UCB W203 - Lab 1: Hypothesis Testing

Kevin Lustig, Rebecca Nissan, Anuradha Passan, Giorgio Soggiu

2/21/2022

Part 1: Foundational Exercises

1.1 Professional Magic

1.1.1 Type I Error of the test

The type I error rate (i.e. false positive) is the probability of rejecting the null hypothesis whe it is correct. The type I error would be the probability of getting 0 or 6, with the assumption that p = 0.5 (null).

This will be the alpha

1.1.2 Power of test given p = 0.75

1.2 Wrong Test, Right Data

In the likert scale the "distance" between the different options is not consistent, thus violating the metric scale assumption needed to run a paired t-test. We could suggest a paired sign test where the only only assumptions needed are: ordinal variables and i.i.d.

1.3 Test Assumptions

1.3.1 World Happiness

Scenario: We have two variables: Life.Ladder and Log.GDP.per.Capita, and we want to see whether countries in high GDP per capita are more or less happy than people in countries with low GDP per capita.

Assumptions for two-sample t-test 1. Normal - CHECK! 2. i.d.d. - yes (assuming same data collection and data transformation procuedures), explain 3. Metric variables - yes, explain

```
wh_data <- read.csv('datasets/happiness_WHR.csv')

# Calculate the mean GDP per capita

# Split the data into low and high gdp per capita countries

# Check out histograms of both variables

# Can do a Shapiro test as an extra source of evidence to test for normalcy
# https://www.datanovia.com/en/lessons/normality-test-in-r/#check-normality-in-r
```

1.3.2 Legislators

Scenario: We want to test whether Democratic or Republic senators are older, with two variables party and age (age needs to be calculated from DOB).

Assumptions for Wilcoxon Rank-Sum test: 1. Ordinal variables - eyes, explain 2. i.i.d. - yes, explain 3. Same shape and spread of the the two variables - box whisker plot, histogram

```
leg_data <- read.csv('datasets/legislators-current.csv')

# Need to calculate the age for each legislator

# Check whether the variables have the similar shape/spread - box whisker plot or histogram
```

1.3.3 Wine and Health

Scenario: We want to test whether these countries have more deaths from heart disease or liver disease.

Assumptions for Wilcoxon Signed Rank Test: 1. Metric Variables - yes, explain 2. i.i.d. - yes, explain 3. Difference is symmetric - need to test for this 4. Paired - explain why/why not

```
# install.packages("wooldridge")
library(wooldridge)
wine_data  <- wine

# test for symmetry
```

1.3.4 Attitudes Toward the Religious

Scenario: We would like to know whether the U.S. population feels more positive towards Protestants or Catholics.

Assumptions for a Paired t-Test 1. Metric Variables - yes, explain 2. i.i.d. - yes, explain 3. Paired - yes, explain 4. Normalcy - Check for this

```
rel_data <- read.csv('datasets/GSS_religion.csv')

# Check for normalcy with histograms, and can do also the Shapiro test
```

Part 2: Statistical Analysis

Data source: 2020 American National Election Studies (ANES) - 2020 Time Series Study

Research question: Did Democratic voters or Republican voters experience more difficulty voting in the 2020 election?

Define the following: 1. Democratic voters 2. Republican voters 3. DFactors contributing to difficulty in voting –> create an index?

Data cleaning and wrangling

Form final dataset

EDA and basic information on the variables and proof we can make assumptions needed for tests we will run

Run tests

Interpret tests

Create visualizations (if not done before)

Write report

```
anes_data <- read.csv('datasets/anes_timeseries_2020_csv_20220210.csv')
```