Day4 Problem Set

Cecilia Sui

1/13/2022

Day 4 Outline:

1. Types of Statistical Data (numerical, categorical, ordinal)
2. Probability distributions: rnorm / dnorm / pnorm / qnorm
3. Descriptive Statistics

Probability Distributions

1. Suppose the height of males at a certain school is normally distributed with a mean of $\mu = 168$ cm and a standard deviation of $\sigma = 9$ cm. Approximately what percentage of males at this school are **taller** than 175 cm?

2. Suppose the weight of a certain species of dogs is normally distributed with a mean of $\mu = 33$ lbs and a standard deviation of $\sigma = 5.5$ lbs. Approximately what percentage of this species of otters weight less than 28 lbs?

3. Suppose the height of flowers in a certain region is normally distributed with a mean of $\mu = 24$ cm and a standard deviation of $\sigma = 5$ cm. Approximately what percentage of flowers in this region are **between 18 and 25 cm** tall?

4. Find the Z-score of the 99th quantile of the standard normal distribution.

5. Find the Z-score of the 95th quantile of the standard normal distribution.

6. Generate a vector of 5 normally distributed random variables with mean = 10 and sd = 2

7. Generate a vector of 1000 normally distributed random variables with mean = 50 and sd = 5

8. Generate a vector of 1000 normally distributed random variables with mean = 50 and sd = 25

9. Generate two histograms using the **hist()** function to view these two distributions.

Descriptive Statistics

This section touches on how to run linear regression briefly. Don't worry if you do not understand the meaning of every item in your summary table for your model. We will learn more about how to interpret them in QPM I and II.

1. Install and load the **faraway** package. Load the **gavote** dataset. Study the dataset using the help() function. Create a new variable called undercount by calculating the percentage of ballots that were not counted into votes. What's the range and quantiles of this new variable? Draw a histogram to illustrate the distribution of this variable.

2. Create a new variable **perGore** for the percentage of votes for Gore. Use **plot(col1, col2)** to create a scatter plot for the columns: **perGore** and **perAA**. What do you see? How can you interpret the plot?

3. Let's run a linear regression with undercount as the response and perAA as the predictor. (Please refer back to the lecture notes for the example we did.) Summarize the regression results and describe what you see. Can you get the coefficients? Use **plot()** to visualize undercount and perAA.

4. Run a linear regression with undercount as the response variable and both perGore and perAA as predictors. Summarize the results and describe what has changed. Can you guess why we observe such change?