Chapter 5: The Importance of Measuring Variability

- Measures of Central Tendency - Numbers that describe what is typical or "central" in a variable's distribution (e.g., mean, mode, median).
- Measures of Variability - Numbers that describe diversity or variability in a variable's distribution (e.g., range, interquartile range, variance, standard deviation).

Why is Variability important?

Example: Suppose you wanted to know how satisfied students are with their living arrangements and you found that the mean answer was "3" on a five point scale where:

1=very unsatisfied, 2=satisfied, 3=neutral, 4= satisfied, 5=very satisfied

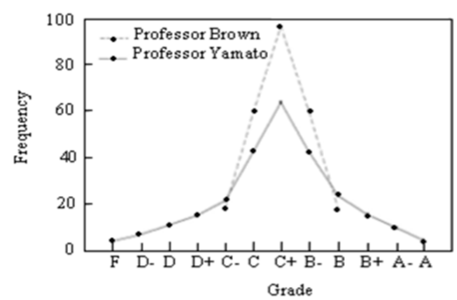
What would you conclude?

Would knowing the variability of the answers help you to understand how satisfied students are with their living arrangements?

Answer: It would help you to see whether the average score of "3" means that the majority of students are neutral about their jobs

or

that there is a split with students either feeling very satisfied (score of 5) or unsatisfied (score of 1) with their living arrangements (average of 1's and 5's = 3).



Another example.

The Range

- Range – A measure of variation in interval-ratio variables.
- It is the difference between the highest (maximum) and the lowest (minimum) scores in the distribution.

Range = highest score - lowest score

Table 1: Level of Ethnic Diversity (IQV) in the State

What is the range for these diversity scores?

Steps to determine: subtract the lowest score ____from the highest _____to obtain the range of IQV scores_____.

State	IQV	State	IQV	State	IQV
California	0.80	Alabama	0.51	Indiana	0.27
New Mexico	0.76	North Carolina	0.51	Utah	0.26
Texas	0.74	Delaware	0.49	Nebraska	0.24
New York	0.66	Colorado	0.45	South Dakota	0.24
Hawaii	0.64	Oklahoma	0.44	Wisconsin	0.24
Maryland	0.62	Connecticut	0.42	Idaho	0.23
New Jersey	0.61	Arkansas	0.40	Wyoming	0.22
Louisiana	0.61	Michigan	0.40	Kentucky	0.20
Arizona	0.61	Tennessee	0.39	Minnesota	0.20
Florida	0.61	Washington	0.37	Montana	0.20
Mississippi	0.61	Massachusetts	0.34	North Dakota	0.17
Georgia	0.59	Missouri	0.31	Iowa	0.13
Nevada	0.57	Ohio	0.31	West Virginia	0.11
Illinois	0.57	Pennsylvania	0.31	New Hampshire	0.08
South Carolina	0.56	Kansas	0.30	Maine	0.06
Alaska	0.56	Rhode Island	0.30	Vermont	0.06
Virginia	0.53	Oregon	0.28		

What is the range for these diversity scores?

Steps to determine: subtract the lowest score _.06___from the highest _____to obtain the range of IQV scores_____.

State	IQV	State	IQV	State	IQV
California	0.80	Alabama	0.51	Indiana	0.27
New Mexico	0.76	North Carolina	0.51	Utah	0.26
Texas	0.74	Delaware	0.49	Nebraska	0.24
New York	0.66	Colorado	0.45	South Dakota	0.24
Hawaii	0.64	Oklahoma	0.44	Wisconsin	0.24
Maryland	0.62	Connecticut	0.42	Idaho	0.23
New Jersey	0.61	Arkansas	0.40	Wyoming	0.22
Louisiana	0.61	Michigan	0.40	Kentucky	0.20
Arizona	0.61	Tennessee	0.39	Minnesota	0.20
Florida	0.61	Washington	0.37	Montana	0.20
Mississippi	0.61	Massachusetts	0.34	North Dakota	0.17
Georgia	0.59	Missouri	0.31	Iowa	0.13
Nevada	0.57	Ohio	0.31	West Virginia	0.11
Illinois	0.57	Pennsylvania	0.31	New Hampshire	0.08
South Carolina	0.56	Kansas	0.30	Maine	0.06
Alaska	0.56	Rhode Island	0.30	Vermont	0.06
Virginia	0.53	Oregon	0.28		

What is the range for these diversity scores?

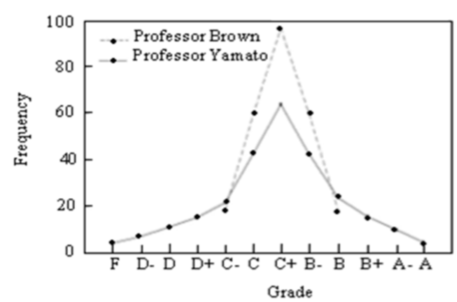
Steps to determine: subtract the lowest score _.06___from the highest __.80___to obtain the range of IQV scores_____.

State	IQV	State	IQV	State	IQV
California	0.80	Alabama	0.51	Indiana	0.27
New Mexico	0.76	North Carolina	0.51	Utah	0.26
Texas	0.74	Delaware	0.49	Nebraska	0.24
New York	0.66	Colorado	0.45	South Dakota	0.24
Hawaii	0.64	Oklahoma	0.44	Wisconsin	0.24
Maryland	0.62	Connecticut	0.42	Idaho	0.23
New Jersey	0.61	Arkansas	0.40	Wyoming	0.22
Louisiana	0.61	Michigan	0.40	Kentucky	0.20
Arizona	0.61	Tennessee	0.39	Minnesota	0.20
Florida	0.61	Washington	0.37	Montana	0.20
Mississippi	0.61	Massachusetts	0.34	North Dakota	0.17
Georgia	0.59	Missouri	0.31	Iowa	0.13
Nevada	0.57	Ohio	0.31	West Virginia	0.11
Illinois	0.57	Pennsylvania	0.31	New Hampshire	0.08
South Carolina	0.56	Kansas	0.30	Maine	0.06
Alaska	0.56	Rhode Island	0.30	Vermont	0.06
Virginia	0.53	Oregon	0.28		

What is the range for these diversity scores?

Steps to determine: subtract the lowest score _.06___from the highest __.80___to obtain the range of IQV scores__.74___.

State	IQV	State	IQV	State	IQV
California	0.80	Alabama	0.51	Indiana	0.27
New Mexico	0.76	North Carolina	0.51	Utah	0.26
Texas	0.74	Delaware	0.49	Nebraska	0.24
New York	0.66	Colorado	0.45	South Dakota	0.24
Hawaii	0.64	Oklahoma	0.44	Wisconsin	0.24
Maryland	0.62	Connecticut	0.42	Idaho	0.23
New Jersey	0.61	Arkansas	0.40	Wyoming	0.22
Louisiana	0.61	Michigan	0.40	Kentucky	0.20
Arizona	0.61	Tennessee	0.39	Minnesota	0.20
Florida	0.61	Washington	0.37	Montana	0.20
Mississippi	0.61	Massachusetts	0.34	North Dakota	0.17
Georgia	0.59	Missouri	0.31	Iowa	0.13
Nevada	0.57	Ohio	0.31	West Virginia	0.11
Illinois	0.57	Pennsylvania	0.31	New Hampshire	0.08
South Carolina	0.56	Kansas	0.30	Maine	0.06
Alaska	0.56	Rhode Island	0.30	Vermont	0.06
Virginia	0.53	Oregon	0.28		



Another example.

Inter-quartile Range

- Inter-quartile range (IQR) – The width of the middle 50 percent of the distribution.
- The IQR helps us to get a better picture of the variation in the data than the range because it focuses on the width of the middle 50% rather than extreme scores in the distribution.
- The shortcoming of the range is that an “outlying” case at the top or bottom can increase the range substantially.

Inter-quartile Range

- Inter-quartile range (IQR) – The width of the middle 50 percent of the distribution.
- It is defined as the difference between the lower and upper quartiles (Q1 and Q3.)
- IQR = q3 – q1
  (e.g., 75th percentile – 25th percentile)

What is the IQR for these Diversity Scores?

State	IQV	State	IQV	State	IQV
California	0.80	Alabama	0.51	Indiana	0.27
New Mexico	0.76	North Carolina	0.51	Utah	0.26
Texas	0.74	Delaware	0.49	Nebraska	0.24
New York	0.66	Colorado	0.45	South Dakota	0.24
Hawaii	0.64	Oklahoma	0.44	Wisconsin	0.24
Maryland	0.62	Connecticut	0.42	Idaho	0.23
New Jersey	0.61	Arkansas	0.40	Wyoming	0.22
Louisiana	0.61	Michigan	0.40	Kentucky	0.20
Arizona	0.61	Tennessee	0.39	Minnesota	0.20
Florida	0.61	Washington	0.37	Montana	0.20
Mississippi	0.61	Massachusetts	0.34	North Dakota	0.17
Georgia	0.59	Missouri	0.31	Iowa	0.13
Nevada	0.57	Ohio	0.31	West Virginia	0.11
Illinois	0.57	Pennsylvania	0.31	New Hampshire	0.08
South Carolina	0.56	Kansas	0.30	Maine	0.06
Alaska	0.56	Rhode Island	0.30	Vermont	0.06
Virginia	0.53	Oregon	0.28		

(Steps are provided on the next slides)

What is the IQR for the Diversity Scores?

<u>Steps to determine the IQR (Q3 – Q1):</u>

1. Order the categories from highest to lowest (or vice versa)
2. To obtain Q1, begin by dividing N (total number of categories or states) by 4 (or alternatively multiply N by .25). This equals______?
3. We now know that Q1 falls between the 12th and 13th category or, in this case, states.
4. To find the exact number for Q1, determine the midpoint between the 12th and 13th states or between .59 and .57)
5. Q1 = ____

What is the IQR for the Diversity Scores?

<u>Steps to determine the IQR (Q3 – Q1):</u>

1. Order the categories from highest to lowest (or vice versa)
2. To obtain Q1, begin by dividing N (total number of categories or states) by 4 (or alternatively multiply N by .25). This equals___12.5___?
3. We now know that Q1 falls between the 12th and 13th category or, in this case, states.
4. The diversity score between these two states is: between .59 and .57 or .58
5. To obtain Q3, multiply the quarter figure (12.5) by 3 = _______ and then locate this category (the 37th and 38th states).

What is the IQR for the Diversity Scores?

<u>Steps to determine the IQR (Q3 – Q1):</u>

1. Order the categories from highest to lowest (or vice versa)
2. To obtain Q1, begin by dividing N (total number of categories or states) by 4 (or alternatively multiply N by .25). This equals___12.5___?
3. We now know that Q1 falls between the 12th and 13th category or, in this case, states.
4. The diversity score between these two states is: between .59 and .57 or **<u>Q1 = .58</u>**
5. To obtain Q3, multiply the quarter figure (12.5) by 3 = __37.5__and then locate this category (the 37th and 38th states).

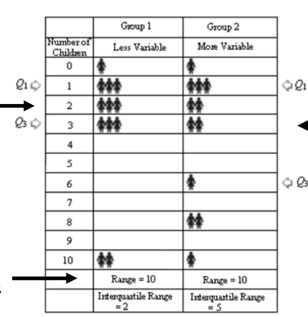
What is the IQR for the Diversity Scores?

<u>Steps to determine the IQR (Q3 – Q1):</u>

6. Based on this number (37.5), Q3 falls between the 37th and 38th states.
7. To find the exact number for Q3 determine the midpoint between the 37th and 38th states or **<u>Q3 = .24</u>**
8. This tells us that 50% of the cases fall between the IQR scores of .58 and .24.
9. **The IQR = .58 – .24 = .34**

The difference between the Range and IQR

These values fall together closely



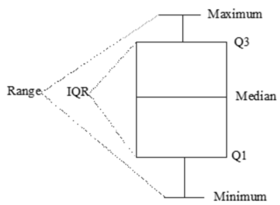
Shows greater variability

Yet the ranges are equal!

Importance of the IQR

The Box Plot

- The Box Plot is a **graphic device** that visually presents the following elements: the **range**, the **IQR**, the **median**, the **quartiles**, amount and direction of **skewness**, the **minimum** (lowest value,) and the **maximum** (highest value.)



Procedures for Creating Box Plots for Variables

- Open SPSS
- Click "graphs"
- Click "legacy dialogs"
- Click "box plot"
- Click "simple" and "summaries for separate variables"
- Click "define"
- Select desired variable and put in "Boxes Represent"
- Click "okay"

Procedures for Creating Box Plots for Groups (for example, Males and Females by Income)

- Open SPSS
- Click "graphs"
- Click "legacy dialogs"
- Click "box plot"
- Click "simple" and "summaries for groups of cases"
- Click "define"
- Select desired dependent variable (such as income) and put in "Variable Box"
- Move desired grouping variable (such as sex) into "Category Axis"
- Click "okay"

Measures of Variability:
Shortcomings of the Range and IQR

- The range is based on only two categories (the highest and lowest)
- Likewise, only two categories are used to calculate the inter-quartile range.
- Neither allows us to know how much variation there is among all the categories.

Measures of Variability: the Variance

- The variance allows us to account for the total amount of variation.
- The variance is an important statistic that is used in most other sophisticated statistics. Therefore, it is important for you to give it particular attention.

Be sure to read the sections of the chapter on variability and standard deviation very carefully.

Determining the Variance in the "Percentage Increase" in the Nursing Home Population, 1980-1990

Nine Regions of U.S.	Percentage
Pacific	15.7
West North Central	16.2
New England	17.6
East North Central	23.2
West South Central	24.3
Middle Atlantic	28.5
East South Central	38.0
Mountain	47.9
South Atlantic	71.7

What statistics have we learned so far to describe the variation above?

Is there a lot of variation between the categories (regions of U.S.)?

Range, Inter-Quartile Range (IQR)
There appears to be a lot of variation between regions.

First Step in Calculating the Variation:
Determine the "**Average**" Between Regions for the percent change in the Nursing Home Population, 1980-1990

Nine Regions of U.S.	Percentage
Pacific	15.7
West North Central	16.2
New England	17.6
East North Central	23.2
West South Central	24.3
Middle Atlantic	28.5
East South Central	38.0
Mountain	47.9
South Atlantic	71.7

The "average" percentage increase in the Nursing Home Population, 1980-1990

Nine Regions of U.S.	Percentage
Pacific	15.7
West North Central	16.2
New England	17.6
East North Central	23.2
West South Central	24.3
Middle Atlantic	28.5
East South Central	38.0
Mountain	47.9
South Atlantic	71.7
$\sum$ Y =	283.1

Average "% increase"

mean = $\bar{Y} = \frac{\sum Y}{N}$ = 31.45

Determining the Variation in the Percentage Change in the Nursing Home Population, 1980-1990

Nine Regions of U.S.	Percentage	$Y - \bar{Y}$
Pacific	15.7	15.7 - 31.5 = -15.8
West North Central	16.2	16.2 - 31.5 = -15.3
New England	17.6	17.6 - 31.5 = -13.9
East North Central	23.2	23.2 - 31.5 = - 8.3
West South Central	24.3	24.3 - 31.5 = - 7.2
Middle Atlantic	28.5	28.5 - 31.5 = - 3.0
East South Central	38.0	38.0 - 31.5 = 6.5
Mountain	47.9	47.9 - 31.5 = 16.4
South Atlantic	71.7	71.7 - 31.5 = 40.2
(mean = 31.5)	$\sum$ Y = 283.1	$\sum (Y - \bar{Y}) = 0$

Next, we can determine the distance between (1) each region and (2) the average (31.5), in order to get the amount of variation from the mean for each region. Then, we can add up the variation scores for each region to get the "total" variation of the scores (but this is not the actual "VARIANCE").

Percentage Change in the Nursing Home Population, 1980-1990

Nine Regions of U.S.	Percentage	$Y - \bar{Y}$
Pacific	15.7	15.7 - 31.5 = -15.8
West North Central	16.2	16.2 - 31.5 = -15.3
New England	17.6	17.6 - 31.5 = -13.9
East North Central	23.2	23.2 - 31.5 = - 8.3
West South Central	24.3	24.3 - 31.5 = - 7.2
Middle Atlantic	28.5	28.5 - 31.5 = - 3.0
East South Central	38.0	38.0 - 31.5 = 6.5
Mountain	47.9	47.9 - 31.5 = 16.4
South Atlantic	71.7	71.7 - 31.5 = 40.2
(mean = 31.5)	$\sum$ Y = 283.1	$\sum (Y - \bar{Y}) = 0$

Problem: when you add up the distances you end up with zero rather than the total variation from all the categories. Why is this?

Percentage Change in the Nursing Home Population, 1980-1990

Nine Regions of U.S.	Percentage	$Y - \bar{Y}$
Pacific	15.7	15.7 - 31.5 = -15.8
West North Central	16.2	16.2 - 31.5 = -15.3
New England	17.6	17.6 - 31.5 = -13.9
East North Central	23.2	23.2 - 31.5 = - 8.3
West South Central	24.3	24.3 - 31.5 = - 7.2
Middle Atlantic	28.5	28.5 - 31.5 = - 3.0
East South Central	38.0	38.0 - 31.5 = 6.5
Mountain	47.9	47.9 - 31.5 = 16.4
South Atlantic	71.7	71.7 - 31.5 = 40.2
(mean = 31.5)	$\sum$ Y = 283.1	$\sum (Y - \bar{Y}) = 0$

- One solution would be to add up the absolute values for each number (ignore the minus signs), or 126.6 and then divide by the number of regions (9) =14.1). Unfortunately, absolute values are very difficult to work with mathematically.
- Fortunately, there is another alternative.

Percentage Change in the Nursing Home Population, 1980-1990

Nine Regions of U.S.	Percentage	$Y - \bar{Y}$	$(Y - \bar{Y})^2$ (squared deviations)
Pacific	15.7	15.7 - 31.5 = -15.8	249.64
West North Central	16.2	16.2 - 31.5 = -15.3	234.09
New England	17.6	17.6 - 31.5 = -13.9	193.21
East North Central	23.2	23.2 - 31.5 = - 8.3	68.89
West South Central	24.3	24.3 - 31.5 = - 7.2	51.84
Middle Atlantic	28.5	28.5 - 31.5 = - 3.0	9.00
East South Central	38.0	38.0 - 31.5 = 6.5	42.25
Mountain	47.9	47.9 - 31.5 = 16.4	268.96
South Atlantic	71.7	71.7 - 31.5 = 40.2	1616.04
(mean = 31.5)	$\sum \bar{Y}$ = 283.1		$\sum (Y - \bar{Y})^2 = 2733.92$

- The best solution is to square the differences before adding them up (when two negative numbers are multiplied the resulting product is a positive number). This eliminates the problem of adding negative and positive numbers.

Measures of Variability: the Variance

The Variance is the average of the squared deviations from the mean.

$$s_Y^2 = \frac{\sum(Y - \overline{Y})^2}{N-1}$$

In our example we would take the sum of the squared deviations (2733.92) and divide this number by the total number of cases minus one (9 – 1 = 8). This would give us _341.74___ or the variance for the Percent Increase in the Nursing Home population by region.

Measures of Variability: The Variance

To Sum Up:

The Variance is the average of the squared deviations from the mean.

The Variance is a measure of variability for interval-ratio variables.

$$s_Y^2 = \frac{\sum(Y - \overline{Y})^2}{N-1}$$

Measures of Variability: Standard Deviation

• One problem with the variance is that the final number obtained is in a squared form

(that is, we squared all the deviations from the mean and so the final number is still "inflated" in this way making it difficult to interpret)

• One solution is to take the square root of the variance so that the number is no longer in a squared form (or "inflated") and it is back to its original form. The square root of the variance is called the Standard Deviation.

Measures of Variability: Standard Deviation

• To obtain the square root of the variance simply enter the number (variance) into your calculator and then push the square root button.

•If the variance is 341.74 the standard deviation would be ___18.49_____. This tells us that the percent of change in the nursing home population for the nine regions is widely dispersed around the mean (mean = 31.45).

• Thus, the standard deviation is a measure of the average amount of variation (or deviation) around the mean.

In Sum

The Standard Deviation is a measure of variation for interval-ratio variables; it is equal to the square root of the variance.

$$s = \sqrt{s_Y^2} = \sqrt{\frac{\sum(Y - \overline{Y})^2}{N-1}}$$

Considerations for Choosing a Measure of Variability

- For ordinal variables, you can calculate the IQR (range and inter-quartile range.)
- For interval-ratio variables, you can use the range, the IQR, the variance or the standard deviation. The variance and standard deviation provide the most information, since they use all of the values in the distribution in their calculations.

Fini