Chapter 2 Notes and elaborations for STAT 141-Introduction to Statistics

Assignment

The material in Chapter 2 is considered high-school review, although you certainly should read it and make sure you are comfortable with all of the terms. I am not terribly concerned with "ogives", "Pareto Charts", or the fussiness about class boundaries and widths. You should know how to read a histogram or other types of graphical data. I tend to use the word "**right-tailed**" instead of "**right-skewed**" and similarly for the left side.

I believe that most people now construct most graphs of data on a computer, so I tend to de-emphasize their construction. However, being able to construct a histogram means that you'll definitely be able to read one so you may wish to try a couple extra odd-numbered problems.

I suspect that most of you know most of the material from this chapter already.

Do the following exercises:

2.2 (page 65): 1, 3 is a good idea, 5, and 20 (Answers to #20 are 0, 14, 10, 16).

2.3 (page 90): 5, 17.

Chapter 3 Notes and elaborations for Math 1125-Introductory Statistics

Assignment

The material in Chapter 3 is of foundational value for this class. Read this chapter well. You will **not** be tested on and do not have to read:

- The mean for grouped data.
- Variance and standard deviation for grouped data.
- The coefficient of variation.
- The range rule of thumb.
- The midrange.
- You should understand what a percentile is, but not worry about computing them in general.
- You will have to be able to compute quartiles, which are specific percentiles.

Do the following exercises:

3.1 (page122): 3, 5, 7 (skip the midrange for those), 25, 27, 28 (answer is 82.67), 29.

3.2 (page143): 1, 3, 4, 7, 9, 11. Be sure to be able to compute the SD and variance by hand, but don't perform calculations without a calculator/computer.

3.2 (page145): 29, 31, 35, 41

3.3 (page159): 1- 7 all. 9 – 15 odd.

3.4 (page172): 1-13 odd

Notes follow on the next page.

This chapter attempts to answer some fundamental questions such as: Where's the "middle" of a data set? How spread out is the data? How does a particular data point compare with the rest of the data set? Before we get to these, though, I will review a couple of the basics.

3.0 Recall of definitions we will need

Recall the following definitions from the Chapter 1 notes:

Definition 1.1.2 A parameter is a numerical fact about a population or distribution.

In most situations, the value of the parameter is not and never will be known. In theory, we can discuss the average age of all people in Europe, but we will never know the true average. This is a parameter.

Definition 1.1.3 A statistic is numerical fact about a sample.

We can, however, take a sample of people in Europe and compute the average. This is a statistic. In general, statistics are used as estimators of parameters, e.g., we can use the average age of the people in the sample to estimate the average age of all people in Europe. Usually the computation of a statistic is done in a similar manner as one would compute a parameter. For example, consider the average. We find averages by adding all the data and dividing by the size of the data set. The average of 4, 6, and 11 is 7. Both population averages and sample averages are computed in this manner.

Example 3.0.0

Suppose a researcher is interested in the strongest wind speed in a hurricane ever. This would be a parameter. It is a numerical fact about the entire population of hurricanes that have occurred throughout history. There is an answer, but nobody knows it, and without a time machine, nobody ever will. One way to estimate it is to consider the strongest wind speed ever measured during a hurricane. This is a statistic, based on the sample of all wind speeds ever measured during hurricanes. It is an example of a biased estimator, as well. An estimator is biased if it tends to be either too big or too small relative to the parameter it is estimating. The strongest wind speed ever during a hurricane cannot be smaller than this (think about this). So, this particular statistic (which is an estimator) is expected to be smaller than the parameter it is estimating. Thus, this is a biased estimator.

Note again that I use the words statistic and estimator interchangeably in the above example.

Throughout the rest of this chapter, we will focus on commonly used descriptive statistics. Descriptive statistics tell us about important features of a data set. Descriptive statistics give us tools to answer such fundamental questions as:

Where's the middle of a data set?
How spread out is the data?

How does a particular data point (or a randomly chosen data point) compare to the rest of the data set?
Is the distribution symmetric? (The same on both sides of the middle.)
Is the distribution peaked or flat near the middle relative to the normal curve?

To answer these questions, we will look at the following:

measures of central tendency: mean, median, mode
measures of variation or dispersion: variance, standard deviation, range, IQR
measures of location: order statistics, percentiles, and z-scores
the empirical rule and Chebyshev's Theorem

3.1 Measures of central tendency

Measures of central tendency give us information on the middle of a data set. We discuss here three commonly used measures of central tendency: mean, median, and mode.

Definition 3.1.0 The mean of a data set is simply the average of the data.

Here is a list of properties of the mean.

• We use the symbol $\overline{X}$ (read this aloud as "X-bar") to represent the statistic, and μ for the parameter. That is, $\overline{X}$ is the sample mean and μ is the population mean.

• The book uses lazy summation notation, and it bothers me. I'll briefly explain it here.

If we have a list of say 10 data points, we could call the first one x_1, the second one x_2, the third one x_3, and so on. The Greek letter sigma, Σ, means "sum". Properly, one writes:

$$\sum_{i=1}^{10} x_i = x_1 + x_2 + \cdots + x_9 + x_{10}$$

The i=1 means start adding with the first data point, and the 10 on top means stop at the 10^{th} one. If we had N data points instead of 10 we would write:

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots + x_{N-1} + x_N$$

Using this notation, the proper way to write the mean (take N numbers each called x with a subscript, add them up and then divide by the total number) is:

$$\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_{N-1} + x_N}{N}$$

The book just uses ΣX to denote "add all the *x*'s."

• The mean uses all the data to calculate the center. This means it has the potential to be much more optimal than other methods that do not.

• The mean tends to vary less than the median from the same population. The book is wrong here for many important types of data sets.

• The mean is unique (every data set has at most one mean) and may not be a value in the data set. Consider that if you have 3 children and your neighbor has 2 children, together you have an average of 2.5 children.

• The sample mean is an unbiased estimator of the population mean μ, provided the population mean exists.

• The mean can be adversely affected by outliers.

• There are distributions which have no mean. I mean the mean does not exist for certain distributions, but you probably don't know what I mean! (Sorry.) You can always take the mean for a list of numbers from a population, but if the population doesn't have a theoretical mean (expected value), then this number will tell you nothing about the middle of the data set.

Example 3.1.0

A student is getting ready to graduate high school and move on to college. This student is researching universities and one piece of information the student is collecting in their study is the average income of graduates. The student is extremely interested in the University of North Carolina at Chapel Hill, because the average income of their graduates is six figures! What this student doesn't know is that Michael Jordan is a graduate of this university, and his (ridiculous) income is used as one of the sample points when computing the average income of graduates. Michael Jordan's income is an outlier. Most people do not make millions of dollars.

So, even though the mean is a good estimator of the central tendency most of the time, there are situations where the mean does not truly capture the central tendency of the data. This is because the mean is not robust.

Definition 3.1.2 A **robust** estimator is not sensitive to outliers.

The median, however, is a robust estimator of the central tendency. In Example 3.1.0, the median income of graduates is essentially unaffected by Michael Jordan's income.

Definition 3.1.3 The **median** is the middle value of a sorted data set.

Example 3.1.1

If there are an odd number of data points in our sample, this is simply the middle value of the sorted data. For example, the median value of {2.3, 4.6, 1.7, 3.2, 1.7} is 2.3. Here's how we find it.

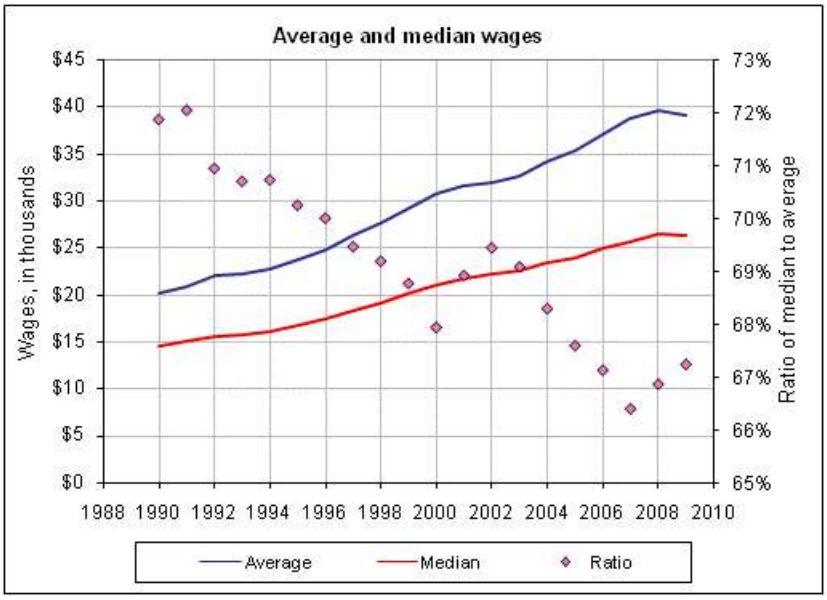
Step 1: Order the data from least to greatest: 1.7, 1.7, 2.3, 3.2, 4.6
Step 2: Find the middle data point. For this data set, it is 2.3.

If there are an even number of data points in our sample, this is the mean of the two middle values of the sorted data. For example, the median value of {2.3, 4.6, 1.7, 3.2, 1.7, 2.9} is 2.6.

Step 1: Order the data from least to greatest: 1.7, 1.7, 2.3, 2.9, 3.2, 4.6
Step 2: Find the two middle data points. For this data set, they are 2.3 and 2.9.
Step 3: Average the two middle data points. We get 2.6.

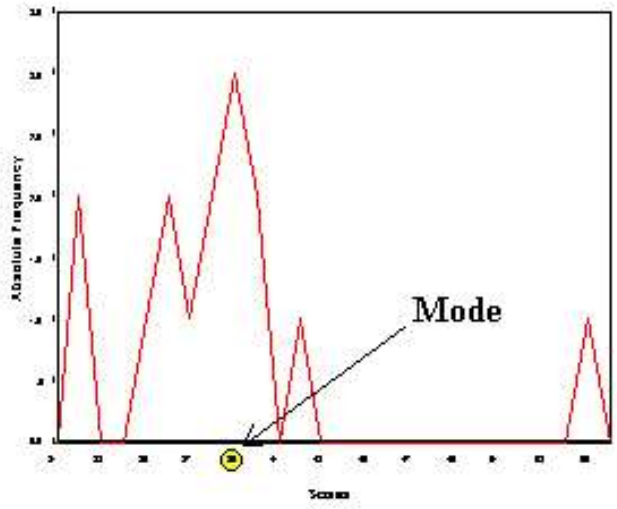
The median is a pretty good measure of central tendency, though it does not use all the data. That is, we lose information when we use the median alone to estimate the population mean μ (which is the true center). And unlike the statistic $\bar{X}$, the median is not always an unbiased estimator of μ. We won't do much with the median at this level other than to discuss skewness and calculate quartiles, but know that it is still quite useful.

It is also useful to look at the mean and median together. We do this a lot. Consider the following graph of average vs. median wages for US workers from 1988 to 2010 (taken from Office of Chief Actuary, US SSA). A drop in the ratio between them is one way of measuring increasing economic inequality. From the graph, we can see that this ratio has actually dropped about 5% over the these 22 years. This means median wages have lowered significantly relative to mean wages, which tells us the distribution of wealth is getting more skewed to the right. Essentially, a few people have quite a bit more money.

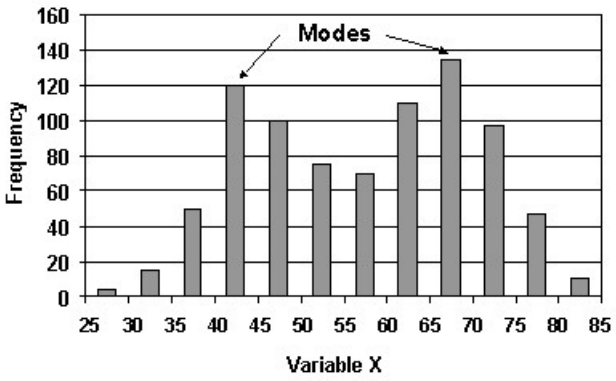


Yet another measure of central tendency is the mode.

Definition 3.1.3. The **mode** of a data set is the value occurring most often in the data set. Oftentimes it is used with histograms to discuss the location of peaks in a data set, although only one of them is truly the mode.



Why does the mode measure central tendency? It is not as simple as follows, but you can think of it like this: if most of the values in your data set are the number 3.2, then we would expect the mean of the data set to be close to 3.2 (unless we have an outlier!). Like the median, the mode is not always an unbiased estimator of μ. In fact, the mode is generally most useful when analyzing qualitative data, and pretty useless otherwise. Notice from the histogram on the right that data sets can have more than one mode.



Just a couple more things I'd like to mention relative to measures of central tendency:

Histogram with skew: See the pictures in the book for this; they do a pretty good job with it. Note that skewness is a measure of symmetry, not central tendency, but we can glean information regarding the mean and the median by looking at the skewness. If you have questions on this, post them or email me.

Weighted mean: Be sure to understand the book's worked-out examples. Here's a bit more that might help.

When data points do not contribute to the mean equally, we must use a weighted mean. Formally, if the data set consists of $x_1, x_2, x_3, \ldots, x_n$ and the corresponding weights are $w_1, w_2, w_3, \ldots, w_n$ then:

$$\text{The weighted mean} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Example 3.1.2

Say you are taking a 4-credit course, three 3-credit courses, and one 2-credit course. Suppose also that an A is four quality points, a B is 3 quality points, and a C is 2 quality points. You get a B in the 4-credit course, two C's and an A in the 3-credit courses, and a B in the 2-credit course. So, in general, your GPA is calculated as follows:

$$\frac{4*3+3*2+3*2+3*4+2*3}{4+3+3+3+2}=2.8$$

Not getting the weighted mean?

Here's a little extra. I'm going to make some non-standard definitions along the way to try and sort this out. Let's say there are 3 ways to take a mean: unweighted, simple weighted, and non-simply weighted.

Let's start with an unweighted mean and look at it in a different way. This is what people intend when they say "take the mean."

Example 3.1.3

Suppose you have 2 exams: x1 is the score on test 1, and x2 is the score on test 2. The mean is computed as $(x_1 + x_2)/2$. I will turn this into what I'll call a simple weighted mean. Here's what I'm calling the unweighted mean:

$$\frac{x_1+x_2}{2}=\frac{x_1}{2}+\frac{x_2}{2}=50\%\left(x_1\right)+50\%\left(x_2\right)$$

So your grade is 50% composed of test 1, and 50% composed of test 2. Pretty easy, right? Notice that 50% + 50% = 100%. These percentages are called the **weights**. There is a cool geometric picture that goes with this as well, but lets forge-on algebraically.

Consider the same scenario, but make it four tests. One adds the four scores, then divides by 4. Lets see the weights:

$$\frac{x_1+x_2+x_3+x_4}{4}=\frac{x_1}{4}+\frac{x_2}{4}+\frac{x_3}{4}+\frac{x_4}{4}=25\%\left(x_1\right)+25\%\left(x_2\right)+25\%\left(x_3\right)+25\%\left(x_4\right)$$

All of the weights turn out to be 25%. The **unweighted mean** could (or should?) be called the equal-weights mean. Each data point has an equal weight. If there were 10 tests, each data point contributes 10% to the mean.

Now lets change to the **simple weighted mean**. Here's how it will work: you are given 100% to distribute to the data points as you see fit. Let's suppose a class has 4 components: participation (p) is 10%, quizzes(q) are 30%, the midterm(m) is 20%, and the final(f) is 40%. The weighted mean is given by the formula:

OverallGrade = 0.1*p + 0.30*q + 0.20*m+ 0.40*f

Instead of giving everything equal weights, we now have put more emphasis on different components.

The **non-simply weighted mean**, what the rest of the world calls the weighted mean, is when we allow the weights to not sum up to 100%. Here's the funny part: we fix it so that the weights sum up to precisely 100% in the process of computing the weighted mean. That is, every weighted mean is transformed into a simply weighted mean while you do the computation. I'll give you two examples.

Example 3.1.4

Let's say a teacher's syllabus reads like this: the midterm is worth 2 tests, there are two "normal" tests, and a final worth 3 tests. Many people (but not me) will compute the grade as below, with the idea that really there are 7 tests: two midterms(copy the score twice), two tests, and three final exams.

OverallGrade = (m + m + t1 + t2 + f + f + f) / 7

Well let's see here:

$$\frac{2m+t_1+t_2+3f}{7}=2\frac{m}{7}+\frac{t_1}{7}+\frac{t_2}{7}+3\frac{f}{7}\approx 0.2857\,m+0.1429*t_1+0.1429*t_2+0.4286*f$$

If they understood a weighted mean, they could just say the midterm is worth (approximately) 28.57%, each test is worth 14.29%, and the final is worth 42.86%. Once you understand, then you realize that this is completely possible as well:

Example 3.1.5

The midterm is worth 1.2 tests, there are two "normal" tests, and a final worth 19.3 tests. Do that out (really, you do it!), and you see the weights are:

$$\frac{1.2\,m+t_1+t_2+19.3\,f}{1.2+1+1+19.3}=\frac{1.2\,m}{22.5}+\frac{t_1}{22.5}+\frac{t_2}{22.5}+\frac{19.3\,f}{22.5}\approx 0.0533\,m+0.0444*t_1+0.0444*t_2+0.8578*f$$

The formula on the far left (above) is exactly the formula for weighted means. The divisor is the number of tests: 1.25+1+1+19.3 = 22.5 tests. Notice at the far right there are four weights, each one giving you the fraction of the final score that each data point is weighted by. They should sum to 100%, and they would except for the rounding error.

So let's review the point of the last two examples. What the world calls weighted means is a broad class of types of means that can be categorized. Unweighted means happen when each of the data points contribute equal rational weights and the sum of the weights is 1. Simple weighted means happen when the data points contribute unequal weights and the sum of the weights is 1. And non-simply weighted means happen when the data points contribute unequal weights and the sum of the weights is not necessarily.

For anyone looking to continue on in mathematics, this idea of weighting different points occurs often. The expected value of a discrete random variable (which we discuss in the next section) is the weighted mean of all possible outcomes where the weights are probabilities per outcome. (Sometimes this is an infinite sum.) This is exactly the dot product of a vector of weights with the vector of outcomes. In fact, it is the projection of a vector onto another vector using the taxicab metric. . .I should stop now. There are continuous analogs as well . . I will stop. Anyway, this concept is very useful.

3.2 Measures of variation/dispersion

Measures of variation (or dispersion) tell us how 'spread out' the data is. Measures of dispersion are essential to the field of statistics. In this section we focus on the commonly used range, interquartile range (IQR), variance and standard deviation.

Definition 3.2.0 The **range** of a data set is the distance between the largest value in the data set and the smallest value in the data set. (Statistically, it is not two numbers separated by a dash!)

Example 3.2.0

Consider the following data sets:

A = {3, -4, 6, 8, -10}, B = {5, 7.2, 0, 9}, and C = {0, -13, -4.2, -9, -13.1}.

Range(A) = 8 -(-10) = 18, Range(B) =9-0 = 9, and Range(C) = 0-(-13.1) = 13.1.

The range is not terribly useful for the types of data we'll deal with in this class, but it is one measure of spread. Where it becomes useful is well beyond the scope of this course.

Another handy measure of dispersion is the IQR.

Definition 3.2.1 The **Interquartile range (IQR)** is the third quartile minus the first quartile, i.e., IQR = $Q_3 - Q_1$.

So, in order to understand the IQR, we must first understand the concept of quartiles. The first quartile of a data set is the data point Q_1 such that one-quarter (25%) of the data points are smaller than Q_1. It is the median of the first half of the ordered data set. The second quartile, Q_2, is the median. Recall that 50% of the data is smaller than the median. The third quartile, Q_3, is the median of the second half of the ordered data and has the property that three-quarters (75%) of the data points are smaller.

Example 3.2.1

Consider the following data set, already sorted for us:

{9, 12, 13, 14, 16, 17, 18, 18, 19, 20, 22, 23, 24}.

There are an odd number of data points in this set, so the median is the middle number. This is our second quartile:

median = Q_2 = 18.

Now to get the first quartile, we remove the median and find the median of the lower half of the data. The lower half of the data is
9, 12, 13, 14, 16, 17.

So, we get $Q_1 = (13+14)/2 = 13.5$. We do the same thing with the upper half of the data to get $Q_3 = 21$. Thus, the interquartile range of this data set is

IQR = 21 - 13.5 = 7.5.
This is a measure of variation.

Boxplots

The largest value (maximum), the smallest value (minimum), and the three quartiles, all taken together, make up what is known as a 5 number summary. Five number summaries are a great way to get a quick feel for a dataset, and even better for comparing datasets. Five number summaries are used to construct box-plots or "box and whiskers" plot. Your book does a pretty good job with these, but here's an extra example to help. I like five number summaries.

_____

Example 3.2.2

Let's construct a box-plot for the data set given in Example 3.2.3. Here's the data again:

{9, 12, 13, 14, 16, 17, 18, 18, 19, 20, 22, 23, 24}.

Our five number summary is: min = 9, Q_1 =13.5, Q_2 =18, Q_3 =21, max = 24.

Step 1: Draw a number line over an interesting range includes your min and max. (If this was data you cared about, you'd know what interesting means here.)

Step 2: Plot your five number from the summary on your number line. Draw a line connecting the two on each edge. I usually do this floating above the number line. Like this.

Step 3: Draw a box with edges at the first and third quartiles and a vertical line through the median.

That's it. That's a boxplot. Note that the middle 50% of the data is in the box. The smallest 25% of the data is in the "left whisker" and the last 25% is in the "right whisker". The median is the vertical bar in the box. Sometimes people put the mean in by using a dashed vertical line.

Definition 3.2.2 The **variance** (**Var**) is a measure of how spread out the data points are relative to each other. The variance is about equal to the mean of the squared deviations. What that means will be explained later.

Definition 3.2.3 The **Standard Deviation** (SD) is the square root of the variance.

These are the two most widely used measures of dispersion. Following is a list of properties of these two measures of dispersion:

• We generally use the symbol s and s^2 to represent the statistic, and σ and σ^2 for the parameter, i.e., s and s^2 are the sample standard deviation and sample variance, respectively, and σ and σ^2 are the population standard deviation and population variance, respectively. The latter symbols are lower-case sigmas from the Greek alphabet. We read them "sigma" and "sigma-squared". I will also use Var and SD.

• The variance is, for many deep mathematical reasons (most of which are well beyond the scope of this course), one of the best ways of measuring variability in data.

• Bigger values indicate more variation, that is, greater dispersion of the data.

• The SD is always the square root of Var. For example, if Var = 1.21, then the SD = 0.11 and if the SD = 2.5, then Var = 6.25.

• The smallest value possible for the Var and for SD is 0. The only way to get a 0 for a variance is by having no variability at all. The following is an example of a data set with a Var = 0: {14, 14, 14, 14, 14}.

• To calculate Var appears intimidating, but if you take it step-by-step, it's not so bad. Just take a data point, subtract the mean, and square it. Do this with every data point. When you're done, sum up all the squares that you have (this is actually called the sum of the squared deviations), and then divide by N, the number of elements in the population. If you have a sample, divide by n -1, instead (recall that n is your sample size, i.e., the number of elements in your sample)*. This is called the sample variance. Here's the formula for the two variances:

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})^2 \qquad\qquad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

*It may seem weird to divide by n-1 instead of n, but we do this so that our sample variance is an unbiased estimator of σ^2.

• To calculate Var and SD by hand, it is useful to make a table. Nowadays, most people just use a computer.

Example 3.2.3

Say we ran an experiment and have the following data set. You are taking an exam and must find the standard deviation. Data set: {9, 12, 13, 14, 16, 17, 18, 18, 19, 20, 22, 23, 24, 27}

Step 1: Find the mean.

Add all the data points and divide by 14 (our sample size).
We get that = 18.

Step 2: Make a table. First, subtract the mean from each data point.

X	X- 18
9	-9
12	-6
13	-5
14	-4
16	-2
17	-1
18	0
18	0
19	1
20	2
22	4
23	5
24	6
27	9

Step 3: Now square the last column.

X	X- 18	$(X- 18)^2$
9	-9	81
12	-6	36
13	-5	25
14	-4	16
16	-2	4
17	-1	1
18	0	0
18	0	0
19	1	1
20	2	4
22	4	16
23	5	25
24	6	36
27	9	81

Step 5: Sum all the squares. In this case, if you add them all you get 326.

Step 6: If you are computing the sample variance, divide by one smaller than the sample size. In this case, that's 13. If you are computing the population variance, divide by the population size. In this example, that would be 14.

So the sample variance is: $326/13 \approx 25.08$ (squiggly equal means approximately equal to).

And so the standard deviation (take the square root) is about 5.008.

We don't always have to actually calculate the variance in order to compare the spread of different data sets. We can see the spread with visual aids such as many of the charts/graphs in Chapter 2 of your text. And if the data sets are simple enough, we can just 'eyeball' the data as follows.

Example 3.2.4

Rank these three data sets in order of increasing variance without calculating anything except the mean:

$$A = \{1,2,3,4,5\},\ B = \{-1,1,3,5,7\},\ \text{and } C = \{1,3,3,3,5\}.$$

First, note that the mean of each set is 3. So now observe each set carefully and find which set contains points that are the least distance from the mean and the furthest from the mean.

It seems a little more obvious once you subtract the mean from each:

$A-3 = \{-2, -1, 0, 1, 2\}$ $B-3=\{-4, -2, 0, 2, 4\}$ $C-3 = \{-2, 0, 0 ,0 2\}$

Here they are smallest to largest variance: {1,3,3,3,5}, {1,2,3,4,5}, and then {-1,1,3,5,7}, i.e., C has the smallest variance and B has the largest variance. You could check this by finding the variance of these sets. I get

$\text{Var}(A) \approx 1.58$, $\text{Var}(B) \approx 3.16$, and $\text{Var}(C) \approx 1.41$.

You could have also constructed box-plots for each of the data sets and compared them visually.

3.3 Measures of position/location

By the term measure of position, we mean where is a data point located with respect to the rest of the data. In this section, we attempt to answer such questions as: is a data value less than the value of the mean, i.e., does it lie to the left of the mean? Is it to the right of the mean? How many standard deviations from the mean is the data point?
We begin with a thorough discussion of the all-important concept of standardized scores, i.e., z-scores.

Definition 3.3.0: A **z-score** (or standard score) corresponding to a data point is the number of standard deviations the data point falls above or below the mean. If the data point is equal to the mean, then its corresponding z-score is 0. If a data point is smaller than the mean, its corresponding z-score is negative, and if a data point is greater than the mean, its corresponding z-score is positive.

Be sure you know how to compute these, i.e., take a data point, subtract the mean of the data set, and divide by the standard deviation of the data set. Recall that the SD is the square root of the variance and both are measures of variation. Let's go over the computation of it all first, and then try to get an intuitive feel for what's going on here.

If the data point is x, the mean of the data set is $\overline{X}$, and the standard deviation of the data set is s, then: the z-score (denoted as z) is computed as: $z = \frac{x - \overline{X}}{s}$

If you have a z-score and need the original data point you reverse this algebraically to get:

$$x = z * s + \overline{X}$$

Example 3.3.0

You are presented with a data set that has a mean of 9 and a variance of 16.

(a) Suppose 6 is a data point. Find its corresponding z-score.

Take the data point, 6, subtract the mean, and then divide by 4 (4 is the square root of 16) to get a z-score of -0.75.

(b) Suppose 15 is a data point. Find its corresponding z-score. You should get 1.5.

(c) Suppose you are told that a data point has a corresponding z-score of 1.8. Find the data point.

Take 1.8, multiply by s (which is 4) and then add 9. We get that x = 16.2.

Don't miss the very important fact that the z-score of a data point is simply the number of standard deviations above or below the mean (reread the definition above). Let us try to expound on exactly

what this means, and we'll also try to provide pictures to complement the text. After we've got a feeling for what a z-score is, then we'll relate it to the normal curve, which we'll explore in Chapter 6.

Example 3.3.1

Recall the data set from Example 9: {9, 12, 13, 14, 16, 17, 18, 18, 19, 20, 22, 23, 24, 27}.
We have that $\overline{X} = 18$ and the SD ≈ 5.008.

(a) Find the new data set created by converting each point to its z-score. Here is a table to help us keep the information organized.

X	X- 18	(X- 18)/ 5.008
9	-9	-1.797
12	-6	-1.198
13	-5	-0.998
14	-4	-0.799
16	-2	-0.399
17	-1	-0.200
18	0	0.000
18	0	0.000
19	1	0.200
20	2	0.399
22	4	0.799
23	5	0.998
24	6	1.198
27	9	1.797

And so we see that the transformed data set is:

{-1.797,-1.198, -0.998, -0.799, -0.399, -0.200, 0, 0, 0.200, 0.399, 0.799, 0.998, 1.198, 1.797}.

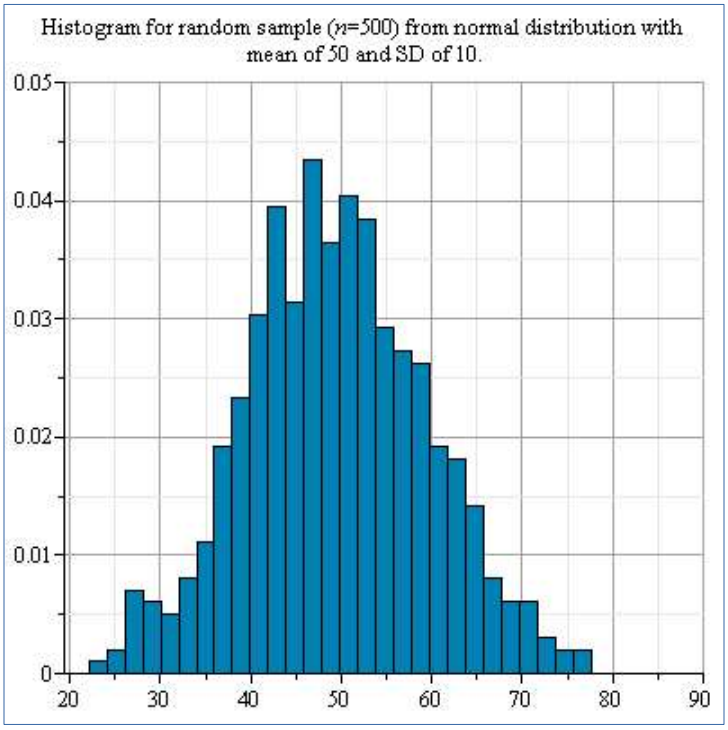
This new data set has mean of 0 and SD of 1. (Calculate them if you want.)

(b) You are told that a data point has a corresponding z-score of 0.2. What is the value of the data point?

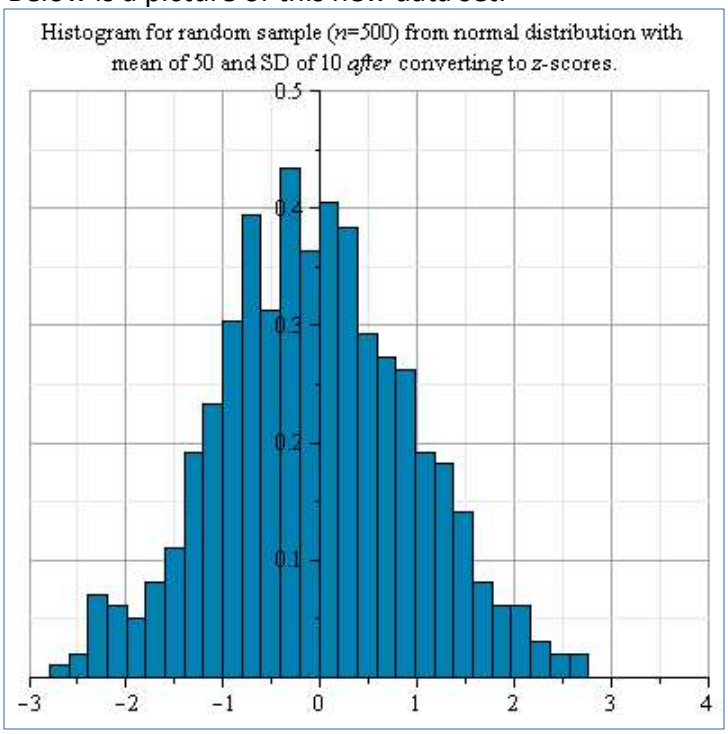
Well, you can just look at the table and see that it is 19 (rounded). But, you can also use algebra as follows:

$$0.2*s + \overline{X} = 0.2*5.008 + 18 = 19.016$$

Say we start with a random sample of size 500 from a bell-shaped with mean of 50 and SD of 10. Here is a picture of this data set.



Here we convert all the data to z-scores. This data set is distributed with the same shape but with mean of 0 and SD of 1. Below is a picture of this new data set.



Compare these two pictures. All we did was slide the histogram to the left, and then multiply by an appropriate scalar (a scalar is just a number that we can multiply by to shrink things or to blow things up) in order to have a variance of 1. The essence of the data remains--the distribution itself remains unchanged except for the location and scale.

The important part of z-score transformations is that the positions of all the data points relative to one another do not change, i.e., it is as if we pick up the picture, shrink/expand it depending on the SD, and place it with its center on zero. For example, if a data point was 2.33 SD's below the mean before the z-score transformation, then it is still exactly 2.33 SD's away from zero after the z-score transformation. In fact, its z-score is 2.33. Let us reiterate, the z-score of a data point is simply the number of standard deviations above or below the mean.

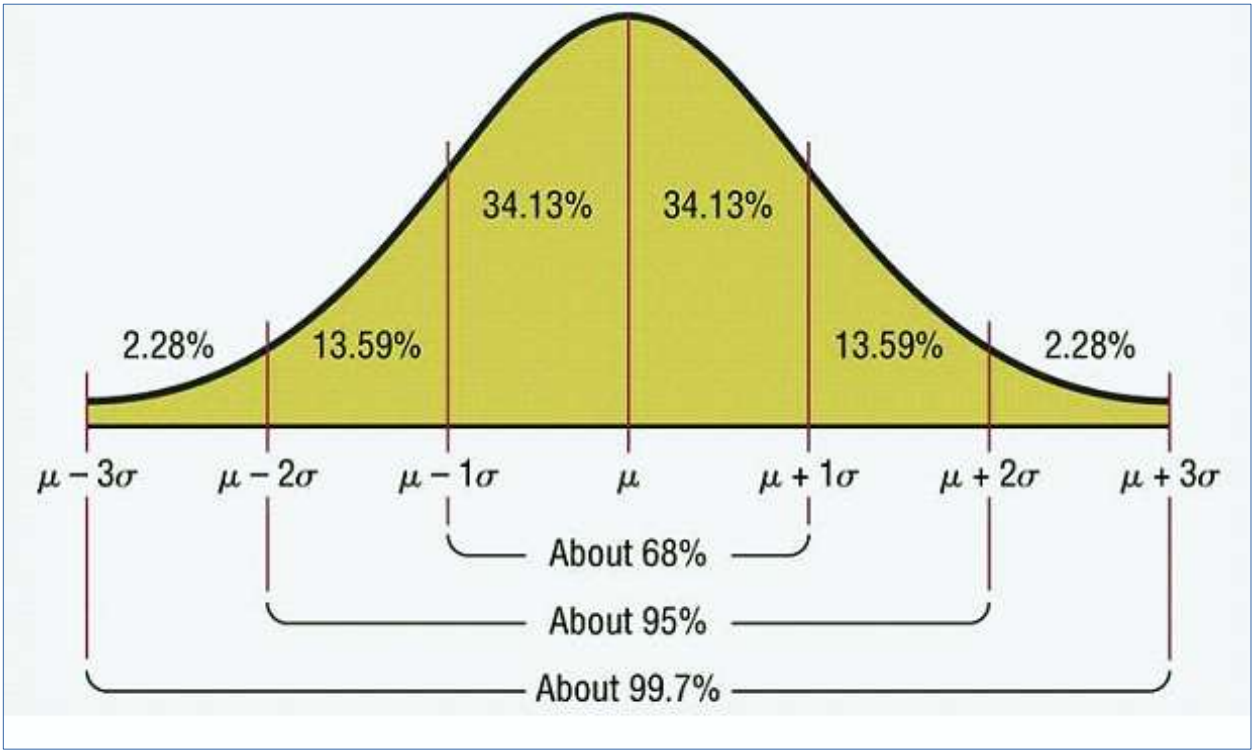
Hopefully, by now you are wondering why we do this. I'm glad you asked. It's easier to compute and compare data with z-scores. Indeed, if we want to compare two points from two different data sets, it makes good sense that we must first get them to the same scale of measurement, doesn't it? These gadgets will be very important to us, so learn them well.

3.4 The Empirical Rule and Chebyshev's Theorem

The Empirical Rule states that, for a sample from a normally distributed population (again, we will explore this concept in Chapter 6 - for now, think bell-shaped), the following properties hold:

- approximately 68% of the data lies within 1 SD from the mean.
- approximately 95% of the data lies within 2 SD's from the mean.
- approximately 99.7% of the data lies within 3 SD's from the mean.

For application purposes with respect to z-scores, this means that after we transform to z-scores, about 68% of the z-scores will fall in the interval (-1, 1), about 95% of the z-scores will fall in the interval (-2, 2), and about 99.7% of the z-scores will fall in the interval (-3, 3).



Example 3.4.0

Consider a random sample of 400 from a population with a mean of 100 and SD of 15.

(a) According to the empirical rule, about 34% of the sample values fall in the interval (100, *x*). Find *x*.

We can reason our way through this. Since (100, x) captures 34% of the data points and 100 is the mean, *x* must be 1 SD above 100 (by the empirical rule). Since the SD is 15, we add 15 to 100 to get 1 SD above 100. And so, *x* = 115.

(b) What percentage of the data points have values less than 70?

70 is 30 below the mean of 100. That is two SDs. So 70 is 2 SD's below the mean. Recall that the empirical rule states about 95% of the data is within 2 SD's of the mean. This means that there is about 5% of the data outside of 2 SDs of the mean (in the tails). Half of that is in the right tail (below 2 SDs) and the other half is in the upper tail. That number is 2.5%.

(c) How many data points in the sample have values above 130?

130 is two SDs above 100. From what we did in (b) we see that 2.5% of the data points have values greater than 130, as this is the right-tail. So, we find 2.5% of 400 (our sample size). Multiply 400 by 0.025 to get 10. This means that about 10 data points are greater than 130.

_____

All that was fun, but what if the data values we're most interested in are not nice multiples of the SD? Well, luckily enough for us, people have already computed all the values we could ever be interested in and put them into a relatively easy to read table called the z-Table. So, we convert to z-scores and use this table; we will begin doing this in Chapter 6.

Chebyshev's Theorem deserves mention because it is actually quite remarkable. It can be thought of as an empirical-like rule that holds for any distribution. Chebyshev proved that, given any distribution, at least 75% of the data values will be within 2 SDs of the mean, at least 89% of the data values will be within 3 SD's of the mean, and at least 94% of the data values will be within 4 SD's of the mean.

This table gives you a guaranteed percentage of the data within a given number of standard deviations according to Chebyshev's rule.

Std Dev	Percentage
1.5	55.56%
2	75%
2.5	84%
3	88.89%
3.5	91.84%
4	93.75%