Estimation

Experimental Design and Data Analysis for Biologists

G. Quinn and M. Keough

Design by M. Logan

2004

About

This presentation is a brief revision of basic statistical concepts.

Links…

Throughout the presentation, marks in the form of (qk2002, …) provide references to sections within the recommended statistical text;

Quinn, G. P. and Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Words and phrases in purple type face provide tooltip-style extra information, while blue type face provide links to popups that contain additional information and or definitions.

Navigation…

Navigation buttons on the right hand side of each page provide (from top to bottom) 'Previous Page', 'Next Page', 'First Page', 'Last Page', 'Go Back' and 'Quit' navigational shortcuts.

Sampling from a population

- Imagine we wanted to know how much an average male koala weighed in a particular Eucalypt forest in southern Australia. In this case, the **statistical population** is defined as all the male koalas in this particular forest. That is, the population is defined as all the possible observations of interest. (qk2002, pg.14)
- The variable in this example is the weight of individual koalas and each koala is an observation.

- Note in this example that the statistical population is also a biological population. This won't usually be the case – the statistical population is likely to be a collection of sampling or experimental units rather than individual organisms.

- Measures used to characterise a population (such as the population mean) are called **population parameters**. What characteristics of the population might we be interested in? Characteristics might include:
  - the range of male koala weights,
  - the shape of the distribution of weights,
  - the average male koala weight,
  - the variability in male koala weights

- There are two potential approaches for obtaining population characteristics:
  - Weigh every single male koala in the forest and calculate the population characteristic directly. This census approach – collecting all possible observations within a population – is usually impossible because most statistical populations have too many observations. A large forest might have 100's of koalas.
  - Weigh a subset (sample) of all the male koalas, and use this sample to **estimate** the population characteristic. This approach – using a sample to estimate characteristics of a population – is the usual statistical approach.

- Hence, measured characteristics of a sample (sample statistics) are used to estimate population parameters.

- It is important that the sample is an unbiased representation of the population. This is achieved by:
  - Sampling randomly so each observation has an equal chance of being selected
  - Collecting a sufficiently large number (*n*) of observations (replicates) in your sample

For example, we could not be sure that sample mean is a good estimate of the population mean if the sample:

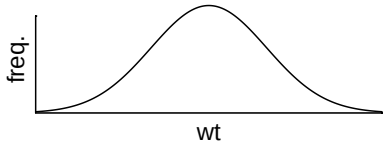
- only comprised koalas collected from one corner of the forest, or
- only comprised koalas that are easy to capture (which may be smaller koalas), or
- comprised only two koalas.

- Other forms of sampling can be used, depending on circumstances (qk2002..):
- stratified random sampling, cluster sampling, systematic sampling

Frequency (probability) distribution of a variable

- Distribution (relative frequency) of the values of a variable in a population

frequency (and long-run probability) of different values of a variable occurring under repeated sampling



For example, the weights of male koalas could follow a distribution like this, whereby the majority of koalas weigh around a certain amount, and progressively heavier and lighter individuals are less and less common.

Type of estimates

Point estimate

A single value estimate of the parameter, e.g. the sample mean (y) is a point estimate of the population mean (μ), the sample standard deviation (s) is a point estimate of the population standard deviation (σ)

Interval estimate

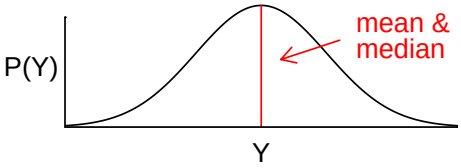
A range within which we have some specified degree of confidence that the true parameter lies, e.g. 95% confidence interval is an interval estimate of the population mean (μ)

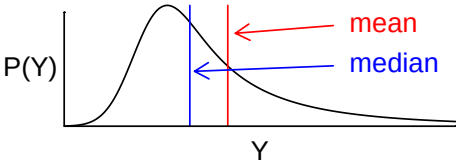
Point estimates

Measures of location

Measures of the average or middle value

- Population mean (μ) – the average value
- Sample mean (y) – estimates (μ)
- Population median – the middle value
- Sample median – estimates population median





Measures of spread

Measures of how variable the observations are

- Population variance (σ^2) – average sum of squared deviations from mean
- Sample variance (s^2) – estimates population variance

$$\frac{\sum(y_i-\bar{y})^2}{n-1}$$

- Standard deviation of sample (s) and population (σ)

square root of variance

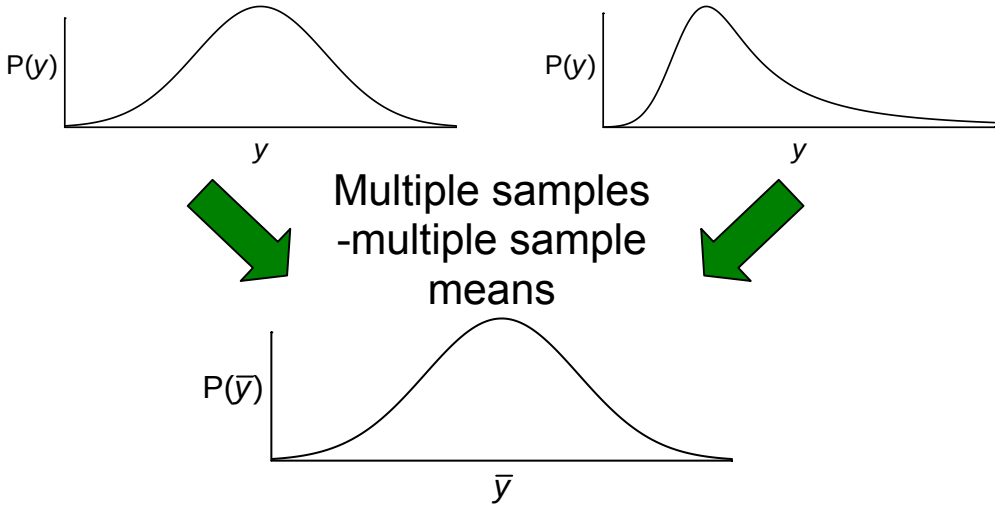
same units as original variable (*c.f.* variance)

Sampling distributions

The frequency (or probability) distribution of a statistic (e.g. sample mean):

- Collect many samples (each with *n* observations) from a population
- Calculate the mean for each sample
- Plot frequency distribution of the sample means (sampling distribution)

Sampling distribution of means



Sampling distribution of means

- The sampling distribution of the sample means approaches a normal distribution as *n* gets larger – this is called the Central Limit Theorem.
- The mean of this sampling distribution is μ, the mean of the original population.
- The standard deviation of this sampling distribution is $\sigma/\sqrt{n}$, the standard deviation of original population divided by the square root of the sample size – the standard error (SE) of the mean.

MONASH UNIVERSITY

Standard error of the mean

Since we usually collect a sample purely for the purpose of estimating some population parameter, we need a measure of how good our estimate is likely to be.

- Population SE is estimated by the sample SE:

$$\frac{s}{\sqrt{n}}$$

- The SE is a measure of the precision of the sample mean. It is a measure of how repeatable a sample statistic is, although is often interpreted as how close the sample mean is likely to be to the true population mean assuming no bias in sampling

Worked example

Lovett *et al*. (2000) measured the concentration of SO_4^{2-} in 39 North American forested streams (qk2002, Box 2.2)

Lovett *et al* (2000)

50.6, 55.4, 56.5, 57.5, 58.3, 63.0, 66.5, 64.5, 63.4, 58.4, 70.6, 56.9, 56.7, 56.0, 60.4, 67.8, 70.8, 58.6, 59.5, 55.5, 63.4, 57.8, 55.1, 65.5, 62.7, 72.1, 63.4, 68.5, 65.8, 69.2, 66.7, 59.3, 61.1, 62.1, 70.4, 62.1, 64.6, 61.4, 56.9

Statistic	Value
Sample mean	61.92
Sample median	62.10
Sample variance	27.46
Sample SD	5.24
SE of mean	0.84

Interval estimates

Interval estimates

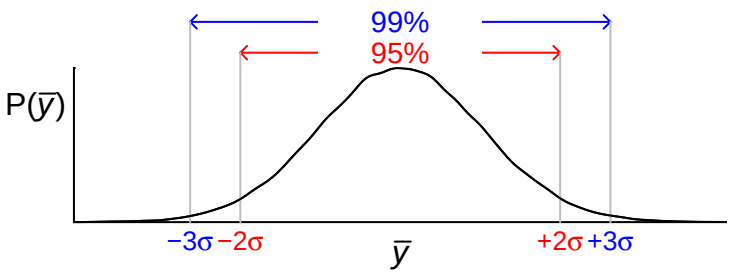
- How confident are we in a single sample estimate of μ, i.e. how close do we think our sample mean is to the unknown population mean?
- Remember that the population mean, μ, is a fixed, yet unknown value
- A confidence interval (CI) is an interval (range of values) within which we are 95% (for example) sure μ occurs
- Strictly, a 95% CI is an interval within which 95% of sample means from repeated sampling are likely to fall

Distribution of sample means

- Based on Central Limit Theorem, we know that the distribution of sample means has a mean of μ and a standard deviation of $\sigma\sqrt{n}$. Based on our single sample, we only estimate σ using s. The distribution of sample means, with $\mu = 0$ and $s/\sqrt{n} = 1$, follows a t-distribution, a well-known probability distribution in statistics.
- Transform any sample mean to its equivalent value (t value) from a t distribution:

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

t distribution



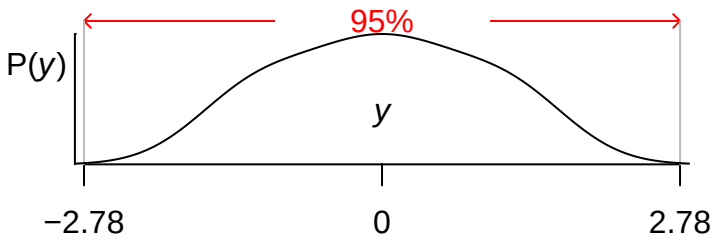
- When $n > 30$, the *t*-distribution is almost the same as a normal distribution, otherwise it is flatter, more spread out than a normal distribution
- There are different *t*-distributions for different $n < 30$ (actually for different degrees-of-freedom, *df*, which is *n*-1)

- The proportions of *t* values between two particular *t* values from *t*-distributions for any given *df* can be calculated

- these proportions are tabulated in many statistics books and programmed into statistical software

t statistic

- For $n = 5$ ($df = 4$), 95% of all t values occur between $t = -2{:}78$ and $t = +2{:}78$.



- 95% probability that t is between -2.78 and +2.78
- 95% probability that $(y-\mu)/(s/\sqrt{n})$ is between -2.78 and +2.78

t statistic

- For 95% CI, use the *t* value between which 95% of all *t* values in the *t*-distribution occur for the specific df (n-1)
- P $\{ \bar{y} - t(S/\sqrt{n}) \leq \mu \leq \bar{y} + t(S/\sqrt{n}) \} = 0.95$
- This is a confidence interval:
  - 95% of CI's from repeated samples of size *n* would contain μ and 5% won't
  - Often interpreted as 95% probability that this interval includes the true population mean

Worked example

Sample mean	61.92
Sample SD	5.24
SE	0.84

(qk2002, Box 2.2)

- The *t* value (95%, 38df) = 2.02 (from a *t* table)
- 2.5% of *t* values are greater than 2.02
- 2.5% of *t* values are less that than -2.02
- 95% of *t* values are between -2.02 and +2.02

$P\{61.92 - 2.02(15.24/\sqrt{39}) \leq \mu \leq 61.92 + 2.02(15.24/\sqrt{39})\} = 0.95$

$P\{\ 61.22 \leq \mu \leq 63.62\ \} = 0.95$

Confidence interval

- The interval 60.22 - 63.62 will contain μ in 95% of repeated samples
- We are 95% confident that the interval 60.22 - 63.62 contains μ
- Note that we **do not** conclude that we are 95% confident that μ is between 60.22 - 63.62, since μ is a fixed, albeit unknown value.

Aim of interval estimation

- Ideally, to obtain a narrow interval (range) for a given level of confidence (e.g. 95%), given s and n.
- From sample (n = 15) with mean of 61.92 and s of 5.24:
- *t* for 95% CI with 38df is 2.02
- 95% CI is 60.22 to 63.62 (as calculated on previous slides), i.e. the interval range is **3.40**

Effect of changing s on CI

- Sample (n=39) with a mean of 61.92 and s of 10.48 (sample data twice as variable):
  - t for 95% CI with 38 df is 2.02
  - 95% CI is 58.53 to 65.31, i.e. the interval range is **6.78** (*c.f.* previous 3.40)

Conclusion: more variability in population (and sample) results in wider CIs – we are 95% confident that a wider interval contains μ. Our sample mean is not as good an estimate of the population mean.

Effect of changing n on CI

- Sample (n=20, i.e. approx half the sample size) with a mean of 61.92 and s of 5.24:
  - t for 95% CI with 19df is 2.09
  - 95% CI is 59.47 to 64.37, i.e. the interval range is **4.90** (*c.f.* original 3.40)

Conclusion: decreasing sample size results in wider CIs – we are 95% confident that a wider interval contains μ. Our sample mean is not as good an estimate of the population mean.

MONASH UNIVERSITY

Effect of level of CI (e.g. 99% vs 95%)

- Sample (n = 39) with a mean of 61.92 and s of 5.24:
  - t for 99% CI with 38df is 2.71
  - 99% CI is 59.65 to 64.20, i.e. the interval range is **4.55** (*c.f.* original 3.40)

Conclusion: requiring greater levels of confidence results in wider interval ranges for given n and s. Increasing levels of confidence require more conservative interval ranges about the sample mean.

Estimating other parameters

- The logic of interval estimation of the population mean using a *t*-distribution can also be applied to other population parameters providing:
  - the exact formula for calculating the standard deviation of the statistic (i.e. standard error) is known
  - the sampling distribution of the statistic divided by its standard error follows a *t*-distribution
- For some parameters, different distributions must be used, e.g. the chi-square distribution for the variance (QK….). For other parameters (e.g. median), there is

no appropriate distribution and resampling methods must be used (QK…)






More general estimation

References

- Lovett, G. M., Weathers, K. C., and Sobczak, W. V. (2000). Nitrogen saturation and retention in forested watersheds of the Catskill Mountains, New York. *Ecological Applications*. **10**:73–84.

Statistical population vs biological population

A **biological population** refers to all the individuals of a particular species and may be further refined by spatial, temporal or other factors. For example, the male, Victorian koala population.

A **statistical population** refers to all the possible observations that could make up a sample and therefore defines the limits of statistical conclusions. For example, Lovett et al. (2000) measured the concentration of SO_4^{2-} from 39 forested streams in Northern America (qk2000, Box 2.2). The statistical population might have been all possible forested streams in Northern America during spring, etc...

Click anywhere to close

Population parameters vs sample statistics

A **population** refers to all the possible observations. Therefore, population parameters refer to the characteristics of the whole population. For example, the population mean.

A **sample** represents a collected subset of the population's observations and is used to represent the entire population. Sample statistics are the characteristics of the sample (e.g. sample mean) and are used to estimate population parameters.

Click anywhere to close

Observations vs variables

Observations: are the actual data (measurements) collected.

Variables: are the actual properties (length, width, count...) that are measured by the individual observations.

Click anywhere to close

Precision vs accuracy

Accuracy: refers to how close an estimate is to the actual value. For example how close an estimate of the population mean is to the actual population mean. Since it is usually impossible to determine the actual population mean (hence the need for statistics), accuracy cannot usually be measured.

Precision: refers to how repeatable an observation or statistic is. For example, how similar are the sample means from repeated samples. If a sample statistic is precise, it suggests that it is a good estimate of the population parameter. However, it is possible that the estimate of the population mean is consistantly offset from the actual population mean.

Click anywhere to close

Degrees of freedom (df)

The degrees of freedom is the number of observations in our sample that are 'free to vary' (not fixed) when we are estimating the variance.

To calculate the variance of a sample, the mean of a sample must be known, and therefore, only n-1 observations are free to vary.

If we know the mean and the value of n-1 observations, then the value of the 1 remaining observation must be known (fixed) and is thus not free to vary. (qk2002, Box 2.1)

Click anywhere to close