Chapter 3: Describing, Exploring & Comparing Data

§3.1 Overview

We will be continuing our exploration of methods of describing data in this chapter. The methods used in this chapter will be mainly methods for mathematically describing quantitative data's **center**, **variation**, to a lesser extent **shape**, and **outliers**. As before, I won't focus on **time**.

§3.2 Measures of Center

Recall in §2.1 we talked about the 5 characteristics of data (CVDOT – Center, Variation, Distribution, Outliers, Time). The 3 most important are center, variation and distribution. In §2.2&2.3 we dealt with descriptive methods for showing the distribution (shape), and now we need to discuss the center (a representative value).

There are 4 major measures of center, each measure being based upon different criteria and some being more appropriate than others. These 4 major measures are:

Mean Median Mode Midrange

The first in our discussion is the mean (also called the arithmetic mean). This measure is only appropriate for ratio or interval data. It only has meaning for numbers that are somehow related on a continuous ordered scale. It is an average. If we let **"x"** be a number from the sample and **"n"** be the total number in the sample. In mathematics and science a Σ (capital sigma – a Greek letter) is used to represent the sum. If we use a capital N then we are talking about the total number in a population. We can compute the mean for a population or for a sample. *It is important to note that English letters are used to describe a sample statistic and Greek letters are used to describe population parameters*.

$\bar{x}$ (read x bar) is the sample mean $\Sigma x / n$

μ (pronounced mu) is the population mean $\Sigma x / N$

It is important to note that population means are rarely known, and the sample mean is usually used as an approximation.

Example: A sample of six school buses in the Carlton District travel the following distances each day:

14.2, 16.1, 7.9, 10.6, 11.2, 12.0

Find the sample mean.

Note: μ would be the sum of all distances traveled by all Carlton District buses and then divided by the total number of buses in the district. This example only uses a sample of 6 buses.

The next measure of center is the median. The median is the middle value of the ranked data. The median can be found for interval, ratio and ordinal data, but not for nominal data. The median is denoted as $\tilde{x}$, pronounced x-tilde. Here is the procedure for finding the median. It is quite easy if there is an odd number of data points but when there are an even number there is a slightly special procedure.

Finding the Median

1. Rank the data in ascending order (a stem-and-leaf is nice for this)
2. a) If odd # there is a number that has an equal number above and an equal number below it. For example if there are 15 points then the 8^{th} is the median since there are 7 above it and 7 below it. It is the middle of the data.
   b) If there is an even number of data points the middle is between two points, so the two points must be averaged. If there are 20 points then the middle is between the 10^{th} and 11^{th}

Example: Find the median of the following ranked data
7.9, 10.6, 11.2, 12, 14.2, 16.1

The median can be used with the same types of data as the mean (ordinal and interval), so why would we need the median instead of the mean? The answer is skewed data and outliers. Outliers can affect the mean and but they do not affect the median. Skewed data also affects the mean, but to a lesser extent the median. So, once the distribution of the data has been observed the decision as to which measure of center to use can be made!

Note: The median is a better measure of center for highly skewed data or data which contains outliers.

The next measure of center we will discuss is the **mode**. The mode is the score that appears most frequently. Ranking the data also helps to find the mode (hence the stem-and-leaf plot has another use). The mode can be found for all for classifications of data, but it is the **only measure of center appropriate for nominal data**! Data can be of three types when considering the mode:

No Mode – Meaning that there is no data point is repeated

Bimodal – Meaning that there are 2 data points that appear with the greatest frequency.

Multimodal – Meaning many data points appear with the greatest frequency

Example: Find the mode(s) if one exists.
Confinement in days: 17, 19, 19, 4, 19, 21, 3, 21, 19

Hourly Incomes: 4, 9, 7, 16, 10

Test Scores: 81, 39, 100, 81, 69, 76, 42, 76

In conclusion, which measure of center is best used depends upon 2 things – first the classification of data and second the presence of outliers. Mean is usually the measure of center that is used but it is not the most appropriate when outliers are present due to the strong influence that outliers can have on this measure.

The midrange of the data is not often that important except when looking the symmetry of data and using it in comparison with other measures of center. It is nothing more than taking the high and low data points' sum and dividing it by two. Please see your text more information (Ed. 9 p. 62-67, Essential Ed. 2 p. 60)

I am going to include an example for the mean of a distribution using a frequency table, but we may not have time to discuss the example in class.

Example: The following frequency table refers to a sample of purse snatchers. The data points represent the ages of the sampled purse snatchers at the time of their arrest.

Class (ages)	**Frequency**
16 – 24	3
25 – 33	4
34 – 42	2
43 – 51	1

In order to calculate the approximate mean using a frequency table we will need to fill in the following table.

Class	f	**Mid-Point (x)**	$f \bullet$ **x**
16-24	3	(24+16)/2 = 20	20•3 = 60
25-33	4		
34-42	2		
43-51	1		
n ➔		Σf•x ➔	

$$\overline{x} = \frac{\Sigma f \bullet x}{n}$$

We will not discuss the weighted mean but you are **expected to read the section** containing this information on p. 91 of Triola (4th ed).

Skewness is something that we need to talk about now that we have discussed the mean and the median. It has to do with the distribution of the data with respect to the mean and the median and mode and therefore must be left for discussion after the mean and the median and mode. Skewness is a measure of symmetry. If a distribution extends more to one side than the other of its central grouping then it is called skewed.

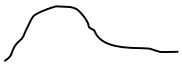
Left Skewed
(negative)
Mean & Median to the left of the Mode

Outlying data is dragging the mean & to less extent median to the left of mode

Symmetric
Mean≈Median≈Mode

Data is in a central grouping & outlying data is evenly spread

Right Skewed
(positive)
Mean & Median Right of Mode



Outlying data is dragging the mean & to less extent median to right of mode

Let's end the section with a recap of the measures of center presented by our book. In this recap we will summarize the type of data for which the measure of center is appropriate and a summary of some key information about the measure of center.

Summary of Measures of Center & Appropriate Type of Data

Mode	All Types of Data	Least informative for quantitative Only choice for nominal
Median	Ordinal, Interval & Ratio	Best for non-symmetric quantitative Only "real" choice for ordinal
Mean	Interval & Ratio	Most common measure Unbiased estimator (more later) Affected by outliers—not a resistant measure

§3.3 Measures of Variation

This section is about the measure of variation, our second characteristic of data. We will be discussing the range and the standard deviation (variance), and how we can use these measures to tell us about our data.

Range is very easy to define. It is how much the data varies from high to low. We find the range by computing the difference between the high and the low data points (high – low). The problem with the range is that it can be affected when there are outliers. Outliers can make the data appear to have a much larger range than it actually does.

Example: Find the range of the test scores:
81, 39, 100, 81, 69, 76, 42, 76

Note: If we look at the distribution of the data we see most of the data is 70 or above with 2 scores that are very different. These 2 scores affect the range of the data drastically. If we compute the standard deviation of these scores it will be less affected by the 2 very low scores, because most of the scores are near the top end of the scale. This is what makes standard deviation better than range in showing variation.

Probably the most important measure of variation for ordinal and interval data is the standard deviation. This is the measure of the variation about the mean. The standard deviation is the square root of the variance, but it is used more often than the variance because of the difficulty in interpreting the units associated with the variance (they are squared, and the units of the mean are not). Let's not be too quick to disregard the variance however as it has a characteristic that is extremely important is more advanced statistics – it is an unbiased estimator (it tends to be a good estimator of the actual population variance). The standard deviation of a population is called sigma and is represented by the Greek lower case letter sigma (σ). The standard deviation of a sample is represented by the lower case "s". If we are talking about population variance it is σ^2 and sample variance s^2. The following is the formula for sample variance. Remember that the sample standard deviation is the square root of the variance.

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

*Note: There are 2 ways to calculate the variance. The 1st formula is much easier and can be called the computational form than the 2nd with the use of a scientific calculator. The 2nd formula truly explains the meaning of the variance, and can be called the defining formula. I should be noted that with a small data set, the 2nd is also a nice formula to use, but the more data, the more cumbersome the formula becomes.

Example: The following are sampled finish times in a bike race (in minutes).
28, 22, 26, 33, 21, 23, 37, 24

a) Find the mean of the data.

b) Complete the following table to calculate the variance using the 2nd

formula given above.

x	x^2	x – x-bar	$(x - \text{x-bar})^2$
28			
22			
26			
33			
21			
23			
37			
24			
Σ =			

*Note: The round-off of any sample statistic should contain 1 more decimal than the original data. Always maintain as many decimals as possible in the calculation process until the final answer is derived. If you can't possibly maintain all decimals, try to keep at least 4, preferably 6.

c) Now use your calculator and the first formula to calculate the variance. Start by inputting all data into the data register of a TI-83/84 (stat➔edit➔enter data in L1) or with a simple scientific TI input data then use Σ+ until all data is entered. After inputting data on a TI-83/84 (stat➔calc➔1varstats➔2^{nd} f(n)#1) or on simple TI 2^{nd} f(n) left parenthesis.

d) Find the standard deviation of the data by taking the square root of the value found in b or c. Remember that those values should be the same!

e) Interpretation of the standard deviation involves the mean. The std. dev. in conjunction with the mean is used to give a range of values in which to find the data. We expect 68% of all symmetric data will fall within one standard deviation of the mean (that is, 1s above and 1s below the mean; it tells us how the data spreads out from the mean).

f) If this data is considered to be symmetric, calculate the range of values where you would expect to find 68% of all bike times to be.

It should be noted that the formula for the **population variance** is slightly different than that of the sample variance. The following are the formulas for the population variance.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N} = \frac{N \Sigma x^2 - (\Sigma x)^2}{N^2}$$

Example: Six families live on Merimac Circle. The number of children in each family is: 1, 2, 3, 5, 3, 4
Since we a using <u>all</u> the families on Merimac Circle this is considered a population.

a) On your own, calculate μ.

b) On your own, complete the table below and calculate σ^2 based upon the table using the 2nd formula above.

x	**x^2**	**$x - \mu$**	**$(x - \mu)^2$**
1			
2			
3			
5			
3			
4			

c) On your own, calculate σ^2 using the 1st formula given above.

d) Calculate the standard deviation of the population (σ).

*Note: In you should get 1.3 when rounded appropriately. On a calculator the pop. Std. dev. is given as σ_{xn} or simply σ_x where as the sample std. dev. is given as s_x or σ_{xn-1}.

The standard deviation can also be calculated using a frequency table. I may not have time in class to cover this calculation. Please recall the following definition from our study of frequency tables.

Class Midpoints (also referred to as Marks) – The point in the middle of each class. This is found by adding the lower and upper limits of the class and dividing by 2, or by subtracting the upper and lower limits and dividing by 2 and then adding that amount to each lower limit. Referred to as ***"x"*** here.

Frequency – The number of data points in each class. Referred to as ***"f"*** here.

Example: The following frequency table refers to a sample of purse snatchers. The data points represent the ages of the sampled purse snatchers at the time of their arrest.

Class (ages)	**Frequency**
16 – 24	3
25 – 33	4
34 – 42	2
43 – 51	1

In order to calculate the approximate standard deviation using a frequency table we will need to fill in the following table.

Class	f	**Mid-Point (x)**	$f \bullet x$	x^2	$f \bullet x^2$
16-24	3	(24+16)/2 = 20	20•3 = 60	$20^2 = 400$	400•3 = 1200
25-33	4				
34-42	2				
43-51	1				
n ➔		Σf•x ➔		$\Sigma f \bullet x^2$ ➔	

After you have finished the table, use the values to calculate the variance of the sample using the following formula:

$$s^2 = \frac{n\Sigma f \bullet x^2 - (\Sigma f \bullet x)^2}{n(n-1)}$$

*Note: This will not give the exact value of the variance, but as the data becomes more symmetric it will give a better and better approximation. The actual variance of this data set is 11.9 $years^2$. Of course, I don't know what a squared year means, so it might be nice to put it in terms of a standard deviation!!! ☺

Even if we are unaware of all data in a data set and can make the assumption that the data is approximately symmetric, we can find an approximate value to use in the place of the standard deviation using something called: **The Range Rule of Thumb**. This comes from the fact that for *symmetric data* nearly all the data (95%) will lie within two standard deviation of the mean, and therefore that the

maximum data point should be approximately the mean plus twice the std. dev.
(x-bar + 2s).
and
minimum data point should be approximately the mean minus twice the std dev
(x-bar – 2s).

Range Rule of Thumb – $s \approx \frac{\text{range}}{4}$

Example: For the bike racing data found on page 10 of these notes, find the value that you would expect for the standard deviation using the range rule of thumb. Compare this value to the actual standard deviation. Based upon the fact that the actual and the approximate should be pretty close if the data is symmetric, do you think the data is symmetric (even knowing nothing about what symmetric really means!)?

Now, let's also test out that idea of the maximum and minimum data points, again remembering that the data should be symmetric for the use of this approximation.

Example: Bookstore sales receipts give a mean of $61.35 and a standard deviation of $28.658. What would you expect the minimum and maximum values in sales to be for the bookstore?

We have been "quoting" an important approximation in this last section called the **Empirical Rule**. This rule has to do with symmetric, bell-shaped data and it tells us the following:

68% of the data will fall within **1 standard deviation of the mean**
95% of the data will fall within **2 standard deviations of the mean**
99.7% of the data will fall within **3 standard deviations of the mean**

Furthermore, we consider it to be **usual** for a data point to be within 2 standard deviations of the mean and unusual to be beyond 2 standard deviations. Once we get beyond 3 standard deviations we consider data points to be possible outliers.

Example: For a sample of blueberry cakes the mean weight was 500 g and the std. dev. was 12 g. In what weight range would you **usually** expect to find blueberry cakes from this population?

We can also use the Empirical Rule in another way. We can use it to give approximate percentages of the population that lie within certain ranges of data. We can do this by using something called the **z-score**. The z-score tells us how many standard deviations a data point is from the mean.

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

Example: Using the Empirical Rule and the z-score, indicate what percentage of the data you would expect to find between the values of 488 and 512 g.

Now, you might ask yourself, what if I know my data is not symmetric? In this case, we have another rule of thumb called **Chebyshev's Theorem**. It is a pretty simple calculation and only works for values beyond 1 standard deviation from the mean. It tells us what percentage of the data to expect with in a specified number of std. dev., based upon that number of standard deviations (k; k > 1).

$$\%\text{ of the Data w/in } \textit{"k"} \text{ std dev of mean} = \{1 - [1/(k^2)]\}\%$$

Example: If we assume that the data for the blueberry cakes is not symmetric, what percent of the data should fall within 2 standard deviation of the mean?

*Note: This value will not change, no matter what the data. It is also a smaller range than that given by the Empirical Rule, since the data is not considered to be symmetric.

Triola also discusses another measure of variation called the **mean absolute deviation (MAD)**. This measure is variation is 1) not algebraic 2) biased. For more on bias, please see p. 103. We will discuss bias in a little more detail at a later time. I will not discuss MAD, but will leave that to your reading. Please see p. 107-108 of Edition 4 of Triola's Essentials.

We talked about a z-score earlier, which is a means of comparing two data points relative to the mean and standard deviation of the sample/population from which they come. This is useful when comparing two similar things, such as your grade in my class with your friend's grade in a different Statistics class. Sometime it is also helpful to compare two sets of data with very different scales, means and standard deviation, and in this case we would want to use a different measure called the **Coefficient of Variation**. These are given as percentages, and are interpreted in the following manner – the higher the value the more the variation about the mean.

The **Coefficient of Variation** has the benefits of:

1) Unitless because the numerator & denominator have the same units
2) Allowance for direct comparison of 2 populations because it is unitless and it is taking into account the variation and mean of the sample/population.

$$\mathbf{CV} = \frac{s}{\bar{x}} \cdot 100 \qquad \text{or} \qquad \mathbf{CV} = \frac{\sigma}{\mu} \cdot 100$$

Example: The following data from Sullivan's 2nd edition, *Statistics*: *Informed Decisions Using Data*, page 131, are ATM fees for a random sample of 8 banks in both New York City and Los Angeles.

LA	2.00	1.50	1.50	1.00	1.50	2.00	0.00	2.00
NYC	1.50	1.00	1.00	1.25	1.25	1.50	1.00	0.00

a) Find the CV for LA & NYC.
b) Compare the ATM fees using the CV values.

§3.4 Measures of Relative Standing & Boxplots

This section considers measure of position which can help us to compare data. The most useful comparison score is a standard score, or a **z-score**. We have already discussed this measure of position in the last section. It measures the number of standard deviations from the mean for a bell-shaped curve and can be used to compare different measurements in an equivalent way. The larger the z-score the more unusual the value. We expect **usual** z-scores to be within 2 standard deviations of the mean and unusual values to be outside 2 standard deviations (an actual calculated value for standard deviation is always preferred to an estimated value using the range rule of thumb).

$$Z = \frac{x - \text{x-bar}}{s} \qquad \text{or} \qquad \frac{x - \mu}{\sigma}$$

Example: For a population of patients with $\mu = 60$ years and $\sigma = 12$ years, would we consider 68 to be an unusual value? Why or why not?

The next important measures of position are **percentiles** and **quartiles**. A percentile is a measure of the percentage of scores below a certain value. Quartiles are special percentiles. The 1st quartile (Q_1) is the same as the 25th percentile (P_{25}). The 2nd quartile (Q_2) is the 50th percentile, also called the median (x-tilde). The 3rd quartile (Q_3) is the 75th percentile (P_{75}); it should be noted that the 3rd quartile is the same as P_{25}, from the upper end of the data.

To find the percentile represented by a data point, the following process should be followed. The data must 1st be ordered and then:

$$\text{Percentile} = \frac{\text{\# of points below}}{\text{Total}} \bullet 100$$

Example: In the purse snatching data what is the percentile of 29?

Stem (x10)	Leaf (x1)
1	6 7
2	1 5 9
3	0 0 9
4	1
5	0

Now, let's consider what needs to be done to find a data point that represents a given percentile. Again, the data must be ordered and then:

Indicator Function: $\mathbf{L_k} = \frac{k \bullet n}{100}$ k = %tile, n = # of data pts.

*If L is whole number average that and next data point. (see flow chart p. 118 of Triola's Ed 4 Essentials) If L is a decimal/fraction then round up to the next whole. The indicator function indicates the position held by the percentile in ordered data – it isn't the percentile itself!

If L is a decimal, round up to the next whole and use that as the %tile.

Example: For the following data that represents the decibels create by an ordinary household item.

Stem (x1)	Leaf (x0.1)
52	0
53	
54	4 5
55	7 8 9 9
56	2 4 4 7 8
57	2 6
58	9
59	4 4 5 8
60	0 2 3 5 6 8
61	0 4 7 8
62	0 1 6 7
63	0 6 8
64	0 6 8 9
65	7
66	2 8
67	0 1 9
68	2 9
69	4
70	
71	
72	
73	
74	
75	
76	
77	1

Find P_{10}.

Find Q_1

Find Q_2

Find Q_3

Another *measure of position* is the **Interquartile Range.** This is referred to as the **IQR**. The *IQR* is nothing more than $Q_3 - Q_1$. The *interquartile range* is used to show where the bulk of the data resides. In symmetric data, seventy-five percent of the data should lie in the *IQR*. As a result, the *IQR* can be used to pinpoint outliers as well. Due to normal theory we know can decide where the "bulk" of our data "should" lie and we have established the following guidelines to set **fences** that tell us where to expect outliers. We can achieve this calculation by taking the *IQR* and multiplying it by 1.5 then subtracting that from Q_1 and adding it to Q_3, to find the range where we should expect most data to lie.

Outliers will lie <u>outside</u> the Range: $(Q_1 - 1.5IQR, Q_3 + 1.5IQR)$

Example: For the above data, should 77.1 be considered an outlier?

Exploratory data analysis (**EDA**) is the first step any Statistician takes when looking at a data set for the first time. It is important to see trends in the data such as shape, center, and variation. Usually the first exploratory analysis conducted is investigation of shape. With consideration of shape and data type, the most appropriate measures of center and variation can be calculated. Once a Statistician has conducted this exploratory analysis they are prepared for further analysis of the data using methods to be discussed in the remainder of the book. BTW your book indicates that Statisticians will use what are called the **hinges** of the data to compute the median rather than the quartiles for EDA. The hinges are the same as the quartiles when the number of data points in the data set is even, but when it is odd, the median itself is used to find the quartiles.

Investigation of Shape
    Historgrams/Stem&Leaf Plots/Dotplots/Box-and-Whisker Plots(Boxplots)
Measures of Center
    Mean/Median/Modes
Measures of Variation
    Variance/Standard Deviation/Range/IQR
Outlier Investigation
    Minimum/Maximum/3xIQR beyond Q_1&Q_3

Of all the above exploratory analysis, the only one not discussed thus far has been the **Box-and-Whisker Plot**, most times simply referred to as the **Boxplot**. The boxplot takes a 5 number summary of the data that shows center, variation, position, spread and shape of the data. It can not be discussed with the other graphical representations of shape because it requires the use of the quartiles.

5 Number Summary

Minimum Q_1 Q_2 Q_3 Maximum

A boxplot uses the 5 number summary in a scaled drawing where the maximum and minimum are represented by a small marking (usually a vertical line), the 1st and 3rd quartiles form a box with the median as a vertical divider of that box, and then "whiskers" are drawn from the central box to the minimum and maximum. If there are outliers present, they are sometimes represented by asterisks (especially in Minitab). The boxplot shows the shape by showing where the "bulk" of the data lies in relation to the minimum and maximum, as well as showing a "swaying" of the data based upon where the median lies within the central box (think of the median as a fulcrum). Because the boxplot is a scaled drawing, it can make a nice comparison tool for multiple data sets. Please note that your book uses a vertical boxplot, but I am more accustomed to drawing my boxplots horizontally, and will therefore continue to draw them horizontally.

Example: For the following data representing a sample of the measures of the diameter (in feet) of Indian dwellings in Wisconsin, create a stem and leaf plot to order the data, find the 5 number summary and then draw a boxplot (I want labels on the boxplot indicating the 5 number summary).

22, 24, 24, 30, 22, 20, 28, 30, 24, 34, 36, 15, 37