#4 Data Summarization, Measures of Central Tendency and Dispersion, Data Visualization II

Iris E. Acquarone

POLI 102

Spring 2023, Rice University

Up-to-date

- Fundamentals, R Data, Data wrangling/cleansing, and Data Visualization ✓
- HW #1 done ✓
  - Hard, doable, easy? Impossible?
- Syllabus change
  - FB on this?

Links to Use

- Canvas chat for attendance
- PollEv for live anonymous comments during lab
- Survey to share topics/phenomena/data you'd like us to work throughout the course when learning R

PollEv.com/ietchacq372

Measures of Central Tendency and Dispersion (MCT&D)

Measures of central tendency: Describe the approximate center of a distribution

- Mean, median, mode

Measures of dispersion/variability: Describe the spread of the data

- Range, upper and lower quartiles, interquartile range, variance, standard deviation

Standard Numeric Summary Built-In Functions

- `mean(x)` : find the mean of a numeric vector `x` .
- `sd(x)` : find the standard deviation of a numeric vector `x` .
- `median(x)` : finds the median of a numeric vector `x` .
- `quantile(x)` : finds the sample quantiles of the numeric vector `x` . Default is min, Q1, M, Q3, and max. Can find other quantiles by using the `probs` argument.
- `range(x)` : finds the range of the numeric vector `x` . Displays `c(min(x), max(x))` .
- `sum(x)` : find the sum of the elements of the numeric vector `x` .

We can apply these to vectors and variables (e.g., `mean(data$varname)`)

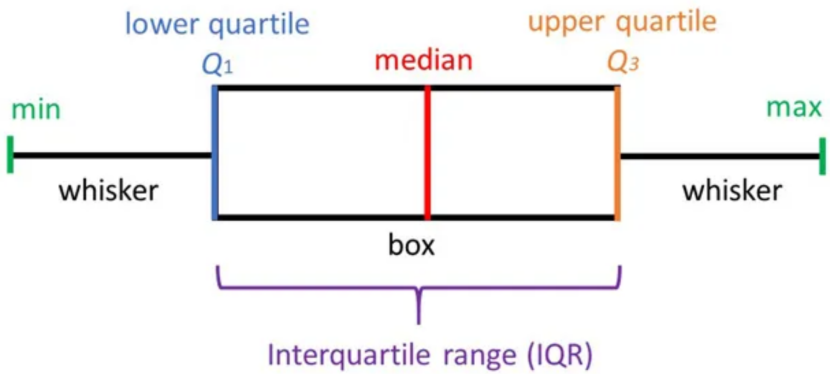
Standard Numeric Summary Built-In Functions

But R also has fast functions built to work on all columns/rows of a data set (e.g., `rowMeans(data)`):

- `rowMeans(x)` : finds the mean of each row of `x`
- `colMeans(x)` : finds the mean of each column of `x`
- `rowSums(x)` : finds the sum of each row of `x`
- `colSums(x)` : finds the sum of each column of `x`
- `summary(x)` : for data.frames, display the quantile information and number of NAs

Data Visualization II

- Histogram, with mean and other values
- Density plot, with mean and SD
- Boxplot



- (Summary based on groups)

Z-score

Measure that shows how much away (below or above) of the mean is a specific value (individual) in a given data set

$$z = \frac{x_i - \mu}{\sigma}$$

In R:

```
z = (x - mean(x))/sd(x)
```

Data Summarization

- Built-in measures to have a glimpse of your data in terms of MCT&D
- Summarizing data by groups
  - Summary based on groups visualization

Other Resources

Potentially useful links from shorter/simpler to lengthier:

- Link 1
- Link 2
  - Some use of `sapply` function; part of base `R apply` family functions
- Link 3
  - Some use of `apply` function; part of base `R apply` family functions

HW #2 posted, next week deadline 2/27

iea@rice.edu

RICE