Statistics and Quantitative Analysis U4320

Segment 2*: *Descriptive Statistics

Prof. Sharyn O'Halloran

Types of Variables

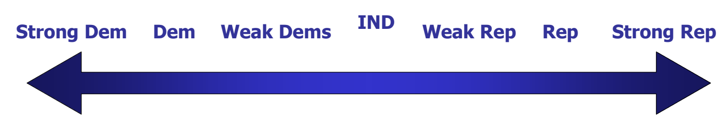
- Nominal Categories
  - Names or types with no inherent ordering:
    - Religious affiliation
    - Marital status
    - Race
- Ordinal Categories
  - Variables with a rank or order
    - University Rankings
    - Academic position
    - Level of education

Types of Variables (con't)

- Interval/Ratio Categories
  - Variables are ordered and have equal space between them
    - Salary measured in dollars
    - Age of an individual
    - Number of children
  - Thermometer scores measure intensity on issues
    - Views on abortion
    - Whether someone likes China
    - Feelings about US-Japan trade issues
  - Any dichotomous variable
    - A variable that takes on the value 0 or 1.
    - A person's gender (M=0, F=1)

Note on thermometer scores

- Make assumptions about the relation between different categories.
  - For example, if we look at the liberal to conservative political spectrum:



- Distance among groups is the same for each category.

Measuring Data

- Discrete Data
  - Takes on only integer values
    - whole numbers no decimals
    - E.g., number of people in the room
- Continuous Data
  - Takes on any value
    - All numbers including decimals
    - E.g., Rate of population increase

Discrete Data

Example: Absenteeism for 50 employees

I	II	III	IV	V
Absences **x**	Frequency **f**	Relative Frequency **f / N**	$x - \overline{X}$	$f\left(x - \overline{X}\right)^2$
0	3	.06	-4.86	70.8
1	2	.04	-3.86	29.8
2	5	.10	-2.86	40.9
3	8	.16	-1.86	27.7
4	7	.14	-0.86	5.2
5	2	.04	0.14	.04
6	13	.26	1.14	16.9
7	2	.04	2.14	9.2
8	4	.08	3.14	39.4
9	0	.00	4.14	0
10	1	.02	5.14	26.4
11	2	.04	6.14	75.4
12	0	.00	7.14	0
13	1	.02	8.14	66.3
	N=50	**1.00**		S = **408.02**

Absenteeism for 50 employees

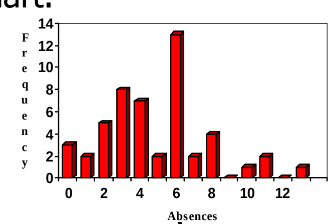
I	II	III	IV	V
Absences **x**	Frequency **f**	Relative Frequency **f / N**	$x - \overline{X}$	$f\left(x - \overline{X}\right)^2$
0	3	.06	-4.86	70.8
1	2	.04	-3.86	29.8
2	5	.10	-2.86	40.9
3	8	.16	-1.86	27.7
4	7	.14	-0.86	5.2
5	2	.04	0.14	.04
6	13	.26	1.14	16.9
7	2	.04	2.14	9.2
8	4	.08	3.14	39.4
9	0	.00	4.14	0
10	1	.02	5.14	26.4
11	2	.04	6.14	75.4
12	0	.00	7.14	0
13	1	.02	8.14	66.3
	N=50	**1.00**		S = **408.02**

Discrete Data (con't)

- Frequency
  - Number of times we observe an event
    - In our example, the number of employees that were absent for a particular number of days.
      - Three employees were absent no days (0) through out the year
- Relative Frequency
  - Number of times a particular event takes place in relation to the total.
    - For example, about a **quarter** of the people were absent exactly 6 times in the year.
  - Why do you think this was true?

Discrete Data: Chart

- Another way that we can represent this data is by a chart.



- Again we see that more people were absent 6 days than any other single value.

Continuous Data

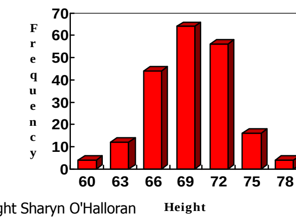
Example: Men's Heights

Height	Midpoint	Frequency (f)	Relative Frequency (f / N)	Percentile (Cumulative Frequency)
58.5-61.5	60	4	.02	.02
61.5-64.5	63	12	.06	.08
64.5-67.5	66	44	.22	.30
67.5-70.5	69	64	.32	.62
70.5-73.5	72	56	.28	.90
73.5-76.5	75	16	.08	.98
76.5-79.5	78	4	.02	1.00
		N=200	**1.00**	

Continuous Data (con't)

- Note: must decide how to divide the data
  - The interval you define changes how the data appears.
    - The finer the divisions, the less bias.
    - ut more ranges make it difficult to display and interpret.
- Chart

Can represent information graphically



ded into
ervals

Continuous Data: Percentiles

- The N^{th} percentile
  - The point such that n-percent of the population lie below and (100-n) percent lie above it.

What interval is the 50^{th} percentile?

Height Example

Height	Midpoint	Frequency (f)	Relative Frequency (f / N)	Percentile (Cumulative Frequency)
58.5-61.5	60	4	.02	.02
61.5-64.5	63	12	.06	.08
64.5-67.5	66	44	.22	.30
67.5-70.5	69	64	.32	.62
70.5-73.5	72	56	.28	.90
73.5-76.5	75	16	.08	.98
76.5-79.5	78	4	.02	1.00
		N=200	**1.00**	

Continuous Data: Income example

- What interval is the 50th percentile?

Average Income in California Congressional Districts

Cells	Freq	Tally	Relative Frequency
$0-$20,708	1	X	0.019
$20,708-$25,232	3	XXX	0.057
$25,232-$29,756	7	XXXXXXX	0.135
$29,756-$34,280	15	XXXXXXXXXXXXXXX	0.289
$34,280-$38,805	8	XXXXXXXX	0.154
$38,805-$43,329	4	XXXX	0.077
$43,329-$47,853	8	XXXXXXXX	0.154
$47,853-$52,378	6	XXXXXX	0.115

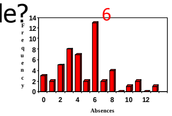
n = 52

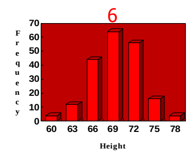
Measures of Central Tendency

- Mode
  - The category occurring most often.
- Median
  - The middle observation or 50th percentile.
- Mean
  - The average of the observations

Central Tendency: Mode

- The category that occurs most often.
  - Mode of the absentee example?





Mode of the height example?

Average Income in California Congressional Districts

Cells	Freq	Tally	Relative Frequency
$0-$20,708	1	X	0.019
$20,708-$25,232	3	XXX	0.057
$25,232-$29,756	7	XXXXXXX	0.135
$29,756-$34,280	15	XXXXXXXXXXXXXXX	0.289
$34,280-$38,805	8	XXXXXXXX	0.154
$38,805-$43,329	4	XXXX	0.077
$43,329-$47,853	8	XXXXXXXX	0.154
$47,853-$52,378	6	XXXXXX	0.115

n = 52

2 - 34

- What is the mode of the income example?

Central Tendency: Median

- The middle observation or 50th percentile.

Median of the height example?

Height Example

Height	Midpoint	Frequency (f)	Relative Frequency (f / N)	Percentile (Cumulative Frequency)
58.5-61.5	60	4	.02	.02
61.5-64.5	63	12	.06	.08
64.5-67.5	66	44	.22	.30
67.5-70.5	69	64	.32	.62
70.5-73.5	72	56	.28	.90
73.5-76.5	75	16	.08	.98
76.5-79.5	78	4	.02	1.00
		N=200	**1.00**	

Median of the absentee example?

Absentee Example

I	II	III	IV	V
Absences **x**	Frequency **f**	Relative Frequency **f / N**	$x - \overline{X}$	$f\left(x - \overline{X}\right)^2$
0	3	.06	-4.86	70.8
1	2	.04	-3.86	29.8
2	5	.10	-2.86	40.9
3	8	.16	-1.86	27.7
4	7	.14	-0.86	5.2
5	2	.04	0.14	.04
6	13	.26	1.14	16.9
7	2	.04	2.14	9.2
8	4	.08	3.14	39.4
9	0	.00	4.14	0
10	1	.02	5.14	26.4
11	2	.04	6.14	75.4
12	0	.00	7.14	0
13	1	.02	8.14	66.3
	N=50	**1.00**		S **= 408.02**

Central Tendency: Mean

- The average of the observations.
  - Defined as:

$$\frac{\text{The sum of observations}}{\text{Number of observations}}$$

  - This is denoted as:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Central Tendency: Mean (con't)

- Notation:
  - $_i$'s are different observations
    - Let $_i$ stand for the each individual's height:
      - $_1$ (Sam), $_2$ (Oliver), $_3$ (uddy), or observation 1, 2, 3
        - stands for height in general
        - $_1$ stands for the height for the first person.
  - When we write $\sum$ (the sum) of $_i$ we are adding up all the observations.

Central Tendency: Mean (con't)

Absentee Example

I	II	III	IV	V
Absences x	Frequency f	Relative Frequency f / N	$x - \bar{X}$	$f\left(x - \bar{X}\right)^2$
0	3	.06	-4.86	70.8
1	2	.04	-3.86	29.8
2	5	.10	-2.86	40.9
3	8	.16	-1.86	27.7
4	7	.14	-0.86	5.2
5	2	.04	0.14	.04
6	13	.26	1.14	16.9
7	2	.04	2.14	9.2
8	4	.08	3.14	39.4
9	0	.00	4.14	0
10	1	.02	5.14	26.4
11	2	.04	6.14	75.4
12	0	.00	7.14	0
13	1	.02	8.14	66.3
	N=50	**1.00**		S **= 408.02**

- Example: Absenteeism
  - What is the average number of days any given worker is absent in the year?
    - First, add the fifty individuals' total days absent (**243**).
    - Then, divide the sum (**243**) by the total number of individuals (**50**).
      - We get 243/50, which equals **4.86**.

Central Tendency: Mean (con't)

- Alternatively use the frequency
  - Add each event times the number of occurrences
    - 0 3 1 2 ..... , which is lect2.xls

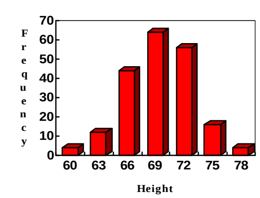
Absentee Example

I	II	III	IV	V
Absences x	Frequency f	Relative Frequency f / N	$x - \bar{X}$	$f\left(x - \bar{X}\right)^2$
0	3	.06	-4.86	70.8
1	2	.04	-3.86	29.8
2	5	.10	-2.86	40.9
3	8	.16	-1.86	27.7
4	7	.14	-0.86	5.2
5	2	.04	0.14	.04
6	13	.26	1.14	16.9
7	2	.04	2.14	9.2
8	4	.08	3.14	39.4
9	0	.00	4.14	0
10	1	.02	5.14	26.4
11	2	.04	6.14	75.4
12	0	.00	7.14	0
13	1	.02	8.14	66.3
	N=50	**1.00**		S **= 408.02**

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i f_i}{N}$$

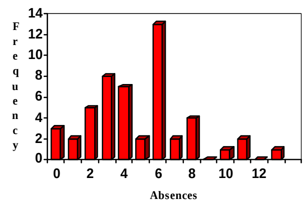
Relative Position of Measures

- For ***Symmetric Distributions***
  - If your population distribution is symmetric and unimodal (i.e., with a hump in the middle), then all three measures coincide.
  - Mode = 6
  - Median =6
  - Mean = 4 3/7= 6



Relative Position of Measures

- For ***Asymmetric Distributions***
  - If the data are skewed, however, then the measures of central tendency will not necessarily line up.



- Mode = 6, Median = 4.5, Mean = 4. 6

Example: International Economic Data

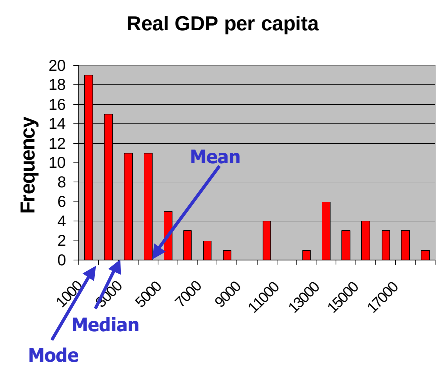
- One source is the World Penn Tables
- Penn World Data Request Form: http://bi ednet.bris.ac.uk: 0 0/dataserv/penn.htm
- Select data format, dates, countries, and indicators
- Read into an Excel sheet
  - World Penn.xls

Example: Descriptive Statistics

STATISTIC	*REAL GPD*	*GOV'T SHARE OF GPD*	*OPENNESS*
Mean	5386.6	19.0	70.7
Standard Error	555.9	1.0	5.2
Median	3073.0	17.6	59.4
Mode	2176.0	17.6	#N/A
Standard Deviation	5331.9	9.1	49.6
Sample Variance	28429209.9	83.2	2464.7
Kurtosis	-0.5	6.7	12.2
Skewness	1.0	1.9	3.0
Range	17577.0	61.2	324.4
Minimum	409.0	4.0	16.6
Maximum	17986.0	65.2	341.0
Sum	495569.0	1748.1	6504.6
Count	92.0	92.0	92.0

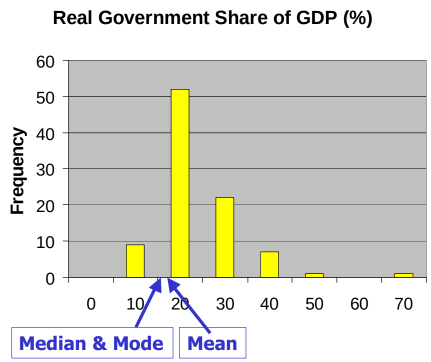
Example: Real DP per capita
(Laspeyres index)

- Mean: 53 7
- Median:3073
- Mode: 2176



Example:
Real overnment Share of DP ( )

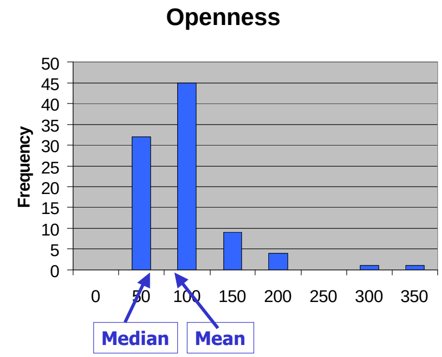
- Mean: 1
- Median: 1
- Mode: 1



25

Example: Openness
(Exports Imports)/Nominal DP

- Mean: 71
- Median: 5
- Mode: N/A



 26

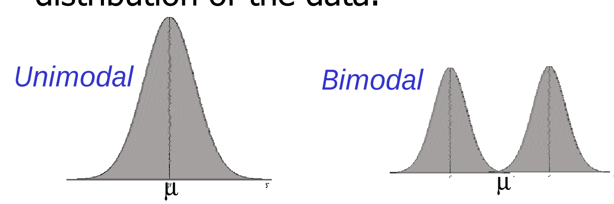
Which Measures for Which Variables?

- *Nominal* variables which measure(s) of central tendency is appropriate?
  - Only the mode
- *Ordinal* variables which measure(s) of central tendency is appropriate?
  - Mode or median
- *Interval-ratio* variables which measure(s) of central tendency is appropriate?
  - All three measures

27

Measures of Deviation

- Need measures of the deviation or distribution of the data.

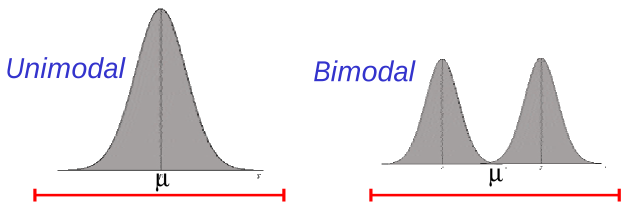


- Every different distributions even though they have the same mean and median.
- Need a measure of dispersion or spread.

2

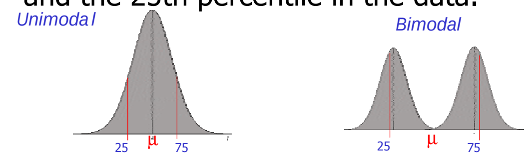
Measures of Deviation (con't)

- Range
  - The difference between the largest and smallest observation.
  - Limited measure of dispersion



Measures of Deviation (con't)

- Inter-Quartile Range
  - The difference between the 75th percentile and the 25th percentile in the data.

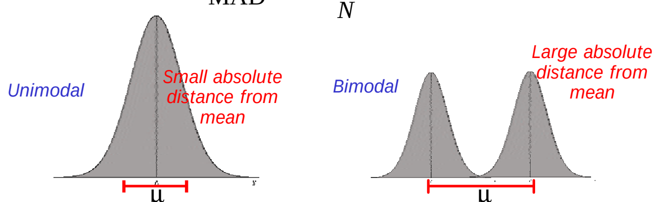


  - etter because it divides the data finer.
  - ut it still only uses two observations
  - Would like to incorporate all the data, if possible.

Measures of Deviation (con't)

- Mean Absolute Deviation
  - The absolute distance between each data point and the mean, then take the average.

$$\text{MAD} = \frac{\sum_{i=1}^{N} \left| X_i - \bar{X} \right|}{N}$$



Measures of Deviation (con't)

- Mean Squared Deviation
  - The difference between each observation and the mean, quantity squared.

$$s^2 = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2}{N}$$

- This is the variance of a distribution.

Measures of Deviation (con't)

- Variance of a Distribution
  - Note since we are using the square, the greatest deviations will be weighed more heavily.
  - If we use a frequency table, as in the absenteeism case, then we would write this as:

$$s^2 = \frac{\sum_{i=1}^{N} f\left(x_i - \overline{X}\right)^2}{N}$$

Measures of Deviation (con't)

- Standard deviation
  - The square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}}$$

  - ives common units to compare deviation from the mean.

Absentee Example: Variances

$$s^2 = \frac{\sum_{i=1}^{N} f\left(x_i - \overline{X}\right)^2}{N} \qquad s = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}}$$

- From above, the mean = 4. 6.
- To calculate the variance, take the difference between each observation and the mean.
- Column 5 then squares that differences and multiplies it by the number of times the event occurred.

$$\text{Variance} = \frac{408.02}{50} = 8.16$$

$$\text{Standard Deviation} = \sqrt{8.16} = 2.86$$

ABSENTEES

I	II	III	IV	V
Absences **x**	Frequency **f**	Relative Frequency **f / N**	$x - \overline{X}$	$f\left(x - \overline{X}\right)^2$
0	3	.06	-4.86	70.8
1	2	.04	-3.86	29.8
2	5	.10	-2.86	40.9
3	8	.16	-1.86	27.7
4	7	.14	-0.86	5.2
5	2	.04	0.14	.04
6	13	.26	1.14	16.9
7	2	.04	2.14	9.2
8	4	.08	3.14	39.4
9	0	.00	4.14	0
10	1	.02	5.14	26.4
11	2	.04	6.14	75.4
12	0	.00	7.14	0
13	1	.02	8.14	66.3
	N=50	**1.00**		S **= 408.02**

Descriptive Statistics for Samples

- Consistent estimators of variance standard deviation:

$$s^2 = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1} \qquad s = \sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1}}$$

- Why n-1?
  - We use n-1 pieces of data to calculate the variance because we already used 1 piece of information to calculate the mean.
  - So to make the estimate consistent, subtract one observation from the denominator.

Descriptive Statistics for Samples (con't)

- Absentee Example:
  - If sampling from a population, we would calculate:

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1} \text{ and } s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1}}$$

$$\text{Variance} = \frac{408.02}{49} = 8.33$$

$$\text{Standard Deviation} = \sqrt{8.33} = 2.88$$

Descriptive Statistics for Samples (con't)

- Sample estimators approximate population estimates:
  - Notice from the formulas, as n gets large:
    - the sample si e converges toward the underlying population, and
    - the difference between the estimates gets negligible.
  - This is known as the law of large numbers.
    - We will discuss its relevance in more detail in later lectures.
  - Mostly dealing in samples
    - Use the (n-1) formula, unless told otherwise.
    - Don't worry the computer does this automatically

Example of Variance: International Economic Data

STATISTIC	REAL GPD	GOV'T SHARE OF GPD	OPENNESS
Mean	5386.6	19.0	70.7
Standard Error	555.9	1.0	5.2
Median	3073.0	17.6	59.4
Mode	2176.0	17.6	#N/A
Standard Deviation	5331.9	9.1	49.6
Sample Variance	28429209.9	83.2	2464.7
Kurtosis	-0.5	6.7	12.2
Skewness	1.0	1.9	3.0
Range	17577.0	61.2	324.4
Minimum	409.0	4.0	16.6
Maximum	17986.0	65.2	341.0
Sum	495569.0	1748.1	6504.6
Count	92.0	92.0	92.0

Real DP

$$s^2 = \frac{(2587058117)^2}{92-1} = 28429210$$

$$s = \sqrt{28429210.08} = 5331.9$$

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1} \text{ and } s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1}}$$

Examples of data presentation

use it, don't abuse it

- Start axes at ero.
- Misleading comparisons
  - overnment spending not taking into account inflation.
    - Nominal vs. Real
- Selecting a particular base years.
  - For example, a university presenting a budget breakdown showed that since 1 6 the number of staff increased only slightly.
  - ut they failed to mention the huge increase during the 5 years before.

HOMEWOR

- Lab
  - Define one variable of each type
    - nominal, ordinal and interval
  - Print out summary statistics on each variable.
    - There will be three statistics that you will not be expected to know what they are:
      - urtosis, SE urtosis, Skewness, SE Skewness.
      - The others though should be familiar.
- Excel