STAT 110: Week 5

University of Otago

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

```
lect4GAG = read.csv('lect4GAG.csv') # it doesn't matter that we are now in week 5
```

- Remember what the data set looks like:

```
head(lect4GAG)
```

```
##    age conc
## 1 0.00 3.14
## 2 0.00 3.17
## 3 0.00 2.83
## 4 0.00 2.92
## 5 0.01 2.88
## 6 0.01 3.25
```

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(lect4GAG$conc)
output
```

```
##
##  One Sample t-test
##
## data:  lect4GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.29 2.44
## sample estimates:
## mean of x
##      2.36
```

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] 2.29 2.44
## 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(lect4GAG$conc, conf.level = 0.9)
output90$conf.int
## [1] 2.30 2.43
## 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

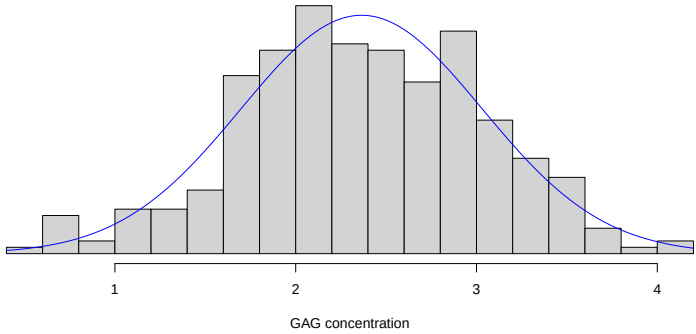
Interpreting the confidence interval

- 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
- STAT 110 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(lect4GAG$conc, conf.level = 0.95)
output95

##
##  One Sample t-test
##
## data:  lect4GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.29 2.44
## sample estimates:
## mean of x
##      2.36
```

Changing confidence level: 99%

```
output99 = t.test(lect4GAG$conc, conf.level = 0.99)
output99

##
##  One Sample t-test
##
## data:  lect4GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  2.27 2.46
## sample estimates:
## mean of x
##      2.36
```

Changing confidence level: 90%

```
output90 = t.test(lect4GAG$conc, conf.level = 0.90)
output90

##
##  One Sample t-test
##
## data:  lect4GAG$conc
## t = 63, df = 313, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  2.30 2.43
## sample estimates:
## mean of x
##      2.36
```

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

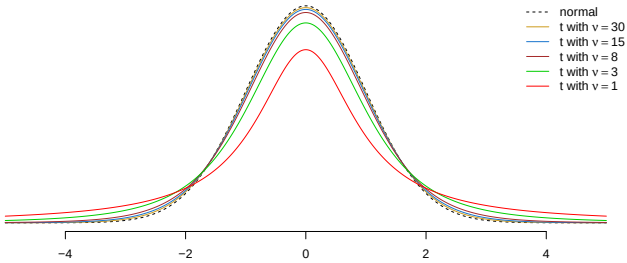
- The statements on the previous slide assume all else is held fixed
  - e.g. the larger the sample size, the narrower the confidence interval, all else held fixed
- In reality: if we took a different (larger) sample, things would not be held fixed
  - $\bar{y}$ varies from one sample to the next
  - 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(lect4GAG$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 STAT 110

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

Outline

- Previous:
  - Learned how to find and interpret confidence intervals
  - Interval estimates of parameter
- Today:
  - Look at hypothesis testing

Hypotheses

- Data are often collected to test a hypothesis
  - e.g. sleep deprivation affects reaction time
  - e.g. survival rates of kākāpō are higher today than they were 10 years ago
  - e.g. calorie values listed on labels of chip packets are not accurate
- Collect data to investigate our hypothesis

Example 1: Shoshone Rectangles

- Shoshone Native Americans used beaded rectangles to decorate their goods

- Native American tribe that originated in the western Great Basin and spread north and east into present-day Idaho and Wyoming
- Anthropologists are interested to know whether there is evidence against the claim that Shoshone Native Americans produced rectangles which conform to the golden ratio
  - What is the golden ratio?

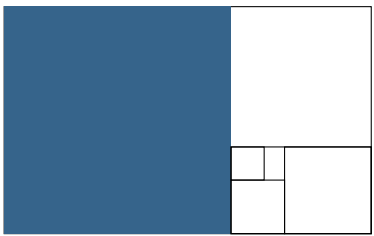
Golden ratio

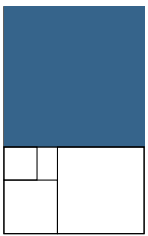
- The golden ratio is a number that appears frequently in geometry
  - First studied by the Greeks
    - Euclid called it the 'extreme and mean ratio'
- A rectangle is 'golden' if the ratio of its long to short side is $\frac{1+\sqrt{5}}{2} \approx 1.618$.

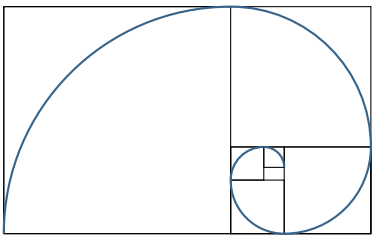
$$\frac{A}{B} \approx 1.618$$

Golden ratio

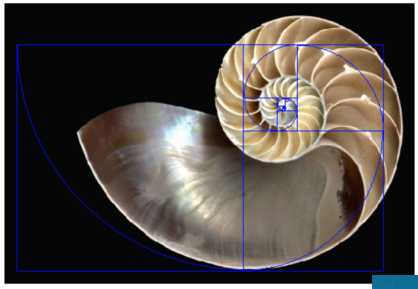
- Rectangles with the golden ratio have some nice mathematical properties
  - e.g. if you take away a square (in blue) you get another golden rectangle
  - Related to a Fibonacci sequence

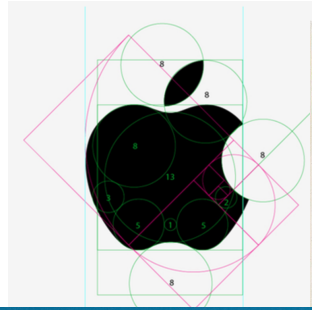


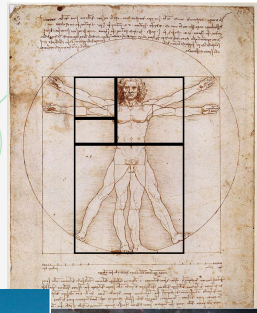


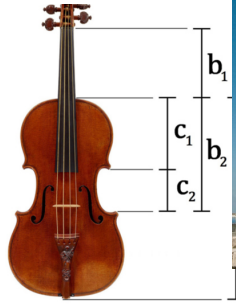


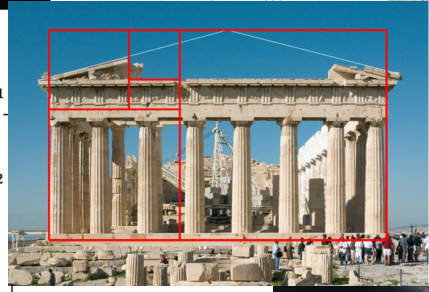
- The golden ratio is apparent in art, architecture and nature




8
8
8
8
13
8
3
5
1
5
8




b1
c1
b2
c2



Example: Shoshone rectangles

- Data of the length-to-width ratios for 20 Shoshone rectangles

```
shoshone = read.csv("shoshone.csv")
```

```
shoshone$ratio
```

```
##  [1] 1.44 1.51 1.45 1.65 1.75 1.33 1.49 1.59 1.64 1.19 1.53 1.63 1.50 1.66 1.74
## [16] 1.49 1.65 1.64 1.81 1.07
```

- How can we investigate how compatible the data are with the golden ratio?

Set up hypotheses

- Two hypotheses: null hypothesis and alternate hypothesis
  - Null: we compare the data to what we expect under the null hypothesis
    - Assess the compatibility of the data to the null hypothesis
    - Often the claim to be tested, the status quo, or assumption of no difference
    - The hypothesis we find evidence against
  - Alternate: alternate claim under consideration
    - Alternate 'state of the world'
    - Hypothesis we want to find evidence in support of

Set up hypotheses: examples I

- Quality control: manufacturing cell phone case
  - To specifications: say mean length $\mu = 6$ inches
  - Collect data to ensure quality
  - $H_0 : \mu = 6$ (status quo)
  - $H_A : \mu \neq 6$
- Collect data from group with specific disease
  - Interested in expression of particular gene
  - Know the expression in the population is 10 TPM
  - $H_0 : \mu = 10$ (claim to be tested)
  - $H_A : \mu \neq 10$

Set up hypotheses: examples II

- Collect data to find evidence that the pH may differ from neutral in some environment
  - $H_0 : \mu = 7$
  - $H_A : \mu \neq 7$
- Collect data on recovery time of a new surgery (for a particular condition)
  - The recovery time for the current surgery is known to average 10 days
  - $H_0 : \mu = 10$
  - $H_A : \mu \neq 10$

Set up hypotheses

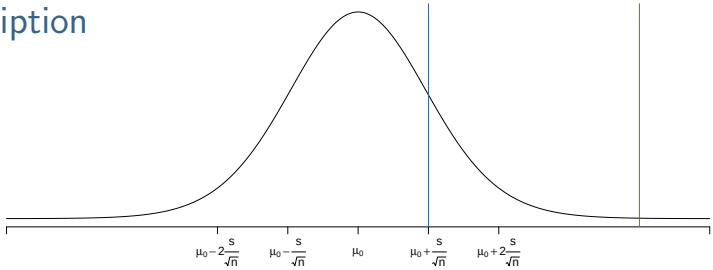
- Hypotheses are statements about parameter values (in our case μ)
- For the Shoshone rectangles we would have:
  - H_0: $\mu = 1.618$ (the golden ratio)
  - H_A: $\mu \neq 1.618$
- The null hypothesis is say that the true mean ratio is the golden ratio
  - Often refer to this as μ_0
  - An individual garment might have a ratio larger or smaller than the golden ratio
  - Mean value (in the population) is given by the golden ratio
- The alternative hypothesis[1] says that the true mean ratio is some other value

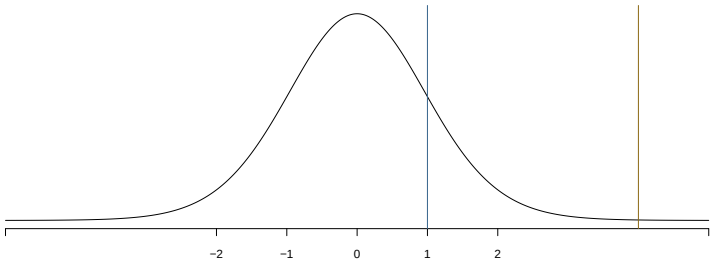
[1] The alternative hypothesis is sometimes referred to as H_1

What if?

- Now we play a 'what if' game:
  - How extreme is the data we observed if the null hypothesis were true?
- Very similar to questions we asked when looking at sampling distributions
  - Only difference: accounting for not knowing σ
- We calculate a test statistic to help us answer the question
  - How many standard errors separate the sample mean from null value ($\mu = 1.618$)
    - The standard error is a measure of how variable the sample mean is
    - If the sample mean is 4 standard errors from the null value: unusual
    - If the sample mean is 1 standard error from the null value: not unusual

Graphical description





Test statistic

- Finding how many standard errors separate the sample mean from null value

$$T = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Find the relevant quantities for the Shoshone example

```
mu0 = 1.618 # null value
ybar = mean(shoshone$ratio)
ybar # sample mean
## [1] 1.54
n = length(shoshone$ratio) # number of samples (20)
se = sd(shoshone$ratio)/sqrt(n) # standard error
se
## [1] 0.041
```

Test statistic

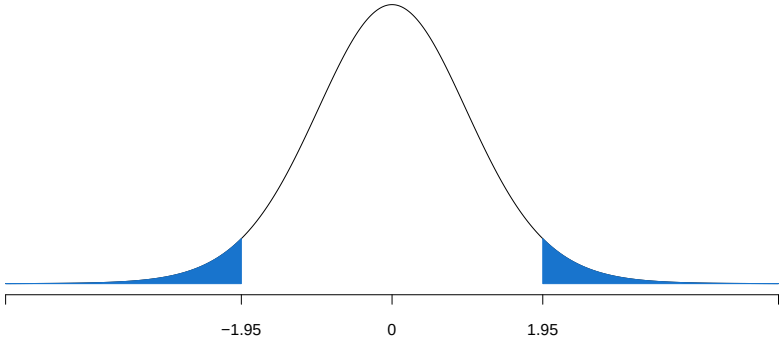
- Find the test statistic

```
Tstat = (ybar - mu0)/se # test statistic
Tstat
```

```
## [1] -1.95
```

- The sample mean is 1.951 standard errors below the null value
- Is that consistent with the null hypothesis?
  - Compare it to a t-distribution with $n - 1$ degrees of freedom

Test statistic



p-value

- The tail areas (on the previous slide) give the p-value
  - Find that with `pt` function in R
    - Remember we need to find both tails (or find one and double it)

```
# Find the lower tail: here we have a negative value
tail_lower = pt(Tstat, df = n-1)
# in general, we would use -abs(Tstat) to ensure it is the lower tail
pval = 2*tail_lower
pval
## [1] 0.0659
```

In R: using `t.test`

- `t.test` assumes the null value is 0 ($\mu_0 = 0$): change with `mu` input

```
out = t.test(shoshone,mu = 1.618)
out
```

```
##
##  One Sample t-test
##
## data:  shoshone
## t = -1.95, df = 19, p-value = 0.066
## alternative hypothesis: true mean is not equal to 1.618
## 95 percent confidence interval:
##  1.4523 1.6238
## sample estimates:
## mean of x
##     1.538
```

In R: using `t.test`

- The R output has all of the features we have discussed today:
  - Test statistic
  - Degrees of freedom
  - p-value
  - Alternative hypothesis (null hypothesis is implicit)

Interpretation

- You would think it should be easy to use and interpret hypothesis tests
  - It is not
- Hypothesis tests are one of the most heavily used statistical 'concepts'
  - Most articles in the (applied science) literature use hypothesis testing in some way
- They are probably the most abused, misunderstood, and misinterpreted concept
  - Controversial: one psychology journal has banned the use of p-values
  - American Statistical Association has published articles on their use
  - We will try to offer a balanced view
    - Further discuss many of the issues later in the semester

What is a p-value?

- The p-value is the probability of observing data as or more extreme than that observed given the null hypothesis is true
- It provides a measure of incompatibility with statistical model
  - Model given by null hypothesis
- The smaller the p-value, the greater the incompatibility between the data and the null hypothesis
  - Often expressed as evidence against the null hypothesis
- A p-value is **not**:
  - The probability the null hypothesis is true
  - The probability that random chance produced the observed data
    - Both of these 'flip' a conditional probability

Hypothesis testing in this course

- If the study / example was (likely) confirmatory
  - Collected data to confirm (or test) a specific hypothesis
  - We will use formal hypothesis testing
    - Compare p-value to α and make a decision
- If the study / example was (likely) exploratory
  - Collect data to try and explore and understand scientific phenomena
  - Use the data to generate hypotheses
  - We will use p-value to assess the incompatibility of the data to null hypothesis
    - Use α as a guide (more details on next slide)
    - Try not to make a decision between between competing hypotheses
  - Often prefer confidence intervals

Formal test

- If the object of the analysis was to test a particular hypothesis
  - Use p-value to help make a 'decision' between H_0 and H_A
- We have a threshold α (significance level) specified in advance
  - Often $\alpha = 0.05$ (or 0.01, etc)
- If the p-value $< \alpha$: reject H_0
  - Evidence in support of H_A
  - Sometimes called 'statistically significant'
- If the p-value $> \alpha$: fail to reject H_0
  - Not enough evidence to reject H_0
    - Not the same as support (or evidence) for H_0
    - Absence of evidence is not evidence of absence
  - Sometimes referred to as 'not statistically significant'

Formal test

- It is very easy to abuse a formal hypothesis testing approach
  - e.g. one of the ASA principles: 'Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.'
- What often happens in practice:
  - Collect data with no clear hypotheses in mind
  - Explore every possible situation trying to find p-value $< \alpha$
  - 'Torture the data until it confesses'
- Return to a discussion about the use of p-values later in the course

Interpretation of exploratory model

- Use α as a guide for incompatibility between the data and null hypothesis
  - If p-value $< \alpha$
    - There is evidence of incompatibility between the data and null hypothesis (relative to α)
    - Investigate further: e.g. look at designing confirmatory study
  - If we obtain a p-value $> \alpha$
    - There is no evidence of incompatibility between the data and null hypothesis (relative to α)
    - The degree of incompatibility between the data and null hypothesis (as quantified by the p-value) is similar to what we would expect if the data came from a model under H_0
    - This is not evidence in support of H_0
- Assessing the incompatibility of the data and the null hypothesis
  - Not making a decision about which hypothesis to adopt based solely on the p-value

Shoshone rectangles

- The Shoshone rectangles: specific hypothesis in mind
  - Confirmatory: use formal hypothesis testing
- Significance value is $\alpha = 0.05$
- p-value is 0.06593
  - No obvious incompatibility with the null hypothesis
  - Formal statement: no evidence to reject H_0
- Recall: $H_0 : \mu = 1.618$
  - There is no evidence that rectangles used by Shoshone do not follow golden ratio

Summary

- Introduced hypothesis testing
- Two hypothesis:
  - Null hypothesis
  - Alternative hypothesis
- Introduced p-value: measure of incompatibility between data and null hypothesis
- Formal hypothesis test
  - Confirmatory study
- Care is needed in interpretation

Outline

- Previous:
  - Confidence interval for μ
  - Hypothesis test
- Today:
  - Explore more of the properties around the hypothesis test
  - Type I and Type II Errors
  - Power of a Test
  - Trade-offs Between Errors and Power

Height of STAT 110 students

- In previous years there was a questionnaire (optional) for STAT 110 students
  - Questions about age, height, sex, ...
- Exploratory study
  - Explore the height of females in STAT 110 relative to national average
    - Average height for NZ female aged 15-24 is 164.7 cm (`figure.nz`)[2]
    - Restrict ourselves to female STAT 110 students aged 15-24

```
STAT110 = read.csv('./data/STAT110_height_f.csv')
head(STAT110$height)
```

```
## [1] 167 153 171 177 161 173
```

- Heights from $n = 451$ female students aged 15-24

[2] Data from New Zealand Health Survey, 2023

Hypothesis test

- Write down the null and alternate hypothesis
  - $H_0 : \mu = 164.7$
  - $H_A : \mu \neq 164.7$
- Use $\alpha = 0.05$
- We can conduct the test in R

```
h_test = t.test(STAT110$height, mu = 164.7)
h_test
##
##  One Sample t-test
##
## data:  STAT110$height
## t = 8.073, df = 450, p-value = 6.32e-15
## alternative hypothesis: true mean is not equal to 164.7
## 95 percent confidence interval:
##  166.891 168.301
## sample estimates:
## mean of x
##   167.596
```

Interpretation

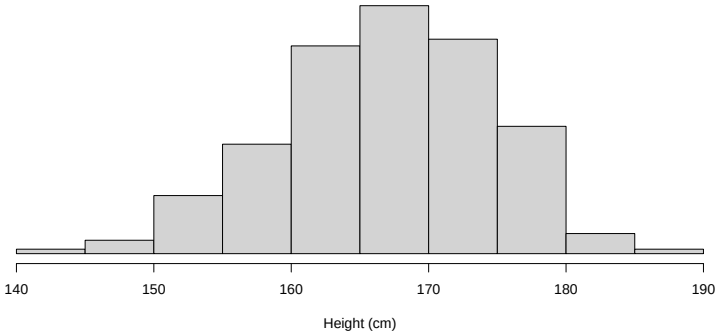
- Exploratory study: interpret p-values (no formal test)
- There is evidence that the data are incompatible with null hypothesis
  - p-value is approximately 1 in a quadrillion[3]
- Evidence that the (mean) height of female STAT 110 students is incompatible with national average
- Do we trust it? It pays to be cautious
  - Students in STAT 110 are not a random selection of 15 – 24 year olds in NZ
  - STAT 110 data are voluntary and heights are self-reported
    - There are also very different rates of left-handedness from national averages
- If this question were of interest
  - There is 'enough' to look into designing a (confirmatory) study

[3] The progression is million (10^6), billion (10^9), trillion (10^{12}), quadrillion (10^{15}), ...

Assumptions

- We have made an assumption that our data are normally distributed
  - Just as we did with confidence intervals
- To check this assumption: looking for serious departures from normality
  - We check visually (histogram)
- As with confidence intervals: if the sample size is large enough
  - p-values are reasonable for non-normal data
  - Discuss more in a few weeks

Histogram



- No obvious departures from normality
- Large sample (~ 450)

Setup

- We want to better understand how hypothesis testing works
- We do this in the context of formal hypothesis test
  - If p-value $< \alpha$ we reject H_0
  - If p-value $> \alpha$ we fail to reject H_0
- There are four possibilities:

	Decision	
	Do not reject H_0	Reject H_0
H_0 true	✓	Type I error
H_0 not true	Type II error	✓

Setup

- Consider a specific gene: GENE-X
  - Reference expression value of 5.0 TPM (transcripts per million) in healthy individuals
- Design a confirmatory study to test if GENE-X is expressed differently in a sample of people with a specific disease
  - $H_0 : \mu = 5$ (the mean expression for the diseased group is the same as the reference)
  - $H_A : \mu \neq 5$
- In this study:
  - We want to find evidence against the null
  - We want to find evidence that gene expression differs in the diseased group
- In the rest of the lecture an effect is defined as:
  - Effect: difference between the mean for the disease group and $\mu_0 = 5$

A tale of two errors

- Type I Error (α): Rejecting H_0 when it is true.
  - Concluding the expression of GENE-X is different for the diseased group, when it isn't
- Type II Error (β): Failing to reject H_0 when H_a is true.
  - Concluding that there is no evidence that expression of GENE-X differs for diseased group, when there is a non-zero effect

Type I error

- Type I error rate is given by α, the significance level
  - Decreasing α from 0.05 to 0.01 will reduce the number of type I errors we make
    - Recall: α is the threshold for incompatibility with null
    - A lower α is applying a higher threshold for incompatibility

Type II error

- The type II error rate is representated as β
- We often refer to the power $= 1 - \beta$
- Power: the probability of rejecting the null hypothesis, given it is incorrect
  - i.e. it is the probability of detecting an effect, given there is one
- All else equal, we want a powerful test
  - More likely to correctly reject H_0
  - More likely to correctly conclude that gene expression differs in diseased group
- We will look at four factors that change the type II error / power

Type I error rate α

- Trade off between type I error rate and power
  - If we decrease α (lower type I error rate)
    - Increase type II error rate β
    - Decrease power
- If we increase α (higher type I error rate)
  - Decrease type II error rate β
  - Increase power

Effect size

- Recall: $\mu_0 = 5$ TPM (transcripts per million)
- Consider two scenarios:
  1. The true mean of the diseased population is $\mu_A = 5.1$ TPM
  2. The true mean of the diseased population is $\mu_A = 12$ TPM
- In which scenario will power be higher (all else equal)?

[4] $|x|$ is the absolute value of x

Effect size

- Recall: $\mu_0 = 5$ TPM (transcripts per million)
- Consider two scenarios:
  1. The true mean of the diseased population is $\mu_A = 5.1$ TPM
  2. The true mean of the diseased population is $\mu_A = 12$ TPM
- In which scenario will power be higher (all else equal)?
- The larger[4] the effect $|\mu_A - \mu_0|$
  - The more powerful the test, all else equal
- The size of the effect is not something we can typically control

[4] $|x|$ is the absolute value of x

Sample size

- For a fixed α and effect size, consider these two scenarios:
  1. The sample size (of diseased participants) is $n = 20$
  2. The sample size (of diseased participants) is $n = 200$
- In which scenario will power be higher?

Sample size

- For a fixed α and effect size, consider these two scenarios:
  1. The sample size (of diseased participants) is $n = 20$
  2. The sample size (of diseased participants) is $n = 200$
- In which scenario will power be higher?
- The larger the sample size
  - The more powerful the test, all else equal
- Scientific research (grant) funding in ecology, food science, global health, etc
  - Typically have to justify your research design
  - Power calculation: determining sample size needed to achieve a certain power

Population standard deviation

- For a fixed n, α, and effect size, consider these two scenarios:
  1. The population standard deviation (of gene expression in the disease group) is $\sigma = 0.1$
  2. The population standard deviation (of gene expression in the disease group) is $\sigma = 1$
- In which scenario will power be higher?

Population standard deviation

- For a fixed n, α, and effect size, consider these two scenarios:
  1. The population standard deviation (of gene expression in the disease group) is $\sigma = 0.1$
  2. The population standard deviation (of gene expression in the disease group) is $\sigma = 1$
- In which scenario will power be higher?
- The smaller the population standard deviation
  - The smaller the standard error
  - The more precise $\bar{y}$ is
  - The more powerful the test, all else equal
- The value of σ is not something we can typically control

p-value

- ASA principle: "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result"
- Suppose we have $p = 0.0000001$. This could be because:
  - This could be because the effect size is large
  - It could occur when the effect size is small (but non-zero) and sample size is large
- Care is needed that we don't confuse a small p-value, with an important result

Relationship with confidence intervals

- If we are testing the hypothesis:
  - $H_0 : \mu = \mu_0$
  - $H_A : \mu \neq \mu_0$
- There is an equivalence between p-value and confidence interval
  - p-value $< \alpha$ is equivalent to μ_0 outside the $(1-\alpha)100\%$ confidence interval
    - e.g. if p-value < 0.05, then μ_0 is outside 95% confidence interval
    - e.g. if p-value > 0.01, then μ_0 is inside 99% confidence interval

Quiz

- It's quiz time!
- Three possible answers for the questions below:
  - (1) increase; (2) decrease; (3) can't tell
- What is the effect on (i) type I error rate, and (ii) power if we:
  - Increase the sample size?
  - Decrease α?
  - Decrease the sample size and increased α?
  - Changed the research design so that the type II error rate β decreased?
  - Collected a sample twice the size for a different gene (GENE-Y) that has a smaller effect and larger σ?

Summary

- Checking assumptions
- Looked more at the properties of hypothesis testing
  - Type I error
  - Type II error
  - Power
- Looked at the effect of
  - Sample size
  - Effect size
  - α
  - σ