Measures of Central Tendency and Variability: Summarizing your Data for Others

I. Measures of Central Tendency:

-Allow us to summarize an entire data set with a single value (the midpoint).

1. Mode : The value (score) that occurs most often in a data set.
   - Mo_x = Sample mode -Mo = Population mode

2. Median : the point (score) which divides the data set in ½ : e.g. ½ of the subjects are above the median and ½ are below the median.
   - Mdn_x= Sample Median -Mdn = Population Median

3. Mean: the arithmetic average: Directly considers every score in a distribution.

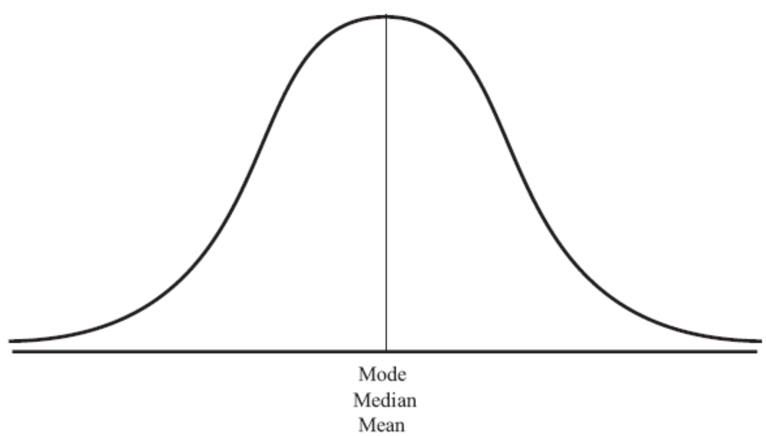
- $\overline{X} = \frac{\sum X}{n}$ = Sample Mean - $\mu = \frac{\sum X}{N}$ = Population Mean - $n\overline{X} = \sum X$

II. Skewed Distributions & the 3M's

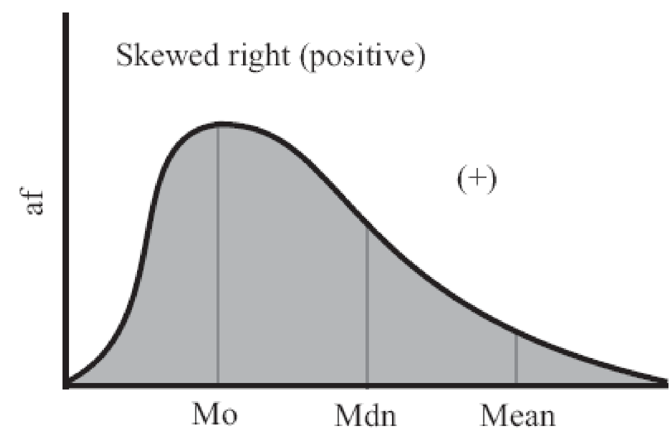
-Skewness refers to the shape of the distribution which can be influenced by extreme scores.

- Skewness is also an estimate of the deviation of the Mean, Median, and Mode.

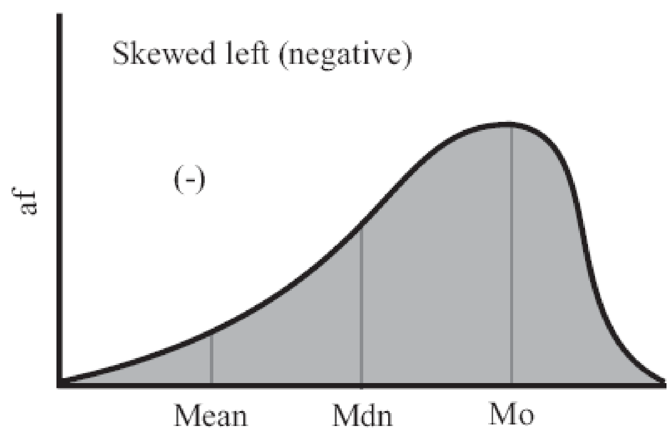
-Symmetrical Dist. = Mean, Median, Mode are all in the same location in the dist.



-Skewed Right (Positively Skewed) = Mode in peak of dist.(left of center), Median in center of distribution, Mean in right tail of distribution.



-Skewed Left (Negatively Skewed) = Mode in peak of dist (right of center), Median in center of distribution, Mean in left tail of distribution.



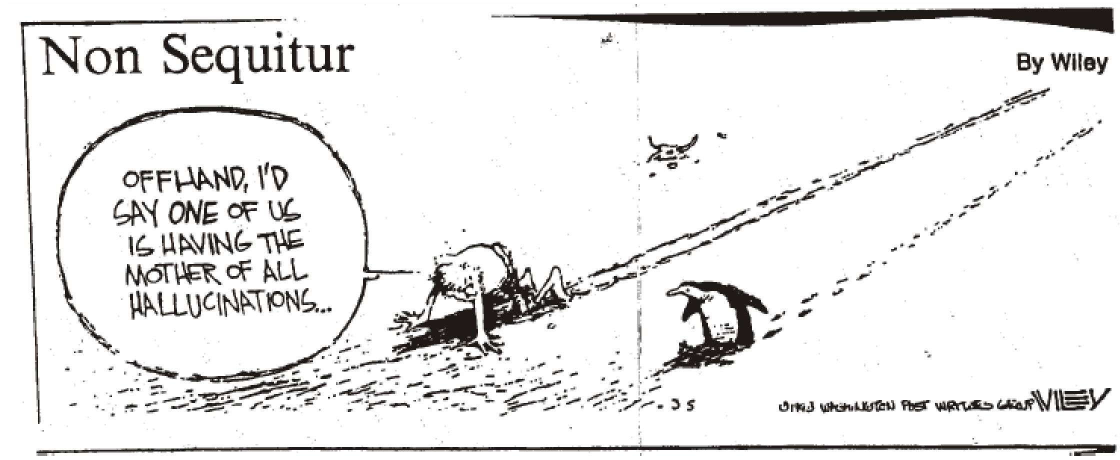
I . Measures of Variability (Dispersion)

-Allow us to summarize our data set with a single value.

-Central Tendency + Variability = a more accurate picture of our data set.

-The 3 main measures of variability: **Range**, **Variance**, and **Standard Deviation**.

- These formulas are the root formulas for many of the statistical tests that will be covered later
  - *t*-test, ANOVA, and Correlation
- Tell us how much observations in a data set ***vary*** (differ from one another)
  - How are they ***dispersed*** within the distribution?


Non Sequitur
By Wiley
OFFHAND, I'D SAY ONE OF US IS HAVING THE MOTHER OF ALL HALLUCINATIONS...
WILEY

-Although measures of central tendency tell summarize some aspects of our data, they don't tell us much about the variance within our data.

Example.

Number of miles traveled before traveling companion appears human n=8

Mean = 5, Mode = 5, Median = 5 for both data sets (They do not differ)

-all zoo penguins hallucinate after traveling 5 mile, while there is much more variability in the distances traveled by South Pole Penguins.

-In order to draw accurate conclusions about our data both central tendency and variability must be considered.

Zoo Penguins	South Pole Penguins
5	2
5	3
5	4
5	5
5	5
5	6
5	7
5	8
$\Sigma X = 40$	$\Sigma X = 40$

II. **<u>Range</u>** : The numerical distance between the largest (***X maximum***) and smallest values (***X minimum***), tells us something about the variation in scores we have in our data, or it tells us the width of our data set.

Range = X maximum - X minimum

- Range for Zoo penguins = 5-5 = 0
- Range for South Pole P's = 8-2 = 6

-Problems with Range:

- Does not directly consider every value in the data set
  - (here only the two extreme numbers; largest and smallest).
- We do not know whether most of the scores occur at the extremes of the distribution or toward the center.

For example:

Zoo Penguins	South Pole Penguins	North Pole Penguins
5	2	2
5	3	5
5	4	5
5	5	5
5	5	5
5	6	5
5	7	5
5	8	8
$\Sigma X = 40$	$\Sigma X = 40$	$\Sigma X = 40$
Range = 0	Range = 6	Range = 6

III. Variance = indicates the total amount of variability (differences between scores) in a data set by directly considering every observation.

-Requires a point from which each observation can be compared to assess the amount they differ.

-The Mean can be used as a point of comparison, since it considers every observation in its calculation.

Mean Deviation = $\Sigma(X - \bar{x})$

Zoo Penguins	X - Mean		South Pole Penguins	X - Mean		North Pole Penguins	X - Mean
5	0		2	-3		2	-3
5	0		3	-2		5	0
5	0		4	-1		5	0
5	0		5	0		5	0
5	0		5	0		5	0
5	0		6	1		5	0
5	0		7	2		5	0
5	0		8	3		8	3
ΣX = 40	Σ(X-Mean) = 0		ΣX = 40	Σ(X-Mean) = 0		ΣX = 40	Σ(X-Mean) = 0
Mean = 5			Mean = 5			Mean = 5	
Range = 0			Range = 6			Range = 6	

The sum of the mean deviation for any data set is always 0. This limits the usefulness of the mean deviation for summarizing different data sets with a single point.

if we square each deviation value then the negative values cancel out and we are left with a more meaningful value.

Zoo Penguins	X - Mean	$(X - Mean)^2$	South Pole Penguins	X - Mean	$(X - Mean)^2$
5	0	0	2	-3	9
5	0	0	3	-2	4
5	0	0	4	-1	1
5	0	0	5	0	0
5	0	0	5	0	0
5	0	0	6	1	1
5	0	0	7	2	4
5	0	0	8	3	9
$\Sigma X = 40$	$\Sigma(X\text{-Mean}) = 0$	$\Sigma(X\text{-Mean})^2 = 0$	$\Sigma X = 40$	$\Sigma(X\text{-Mean}) = 0$	$\Sigma(X\text{-Mean})^2 = 28$
Mean = 5			Mean = 5		
Range = 0			Range = 6		

North Pole Penguins	X - Mean	$(X - Mean)^2$
2	-3	9
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
8	3	9
$\Sigma X = 40$	$\Sigma(X\text{-Mean}) = 0$	$\Sigma(X\text{-Mean})^2 = 18$
Mean = 5		
Range = 6		

-If we sum these values we no longer get 0, but a number that reflects the total variance for this data set,

-if we divide that number by N or n we get the average variance for this data set.

Definitional Population Formula = $\sigma^2 = \frac{\Sigma(X - \text{Mean})^2}{N}$

Definitional Sample Formula = $s^2 = \frac{\Sigma(X - \text{Mean})^2}{n-1}$

Note sample variance uses n-1 rather than N because it is an estimate of the population variance.

Due to the smaller denominator, the sample variance will always be slightly larger than the population variance.

Zoo Penguins	X - Mean	(X - Mean)^2
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
ΣX = 40	Σ(X-Mean) = 0	Σ(X-Mean)^2 = 0

Mean = 5
Range = 0

Zoo Penguins =

$$\sigma^2 = \frac{\sum\left(X - \overline{X}\right)^2}{N} = \frac{0}{8} = 0$$

$$s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n-1} = \frac{0}{8-1} = \frac{0}{7} = 0$$

South Pole Penguins	X - Mean	$(X - Mean)^2$
2	-3	9
3	-2	4
4	-1	1
5	0	0
5	0	0
6	1	1
7	2	4
8	3	9
ΣX = 40	Σ(X-Mean) = 0	$\Sigma(X\text{-Mean})^2 = 28$

Mean = 5
Range = 6

South Pole Penguins =

$$\sigma^2 = \frac{\sum \left(X - \overline{X}\right)^2}{N} = \frac{28}{8} = 3.5$$

$$s^2 = \frac{\sum \left(X - \overline{X}\right)^2}{n-1} = \frac{28}{8-1} = \frac{28}{7} = 4$$

North Pole Penguins	X - Mean	$(X - Mean)^2$
2	-3	9
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
8	3	9
ΣX = 40	Σ(X-Mean) = 0	$\Sigma(X\text{-Mean})^2 = 18$

Mean = 5
Range = 6

North Pole Penguins =

$$\sigma^2 = \frac{\sum \left(X - \overline{X}\right)^2}{N} = \frac{18}{8} = 2.25$$

$$s^2 = \frac{\sum \left(X - \overline{X}\right)^2}{n - 1} = \frac{18}{8 - 1} = \frac{18}{7} = 2.5714$$

- Definitional Formula is time consuming for large data sets.
- We have developed mathematically identical (algebraically equivalent) formulas that are easier to calculate.

Computational Formulas=

Population Variance $= \sigma^2 = \dfrac{\Sigma X^2 - (\Sigma X)^2/N}{N}$

Sample Variance $= s^2 = \dfrac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}$

-Note sample variance uses n-1 rather than N because it is an estimate of the population variance. Due to this reduced denominator the sample variance will always be slightly larger than the population variance.

	Zoo Penguins	X^2	South Pole Penguins	X^2	North Pole Penguins	X^2
	5	25	2	4	2	4
	5	25	3	9	5	25
	5	25	4	16	5	25
	5	25	5	25	5	25
	5	25	5	25	5	25
	5	25	6	36	5	25
	5	25	7	49	5	25
	5	25	8	64	8	64
ΣX	40		40		40	
ΣX^2		200		228		218

Zoo Penguins:

$$\sigma^2 = \frac{\sum X^2 - \frac{\left(\sum X\right)^2}{N}}{N} = \frac{200 - \frac{(40)^2}{8}}{8} = \frac{200 - \frac{1600}{8}}{8} = \frac{200 - 200}{8} = \frac{0}{8} = 0$$

$$s^2 = \frac{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}{n - 1} = \frac{200 - \frac{(40)^2}{8}}{8 - 1} = \frac{200 - \frac{1600}{8}}{7} = \frac{200 - 200}{7} = \frac{0}{7} = 0$$

	Zoo Penguins	X^2	South Pole Penguins	X^2	North Pole Penguins	X^2
	5	25	2	4	2	4
	5	25	3	9	5	25
	5	25	4	16	5	25
	5	25	5	25	5	25
	5	25	5	25	5	25
	5	25	6	36	5	25
	5	25	7	49	5	25
	5	25	8	64	8	64
ΣX	40		40		40	
ΣX^2		200		228		218

South Pole Penguins:

$$\sigma^2 = \frac{\sum X^2 - \frac{\left(\sum X\right)^2}{N}}{N} = \frac{228 - \frac{(40)^2}{8}}{8} = \frac{228 - \frac{1600}{8}}{8} = \frac{228 - 200}{8} = \frac{28}{8} = 3.5$$

$$s^2 = \frac{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}{n-1} = \frac{228 - \frac{(40)^2}{8}}{8-1} = \frac{228 - \frac{1600}{8}}{7} = \frac{228 - 200}{7} = \frac{28}{7} = 4$$

	Zoo Penguins	X^2	South Pole Penguins	X^2	North Pole Penguins	X^2
	5	25	2	4	2	4
	5	25	3	9	5	25
	5	25	4	16	5	25
	5	25	5	25	5	25
	5	25	5	25	5	25
	5	25	6	36	5	25
	5	25	7	49	5	25
	5	25	8	64	8	64
ΣX	40		40		40	
ΣX^2		200		228		218

$$\sigma^2 = \frac{\sum X^2 - \frac{\left(\sum X\right)^2}{N}}{N} = \frac{218 - \frac{(40)^2}{8}}{8} = \frac{218 - \frac{1600}{8}}{8} = \frac{218 - 200}{8} = \frac{18}{8} = 2.25$$

North Pole Penguins:

$$s^2 = \frac{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}{n - 1} = \frac{218 - \frac{(40)^2}{8}}{8 - 1} = \frac{218 - \frac{1600}{8}}{7} = \frac{218 - 200}{7} = \frac{18}{7} = 2.5714$$

- -Problems: This formula is the base for many other statistical formulas, however as a single summary measure it has little numerical meaning until it is converted to a standardized score.
  – Right now it represents the average distance each penguin is from the mean, in squared mile units.

3. **Standard Deviation**= The square root of a variance.
   - The standardized variance value.
   - It provides us with a numerically meaningful measure of variance:
     -The average distance each observation is from the mean.
   -This value (when combined with other stats methods) allow us to infer what percentage of our observations are a certain distance from the mean.

Standard Deviation (based on computational formula of variance)

Population St. Dev. :

$$\sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}}$$

$$\sigma = \sqrt{\sigma^2}$$

Sample St. Dev. :

$$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}}$$

$$s = \sqrt{s^2}$$

$$\text{Zoo Penguins}: \ \sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{0} = 0$$

$$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)}{n}}{n-1}}$$

$$s = \sqrt{s^2} = \sqrt{0} = 0$$

With respect to sample standard deviations (*s*), we can say:

-Zoo penguins are an average of 0 miles from the mean number of miles walked before hallucinating

South Pole Penguins : $\sigma = \sqrt{\dfrac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}}$

$$\sigma = \sqrt{\sigma^2} = \sqrt{3.5} = 1.8708$$

$$s = \sqrt{\dfrac{\Sigma X^2 - \frac{(\Sigma X)}{n}}{n - 1}}$$

$$s = \sqrt{s^2} = \sqrt{4} = 2$$

With respect to sample standard deviations (*s*), we can say:

- South Pole penguins are an average of 2 miles from the mean number of miles walked before hallucinating.

$$\text{North Pole Penguins : } \sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.25} = 1.5$$

$$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)}{n}}{n - 1}}$$

$$s = \sqrt{s^2} = \sqrt{2.5714} = 1.6935$$

With respect to sample standard deviations (*s*), we can say:

- North Pole penguins are an average of 1.69 miles from the mean number miles walked before hallucinating.