Variability

2.2

Frequency Distributions

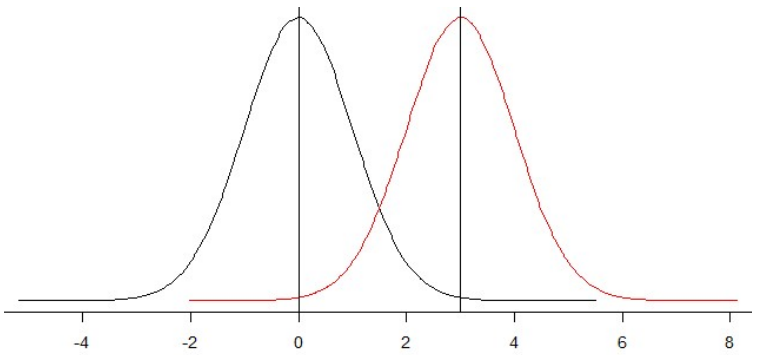
- In the previous video I mentioned that you can summarize frequency distributions (and therefore data), using:

1. A measure of central tendency

Frequency Distributions

- In the previous video I mentioned that you can summarize frequency distributions (and therefore data), using:

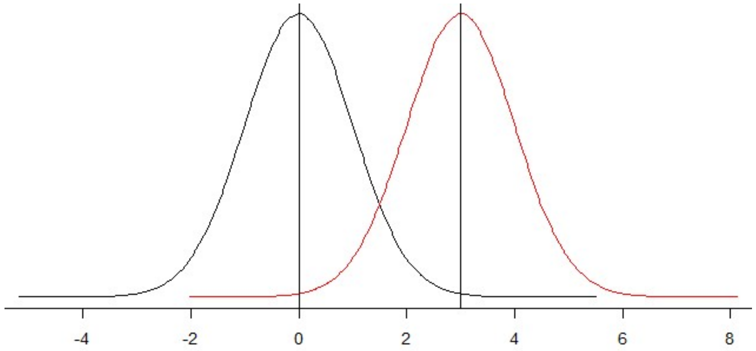
1. A measure of central tendency

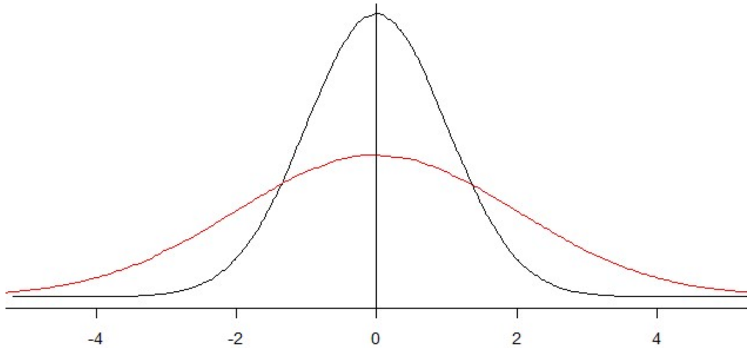


Frequency Distributions

- In the previous video I mentioned that you can summarize frequency distributions (and therefore data), using:

1. A measure of central tendency.
2. A measure of spread.





Variability

- In the previous videos we looked at three measures for central tendency:
  - The mean, median, and the mode
- In this video we will look at measures of spread, or variability.
  - Specifically, variance, standard deviation, and range, although…
  - …variance and the standard deviation are actually the same statistic, just on a different scale (much like length in meters and centimeters)

Range (discrete)

- For discrete variables: The range is simply the distance between the largest and the smallest number.

$$range = X_{max} - X_{min}$$

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

Range (discrete)

- For discrete variables: The range is simply the distance between the largest and the smallest number.

$$range = X_{max} - X_{min}$$

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$$range = 5 - 2 = 3$$

Range (continuous)

- For continuous variables: We need to take the real limits into account again!
- The range is the distance between the upper real limit of the largest number and the lower real limit of smallest number.

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$$range = URL\ for\ X_{max} - LRL\ for\ X_{min}$$

Range (continuous)

- For continuous variables: We need to take the real limits into account again!
- The range is the distance between the upper real limit of the largest number and the lower real limit of smallest number.

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$$range = URL\ for\ X_{max} - LRL\ for\ X_{min}$$

$$range = 5.5 - 1.5 = 4$$

Variance

- The variance is:
  - Only used for continuous variables.
  - The average distance between scores and the mean.

- The formula for the variance may look a little daunting, so let's look at where it comes from.

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

Variance

- Since the variance is about the distance between scores and the mean, you would initially just look at the differences between each score and the mean right?
- And then add all those differences to get the total distance to the mean.

Population:

$$total\ deviation = \sum X - \mu$$

Sample:

$$total\ deviation = \sum X - M$$

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$M = 3.57$

$$(4 - 3.57) + (5 - 3.57) + (4 - 3.57) + (4 - 3.57) + (3 - 3.57) + (3 - 3.57) + (2 - 3.57)$$

Variance

- Since the variance is about the distance between scores and the mean, you would initially just look at the differences between each score and the mean right?

- And then add all those differences to get the total distance to the mean.
    - There is a problem with this!

Population:

$$total\ deviation = \sum X - \mu$$

Sample:

$$total\ deviation = \sum X - M$$

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$M = 3.57$

$$(4-3.57)+(5-3.57)+(4-3.57)+(4-3.57)+(3-3.57)+(3-3.57)+(2-3.57) = \mathbf{0}$$

Variance

- Remember that the mean sees the middle such that the total distance between the scores below the mid-point and the mid-point, AND the scores above the mid-point and the mid-point are the same.

- So, the positive and negative distances add to 0, making it seem like there is no variability…this is obviously wrong!

Population:

$$total\ deviation = \sum X - \mu$$

Sample:

$$total\ deviation = \sum X - M$$

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$M = 3.57$

$(4 - 3.57) + (5 - 3.57) + (4 - 3.57) + (4 - 3.57) + (3 - 3.57) + (3 - 3.57) + (2 - 3.57) = \mathbf{0}$

Variance

- Ok....so we got to make sure positive and negative do not cancel each other out!...

- Then let's square the distances between the scores and the mean before we add them.

Population:

$$SS = \sum (X - \mu)^2$$

Sample:

$$SS = \sum (X - M)^2$$

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$M = 3.57$

$$(4-3.57)^2 + (5-3.57)^2 + (4-3.57)^2 + (4-3.57)^2 + (3-3.57)^2 + (3-3.57)^2 + (2-3.57)^2 = 5.71$$

Variance

- Ok....so we got to make sure positive and negative do not cancel each other out!...
- Then let's square the distances between the scores and the mean before we add them.
- Better...but still a problem

Participant	# of shoes
Lennie	4
Joran	5
Leonie	4
Zaïra	4
Vince	3
Natascha	3
Sacha	2

$M = 3.57$

$$(4-3.57)^2 + (5-3.57)^2 + (4-3.57)^2 + (4-3.57)^2 + (3-3.57)^2 + (3-3.57)^2 + (2-3.57)^2 = 5.71$$

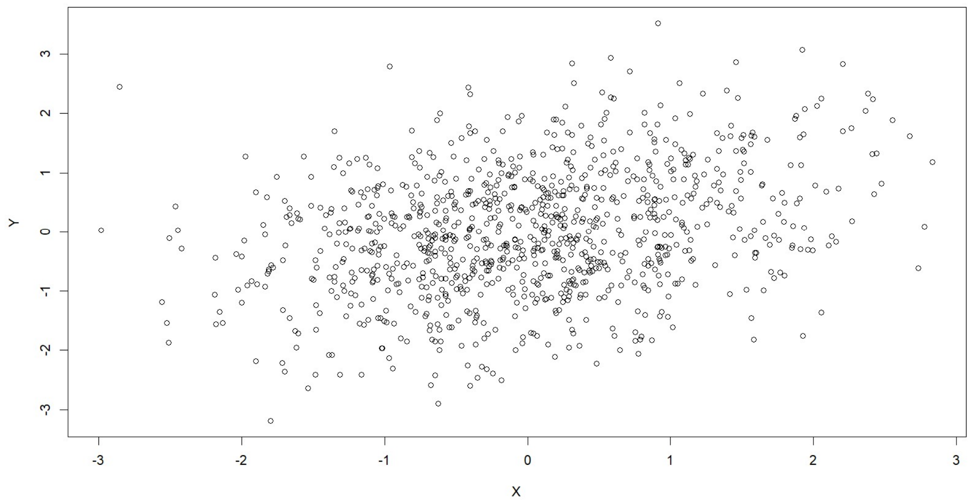
Population:

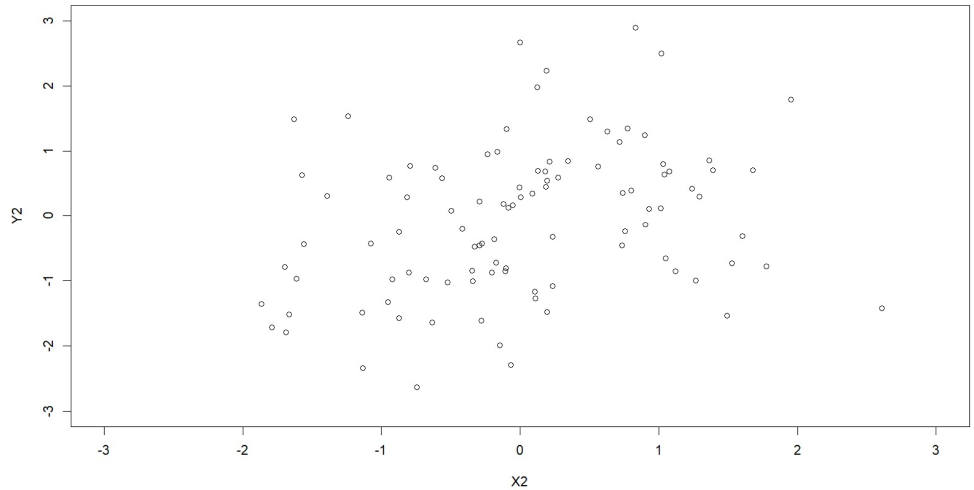
$$SS = \sum (X - \mu)^2$$

Sample:

$$SS = \sum (X - M)^2$$

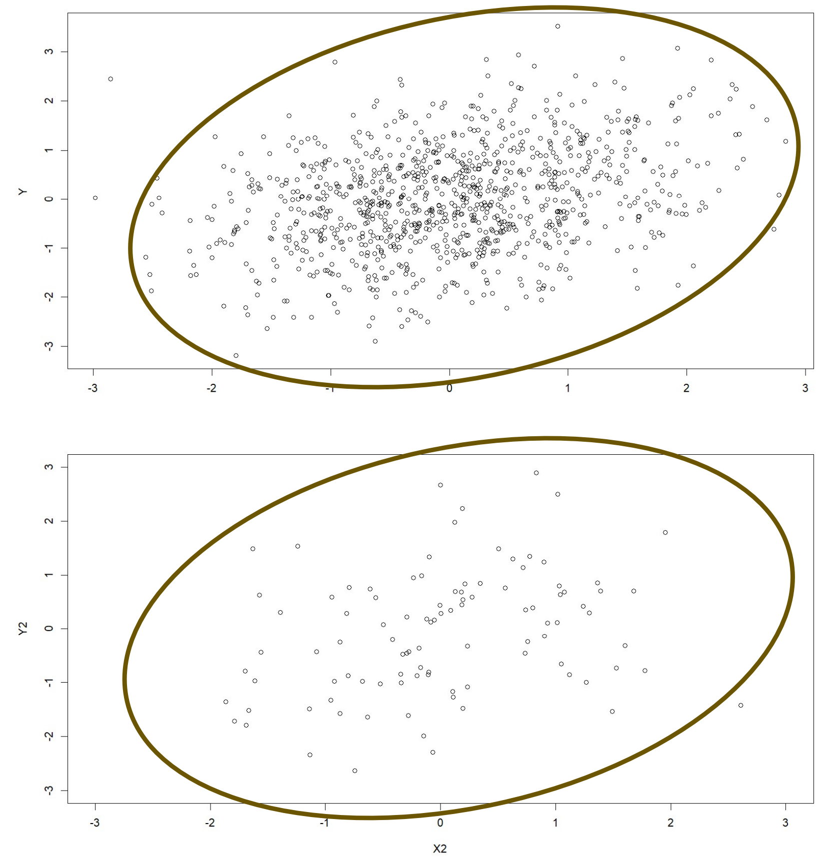
Variance





Variance

- In both figures there is the same amount of variance.
- The top just has more points/observations
- No what would happen of we sum all squared differences for the data in both figures?



Population:

$$SS = \sum (X - \mu)^2$$

Sample:

$$SS = \sum (X - M)^2$$

Variance

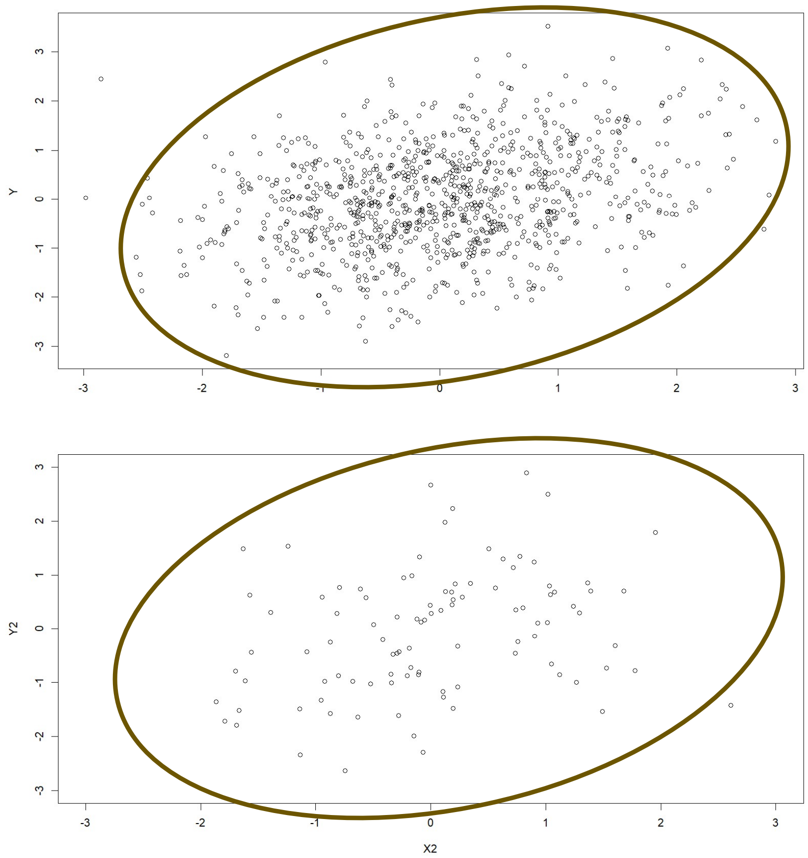
- In both figures there is the same amount of variance.
- The top just has more points/observations
- No what would happen of we sum all squared differences for the data in both figures?
  - The top data would get a higher score simply because we add more things together (1000 numbers vs 100 numbers below)…even though the variance is the same!!

Population:

$$SS = \sum(X - \mu)^2$$

Sample:

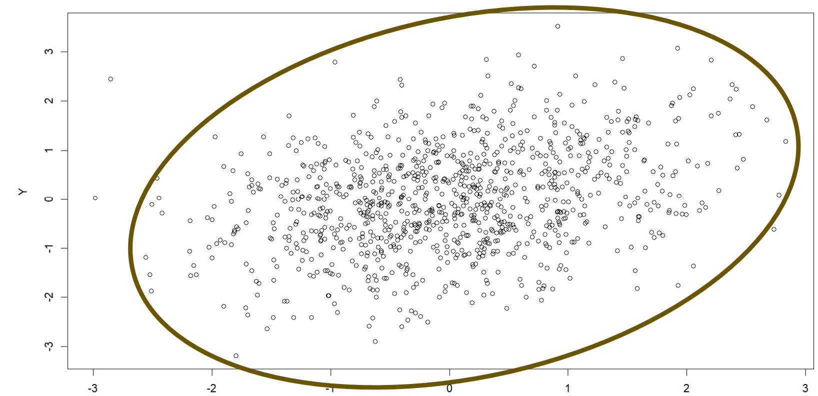
$$SS = \sum(X - M)^2$$

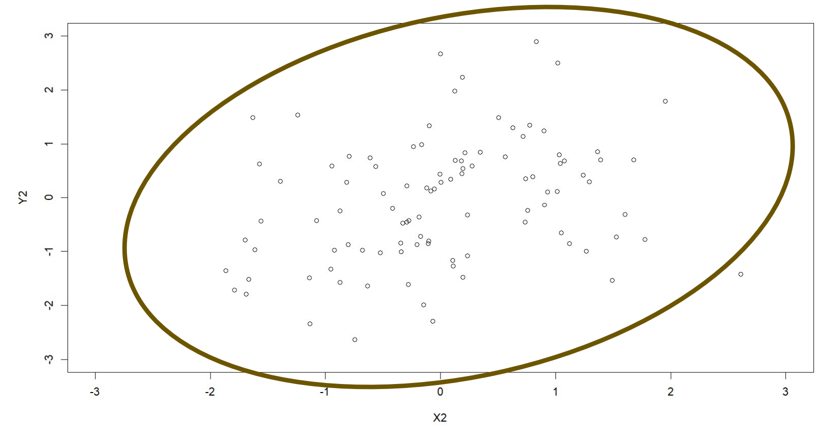


Variance

- No what would happen of we sum all squared differences for the data in both figures?
  - The top data would get a higher score simply because we add more things together (1000 numbers vs 100 numbers below)…even though the variance is the same!!

- So, we need to correct for the number of observations





Population:

$$variance = \frac{\sum(X - \mu)^2}{N}$$

Sample:

$$variance = \frac{\sum(X - M)^2}{n - 1}$$

Variance

- These variables for the variance give us what they need to, but there is still one small problem…not a big issue…just an annoyance.
  - Because we square (differences in) scores the scale of the variance is not the same as the scale of the variables.
  - Like our scores are in centimeters but the variance is in inch.

- This is not nice, but we can easily solve it by taking the square root of the variance, which is the standard deviation!

Population:

$$variance = \frac{\sum(X-\mu)^2}{N}, \quad sd = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

Sample:

$$variance = \frac{\sum(X-M)^2}{n-1}, \quad sd = \sqrt{\frac{\sum(X-M)^2}{n-1}}$$

Variance & Standard deviation

- These variables for the variance give us what they need to, but there is still one small problem…not a big issue…just an annoyance.
  - Because we square (differences in) scores the scale of the variance is not the same as the scale of the variables.
  - Like our scores are in centimeters but the variance is in inch.

- This is not nice, but we can easily solve it by taking the square root of the variance, which is the standard deviation!

Population:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}, \quad \sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

Sample:

$$s^2 = \frac{\sum(X-M)^2}{n-1}, \quad s = \sqrt{\frac{\sum(X-M)^2}{n-1}}$$

Variance & Standard deviation

- The formulas presented so far are the default ones that always work.

- But when the mean is not a whole number they are more cumbersome to use.

Variance & Standard deviation

- When the mean is not a whole number, easier to replace....

Population:

$$SS = \sum (X - \mu)^2$$

Sample:

$$SS = \sum (X - M)^2$$

- ...with

Population:

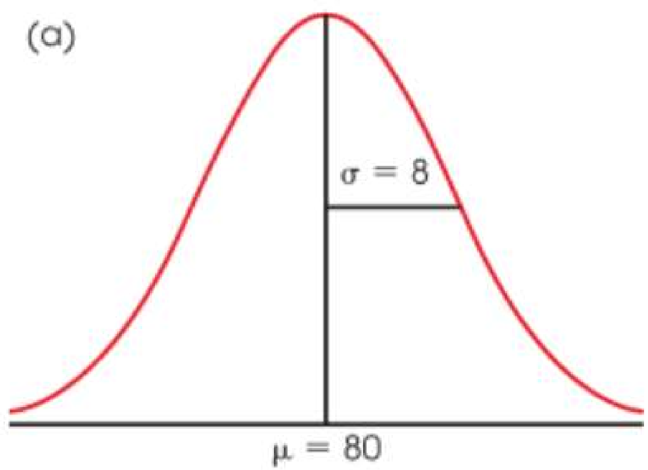
$$total\ deviation = \sum X^2 - \frac{(\sum X)^2}{N}$$

Sample:

$$total\ deviation == \sum X^2 - \frac{(\sum X)^2}{n}$$

Mean and Variance

- Now we can summarize data even more parsimoniously.
- And calculate percentiles using just the mean and standard deviation too!
    - Hypothesis testing is based entirely on this.



The End

Any *questions* ?

Go to Canvas Discussions and ...

- ask your question
- like relevant questions of others