Chapter 2 - Describing Data with Numerical Measures

1. Motivation for using numerical measures.
2. Restrictions on what can be measured numerically.

Measure of Center

(A) Arithmetic Mean (Interval Variables)

1. Population mean (μ) (mu)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

2. Sample mean $\bar{x}$ (x bar)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measure of Center

(B) Median (Interval Variables)

1. Order values from lowest to highest.
2. Position of the median → 0.5*(N+1) (or 0.5*(n+1) for a sample)
3. If the position of the median is a number that ends in 0.5, you need to average the two adjacent points.

Comparing the mean and median values of the data

1. Median is less sensitive to outliers. Comparing the two values tells you if the data are skewed.
2. Median is a good measure of central tendency if the data are skewed.

Experience	Firm #1	Firm #2	Firm #3
1	28000	28000	28000
2	30500	30500	30500
3	33000	33000	33000
4	*35500*	*35500*	*50000*
5	*38000*	*50000*	*75000*

	Experience	Firm #1	Firm #2	Firm #3
Min	1	28000	28000	28000
Max	5	38000	50000	75000
Median	3	33000	33000	33000
Mean	3	33000	35400	43300

Measure of Center

(C) Mode (All Variables)

1. It is the number (or the category) that appears most frequently in the data set.
2. If a data set has two number (or categories) that appear the same number of times and more than any other number, then the data set is described as **bimodal**.

Valid numeric measures for the three main variable types.

Variable Type	Mean	Median	Mode
Categorical	No	No	Yes
Ordinal	No	No	Yes
Interval	Yes	Yes	Yes

Measure of variability (or measure of the variation or dispersion around the center of the distribution)

1. **Range** - the difference between the largest (MAX) and smallest (MIN) values.
2. **Deviation** - difference between a given value in the sample or population distribution (x_i) and its mean ($\bar{x}$ or μ) value.
3. **Variance** - average of the squared deviation from the mean.

Population variance (σ^2) (sigma squared)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample variance (s^2)

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Why are we concern about using the sum of the squared deviations (versus the sum of the deviation or other measures)

1. Why can't we just add up all of the deviations from the mean or use

   $$\sum_{i=1}^{n}(x_i - \bar{x})$$

   The reason we can't is that the sum of the deviation is always zero. Here is why

   $$\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$$

   The rule of summation operator tells you that

   $$\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{n-1}$$

The definition of the mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Replace $\bar{x}$ with the function in the variance equations.

$$\frac{\sum_{i=1}^{n} x_i - n\frac{\sum_{i=1}^{n} x_i}{n}}{n-1}$$

The two values of n cancel out in the numerator, leaving you with

$$\frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i}{n-1} = 0$$

Ok, what is the variance trying to tell us? We are trying to find single value that measures the difference of each value of X (each x_i) from the mean. Some of these differences are negative, some are positive. Squaring the difference is one way to ensure that the value is positive. This ensures that each difference calculated gives same kind of weight to calculation.

This logic is used to compute deviations from other statistical values.

Finally, why do we use (n-1) in the denominator of the sample variance? The basic reasoning is that the sample is an approximation of the population. The smaller the sample (the smaller the value of n), the greater the chance that our sample estimate of the variance s^2 is a poor estimate. One way of ensuring that the sample variance is not an underestimate of the population estimates is to reduce the denominator by 1.

Standard Deviation - Square root of the variance

1. Population standard deviation (σ) (sigma)
2. Sample variance (s)

Points regarding Variance and Standard Deviations

1. s and σ are always greater than or equal to zero.
2. The larger the value of the variance or standard deviation (population or sample), the greater the dispersion.
3. If the variance is equal to zero, all values are equal to the mean.

To measure or describe the variation or dispersion the mean in terms of the units of analysis, use the standard deviation value.

Example -First Column - $\bar{x}$ = 33,000

x	x-$\bar{x}$	(x-$\bar{x}$)^2
28000	-5000	25000000
30500	-2500	6250000
33000	0	0
35500	2500	6250000
38000	5000	25000000
	0	62500000

$$s^2 = \frac{62500000}{n-1} = \frac{62500000}{4} = 15625000$$

$$s = \sqrt{15625000} = 3952.85$$

Example -Third Column - $\bar{x}$ = 43,300

x	x-$\bar{x}$	(x-$\bar{x}$)^2
28,000	-15,300	234,090,000
30,500	-12,800	163,840,000
33,000	-10,300	106,090,000
50,000	6,700	44,890,000
75,000	31,700	1,004,890,000
	0	1,553,800,000

$$s^2 = \frac{1,553,800,000}{n-1} = \frac{1,553,800,000}{4} = 388,450,000$$

$$s = \sqrt{388,450,000} = 19,709.13$$

Tchebysheff's and the Empirical Rule

Seriously useful tools for describing the distribution of values in the variable. It takes mean and the standard deviation and allows you to describe the distribution in non-technical terms. You only need two values and two functions to go beyond the description of the mean and measure of dispersion.

Tchebysheff's Theorem

Theorem - Given a number k greater than or equal to one, and a set of n measurements, at least $(1-(\frac{1}{k^2}))$ of the distribution lies in the interval within k*(standard deviation) from the mean.

What is k?

There is a relationship between the value of k and measured interval under consideration. Here is the table from the textbook

k	**1-(1/k^2)**
1	0%
2	75%
3	88.9%

If you wanted to know that range of values lying in a given proportion (e.g., 65%), you would use this formula to find the value of k.

$$k = \sqrt{(1/(1-P))}$$
$$P \neq 1$$

To determine the interval where 65% of the measurement lies, you will need a k of around 1.69.

To determine the interval where p% of the measurement lies, use the following formula

$$\mu \pm k\sigma$$

or

$$\bar{x} \pm ks$$

Example
Mean starting wage for persons with undergraduate degrees in economics is \$ 32,900. Standard deviation is \$3,850.

1. Determine the interval where 50% of the wages lie.
2. Determine the interval where 75% of the wages lie.

Answer to 1

$$\text{Mean} = \$32,900$$
$$\text{Standard Deviation} = \$3,850$$
$$k = \sqrt{(1/(1-0.5))} = \sqrt{2}$$
$$\text{Lower} = 32,900 - 3,850 * \sqrt{2} = 27,450$$
$$\text{Upper} = 32,900 + 3,850 * \sqrt{2} = 38,350$$

50 % of the starting wages for econ undergrads lies between $27,400 and $38,350.

Answer to 2

$$\text{Mean} = \$32,900$$

$$\text{Standard Deviation} = \$3,850$$

$$k = \sqrt{(1/(1-0.75))} = 2$$

$$\text{Lower} = 32,900 - 3,850 * 2 = 25,200$$

$$\text{Upper} = 32,900 + 3,850 * 2 = 40,600$$

75 % of the starting wages for econ undergrads lies between \$25,200 and \$40,600.

Empirical Rule

Holds for distributions that are symmetrical and mound shaped (unimodal at the center of the distribution).

$(\mu \pm \sigma) \rightarrow$ Contains around 68% of the distributions.

$(\mu \pm 2\sigma) \rightarrow$ Contains around 95% of the distributions.

$(\mu \pm 3\sigma) \rightarrow$ Contains almost all of the values in the distributions.

Measure of Relative Standing

Suppose you were offered a starting salary of $45,000. How would you view this offer compared to the distribution of wages offered to graduating seniors. The z score function helps you do this comparison.

Z scores are similar to the k values in Tchebysheff's theorem and the numeric weights used in the Empirical Rule.

Z score function

$$\text{Z score} = \frac{x - \bar{x}}{s}$$

Range of values: 95% of the values lie within a z score of ± 2. (In Chapter 5, we will learn to use the z scores to calculate the probability of obtaining a starting values of $45,000 given the mean and standard deviation values given.)

Computing the z score for a starting wage of $45,000

$$\text{Mean} = \$32,900$$

$$\text{Standard Deviation} = \$3,850$$

$$\text{Z score} = \frac{45000 - 32900}{3850} = 3.143$$

Using the empirical rule, one expects that nearly all of the values in the distribution of wages lie 3 standard deviations away from the mean value. A z score of 3.143 implies that this wage offer is an extreme or an outlier.

Another method of viewing dispersion - Quartiles, interquartile range and outlier boundaries.

Using the following set of formulas, you can determine ranges of values that fall within proportions of the distribution and the values that are outliers. Here are the tools and the techniques to work with the tools.

Quartiles Recall that the median was calculated by ordering the values of a variable from lowest to highest and finding the value where 50% of the values lie above it and 50% below it. Quartiles are an extension of this ideas.

1. Q1 (Quartile 1) - 25% below and 75% above

$$0.25(n+1)$$

2. Q2 (Quartile 2 - median) - 50% below and 50% above

$$0.50(n+1)$$

3. Q3 (Quartile 3) - 75% below and 25% above

$$0.75(n+1)$$

If the result is not an integer, you need to use interpolation methods to find the number.

Interquartile range (IQR) This value is the difference between Q3 and Q1. It is the area plus or minus 50% around the mean.

Fence Values

Finds the presence of suspected and actual outliers.

1. Inner fence (left) Q1 - 1.5 IQR
2. Inner fence (right) Q3 + 1.5 IQR
3. Outer fence (left) Q1 - 3.0 IQR
4. Outer fence (right) Q3 + 3.0 IQR

Technique - paper and pencil method

1. Sort data in order from lowest value to highest.
2. Find Q1, Q2, and Q3.
3. Calculate IQR
4. Calculate inner and outer fence values.

Technique - Mathematica

Conclusion

In this section, we are introduced to a set of functions that allow us to describe and analyze a data set. This is the first stage in analysis, when we get into probability theory, we will find that these functions have stronger interpretations.