

Measures of Spread – Part 1 (Measures of Spread Related to the Mean)

When one uses a measure of central tendency (such as a mean and mode) the statistic has little validity and certainly lacks predictive ability without also considering how dispersed or spread out the data is around the mean.

Consider these two data sets ...

eg – 1, 10, 1, 20

$$\bar{x} = \frac{32}{4} = 8$$

eg – 8, 9, 8, 7

$$\bar{x} = \frac{32}{4} = 8$$

Given the above data sets ... which mean of 8 is more valid? This situation certainly requires us to create a measurement that assures that a mean of 8 is either wildly unpredictable or almost a certain result. We approach a situation where a measure of spread must accompany the mean.

Deviation

This is the degree or amount to which a value lies above or below the mean. Obviously the lower the deviation the more predictive (and valid) the mean.

$$\text{deviation} = x_i - \bar{x}$$

x_i	deviation
1	$1 - 8 = -7$
10	$10 - 8 = 2$
1	$1 - 8 = -7$
20	$20 - 8 = 12$

Sum of deviations = 0

x_i	deviation
8	$8 - 8 = 0$
9	$9 - 8 = 1$
7	$7 - 8 = -1$
8	$8 - 8 = 0$

Sum of deviations = 0

Clearly the lesser deviation is seen in the second example however if we sum the degree of deviation to quantify the variability we still seem to get a result of zero. This is largely due to the negative and positive deviations that sum to zero so perhaps taking absolute value of these results would be more effective.

Sum of deviations (all positive) = 28

Sum of deviations (all positive) = 2

This statistic (however effective) can certainly be affected by the number of data values. Deviations (even in closely spaced data as in eg 2) would add up quickly if we had 20 or 30 data values (especially in variable data as in eg 1). This leads to a more evolved version called the average deviation where essentially the mean deviation is compiled.

$$\text{mean deviation} = \frac{\sum |x_i - \bar{x}|}{n \text{ or } N}$$

— Scaled value
(again ... typical statistic)

Variance

Another measure of variability is called the **variance** (σ^2 or s^2).

population

sample

$$\sigma^2 = \frac{\sum(x_i - \bar{\mu})^2}{N}$$

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

note diff

Instead of using the absolute value function the deviation is squared which accomplishes two tasks quite well ...

- 1) The squaring makes positives positive and negatives positive (like absolute value)
- 2) The squaring **amplifies** large deviations making their deviations HUGE and takes very small deviations and makes them MINISCULE. This accentuates deviations to create a more pronounced measure of spread as a statistic.

In addition we see a subtle **variation between the population and the sample** variance formulas.

The sample variance is always divided by a smaller number ($n - 1$ rather than N) thereby keeping the variance larger. This difference is unimportant if the sample size is large and therefore large enough to properly account for the variation of the sample however for very small samples this significance is large enough to overestimate the variance and create a less valid measure of central tendency. This is to account for the lost or unaccounted for variation in the missing members of the population. If the data set IS the population ... there is no variation and as such an overestimation is unnecessary.

The variance is an exceptionally important measure in the statistics of variability however in a measure of deviation from a mean it is much too large of a value to effectively communicate how far a value is from the mean on average ... for this we use the standard deviation from the mean.

Raw Standard Deviation

This statistic quantifies how far a point is expected (on average) to be above and below the mean (like a +/- error bar of sorts). It will be used as a measure of spread as well as later on as a component of the z-score and the normal distribution of continuous data.

population

sample

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{\mu})^2}{N}}$$

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

As you can see the standard deviation is simply the square root of the variance!

Watch order! ① → ⑦

ORGANIZE IN A TABLE!!!

Consider the following data set ... (rushing yards for M. Lynch in 2015 season ... 1st six games)

110, 76, 150, 103, 54, 125

Game	Yards, x_i	Deviation, $x_i - \bar{x}$	$(x_i - \bar{x})^2$
v. Falcons	110	7	49
v. Chargers	76	-27	729
v. Bears	150	47	2209
v. Steelers	103	0	0
v. Cardinals	54	-49	2401
v. Vikings	125	22	484
SUM, Σ	618	0	5872

For the mean ...

$$\bar{x} = \frac{618}{6} = 103$$

For the Standard Deviation ...

$$s = \sqrt{\frac{5872}{6-1}} = \sqrt{1174} \approx 34.3$$

So ... The mean yards per game for Marshawn Lynch is 103 +/- 34.3 yards!

Grouped Standard Deviation

population

$$\sigma = \sqrt{\frac{\sum f_i(m_i - \bar{\mu})^2}{N \text{ or } f_i}}$$

sample

$$s = \sqrt{\frac{\sum f_i(m_i - \bar{x})^2}{n-1}}$$

Remember our mark set ...

Again ① →

Group or Interval	m_i	f_i	$f_i m_i$	$m_i - \bar{x}$	$(m_i - \bar{x})^2$	$f_i(m_i - \bar{x})^2$
39.5 to 49.5	44.5	1	44.5	-28.125	791.0 ...	791.0 ...
49.5 to 59.5	54.5	4	218	-18.125	328.5 ...	1314.0 ...
59.5 to 69.5	64.5	7	451.5	-8.125	66.0 ...	462.1 ...
69.5 to 79.5	74.5	11	819.5	1.875	3.5 ...	38.6 ...
79.5 to 89.5	84.5	6	507	11.875	141.0 ...	846.0 ...
89.5 to 99.5	94.5	3	283.5	21.875	478.5 ...	1435.5 ...
Sum		32	2324	n/a	n/a	4887.5

For the mean ...

$$\bar{x}_g = \frac{2324}{32} = 72.625$$

For the Standard Deviation ...

$$s = \sqrt{\frac{4887.5}{32-1}} = \sqrt{157.6} \approx 12.6$$

So, The mean test score for the class grouped data is 72.6 +/- 12.6 %!

z – score

The z-score is calculated for any and all data values. Each datum can be used to calculate its own z-score and this value represents the distance that the point is away from the mean scaled to (or in proportion to) the standard deviation.

That means that the z-score communicates the deviation an individual value is from the mean in units that are standard deviations.

Population

$$Z = \frac{x_i - \mu}{\sigma}$$

Sample

$$Z = \frac{x_i - \bar{x}}{s}$$

eg – from our ungrouped data ... recall that $\bar{x} = 103$ and $s = 34.3$

a) For $x_i = 76$

$$Z = \frac{x_i - \bar{x}}{s}$$

$$Z = \frac{76 - 103}{34.3}$$

$$z = -0.79$$

So, the value of 76 yards is a game that is 0.79 standard deviations **below** the mean.

b) For $x_i = 150$

$$Z = \frac{x_i - \bar{x}}{s}$$

$$Z = \frac{150 - 103}{34.3}$$

$$z = +1.37$$

So, the value of 150 yards is a game that is 1.37 standard deviations **above** the mean.

HW - pg 148 #'s 1, 5, 6a, 7a, 9a (mean and sd only), 9c, 9d, 19ab

Interquartile Range

$$Q_2 = \frac{1}{2}(11+1) = 6 \rightarrow 6^{\text{th}}$$

The measure of spread that relates to the median is the interquartile range. This stat relies on dividing the data set into four equal groups called quartiles.

eg - 1, 1, 2, 3, 4, 4, 7, 8, 11, 12, 12

Orders ... $Q_1 = \frac{1}{4}(n+1)$

$$Q_3 = \frac{3}{4}(n+1)$$

$$Q_2 = \frac{1}{2}(n+1)$$

Since there are 11 data values the median is @ the 6th datum having a value of 4

This divides the data set into 5 below and 5 above the median. The median of the first half of the data set is the first quartile ...

$$Q_1 = \frac{1}{4}(11+1) = \frac{1}{4}(12) = 3 \rightarrow 3^{\text{rd}}$$

1, 1, 2, 3, 4 ... 4, ... and is found at the 3rd datum. The value of the first quartile (Q_1) is 2.

$$Q_3 = \frac{3}{4}(11+1) = 9 \rightarrow 9^{\text{th}}$$

The median of the second half of the data set is the 3rd quartile ...

... 4, ... 7, 8, 11, 12, 12 ... and is found at the 9th datum. The value of the third quartile (Q_3) is 11

Q_1 Q_3

The **interquartile range** (IQR) is the range between the third and first quartiles.

$$IQR = Q_3 - Q_1 = 11 - 2 = 9$$

So, the interquartile range is 9 units.

The **semi-quartile range** (SQR) is half of the IQR ... so the SQR is 4.5

$$SQR = \frac{1}{2} IQR$$

If we alter the data set slightly we introduce a few new ideas ...

eg - 1, 1, 2, 3, 4, 4, 7, 8, 11, 12, 12, 40

$$Q_1 = \frac{2+3}{2} = 2.5$$

$$\text{median}, Q_2 = \frac{4+7}{2} = 5.5$$

$$Q_3 = \frac{11+12}{2} = 11.5$$

$$IQR = Q_3 - Q_1 = 11.5 - 2.5 = 9$$

$$SQR = \frac{IQR}{2} = 4.5$$

Box and Whisker Plots

We can graph this data distribution in a visual display called a modified box-and-whisker plot.

- 1) Create a number line that spans all data.
- 2) Draw a box Q_3 at one end and Q_1 at the other.
- 3) Place Q_2 as a vertical line bisecting the box at the appropriate location on the number line.
- 4) Determine the acceptable data points and the outliers.

An outlier is a data value that is ...

$> 1.5 \times IQR$ larger than Q_3 OR
 $< 1.5 \times IQR$ smaller than Q_1

defines an acceptable range

This creates an acceptable range where data points can lie as follows:

$$Q_1 - 1.5 \times IQR \leftarrow \text{Acceptable Data values} \rightarrow Q_3 + 1.5 \times IQR$$

does not define whiskers

Anything outside this value is considered an **outlier**.

- 5) Create one whisker that reaches from the first acceptable value to the box and from the box to the last acceptable value. Outliers are plotted as individual points.

eg - 1, 1, 2, 3, 4, 4, 7, 8, 11, 12, 12, 40

Recall,

$$Q_1 = \frac{2+3}{2} = 2.5$$

$$\text{median, } Q_2 = \frac{4+7}{2} = 5.5$$

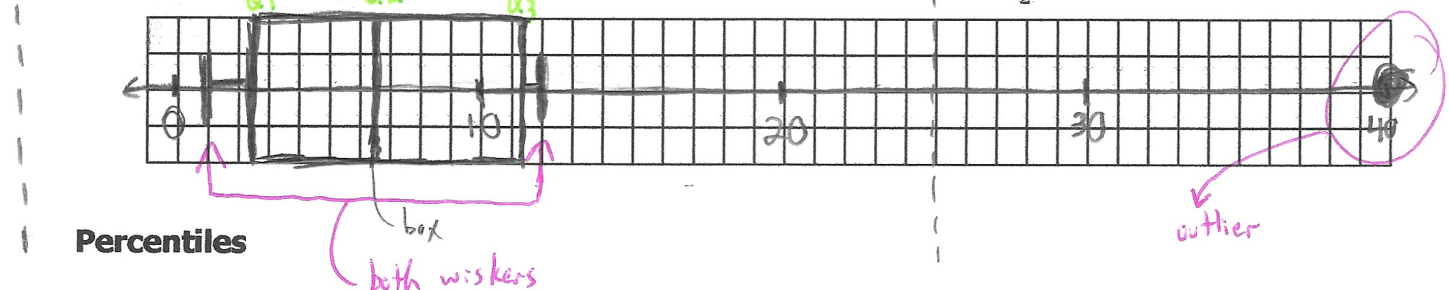
$$Q_3 = \frac{11+12}{2} = 11.5$$

$$\text{Top of range} = Q_3 + 1.5 \times IQR = 11.5 + 1.5(9) = 25$$

$$\text{Bottom of range} = Q_1 - 1.5 \times IQR = 2.5 - 1.5(9) = -11$$

$$IQR = Q_3 - Q_1 = 11.5 - 2.5 = 9$$

$$SQR = \frac{IQR}{2} = 4.5$$



Percentiles

A percentile is basically the same as a quartile where Q_1 is the 25th percentile, the median is the 50th percentile and Q_3 is the 75th percentile. A percentile calculation communicates the value where n% of the data is that value or less.

To calculate the percentile, p , of a particular datum, x_i , we use the following formula:

$$p = 100 \left(\frac{(L + 0.5E)}{n} \right)$$

where

- L = the **number** of values **below** x_i
- E = the **number** of values **equal** to x_i
- n = the total number of data values in the entire sample

To calculate the percentile rank (position of value that defines the percentile), R , we use the following formula:

$$R = \frac{p}{100}(n + 1)$$

where p = the percentile you are looking for
 n = the total number of data values in the entire sample

NOTE – whatever value you obtain ... you round down and the percentile is between that value and the next value greater.

Consider the data set below:

10	28	44	66	85
12	30	45	67	87
15	31	47	70	89
17	33	50	75	90
20	34	55	76	92
21	35	56	77	93
23	37	57	78	94
24	39	60	80	94
25	40	62	82	98
27	42	63	84	99

a) Find the value corresponding to the 60th percentile.

The data set has 50 values so the 60th percentile is located @ ...

$$R = \frac{60}{100}(50 + 1) = 30.6$$

So the 60th percentile is located between the 30th and the 31st datum ... so between the values 60 and 63. The 60th percentile is located at 61.5.

$$\frac{60 + 63}{2} = 61.5$$

b) A measurement of 80 is obtained. To find the percentile score ...

$$L = 37 \quad E = 1 \quad n = 50$$

$$p = 100 \left(\frac{(37 + 0.5(1))}{50} \right) = 75 \text{ or the } 75^{\text{th}} \text{ percentile}$$

Homework - (Handout from Day 3) #2, 3, 4, 6bcd, 7b, 9ab, 14