STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 14: Understanding Confidence Intervals

Outline

- Previous lecture:
  - Confidence interval for population mean μ
- Today: understand more about the confidence interval
  - How to find the confidence interval
  - How to interpret the confidence interval
  - Understanding the properties of the confidence interval
  - How large of a sample do we need?

Data: GAG concentration

- Call in the data

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

- Remember what the data set looks like:

```
head(GAG)

##    age  conc
## 1 0.00 3.135
## 2 0.00 3.170
## 3 0.00 2.827
## 4 0.00 2.923
## 5 0.01 2.885
## 6 0.01 3.254
```

Recall: GAG concentration

- Data from urine tests of $n = 314$ children (aged 0 – 17 years)
  - Interest in estimating the mean (log) concentration of glycosaminoglycan (GAG)
- In the last lecture we found a confidence interval
  - Quite an involved process
- Several steps

1. Call the data into R
2. Find the sample mean: $\bar{y}$
3. Find the sample standard deviation: s
4. Find the sample size: n
5. Find the standard error: $s_{\bar{y}} = \frac{s}{\sqrt{n}}$
6. Find the multiplier: $t_{\nu,1-\alpha/2}$
7. Find the confidence interval: $\bar{y} \pm t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}$

That's a lot of steps!

- That's not how we find a confidence interval in practice
  - R function that finds it for us: `t.test`
- So why did we go through those steps?
  - Important for our understanding of what a confidence interval is
    - We will be exploring 'properties' of confidence intervals that use this information
  - To use any tool well, it helps to know how it works
    - What its limitations are

Finding confidence interval: in practice

- We can find a confidence interval for μ with `t.test`

```
output = t.test(GAG$conc)
output
```

```
##
##  One Sample t-test
##
## data:  GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.290 2.439
## sample estimates:
## mean of x
##     2.364
```

Output of `t.test`

- We can understand some of the output
  - df = degrees of freedom for the multiplier
  - sample mean
  - 95% confidence interval
  - We will be learning about the other things soon
- We can isolate the confidence interval

```
output$conf.int

## [1] 2.290 2.439
## attr(,"conf.level")
## [1] 0.95
```

Using t.test

- The input to `t.test` is the full data set
  - No need to summarize data in terms of $\bar{y}$ and s
  - No need to find the multiplier

Changing the confidence level

- The function `t.test` has optional arguments
  - These are arguments that have some default, but we can choose to change them
  - One of these is `conf.level`
    - Defaults to 0.95 (95% confidence interval)
- For a 90% confidence interval:

```
output90 = t.test(GAG$conc, conf.level = 0.9)
output90$conf.int
```

```
## [1] 2.302 2.427
## attr(,"conf.level")
## [1] 0.9
```

- How would we find a 99% interval?

Diversion: R help

- How would you figure out that `conf.level` changes the confidence level?
- Many answers:
  - In this course: we will show you how to make changes like this
  - Outside this course: you can consult the R help
    - Surprisingly, not really the recommended first option
  - This is where chatGPT (or equivalent) can be really helpful
    - e.g. ask "how do I find a 90% confidence interval when using t.test in R?"
    - Not always 100% accurate, but it is pretty good
  - Google can also be very helpful

Interpreting the confidence interval

- What do we do with the confidence interval: (2.29, 2.44)?
  - We are 95% confident that mean GAG concentration is between 2.29 and 2.44
- What does 95% confident mean?
  - Recall the definition of a confidence interval
  - It does not guarantee that the true mean GAG concentration is inside the interval
    - Across many samples, the true mean should be in the interval 95% of the time
  - Confidence in the procedure: long-term performance

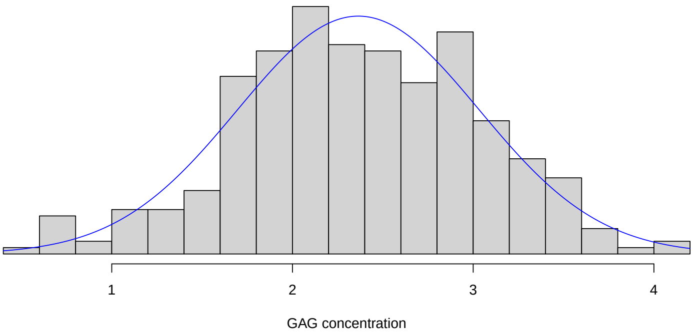
Interpreting the confidence interval

- What do we do with the confidence interval: (2.29, 2.44)?
  - We are 95% confident that mean GAG concentration is between 2.29 and 2.44
- This is a statement about the population parameter
  - Average GAG concentration for children aged 0-17
  - Population isn't well defined
    - Geographical area?
    - It isn't clear how the data were collected
    - Important factors in determining whether the confidence interval tells us anything useful
    - We will be talking more later in the course about the importance of data collection

Checking model assumptions

- Recall: it is important to check model assumptions
- We have assumed the data came from a normal distribution
- STAT115 approach: check visually
  - Histogram
  - Looking for major departures from normality
    - Obvious skew
    - Large outliers
- If the sample size is large enough
  - Confidence intervals for μ are suitable for non-normal data
  - $n > 30$ is rule of thumb often used
    - If there are major departures from normality, we may need a much larger n
  - Discuss more in a few weeks

Model fit: GAG



- No obvious departures from normality
  - Blue curve: normal density using the sample mean and sd

Width of the confidence interval

- The width of the confidence interval is important
  - Tells us how precise the estimate is
- The CI we found is (2.29, 2.44)
  - An example of a wider (less precise) interval: (2.22, 2.51)
  - An example of a narrower (more precise) interval: (2.34, 2.39)
- The width of a confidence interval is given by upper limit - lower limit
  - Width: $2.44 - 2.29 = 0.15$
- We often refer to the margin of error: half of the interval width
  - Recall our confidence interval formula:

$$\bar{y} \pm \underbrace{t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}}}_{\text{margin of error}}$$

Changing confidence level

- What happens to interval width if we increase the confidence level, say from 95% to 99%? Why?

Changing confidence level

- What happens to interval width if we increase the confidence level, say from 95% to 99%? Why?
  - The interval gets wider (margin of error gets larger)
    - Confidence level increases, α decreases
    - Multiplier $t_{\nu,1-\alpha/2}$ increases
    - Can be seen graphically
- This makes sense:
  - Making the interval wider: increasing the confidence that parameter (μ) is in interval
  - If we have a wider interval, the true mean will be in the interval a higher percentage of the time
- The opposite also holds:
  - If we decrease the confidence level: interval gets narrower

Changing confidence level: 95%

```
output95 = t.test(GAG$conc, conf.level = 0.95)
output95

##
##  One Sample t-test
##
## data:  GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.290 2.439
## sample estimates:
## mean of x
##     2.364
```

Changing confidence level: 99%

```
output99 = t.test(GAG$conc, conf.level = 0.99)
output99

##
##  One Sample t-test
##
## data:  GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  2.267 2.462
## sample estimates:
## mean of x
##     2.364
```

Changing confidence level: 90%

```
output90 = t.test(GAG$conc, conf.level = 0.90)
output90

##
##  One Sample t-test
##
## data:  GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  2.302 2.427
## sample estimates:
## mean of x
##     2.364
```

Standard error

- The standard error is a critical part of the calculation of a confidence interval:

  $s_{\bar{y}} = \frac{s}{\sqrt{n}}$
- Recall: tells us how variable the statistic $\bar{y}$ is
  - Quantifies how much we expect $\bar{y}$ to vary
    - If we took multiple samples of size n from the population
- It has two components
  1. s: sample standard deviation
     - The larger the variation in the data, the larger the standard error
     - The larger the variation in the data, the wider the confidence interval for μ
  2. n: sample size
     - The larger the sample size, the smaller the standard error
     - The larger the sample size, the narrower the confidence interval for μ

Caution

- The statements on the previous slide assume all else is held fixed
  - e.g. the larger the sample size, the narrower the confidence interval, all else held fixed
- In reality: if we took a different (larger) sample, things would not be held fixed
  - $\bar{y}$ varies from one sample to the next
  - s also varies from one sample to the next
  - On average: $\bar{y}$ from a larger sample will be closer to the true mean
- We cannot (and should not) use the $\bar{y}$ and s we observe and pretend we had a larger sample size to find a narrower confidence interval
  - Fabricating (or falsifying) data
  - Unethical
  - Scientific misconduct

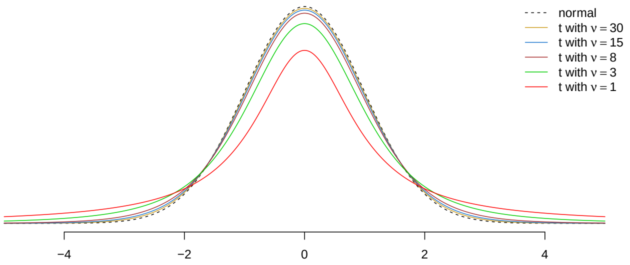
Sample size calculation

- The GAG data appear to be from the UK
- We may choose to replicate the study here in NZ
  - We want the study to be accurate: margin of error of 0.04
  - How large of a sample should we take?
- We want to find value n such that the margin of error is 0.04
- This is a common scenario when designing research studies
  - Too few samples: imprecise estimates of limited value
  - Too many samples: poor use of precious resources (time and money)

Sample size calculation

- This is an approximate process (we'll see why as we go)
- Recall: the margin of error is $t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}$
  - Find n so that the margin of error has a desired level of accuracy
- This is problematic for two reasons:

1. The multiplier $t_{\nu,1-\alpha/2}$ depends on n $(\nu = n - 1)$
   - Approximate it with $z_{1-\alpha/2}$



Sample size calculation

- We want to find n so that the margin of error has a desired level of accuracy
- This is problematic for two reasons:

  2. The standard deviation s is an estimate that will change from one sample to the next
     - Take s as our best estimate of σ

- To find n, we use an approximate margin of error $\approx z_{1-\alpha/2}\frac{s}{\sqrt{n}}$
- If the desired level of accuracy (in our case 0.04) is given by the symbol ξ, we want to find the value of n such that

$$z_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \xi$$

Sample size calculation

- We rearrange the formula to get:

$$n \geq \left(\frac{z_{1-\alpha/2}\, s}{\xi} \right)^2$$

- In our case

```
alpha = 0.05 # 95% confidence interval
z = qnorm(1-alpha/2) # approximate multiplier: normal distribution
s = sd(GAG$conc) # best guess as to the sigma
xi = 0.04 # desired margin of error
n = ceiling((z * s / xi)^2) # sample size; ceiling rounds up
n

## [1] 1073
```

Sample size calculation

- This is an approximate process
  - Approximated the multiplier
  - Used an estimate of standard deviation
- Always ‘round up’ (R command `ceiling` rounds up)
- We tend to be conservative
  - It’s better to have a few more observations than you need, than too few.
    - Often round up further, to say $n = 1100$ or $n = 1200$ participants, or
  - In practice, we often find a confidence interval for σ
    - Use the upper limit of the CI in the calculation (in place of s)
    - Outside the scope of STAT115

Summary

- Looked at more detail into calculation and use of confidence intervals
  - How to find them in R: `t.test`
    - Changing confidence level
  - Interpreting the confidence interval
  - Width and margin of error
  - Sample size calculation
  - Tomorrow: hypothesis testing