Chapter 3 Notes - Numerical Summary Measures

Matt Warner

1 Important features of a numerical data set

- Shape of the distribution → see Chapter 2
- Center of the distribution
- Spread of the distribution

Measure of Center

Goal - calculate a number that identifies the middle of a data set or identifies a typical data value

- **Sample mean**
  - Let $x_1, x_2, x_3, \ldots, x_n$ denote the n data values in a sample.
  - $\bar{x} = \frac{x_1+x_2+x_3+\ldots+x_n}{n} = \frac{\sum x_i}{n}$
- **Sample median**
  - Sort the data values from smallest to largest
  - $\tilde{x} = \begin{cases} \text{middle value} & \text{if } n \text{ is odd} \\ \text{average of middle pair} & \text{if } n \text{ is even} \end{cases}$
- **Sample mode**
  - M = data value occurring most often.
  - if two (or three) values each occur most often the dist'n is bimodal (or trimodal).
  - If every value occurs equally often then the mode does not exist.

Example 1.1

A random sample of $n = 12$ tractor trailers was selected from a particular stretch of highway and their speeds recorded. Here are the data.

$$65, 79, 60, 67, 71, 83, 61, 67, 64, 77, 69, 74$$

a) Calculate the sample mean ($\bar{x}$), sample median($\tilde{x}$), and the sample mode (M).

$$\begin{aligned} \text{Sample mean} &= (65 + 79 + 60 + 67 + 71 + 83 + 61 + 67 + 64 + 77 + 69 + 74)/2 \\ &= 69.75 \end{aligned}$$

Sample median = 68

Sample mode = 67

b) Remove 83 and re-calculate $\tilde{x}$

$\tilde{x} = 67$

c) Change 60 to 61 and re-calculate M.

$M = 67$ and 61

d) Use the original data and change one copy of 67 to 68 and re-calculate M.

$M = \textit{No Mode}$

Note:-

If there are outliers, the mean should not be used to obtain a measure of center since it can pull the mean up or down. Instead, consider using the median as it is less sensitive to outliers.

What does the shape of a distribution possibly tell us about the mean and median?

- When a distribution is **left skewed** the mean is likely **less** than the median.
- When a distribution is **right skewed** the mean is likely **greater** than the median.
- When a distribution is **symmetric** the mean is likely **roughly the same** as the median.

Question - The distribution of tractor trailer speeds has possibly what shape?

Solution: The data creates a shape on the histogram that is skewed to the right.

Note:-

- $\bar{x}$ = mean of a sample ($\bar{x}$ is an example of a samplestatistic)

μ = mean of the entire population (μ is an example of a population parameter)

When μ is unknown we often use $\bar{x}$ as an estimate (or prediction) of μ.

1.1 Measures of Spread

Goal - calculate a number that measures the variability among the data values. or measures how far apart the data values are from each other.

- For measures of spread, if the answer is:
  - large, then the data is more spread out (has more variability)
  - small, then the data is less spread out (has less variability)

- **Sample range** R = max - min
- **sample variance**
  - Definition $S^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1} = \frac{\sum(x_i-\bar{x})^2}{n-1}$
  - (Computational or short cut) $s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$
- **Sample standard deviation** $s = \sqrt{s^2}$

1.2 Interquartile Range

$$\text{IQR} = Q_3 - Q_1$$

- $Q_3 = 3^{rd}$ quartile $= 75^{th}$percentile
- the number that has 75% of the data set at or below it
- the median of the upper half of the data.

- $Q_1 = 1^{st}$ quartile $= 25^{th}$ percentile
- the number that has 25% of the data set at or below it
- the median of the lower half of the data.

Example 1.2

A random sample of $n = 12$ tractor trailers was selected froma a particular strectch of highway and their speeds recorded. Here are the data.

$$60, 61, 64, 65, 67, 67, 69, 71, 74, 77, 79, 83$$

a) Find the sample range

$$83 - 60 = 23$$

b) Q_1 = median of the first half $= (64 + 62)/2 \rightarrow 64.5$

$M = (67 + 69)/2 = 68$

Q_3 = median of second half $= 75.5$

IQR $= 11$

c) Suppose ther was a 13^{th} tractor trailer in the sample with $x_{13} = 85$. Re-do part **(b)**

$Q_1 = 65$ $\quad M = 69$ $\quad Q_3 = 77$ $\quad$ IQR $= 12$

Example 1.3

find Q_1 and Q_3 for the following data values: 8, 3,7, 4, 1, 10, 5, 2, 11, 9, 6

a) $Q_1 = (3 + 4)/2 = 3.5$

b) $Q_3 = (8 + 9)/2 = 8.5$

Note:-

When finding range, do not forget to order the data from smallest to largest.

Example 1.4

Consider the following data

$$2, 5, 6, 9, 13 \text{ (in inches)}$$

a) Find the sample variance using the definition

First we need to find the mean.

$$(2 + 5 + 6 + 9 + 13)/5 = 7$$
$$\text{Mean} = 7$$

So,

$$S^2 = \frac{(2-7)^2 + (5-7)^2 + (6-7)^2 + (9-7)^2 + (13-7)^2}{5-1}$$

$$S^2 = 17.5$$

b) Find the sample variance using the computational (or short cut) formula.

Computational Formula:

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$$

So,

$$s^2 = \frac{\sum(2^2 + 5^2 + 6^2 + 9^2 + 13^2) - \frac{35}{5}}{4}$$

$$s^2 = \frac{315 = 245}{4}$$

$$s^2 = 17.5$$

c) $S = \sqrt{S^2} = \sqrt{17.5} = 4.18$

d) The units are measured in inches.

1.3 Important Facts

- Standard deviation is often thought of as measuring the typical distance that an observation is from the mean.
- Standard deviation is often used as a ruler for judging distances.
- s^2 = variance of a **sample** (s^2 is an example of a sample statistic)

  σ^2 = variance of the **entire population** (σ^2 is an example of a population parameter)

  When σ^2 is an unknown we often use s^2 as an estimate (or prediction) of σ^2

  The mean and the standard deviation can be used together to describe the distrobution of a data set more precisely.

Example 1.5

Suppose that Bob took exams in his English and Math classes. His scores together with the class summaries are below. In which class did Bob do better, relative to his classmates?

	English	Math
Bob	80	85
Mean	75	80
Standard Deviation	2.5	5

Solution: English is the correct answer due to the standard deviation being lower.

Bob scored 2 standard deviations above the mean in english, in math he was only 1 standard deviation above the mean

1.4 Empirical Rule

- Applies **only** when the shape of the distribution is **approximately normal** (bell shaped)
- Approximately 68% of the observations are within 1 standard deviation of the mean
- Approximately 95% of the observations are within 2 standard deviations of the mean
- Approximately 99.7% of the observations are within 3 standard devations of the mean

Example 1.6

In a random sample of people, the mean height was $\bar{x} = 66$ inches with a standard deviation of $s = 4$ inches. Assuming the distribution of heights is normal, answer the following questions:

(a) 68% of the people have heights between 62 & 70.

(b) 99.7% of the people have heights between 54 and 78.

(c) 95% of the people have heights between 58 & 74.

(d) What percent of the heights are between 66 and 70?

34% of the data is between 66 and 70. This is because it is only the right side of the middle 68%

(e) What percent of the heights are between 58 and 66?

47.5% of the data. rhis is because it is only the left side of the 2 standard deviation range (95%)

(f) What percent of the heights are greater than 74?

100 - 95 = both tails. Here we only want 1 tail. So, 5 / 2 = 2.5%

(g) What percent of the heights are either less than 54 or greater than 78?

100 - 99.7 = .3%

1.5 Chebyshev's Rule

- Applies to any data set, regardless of shape.
- Provided $k > 1$, at least $100\left(1 - \frac{1}{k^2}\right)\%$ of the data set is within k standard deviations of the mean. In other words, the data lie in the interval from $\bar{x} - (k \cdot s)$ to $\bar{x} + (k \cdot s)$.

Example 1.7

– In a random sample of people the mean height was $\bar{x} = 66$ inches with a standard deviation of $s = 4$ inches. Without assuming anything about the shape of the distribution of heights, answer the following.

(a) What percent of heights are between 58 and 74? How does this compare to using the Empirical Rule?

$k = 2$ at least $100(1 - \frac{1}{2^2})\% \rightarrow$ at least 75%

(b) What percent of heights are either less than 58 or greater than 74?

(c) What percent of heights are between 54 and 78?

(d) What percent of heights are between 60 and 72?

(e) At least 84% of the heights are between ______ & ______.