STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 13: Introduction to Confidence Intervals

Outline

- Previous:
  - Introduction to (normal) statistical model
  - Sampling distributions
    - Describe variation in the sample mean $\bar{y}$ (or any other statistic) from one sample to another
    - Relies on us knowing σ
- Today:
  - Use that to find confidence interval
    - Interval estimate for the parameter value
  - Look at what happens when σ is unknown

Example

- Continue using the GAG concentration data
  - Data from urine tests of $n = 314$ children (aged 0 – 17 years)
  - (log) concentration of glycosaminoglycan (GAG)
- Asking: what is the expected (or mean) GAG concentration?

Sampling distribution

- Recall we have a normal model for the data
  - Data come from a normal distribution with mean μ and standard deviation σ
- Last lecture we found the sampling distribution for $\bar{y}$
  - Distribution that describes how $\bar{y}$ will vary from one sample to another
  - Sampling distribution is normally distributed (for a normal model)
    - Mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Cool result!

- We know about what will happen in repeated samples
  - Without having to take repeated samples!
- If we know the data distribution (i.e. we know μ and σ):
  - We know how variable we expect $\bar{y}$ to be without even sampling from the population
- If we know σ (but don't know μ):
  - Can we use a single sample to tell us about a range of plausible values of μ?

Cool result!

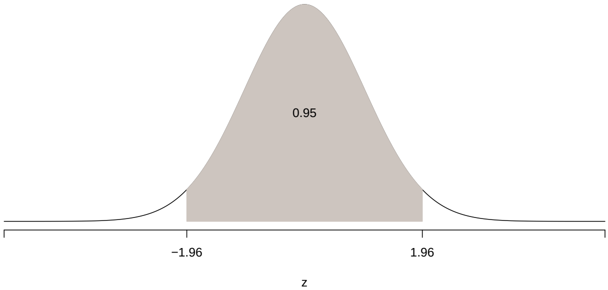
- We know about what will happen in repeated samples
  - Without having to take repeated samples!
- If we know the data distribution (i.e. we know μ and σ):
  - We know how variable we expect $\bar{y}$ to be without even sampling from the population
- If we know σ (but don't know μ):
  - Can we use a single sample to tell us about a range of plausible values of μ?
- Yes!

Excursion: standard error

- Over the past few lectures, we have seen:
  - Population standard deviation σ
  - Sample standard deviation s
  - Standard deviation of sampling distribution of $\bar{y}$
    - It is $\frac{\sigma}{\sqrt{n}}$
    - Has a special name: standard error
    - Can be represented with notation $\sigma_{\bar{y}}$
  - Estimate of the standard deviation of the sampling distribution of $\bar{y}$
    - It is $\frac{s}{\sqrt{n}}$
    - It is often also called the standard error
    - Can be represented with notation $s_{\bar{y}}$

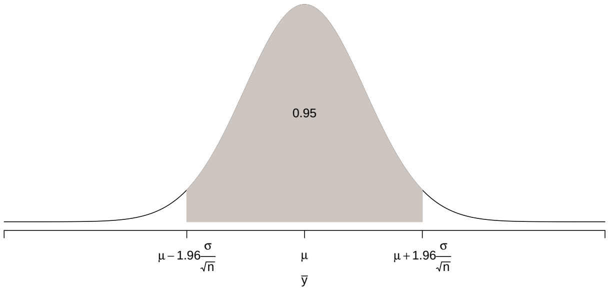
Previous knowledge

- Want to determine an interval estimate of μ from $\bar{y}$
- From our knowledge of normal distribution:
  - 95% of observations will fall within (approx) ± 2 standard deviations of mean
    - More precisely it is ± 1.96
    - In R: `qnorm(0.025)` and `qnorm(0.975)`
  - $\Pr(-1.96 < Z < 1.96) = 0.95$



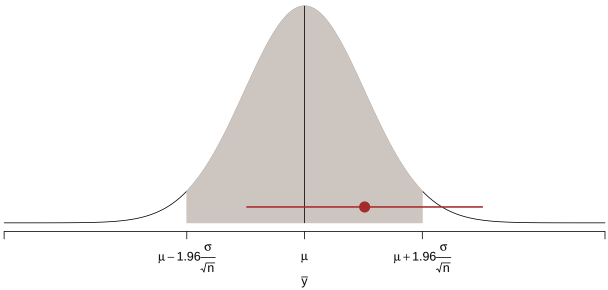
Sampling distribution

- Applying this to the sampling distribution we have:
  - 95% of sample means ($\bar{y}$) are between ± 1.96 standard errors ($\frac{\sigma}{\sqrt{n}}$) of the mean
- 95% of samples we collect will have sample means in the grey area
  - Given by $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$



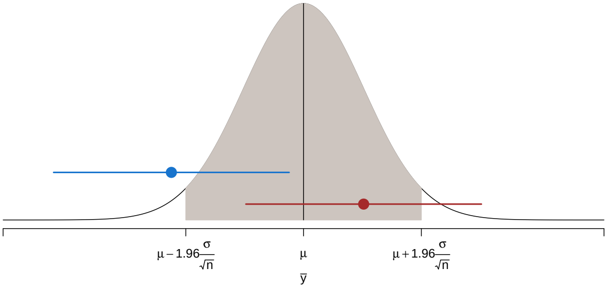
Flipping things I

- Consider any sample mean that is **inside** the shaded grey area
  - We've plotted one in brown on plot below
- Here's the magic:
  - If $\bar{y}$ is inside the grey area $(\mu \pm 1.96\frac{\sigma}{\sqrt{n}})$ (brown point)
  - Then μ (vertical black line) is inside the interval $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ (brown interval)



Flipping things II

- Consider any sample mean that is **outside** the shaded grey area
  - We've plotted one in blue on plot below
- Here's the magic:
  - If $\bar{y}$ is outside the grey area $(\mu \pm 1.96\frac{\sigma}{\sqrt{n}})$ (blue point)
  - Then μ (vertical black line) is outside the interval $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ (blue interval)



Confidence interval

$$\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- This is a 95% confidence interval for μ
  - Interval estimate of μ
  - Quantifies how precise the estimate of μ is
- On average, 95% of sample means will lie in shaded grey area (established above)
  - That means that our confidence interval should contain the true μ in 95% of samples
  - Gives us confidence in the procedure (hence the name)
    - Care is needed: we cannot say that there is a probability of 0.95 that μ is in the interval

A few notes on confidence intervals

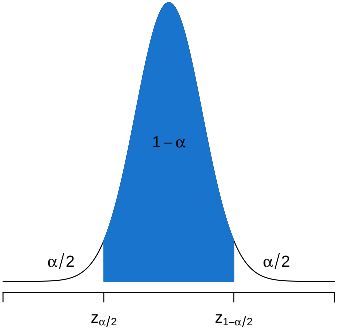
- The confidence interval is in a general form:

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

- estimate: $\bar{y}$
- multiplier:
  - 1.96 for 95% confidence interval
  - More generally, we write $z_{1-\alpha/2}$
    - More details on next slide
- Standard error: $\frac{\sigma}{\sqrt{n}}$

Multiplier

- Multiplier: $z_{1-\alpha/2}$
  - Also referred to as the critical value
- α: significance level
  - significance level $=$ 1 - confidence level
    - 95% interval: $\alpha = 1 - 0.95 = 0.05$
    - 90% interval: what is α?
- $\Pr(Z < z_{1-\alpha/2}) = 1 - \alpha/2$
  - Find z-value so that tails have probability $\alpha/2$



Multiplier

- For a 95% interval
  - $\alpha = 0.05$
  - $1 - \alpha/2 = 0.975$
  - We want to find $z_{0.975}$

```r
qnorm(0.975)

## [1] 1.96
```

- How do we find the multiplier for a 90% interval?

Multiplier

- For a 95% interval
  - $\alpha = 0.05$
  - $1 - \alpha/2 = 0.975$
  - We want to find $z_{0.975}$

```r
qnorm(0.975)

## [1] 1.96
```

- How do we find the multiplier for a 90% interval?
  - $\alpha = 0.10$
  - $1 - \alpha/2 = 0.95$
  - We want to find $z_{0.95}$

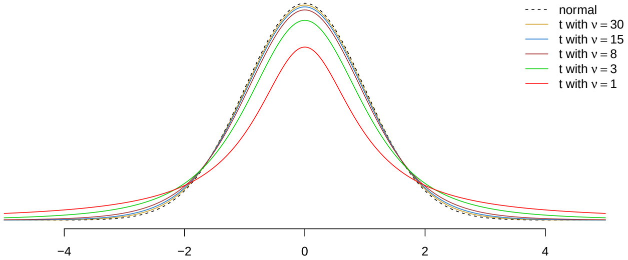
```r
qnorm(0.95)

## [1] 1.645
```

GAG concentrations

- Let's find an interval estimate for mean GAG concentration!
- We can't... we don't know σ
  - Population standard deviation
- Can we just replace σ with s?
  - No, the sampling distribution is no longer normal
    - All is not lost: most of the reasoning we worked through remains the same
- Replacing σ by s introduces additional noise (variability)
  - Sampling distribution no longer normally distributed
  - We need to use a t-distribution instead

t-distribution

- A t-distribution looks a lot like a (standard) normal distribution
  - Has fatter tails
- Additional parameter $\nu > 0$, called the degrees of freedom
  - This defines how fat the tails are



Historical excursion: William Gosset (1876 – 1937)

- Head Brewer of Guinness who 'discovered' the t-distribution
- Running experiments on yield of barley varieties and did not have statistical tools he needed to analyze the data
  - Statistical methodology developed due to applications in food science, agriculture
- The t-distribution is commonly known as Student's t-distribution
  - Gosset published under the pseudonym 'Student'
  - Guinness allowed its scientists to publish research if they did not mention:
    - Beer
    - Guinness
    - Their own surname

Confidence interval: unkonwn σ

- Replacing σ by s leads to the confidence interval

$$\bar{y} \pm t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}}$$

- $t_{\nu,1-\alpha/2}$: multiplier for the t-distribution
  - Significance level α
  - Degrees of freedom ν
- When finding confidence interval for μ
  - Degrees of freedom $\nu = n - 1$
- Find multiplier in R: for 95% interval when $n = 30$

```
n = 30
qt(0.975, df = n-1)

## [1] 2.045
```

GAG concentrations

- We are now ready to find an interval estimate for mean GAG concentration
- We need to get a few bits and pieces together:
  - Call in the data:

```
GAG = read.csv('GAG.csv')
```

  - Find the sample mean: $\bar{y}$

```
ybar = mean(GAG$conc)
ybar
## [1] 2.364
```

  - Find the sample standard deviation: s

```
s = sd(GAG$conc)
s
## [1] 0.6682
```

GAG concentrations

- Find the sample size: n

```
n = length(GAG$conc) # length() tells us the number of values
n

## [1] 314
```

- Find the standard error: $s_{\bar{y}} = \frac{s}{\sqrt{n}}$

```
se = s/sqrt(n)
se

## [1] 0.03771
```

- Find the multiplier: 95% confidence interval

```
alpha = 0.05
tcrit = qt(1-alpha/2, df = n-1)
tcrit

## [1] 1.968
```

GAG concentrations

- Put it all together

```
lower = ybar - tcrit * se # lower confidence limit
upper = ybar + tcrit * se # upper confidence limit
ci = c(lower, upper)
ci

## [1] 2.290 2.439
```

- The 95% confidence interval for μ is (2.29, 2.44)
  - Interval estimate for μ

- Spend some time interpreting the interval in the next lecture

Summary

- Found confidence interval for μ
  - Interval that quantifies how precise our estimate of μ is
- Found confidence interval if σ is known
  - Useful for understanding
  - Not practically useful
- Found confidence interval if σ is unknown
  - Introduced the t-distribution
- Looking forward:
  - More about confidence intervals