STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 15: Introduction to 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?

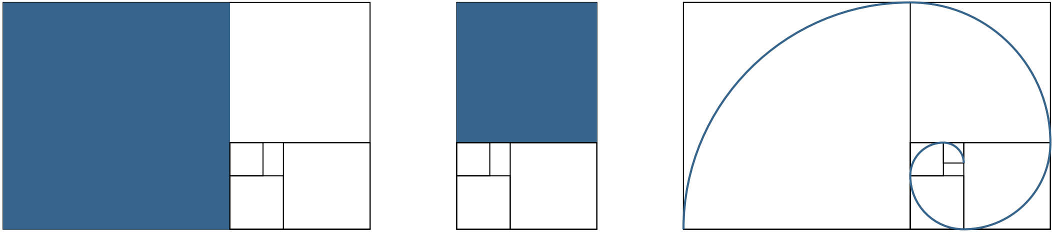
Golden ratio

- 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 $\frac{1+\sqrt{5}}{2} \approx 1.618$.

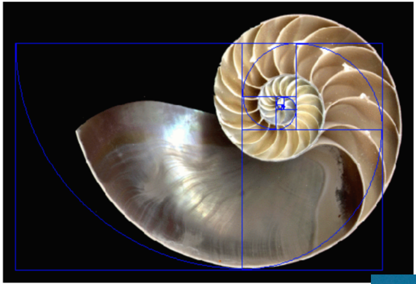
$$\frac{A}{B} \approx 1.618$$

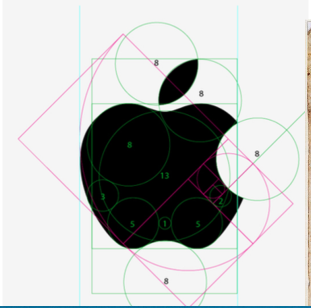
Golden ratio

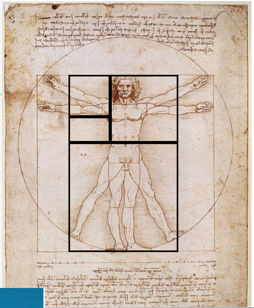
- Rectangles with the golden ratio have some nice mathematical properties
  - e.g. if you take away a square (in blue) you get another golden rectangle
  - Related to a Fibonacci sequence

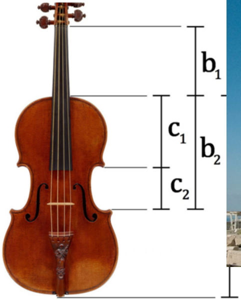


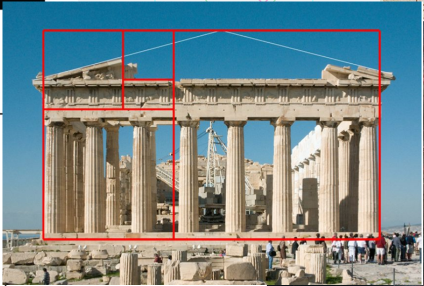
- The golden ratio is apparent in art, architecture and nature

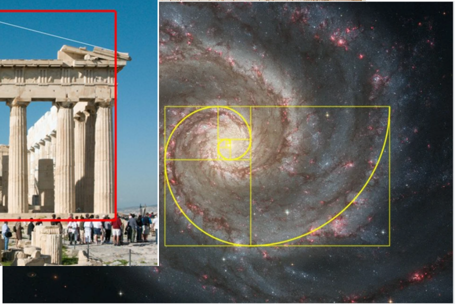







b_1
c_1
b_2
c_2





Example: Shoshone rectangles

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

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

```r
shoshone$ratio
```

```
##  [1] 1.443 1.511 1.449 1.650 1.754 1.335 1.488 1.592 1.642 1.185 1.529 1.626
## [13] 1.497 1.664 1.736 1.493 1.650 1.637 1.808 1.072
```

- 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
  - Alternative: other 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
  - $\mathrm{H}_0 : \mu = 6$ (status quo)
  - $\mathrm{H}_A : \mu \neq 6$
- Collect data from group with specific disease
  - Interested in expression of particular gene
  - Know the expression in the population is 10 TPM (transcripts per million)
  - $\mathrm{H}_0 : \mu = 10$ (claim to be tested)
  - $\mathrm{H}_A : \mu \neq 10$

Set up hypotheses: examples II

- Collect data to find evidence that the pH may differ from neutral in some environment
  - $\mathrm{H}_0 : \mu = 7$
  - $\mathrm{H}_A : \mu \neq 7$
- 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
  - $\mathrm{H}_0 : \mu = 10$
  - $\mathrm{H}_A : \mu \neq 10$

Set up hypotheses

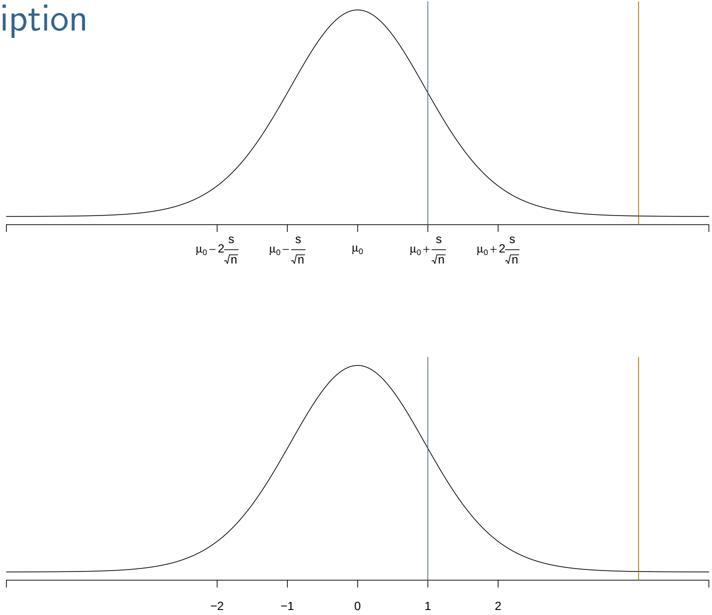
- Hypotheses are statements about parameter values (in our case μ)
- For the Shoshone rectangles we would have:
  - H_0: $\mu = 1.618$ (the golden ratio)
  - H_A: $\mu \neq 1.618$
- The null hypothesis is say that the true mean ratio is the golden ratio
  - Often refer to this as μ_0
  - An individual 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] 1.538

n = length(shoshone$ratio) # number of samples (20)
se = sd(shoshone$ratio)/sqrt(n) # standard error
se

## [1] 0.04097
```

Test statistic

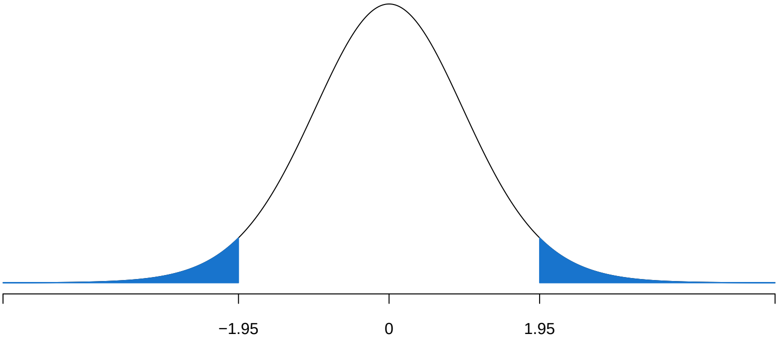
- Find the test statistic

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

```
## [1] -1.951
```

- The sample mean is 1.9513 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] 0.06593
```

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: $\mathrm{H}_0 : \mu = 1.618$
  - 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