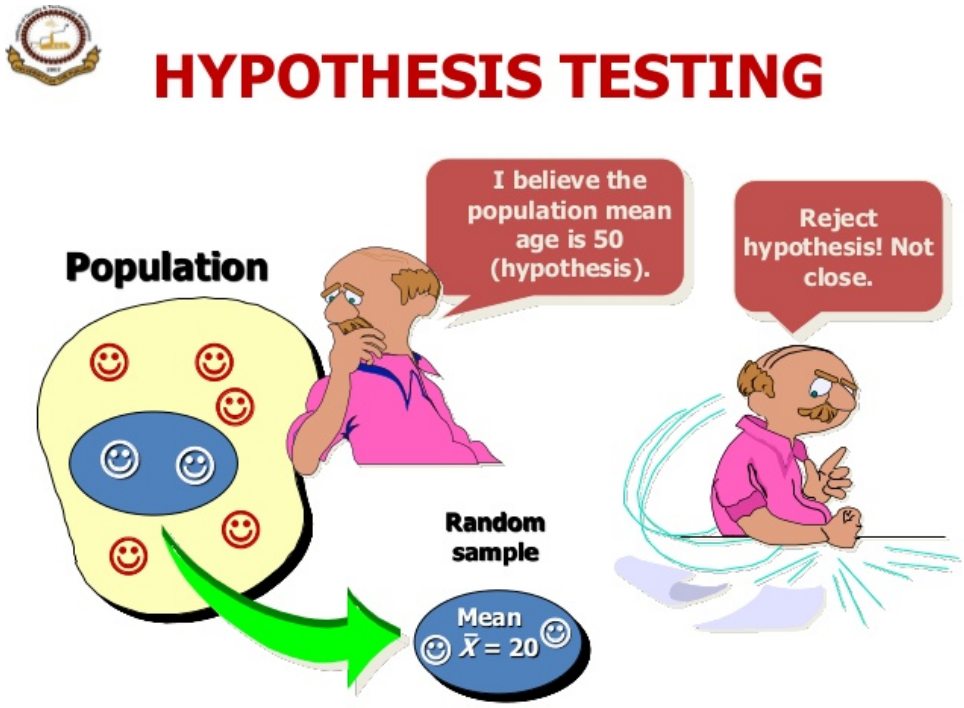
Practical: Statistics

BaselRBootcamp 2017



Source: https://www.slideshare.net/hakeemrehman/8-testing-of-hypothesis-for-variable-amp-attribute-data (https://www.slideshare.net/hakeemrehman/8-testing-of-hypothesis-for-variable-amp-attribute-data)

Slides

- Here are the introduction slides for this practical on statistics! (https://therbootcamp.github.io/_sessions/D2S2_Statistics/Statistics.html)

Overview

In this practical you'll conduct hypothesis tests. By the end of this practical you will know how to:

1. Calculate basic descriptive statistics.
2. Conduct a hypothesis test on complete datasets.
3. Conduct a hypothesis test on subsets of datasets.

Glossary and Packages

Here are the main descriptive statistics functions we will be covering.

Function	Description
`table()`	Frequency table
`mean(), median(), mode()`	Measures of central tendency
`sd(), range(), iqr(), var()`	Measures of variability
`max(), min()`	Extreme values
`summary()`	Several summary statistics

Here are the main hypothesis test functions we will be covering.

Function	Hypothesis Test	Additional Help
`t.test()`	One and two sample t-test	https://bookdown.org/ndphillips/YaRrr/htests.html#t-test-t.test (https://bookdown.org/ndphillips/YaRrr/htests.html#t-test-t.test)
`cor.test()`	Correlation test	https://bookdown.org/ndphillips/YaRrr/htests.html#correlation-cor.test (https://bookdown.org/ndphillips/YaRrr/htests.html#correlation-cor.test)
`chisq.test()`	Chi-Square test	https://bookdown.org/ndphillips/YaRrr/htests.html#chi-square-chsq.test (https://bookdown.org/ndphillips/YaRrr/htests.html#chi-square-chsq.test)
`aov(), TukeyHSD()`	ANOVA and post-hoc test	https://bookdown.org/ndphillips/YaRrr/anova.html#full-factorial-between-subjects-anova (https://bookdown.org/ndphillips/YaRrr/anova.html#full-factorial-between-subjects-anova)

Examples

- The following examples will take you through the steps of doing basic hypothesis tests. Follow along and try to see how piece of code works!

```r
# ------------------------------------------------
# Examples of hypothesis tests on the ChickWeight data
# ------------------------------------------------
library(tidyverse)

chick <- as_tibble(ChickWeight)   # Save a copy of the ChickWeight data as a tibble c
alled chick

# -----
# Descriptive statistics
# -----

mean(chick$weight)   # What is the mean weight?
median(chick$Time)   # What is the median time?
max(chick$weight)    # What is the maximum weight?
table(chick$Diet)    # How many observations for each diet?

# -----
# 1-sample hypothesis test
# -----

# Q: Is the mean weight of chickens different from 110?

htest_A <- t.test(x = chick$weight,     # The data
                  alternative = "two.sided",  # Two-sided test
                  mu = 110)                   # The null hyopthesis

htest_A            # Print result
names(htest_A)     # See all attributes in object
htest_A$statistic  # Get just the test statistic
htest_A$p.value    # Get the p-value
htest_A$conf.int   # Get a confidence interval

# -----
# 2-sample hypothesis test
# -----

# Q: Is there a difference in weights from Diet 1 and Diet 2?

htest_B <- t.test(formula = weight ~ Diet,     # DV ~ IV
                  alternative = "two.sided",  # Two-sided test
                  data = chick,         # The data
                  subset = Diet %in% c(1, 2)) # Compare Diet 1 and Diet 2

htest_B  # Print result

# -----
# Correlation test
# ------

# Q: Is there a correlation between Time and weight?

htest_C <- cor.test(formula = ~ weight + Time,
                    data = chick)

htest_C
```

```
# A: Yes. r = 0.84, t(576) = 36.7, p < .001


# Q: Does the result hold when ONLY considering Diets 1 and 2?

htest_D <- cor.test(formula = ~ weight + Time,
                    data = chick,
                    subset = Diet %in% c(1, 2))     # Only take data where Diet is 1
 or 2

htest_D

# A: Yes. r = 0.81, t(339) = 25.08, p < .001


# -----
# Chi-Square test
# ------

# Q: Are there more observations from chicks on one diet versus another?

htest_E <- chisq.test(x = table(chick$Diet))  # Input is a table of values

htest_E

# A: Yes, some diets are observed more than others. X2(3) = 52.6, p < .001

# -----
# ANOVA
# -----

# Q: Is there an overall effect of diet on weight?

Diet_aov <- aov(formula = weight ~ factor(Diet),   # Run the anova
                data = chick)

summary(Diet_aov)       # Look at summary for overall test results
TukeyHSD(Diet_aov)      # Conduct post-hoc tests

# A: Yes, there is an overall effect of diet on weight, F(3, 574) = 10.81, p < .001
# Furthermore, we find significant differences between diets 1-3, and diets 1-4 at th
e 0.05 level.
```

Tasks

Gettting started

A. For this practical, we'll use the `ACTG175` dataframe from the `speff2trial` package, load the package with the `library()` function. Also load the `tidyverse` as always!

```
library(tidyverse)
library(speff2trial)
```

B. Convert the data to a tibble (Hint, use assignment and `as_tibble()`)

```
ACTG175 <- as_tibble(ACTG175)
```

C. First thing's first, take a look at the data by printing it. It should look like this

```
# A tibble: 2,139 x 27
   pidnum   age    wtkg  hemo  homo drugs karnof oprior   z30 zprior
    <int> <int>   <dbl> <int> <int> <int>  <int>  <int> <int>  <int>
 1  10056    48 89.8128     0     0     0    100      0     0      1
 2  10059    61 49.4424     0     0     0     90      0     1      1
 3  10089    45 88.4520     0     1     1     90      0     1      1
 4  10093    47 85.2768     0     1     0    100      0     1      1
 5  10124    43 66.6792     0     1     0    100      0     1      1
 6  10140    46 88.9056     0     1     1    100      0     1      1
 7  10165    31 73.0296     0     1     0    100      0     1      1
 8  10190    41 66.2256     0     1     1    100      0     1      1
 9  10198    40 82.5552     0     1     0     90      0     1      1
10  10229    35 78.0192     0     1     0    100      0     1      1
# ... with 2,129 more rows, and 17 more variables: preanti <int>,
#   race <int>, gender <int>, str2 <int>, strat <int>, symptom <int>,
#   treat <int>, offtrt <int>, cd40 <int>, cd420 <int>, cd496 <int>,
#   r <int>, cd80 <int>, cd820 <int>, cens <int>, days <int>, arms <int>
```

Descriptive statistics

D. What was the mean age of all patients?

E. What was the median weight of all patients?

F. What was the mean CD4 T cell count at baseline? What was it at 20 weeks?

G. How many patients have a history of intraveneous drug use and how many do not? (Hint: use `table()`)

T tests with t.test()

1. Conduct a one-sample t-test comparing the age of the patients versus a null hypothesis of 40 years. What is the test statistic? What is the p-value? Do you accept or reject the hull hypothesis?
2. Now, compare the mean age to a null hypothesis of 35 years. What has changed?
3. A researcher wants to make sure that men and women in the clinical study are similar in terms of age. Conduct a two-sample t-test comparing the age of men versus women to test if they are indeed similar or not.

- Women are coded as 0 in `gender`, and men are coded as 1.
- Be sure to use the formula notation `formula = age ~ gender`

4. Conduct a two-sample t-test comparing the number of days until the first occurrence of a major negative event (`days`) between those with a history of intravenous drug use (`drugs`) and those without a history of intravenous drug use

Correlation test with cor.test()

5. Do older people tend to weigh more? Conduct a correlation test between weight (`wtkg`) and age (`age`). What is your conclusion?

6. We would expect a correlation between CD4 T cell count at baseline (`cd40`) and at 20 weeks (`cd420`). But how strong is the correlation? Answer this question by conducting a correlation test between CD4 T cell count at baseline (`cd40`) and CD4 T cell count at 20 weeks (`cd420`).
7. Is there a relationship between CD4 T cell count at baseline (`cd40`) and the number of days until the first occurrence of major negative event (`days`)?
8. Only considering men, is there a correlation between CD4 T cell count at baseline (`cd40`)and CD8 T cell count at baseline (`cd80`)?

- Include the argument `subset = gender == 0` to restrict the analysis to men

9. Now, repeat the previous test, but only for women

Chi-square test with chisq.test()

10. Do men and women (`gender`) have different distributions of race (`race`)? That is, is the percentage of women who are white differ from the percentage of men who are white?

- Be sure to create a table of gender and race values with `table(ACTG175$gender, ACTG175$race)`

11. Is there a relationship between a history of intravenous drug use (`drugs`) and hemophilia (`hemo`)?
12. Is there a relationship between homosexual activity (`homo`) and gender (`gender`)
13. Only for patients older than 40, is there a relationship between antiretroviral history (`str2`) and race (`race`)?

- Create a new dataframe called `ACTG175.o40 <- subset(ACTG175, age > 40)` and then do your analysis on this new dataframe.

14. Now repeat the previous analysis, but only for male patients

- Create a new dataframe called `ACTG175.male <- subset(ACTG175, gender == 0)` and then do your analysis on this new dataframe.

ANOVA with aov()

15. One of the main research hypotheses might be that there is an effect of treatment on CD8 T cell count at 20 weeks of treatment. Test this hypothesis to see if there an effect of treatment arms (`arms`) on CD8 T cell count at 20 weeks (`cd820`). If there is a significant effect, conduct post-hoc tests to see which treatment arms differed.
16. A researcher might be concerned that certain treatments might lead to substantial weight-loss or weight-gain. Answer this question by testing if there an effect of treatment arms (`arms`) on weight (`wtkg`). If the effect is significant, conduct post-hoc tests.
17. The main variable of interest is if there is an effect of treatment arms (`arms`) on the number of days until the occurrence of a major negative event (`days`). Answer this by conducting the appropriate ANOVA (with post-hoc tests if necessary).
18. Does the previous result hold if you only consider patients with a history of intravenous drug use (`drugs`)? Answer this by conducting the same ANOVA *only* on these patients.

- Create a new dataframe called `ACTG175_drugs = subset(ACTG175, drugs == 1)` and run your analysis on this dataframe

Extras and Challenges

Generating random samples from distributions

19. You can easily generate random samples from statistical distributions in R. To see all of them, run `?distributions`. For example, to generate samples from the well known Normal distribution, you can use `rnorm()`. Look at the help menu for `rnorm()` to see its arguments.

20. Using `rnorm()`, create a new object `samp_10` which is 10 samples from a Normal distribution with mean 10 and standard deviation 5. Print the object to see what the elements look like. What should the mean and standard deviation of this sample? be? Test it by evaluating its mean and standard deviation directly using the appropriate functions. Then, do a one-sample t-test on this sample against the null hypothesis that the true mean is 12. What are the results?

21. Evalaute your code for the previous question *exactly* as it is – that is, don't change *anything*. What are the new values in `samp_10` and the new mean, standard deviation, and t-test result. Why are the new results different?

22. Now, create a new object called `samp_1000` which is 1,000 samples from a Normal distribution (again with mean 12 and standard deviation 5). Print this object to see what it looks like. What should the mean and standard deviation of this sample be? Do the same hypothesis test as you did in the previous question. What is your new p-value?

2 - Way ANOVA

23. Conduct a two-way ANOVA testing the effects of *both* hemophilia (`hemo`) and drug use (`drugs`) on the number of days until a major negative event.

- To include multiple factors in an anova, just include both in the formula such as: `formula = dv ~ factor(x) + factor(y) + ...`. See https://bookdown.org/ndphillips/YaRrr/anova.html#ex-two-way-anova (https://bookdown.org/ndphillips/YaRrr/anova.html#ex-two-way-anova) for an example

24. Repeat the previous ANOVA, but now test if there is an *interaction* between hemophilia and drugs on the number of days until a major negative event.

- To include interactions in an ANOVA, just include both in the formula using the `*` operator: `formula = dv ~ factor(x) * factor(y)`. See https://bookdown.org/ndphillips/YaRrr/anova.html#ex-two-way-anova (https://bookdown.org/ndphillips/YaRrr/anova.html#ex-two-way-anova) for an example

You choose the test!

25. Is there a difference in the CD4 T cell count at baseline between whites and non-whites? Answer this by conducting the appropriate hypothesis test.

26. A researcher is particularly interested in whether or not there is a difference in the number of days until the first occurrence of a major negative event between patients taking zidovudine and those taking didanosine. Conduct the appropriate test to answer this question.

27. A researcher wants to know if the relationship between CD4 T cell count at baseline and age is similar for whites and non-whites. Specifically, she wants to know if both correlations are significant (and in the same direction!) or not. Conduct the appropriate statistical test separately for both groups. Are the conclusions the same or different?

28. A researcher is concerned that patients were not properly randomly assigned to the different treatment arms. Using the appropriate test(s), see if there is a significant imbalance between treatment arms in terms of gender, drug use, race, and homosexual activity. Do you find evidence for a significant imbalance in any of these domains?

Additional reading

- For more details on hypothesis tests in R, check out the chapter on hypothesis tests in YaRrr! The Pirate's Guide to R YaRrr! Chapter Link (https://bookdown.org/ndphillips/YaRrr/htests.html)
- For more advanced mixed level ANOVAs with random effects, consult the `afex` and `lmer` packages.
- To do Bayesian versions of common hypothesis tests, try using the `BayesFactor` package. BayesFactor Guide Link (https://cran.r-project.org/web/packages/BayesFactor/vignettes/manual.html)