i1}, x_{i2}) = (1, 0)$ for $i = 1, \ldots, n$ and $(x_{i1}, x_{i2}) = (0, 1)$ for $i = n+1, \ldots, n+m$.
- Observed data: (y_i, x_{i1}, x_{i2})
- Then $Y_1, \ldots, Y_n$ each have mean $1 \times \beta_1 + 0 \times \beta_2 = \mu_1$ and $Y_{n+1}, \ldots, Y_{n+m}$ each have mean $0 \times \beta_1 + 1 \times \beta_2 = \mu_2$.
- This is in the form a multiple regression model.
- The *general linear model* unifies many different types of models together into a common framework. The subject MAST30025 covers this in more detail.

5 Correlation

5.1 Definitions

Correlation coefficient

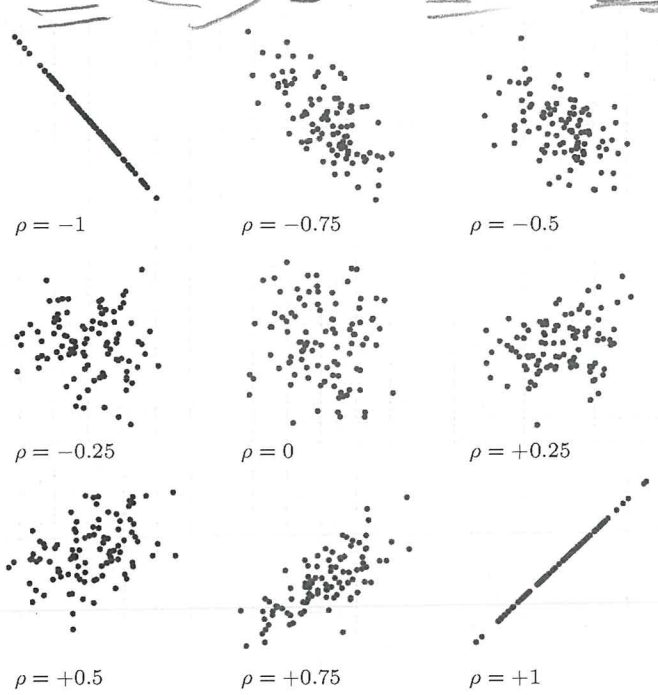
(Revision) for two rvs X and Y, the *correlation coefficient*, or simply the *correlation*, is defined as:

$$\rho = \rho_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}\, X \,\text{var}\, Y}} = \frac{\sigma_{XY}}{\sigma_X \sigma_y}$$

This is quantitative measure of the strength of relationship, or *association*, between X and Y.

We will now consider inference on ρ, based on an iid sample of pairs (X_i, Y_i).

Note: unlike in regression, X is now considered as a random variable.



5.2 Point estimation

Sample covariance

To estimate $\text{cov}(X,Y)$ we use the *sample covariance*:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right)$$

You can check that this is unbiased, $\mathbb{E}(S_{XY}) = \sigma_{XY} = \text{cov}(X,Y)$.

Sample correlation coefficient

To estimate ρ we use the *sample correlation coefficient* (also known as *Pearson's correlation coefficient*):

$$R = R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

You can check that $|R| \leqslant 1$, just like $|\rho| \leqslant 1$.

This gives a point estimate of ρ.

For further results, we make some more assumptions...

5.3 Relationship to regression

Bivariate normal

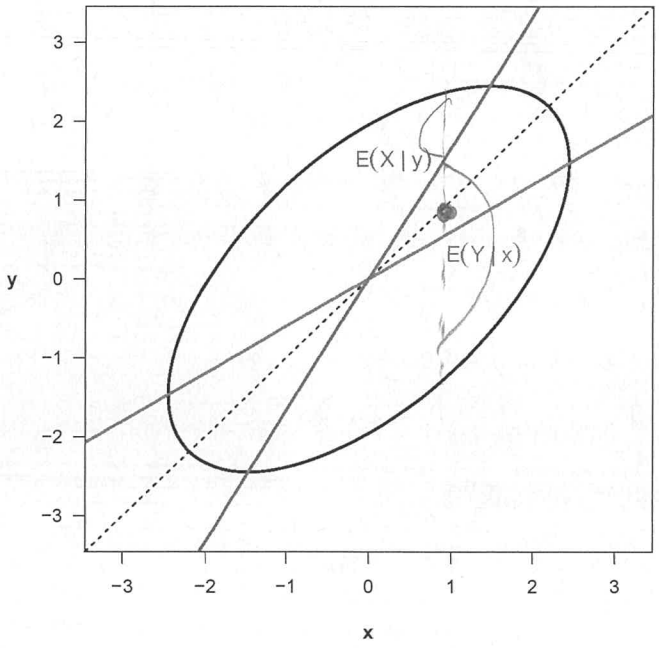
Assume X and Y have correlation ρ and follow a bivariate normal distribution,

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}_2\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right)$$

In this case, the regressions are linear,

$$\mathbb{E}(X \mid y) = \mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y) = \alpha' + \beta' y$$
$$\mathbb{E}(Y \mid x) = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) = \alpha + \beta x$$

Note: $\beta' \neq 1/\beta$



Variance explained

An alternative analysis of variance decomposition:

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 + \hat{\beta}^2 \sum (X_i - \bar{X})^2 \\ &= (1 - R^2) \sum (Y_i - \bar{Y})^2 + R^2 \sum (Y_i - \bar{Y})^2 \end{aligned}$$

This implies that R^2 is the proportion of the variation in Y 'explained' by x.

In this usage, R^2 is called the *coefficient of determination*.

Remarks

- For simple linear regression, the coefficient of determination is the same as the square of the sample correlation, with both being denoted by R^2.

- Also, the proportion of Y explained by x is the same as the proportion of X explained by y. Both are equal to R^2, which is a symmetric expression of both X and Y.
- For more complex models, the coefficient of determination is more complicated: it needs to be calculated using all predictor variables together.

5.4 Confidence interval

Approximate sampling distribution

Define:

$$g(r) = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

This function has a standard name, $g(r) = \operatorname{artanh}(r)$, and so does it's inverse, $g^{-1}(r) = \tanh(r)$. The function $g(r)$ is also known as the *Fisher transformation*.

The following is a widely used approximation:

$$g(R) \approx \mathrm{N}\left(g(\rho), \frac{1}{n-3} \right)$$

We can use this to construct approximate confidence intervals.

Example: correlation

For our data:

$$r = 0.91$$
$$r^2 = 0.84$$

An approximate 95% CI for $g(\rho)$ is:

$$g(r) \pm \frac{c}{\sqrt{n-3}} = 1.56 \pm 1.96 \times 0.378 = (0.819, 2.30)$$

where $c = \Phi^{-1}(1 - \alpha/2)$. Transforming this to an approximate 95% CI for ρ:

$$(\tanh(0.819), \tanh(2.30)) = (0.67, 0.98)$$

5.5 R example

```
> cor(x, y)
[1] 0.9148421

> cor(x, y)^2
[1] 0.836936

> cor.test(x, y)

        Pearson's product-moment correlation

data:  x and y
t = 6.4078, df = 8, p-value = 0.0002074
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6726924 0.9799873
sample estimates:
      cor
0.9148421
```

```
> model1 <- lm(y ~ x)
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01970 -1.05963  0.02808  1.04774  1.80580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9114     0.8479   3.434 0.008908 **
x             2.5897     0.4041   6.408 0.000207 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.419 on 8 degrees of freedom
Multiple R-squared:  0.8369,Adjusted R-squared:  0.8166
F-statistic: 41.06 on 1 and 8 DF,  p-value: 0.0002074
```