STAT 350 Notes

Josh Park

Summer 2024

Chapter 1

An Introduction to Statistics and Statistical Inference

Chapter 2

Summarizing Data Using Graphs

Chapter 3

Numerical Summary Measures

3.1 Center of a distribution

3.1.1 Notation

x = random variable
x_i = specific observation
n = sample size

3.1.2 Sample mean

$$\bar{x} = \frac{sum\ of\ observations}{n} = \frac{1}{n}\sum x_i$$

R command: mean(variable)

3.1.3 Sample median

$\tilde{x} = centermost\ value\ in\ ordered\ dataset$

R command: median(variable)

3.2 Spread or variability of the data

three common ways to measure spread:

1. sample range
2. sample variance (or stdev)
3. interquartile range (IQR)

3.2.1 Range

range = $\max(x) - \min(x)$
completely depends on extreme values, so not very reliable
no R command for this

3.2.2 Sample Variance (sample standard deviation)

Variance

$variance = s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
R command: var(variable)

Standard Deviation

standard deviation $= s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$
R command: sd(variable)

if var = sd = 0, there is no spread (all data is the same)

3.2.3 Interquartile range (IQR)

Quartile

quartile = 1/4 of the data
R command = quantile(variable)
R command for % = quantile(variable, prob=c (p1, p2))

IQR

IQR $= Q_3 - Q_1$

3.3 Boxplots

fast way to vizualize five-number summary
five number summary: minimum, first quartile, median, third quartile, maximum

3.3.1 Outliers

IF = inner fence
OF = outer fence
subscript L = lower bound
subscript H = higher bound

$$IF_L = Q_1 - 1.5(IQR) \qquad IF_H = Q_3 + 1.5(IQR) \qquad \text{mild} \tag{3.1}$$

$$OF_L = Q_1 - 3(IQR) \qquad OF_H = Q_3 + 3(IQR) \qquad \text{extreme} \tag{3.2}$$

3.4 Choosing Measures of Center and Spread

if data is skewed, use median and IQR.
if symmetric, use mean and standard deviation.

3.5 z-score

3.5.1 z-score

the z-score of a data point x_i quantifies distance from the mean value in terms of standard deviations.

$$z_i = \frac{x_i - \bar{x}}{s}$$