STAT 201: Problem Set 6 (R)

Your name

2025-10-27

In every code chunk where you perform random sampling, set a seed at the top of the chunk. I don't care what seed you choose so long as you set a seed!

Make your code as reproducible as possible. You should avoid "hard-coding" values. Instead, store values as variables for future use. Additionally, try using in-line code when reporting p-values!

Note: you will practice typing in mathematical notation in the narrative using Latex. This is done using two dollar signs, and then typing Latex code between the dollar signs. To write the Greek letter mu, type μ in the narrative. You can add subscripts with the underscore like this: μ_H or H_0. To write "does not equal" sign, type $\neq$ into the narrative. As an example: $H_0 : \mu = 0$ or $H_A : p_1 \neq 0.5$.

The dataset is adapted from Little et al. (2007), and contains voice measurements from individuals both with and without Parkinson's Disease (PD), a progressive neurological disorder that affects the motor system. The aim of Little et al.'s study was to examine whether they could diagnose PD by examining the spectral (sound-wave) properties of patients' voices.

147 measurements were taken from patients with PD, and 48 measurements were taken from healthy controls. For the purposes of this lab, you may assume that measurements are representative of the underlying populations (PD vs. healthy).

The variables in the dataset are as follows:

- `clip`: ID of the recording
- `jitter`: a measure of variation in fundamental frequency
- `shimmer`: a measure of variation in amplitude
- `hnr`: a ratio of total components vs. noise in the voice recording
- `status`: PD vs. Healthy
- `avg.f.q`: 1, 2, or 3, corresponding to average vocal fundamental frequency (1 = low, 2 = mid, 3 = high)

The data are stored in a variable called `parkinsons`. Run the following code chunk and take a look at the data before getting started.

```r
library(tidyverse)
library(readr)
knitr::opts_chunk$set(fig.width=12, fig.height=4)

# load data here
parkinsons <- read_csv("https://raw.githubusercontent.com/midd-stat201-fall2024/midd-stat201-
```

Part 1

Researchers suspect that patients with PD are less able to control their vocal muscles, and thus may have a greater voice jitter compared to healthy volunteers. Thus, they are interested in whether the mean jitter in voice recordings among patients with PD is greater than the mean jitter in voice recordings among healthy patients. The researchers select the 0.01 significance level.

1. Write out the null hypothesis and alternative hypotheses for the question in statistical notation, defining quantities when necessary. (See the note at the top about how to type in math mode!)

Answer:

2. Describe in words how you would obtain 5000 simulations from the null distribution. (*Hint: this should be very similar to the sex discrimination or CPR examples from class.*) Then, describe how you would estimate the p-value using this null distribution. Be as specific as possible (this might include referencing specific values from your data)!

Answer:

3. It will be helpful to define some variables here. To make your life easier, create the following objects/variables in R that store the following quantities:

- `n`: The number of total patients
- `n_h`: The number of healthy patients
- `n_pd`: The number of PD patients
- `jitters`: The vector of voice jitters

4. Simulate the null distribution according to your description above. Store the results into a vector called `null_dist_pt1`.

5. What is your p-value, decision, and conclusion in the context of the research question? Is it possible that you've made an error? If so, which one?

Answer:

Part 2

We will now answer the following question: is there enough evidence to suggest that the mean HNR in the voice recordings of PD patients is significantly different from 21.5 at the $\alpha = 0.10$ significance level?

6. Write out the null and alternative hypotheses for this question using symbols, defining any quantities as necessary.

Answer:

7. Describe in words how you would obtain 5000 simulations from the null distribution. Then, describe how you would estimate the p-value using this null distribution. Be as specific as possible (this might include referencing specific values from your data)!

Answer:

8. To make your life easier, create the following objects/variables.

- `hnr_pd`: a vector of HNR for the PD patients only (you may want to use some combination of `filter()` and `pull()` to obtain this)
- `n_pd`: the number of PD patients (though you should have already defined this in problem 3)
- `xbar_pd`: the observed/sample mean HNR of PD patients
- `mu_h0`: the null hypothesized value for the population parameter

9. Simulate the null distribution according to your description above. Store the results into a vector called `null_dist_pt2`.

10. Visualize your null distribution. Make sure your visualization has informative axis labels and title.

11. Estimate the p-value. Then answer the following: what is your p-value, decision, and conclusion in the context of the research question?

Answer: