Workshop 2: Distribution and its Tendency

Bolun

2021-01-20

Review

- Hope everyone successfully installed R and Rstudio, otherwise, please still use Rstudio Cloud today and contact me after the workshop to figure it out.
- Frequency Distribution: A frequency distribution is an overview of all distinct values in some variable and the number of times they occur.
- Relative Frequency Distribution: A relative frequency distribution shows the proportion of the total number of observations associated with each value or class of values
- Example, transform a histogram to relative frequency distribution, how you could do that?.

Warming up

Before we move into today's topic, let's warm up by loading some data sets into R

- go to https://gss.norc.org/get-the-data/stata and download data from 2008 and 2016
- set your working directory.
- using the following code to load the data.

```
install.packages("haven")
library(haven)
gss2008 <- read_dta("[replace your path here]/GSS2008.dta")
gss2016 <- read_dta("[replace your path here]/GSS2016.dta")
```

Shape of the Distribution 1

Group Activity: More visualization, but use ggplot2. You can refer to https://ggplot2.tidyverse.org/reference/geom_histogram.html and the example code.

- Using the GSS data to plot histograms for income, age and number of children.
- Turn the histogram into relative frequency distribution (or density distribution)
- What can you read out of the plots? To which direction the skewness is towards?

Shape of the Distribution 2

Briefing your findings:

- To which direction the skewness is towards?
- Is there any difference across years? What does the difference imply?

Tendency and the shape, part 1

Group Activity: Adding vertical lines of mean, median and mode to your previous visualization.

You can use the following code to get the mean, median and mode.

```
data <- c(1:10, 5, NA)
data_median <- median(data, na.rm = TRUE)
data_mean <- mean(data, na.rm = TRUE)
data_mode <- mode(data, na.rm = TRUE)
```

You can add a vertical line by using

```
ggplot(data, aes(x = x)) +
  geom_histogram() +
  geom_vline(xintercept = value)
```

Tendency and the shape, part 2

Group Discussion:

Based on the visualization, each group discuss a variable from the three.

- What do you notice?
- How do the three different lines locate in the graph? Why is it the case?
- Optional: can you tell the positions of mean, median and mode and their relation with the skewness?

Probability Distribution 1

Probability distribution: you can see probability distribution as a special kind of relative frequency distribution. Instead of being created out of a sample/dataset, it is generated via a formula.

- Using binomial probability distribution as an example.
- Imagine that you are really good at flipping coin, so in a Sunday afternoon you flip the coin for 40 times. How many heads would you find in these 40 trials?
- Simulating the situation: you are stuck at home because of COVID, you decide to do the same trial for 500 times.
- More math please refer to https://en.wikipedia.org/wiki/Binomial_distribution

Probability Distribution 2

Group Activity: Adjust the p in the simulation, what do you find the difference in the probabilty distribution?

Closer

Next time, we will discuss inference, that is, how much we can tell about a population from a sample?

- Review what is sample and what is population.