CHAPTER THREE: CENTRAL TENDENCY AND VARIATION — Review October 12 ~ Test October 19

The two primary descriptors of data sets are central tendency and variation. For central tendency, mean is most commonly used, but other measures such as median and trimmed mean are sometimes used as well. Unlike mean, these other measures are resistant, meaning they are not significantly affected by outliers. For variation, the most common are variance or its square root, standard deviation. These measures are fundamental in most statistical formulas. Other measures of variation include sum of squares, coefficient of variation, range, and interquartile range, the last two of which can be displayed visually in a box-and-whisker plot.

3-A Central Tendency — Monday • 10/3

sigma • mode • median • mean • trimmed mean

1. Calculate sums for data sets.
2. Find the median of a data set.
3. Find the p% trimmed mean of a data set.

3-B Variation — Wednesday • 10/5

sum of squares • variance • standard deviation • coefficient of variation

1. Calculate standard deviation, variance, and coefficient of variation.
2. Enter data into lists.
3. Calculate mean, standard deviation, and other statistics.
4. Distinguish between σ and s.

3-C Mean and Standard Deviation of Grouped Data — Friday • 10/7

weighted average

1. Calculate the mean and standard deviation of grouped data by hand.
2. Calculate the mean and standard deviation of grouped data by using a calculator table.
3. Calculate a weighted average.

3-D Percentiles and Quartiles — Wednesday • 10/12

percentile • quartile • range • interquartile range • box-and-whisker plot • outlier • resistant measure

1. Find the p^{th} percentile of a data set.
2. Find the quartiles and interquartile range of a data set.
3. Make a box-and-whisker plot.
4. Identify outliers in a data set.
5. Identify whether or not a measure is resistant.

3-A Central Tendency

Σ is the capital Greek letter SIGMA. It means *sum of*.

❶ Calculate sums for data sets.

1. Plug in each data value for x.
2. Add the resulting values.

❶ Calculate $\Sigma(x-10)^2$ for the data set { 15, 12, 3 }.

1. $(15-10)^2 = 5^2 = 25$
   $(12-10)^2 = 2^2 = 4$
   $(3-10)^2 = (-7)^2 = 49$
2. $25 + 4 + 49 = 78$

The MODE of a data set is the most common result (value, range, response, etc.). There may be more than one mode.

The MEDIAN of a data set is the middle value (or average of the two middle values).

The MEAN (average) of a data set is the sum of the data values divided by the sample size. $\bar{x} = \frac{\Sigma x}{n}$

The p% TRIMMED Mean of a data set is the mean after the highest p% and lowest p% of the data have been removed.

❷ Find the median of a data set.

1. Put the data values in order from lowest to highest
2. Identify the sample size n.
3. Calculate $i = \frac{n}{2}$. Round up.
4. If n is odd, the median is the i^{th} term in the data set.
   If n is even, the median is midway between term i and term i+1.

❷ Find the median of the data set { 40, 6, 6, 15, 24, 10 }.

1. 6, 6, 10, 15, 24, 40
2. $n = 6$
3. $i = \frac{6}{2} = 3$
4. median $= \frac{10+15}{2} = 12.5$

❸ Find the p% trimmed mean of a data set.

1. Identify the sample size n.
2. Calculate $T = p$% of n, and round it to the nearest integer.
3. Remove the highest T data values and the lowest T values from the data set.
4. Calculate the mean of the remaining data values.

❸ Find the 10% trimmed mean of the data set { 1, 3, 3, 5, 9, 10, 12, 12, 18, 21, 23, 30, 34, 40, 42, 50, 99 }.

1. $n = 17$
2. $T = .10(17) = 1.7 \approx 2$
3. Take out the 1, one of the 3's, the 50, and the 99: { 3, 5, 9, 10, 12, 12, 18, 21, 23, 30, 34, 40, 42 }.
4. $\bar{x} = \frac{259}{13} \approx 19.9$

3-B Variation

Four common measures of variability within a data set are sum of squares, variance, standard deviation, and coefficient of variation.

SUM OF SQUARES (SS) is the sum of the squared differences between each data value and the mean.

VARIANCE (σ^2 or s^2) is the average of the squares, that is, $SS \div n$ (but see below).

STANDARD DEVIATION (σ or s) is the square root of variance. σ is the standard deviation of an entire population. When only a sample is known, the sample standard deviation s can be found by dividing SS by $n - 1$ instead of by n. Though called *sample* standard deviation, s is not the standard deviation of a sample. Instead it is the best estimate of the standard deviation of the population based on data in a sample.

COEFFICIENT OF VARIATION (CV) is standard deviation divided by mean.

	Sum of Squares	Variance	Standard Deviation	Coefficient of Variation
Population:	$SS = \Sigma(x-\mu)^2$	$\sigma^2 = \frac{SS}{n}$	$\sigma = \sqrt{\sigma^2}$	$CV = \frac{\sigma}{\mu}$
Sample:	$SS = \Sigma(x-\bar{x})^2$	$s^2 = \frac{SS}{n-1}$	$s = \sqrt{s^2}$	$CV = \frac{s}{\bar{x}}$

❶ Calculate standard deviation, variance, and coefficient of variation.

1. Identify whether the data represent a sample of the population or the entire population.
2. Find the sample mean $\bar{x}$ or the population mean μ.
3. Subtract $\bar{x}$ or μ from each data value.
4. Square each result in step 3.
5. To get the sum of squares, add the squares in step 4.
6. To get the variance, divide the sum of squares in step 5 by n for a population or by $n - 1$ for a sample.
7. To get the standard deviation, take the square root of the variance in step 6.
8. To get the coefficient of variation, divide the standard deviation in step 7 by the mean.

❶ Find σ^2, s^2, σ, s, and the population and sample coefficient of variation for the data set { 11, 12, 14, 14, 19, 29, 34 }.

x	$x - \bar{x}$	$(x-\bar{x})^2$
11	-8	64
12	-7	49
14	-5	25
14	-5	25
19	0	0
29	10	100
34	15	225
$\Sigma x = 133$		$SS = \Sigma(x-\bar{x})^2 = 488$

$\bar{x} = 133 \div 7 = 19$

	Variance	Standard Deviation	Coefficient of Variation
Population:	$\sigma^2 = 488 \div 7 \approx 69.7$	$\sigma \approx \sqrt{69.7} \approx 8.35$	$CV \approx 8.35 \div 19 \approx .439$
Sample:	$s^2 = 488 \div 6 \approx 81.3$	$s \approx \sqrt{81.3} \approx 9.02$	$CV \approx 9.02 \div 19 \approx .475$

The TI-83 can calculate standard deviation and other statistics.

❷ Enter data into lists.

1. Push [STAT] and choose `EDIT`.
2. If needed, clear any previous data by moving the cursor onto `L1` and pushing [CLEAR] and [ENTER].
3. Type each data value followed by [ENTER].
4. If there is more than one list, move the cursor right and repeat steps 2 and 3 in `L2` and up to four additional lists.

❸ Calculate mean, standard deviation, and other statistics.

1. Enter the data into a list (see ❷).
2. Push [STAT] and choose `CALC` and `1-Var Stats`.
3. Push [2nd] and the list number. (You can skip this step if the list is `L1`.)
4. Push [ENTER].

Note that the calculator says `Sx` and σ`x` instead of s and σ.

❹ Distinguish between σ and s.

1. To find the standard deviation of the available data, find σ.
   To estimate the standard deviation of all possible data based on a sample, find s. This is much more common in a statistics course.

❹ Smith Safety Supplies manufactures supplies. Jackson times the burn duration of 5 emergency flares (in seconds): 812, 995, 909, 844, 960.

a) Calculate σ and s

$\sigma \approx 68.6$, $s \approx 76.7$ (see ❸)

b) Identify what σ and s represent

σ is the standard deviation of these five flares' burn times.

s is the best estimate of the standard deviation of all Smith Supplies flares' burn times.

c) State which value is relevant in this context.

Jackson wants to know about all Smith Supplies flares, not just the five he burned, so he should use $s \approx 76.7$ as an estimate.

3-C Mean and Standard Deviation of Grouped Data

When a data set has multiple instances of a single value, it is easiest to calculate the mean by first multiplying each value x by its frequency f. Then the mean can be calculated as $\mu = \frac{\Sigma fx}{\Sigma f}$, and the sum of squares can be calculated multiplying each squared difference by f: $SS \approx \Sigma f(x - \bar{x})^2$. A common occurrence of this is when a distribution of grouped data is known, such as from a histogram, but the individual data are not known.

❶ Calculate the mean and standard deviation of grouped data by hand.

1. Identify the value, average value, or midpoint x for each category.
2. Count the frequency f of each value x.
3. To get the sample size, find the total frequency by adding the f values: $n = \Sigma f$.
4. For each group, multiply x by f to calculate fx.
5. Add up the fx values to calculate Σfx.
6. To get the mean, divide Σfx by n.
7. Subtract the mean from each x value.
8. Square each difference in step 7.
9. Multiply each square in step 8 by f for that value.
10. To get the sum of squares, add the products in step 9.
11. To get the variance, divide SS from step 10 by n if you have the whole population, or by $n - 1$ if you have a sample.
12. To get the standard deviation, take the square root of the variance in step 11.

❶ In PreCalculus there were 4 D's, 9 C's, 7 B's, and 16 A's. Calculate an estimate of the class average and standard deviation, using 65 as the average D grade, 75 as the average C grade, etc.

category	f	x	fx	$x - \mu$	$(x - \mu)^2$	$f(x - \mu)^2$
D	4	65	260	-19.7	388.1	1552.4
C	9	75	675	-9.7	94.1	846.9
B	7	85	595	0.3	0.1	0.7
A	16	95	1520	10.3	106.1	1697.6
	$n = \Sigma f = 36$		$\Sigma fx = 3050$			$SS \approx 4097.6$
			$\mu \approx \frac{3050}{36} \approx 84.7$			$\sigma \approx \sqrt{\frac{4097.6}{36}} \approx 10.7$

For large tables, it is faster to do the calculations in calculator lists, rather than one value at a time. This can be done by using the variable list names as variables in later list names. If you know how to use formulas in spreadsheets, statistical tables are much easier to calculate using spreadsheets than by hand or with a calculator, and they can be easily reused for new problems.

❷ Calculate the mean and standard deviation of grouped data by using a calculator table.

1. Enter the frequencies in `L1` and the values in `L2` (see ❷ in 3-B).
2. Change `L3` to `L1L2`.
3. Calculate the mean μ for `L3` (see ❸ in 3-B).
4. Change `L4` to `L2`$-\mu$, using the μ calculated in step 3.
5. Change `L5` to `L4`^{2}.
6. Change `L6` to `L1L5`.
7. Find the sum n for `L1` and the sum SS for `L6` (see ❸ in 3-B).
8. Calculate $\sigma = \sqrt{\frac{SS}{n}}$.

In a WEIGHTED Average, some values are given more importance than others. It is calculated the same as the mean in ❶, except that f represents weightings (proportions, such as 20%) rather than frequencies (counts, such as 5), and since the total weighting must be 100%, $\Sigma f = 1$ and therefore $\mu = \Sigma fx$.

❸ Calculate a weighted average.

1. List the weighting f and the value x for each category.
2. Calculate fx for each category.
3. Calculate the sum $\mu = \Sigma fx$.

❸ A calculus professor weights homework 10%, the midterm 40%, and the final 50%. Calculate Casey's grade if she scores 82 on the homework, 74 on the midterm, and 95 on the final.

category	f	x	fx
homework	.10	82	8.2
midterm	.40	74	29.6
final	.50	95	47.5
	$\Sigma f = 1.00$		$\mu = \Sigma fx = 85.3$

3-D Percentiles and Quartiles

Conceptually, the p^{th} PERCENTILE is a value below which are p% of the data in a data set. One definition is the i^{th} data value, where $i = \frac{1}{2} + p\%$ of n.

❶ Find the p^{th} percentile of a data set.

1. Put the data in order and count them.
2. Calculate $i = \frac{1}{2} + p\%$ of n.
3. Count to the i^{th} data value in the data set.
4. If i is not an integer, add rd to the lower data value, where r is the decimal remainder and d is the difference between the two data values.

❶ The scores on an Algebra II test were 30, 43, 50, 59, 60, 62, 62, 65, 70, 70, 72, 72, 74, 80, 81, 82, 86, 86, 86, 87, 90, 90, 91, 93, 94, 98, 101, 102, and 106. Find the 10^{th} percentile.

1. $n = 29$
2. $i = \frac{1}{2} + .10(29) = 3.4$
3. The 10^{th} percentile is between 50 and 59.
4. 10^{th} percentile $= 50 + .4(59 - 50) = 53.6$

QUARTILES divide a data set into four equal groups.

The first quartile (Q_1) is the 25^{th} percentile.

The second quartile (Q_2) is the 50^{th} percentile (the median).

The third quartile (Q_3) is the 75^{th} percentile.

The RANGE of a data set is the total spread, from highest to lowest: range = highest value – lowest value.

The INTERQUARTILE Range (IQR) is the range of the middle half of the data: $IQR = Q_3 - Q_1$.

❷ Find the quartiles and interquartile range of a data set.

1. Put the data in order.
2. Q_2 is the median.
3. Q_1 is the median of the data values below Q_2.
4. Q_3 is the median of the data values above Q_2.
5. Subtract Q_1 from Q_3 to get the interquartile range.

❷ { 4, 6, 6, 8, 9, 9, 9, 9, 10, 12, 12, 19, 25 }

2. Q_2 is the median of { 4, 6, 6, 8, 9, 9, 9, 9, 10, 12, 13, 19, 25 }, which is 9.
3. Q_1 is the median of { 4, 6, 6, 8, 9, 9 }, which is 7.
4. Q_3 is the median of { 9, 10, 12, 13, 19, 25 }, which is 12.5.
5. $IQR = 12.5 - 7 = 5.5$

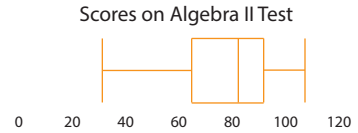
A BOX-AND-WHISKER Plot shows the median in a box from Q_1 to Q_3, and shows whiskers below and above it to the lowest and highest data values.

❸ Make a box-and-whisker plot.

1. Identify the quartiles and the highest and lowest value of the data set (see ❷).
2. Draw and scale an axis. It can be horizontal or vertical.
3. Draw a line at each of the five values identified in step 1.
4. Make a box between Q_1 and Q_3.
5. Title the plot.

❸ Make a box-and-whisker plot for the Algebra II test scores in ❶.

1. $Q_2 = 81$, $Q_1 = 63.5$, $Q_3 = 90.5$, low = 30, high = 106



An OUTLIER is a value that is much further from the mean than most or all other data. One definition of outlier is a value below Q_1 or above Q_3 by more than 1.5 times the interquartile range.

❹ Identify outliers in a data set.

1. Identify Q_1, Q_3, and IQR (see ❷).
2. Calculate $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.
3. Outliers are any data points not between the answers in step 2.

❹ Identify any outliers in the following Algebra II chapter one homework scores: 6, 11, 21, 25, 26, 27, 33, 34, 34, 35, 36, 37, 38, 43, 43, 47, 47, 48, 48, 49, 50, 51, 52, 52, 52, 53, 54, 54, 55.

1. $Q_1 = 33.5$, $Q_3 = 51.5$, $IQR = 51 - 34 = 18$
2. $33.5 - 1.5(18) = 6.5$
   $51.5 + 1.5(18) = 78.5$
3. The score of 6 is not between 6.5 and 78.5 so it is an outlier.

A RESISTANT Measure is one that is not significantly affected by outliers.

❺ Identify whether or not a measure is resistant.

1. If the calculation uses the outlier, the measure is not resistant.

❺ Use the data set { 3, 5, 5, 9, 10, 11, 12, 12, 13, 80 } to explain why 10% trimmed mean is resistant but mean is not.

The outlier 80 is part of the calculation for the mean: $(3 + 5 + 5 + 9 + 10 + 11 + 12 + 12 + 13 + 80) \div 10 = 16$, but not for the trimmed mean: $(5 + 5 + 9 + 10 + 11 + 12 + 12 + 13) \div 8 = 9.625$ or for the median: $(10 + 11) \div 2 = 10.5$.