Unit 2: Descriptive Statistics

Review

- Sampling design
  - Probability sampling
  - Nonprobability sampling
  - Sampling geographic data
  - Sample size, bias
- Measurement scale
  - Nominal
  - Ordinal
  - Interval/ratio
- Dataset creation, recoding, SPSS

RATs

- iRAT: 15 minutes
- tRAT: 15 minutes
- Missed questions/concepts?
- Appeal?

Ways to Summarize Data

- Tabulation
- Graph
- Mapping
- Summary statistics

Tabulations

- One variable: frequency table
- Two+ variables: cross-tabs

Group activity #1: Tabulations

- Frequency table is appropriate for
  - A. Nominal, ordinal, ratio
  - B. Nominal, ordinal
  - C. Ratio, ordinal
  - D. Ratio: continuous
  - E. Ratio: discrete

Tabulations

- Frequency table
  - Table that shows *how many* cases take on a particular value, or fall in an interval
  - For nominal, ordinal, even ratio variables
  - Frequency (cases), percentage (relative frequency) , cumulative percentage
  - Valid percentage
    - Excluding missing values

Frequency Table: Categorical

TABLE 3.4 Family Structure, U.S. Families, 1994

Type of Family	Number (millions)	Percentage
Married couple with children	25.1	36.6
Married couple, no children	28.1	41.0
Single mother with children	7.6	11.1
Single father with children	1.3	1.9
Other families	6.4	9.3
Total	68.5	99.9

Source: U.S. Bureau of the Census, *Current Population Reports.*

Frequency Table: Ratio (Continuous)

TABLE 3.3 Relative Frequency Distribution and Percentages for Murder Rates

Murder Rate	Frequency	Relative Frequency	Percentage
0.0–2.9	5	.10	10.0
3.0–5.9	16	.32	32.0
6.0–8.9	12	.24	24.0
9.0–11.9	12	.24	24.0
12.0–14.9	4	.08	8.0
15.0–17.9	0	.00	0.0
18.0–20.9	1	.02	2.0
Total	50	1.00	100.0

Table Style by Taylor & Francis

Table 4.1 Age and sex of population of Banbury Municipal Borough

Age (years)	Male		Female	
	no.	%	no.	%
0–4	1,340	11.0	1,250	9.8
5–14	1,670	13.7	1,750	13.7
15–19	950	7.8	1,120	8.7
20–4	830	6.8	820	6.4
25–44	3,420	28.1	3,310	25.9
45–59	2,200	18.1	2,260	17.7
60–4	600	4.9	700	5.5
65+	1,140	9.4	1,580	12.3
Total	12,150	99.8	12,790	100.0

Sources: GRO 1966, 10 per cent sample census.
Note: It is not possible to extract the Banbury and District survey area from the 1966 sample census.

Frequency Distribution of Number of Children: Ratio (discrete)

# of kids	Frequency	Percentage	Cumulative Percentage
0	3	15	15
1	5	25	40
2	4	20	60
3	4	20	80
4	2	10	90
5	1	5	95
6	1	5	100
	n = 20		

Source: Hypothetical Data

Exhibit 2.4: A Grouped Frequency Distribution

Frequency Distribution of Number of Children

# of kids	Frequency	Percentage	Cumulative Percentage
0-2	1	5	5
3-5	6	30	35
6-8	5	25	60
9-11	3	15	75
12-14	3	15	90
15-17	1	5	95
18-21	1	5	100
	n = 20		

Source: Hypothetical Data

Frequency Tables

- Pros:
  - Also easy
  - Useful for large datasets
  - Fairly rich description of data
- Cons:
  - Unlike a list, you can't see which case is which or compare with other variables
  - Not useful if all values are unique

Group activity #1: Tabulations

- Cross-tab is appropriate for
  - A. Nominal, ordinal, ratio
  - B. Nominal, ordinal
  - C. Ratio, ordinal
  - D. Ratio: continuous
  - E. Ratio: discrete

Cross-tab (Contingency table)

- 2 or 3 categorical variables
- For ratio variables, convert into categorical/ordinal
- Controlling one or two variables
- Frequency/relative frequency of subjects in each combined category
  - 2 variables: gender, party affiliation
  - Combined categories:
    - democratic women, democratic men, republican women, Republican men

Cross-tabulation

- Resulting table ("crosstab" or "joint contingency table"):

	Women	Men
Democrats	27	10
Republicans	16	15

Each box with a value is a "cell"

This is a table row

This is a table column

Crosstabulation

- Tables may also have row and column **marginals** (i.e., totals)

	Women	Men	Total
Dem	27	10	37
Rep	16	15	31
Total	43	25	68

This is the total N

Group Activity#2 Crosstabulation

	Women	Men	Total
Dem	27	10	37
Rep	16	15	31
Total	43	25	68

- Women are more likely to be democrats than men.
  - A) True
  - B) False

Group Activity#2 Crosstabulation

	Women	Men	N
Dem	39.7%	14.7%	37
Rep	23.5%	22.1%	31
N	43	25	68

- Women are more likely to be democrats than men.
  - A) True
  - B) False

Group Activity#2 Crosstabulation

- If we follow the placement rule, what kind of percentages should we have in cross-tabs?
  - A) row percentages
  - B) column percentages
  - C) total percentages
  - D) row and column percentages
  - E) row, column, and total percentages

Cross-tab: three rules

- The placement rule
  - DV on the side (row), IDV on the top (column)
- The percentage rule
  - Column percentage, with the placement rule
- The title rule:
  - DV by IDV

TABLE 8.1 Party Identification and Gender

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	279	73	225	577
Males	165	47	191	403
Total	444	120	416	980

Note: Data from 1991 General Social Survey.

- We are interested in demonstrating the relationship between gender and party identification using the above table.
- Is this an effective cross-tab? Why?
- Please list aspects that need to be revised

TABLE 8.2 Party Identification and Gender: Percentages Computed Within Rows of Table 8.1

	Party Identification				
Gender	Democrat	Independent	Republican	Total %	*n*
Females	48.3	12.7	39.0	100.0	577
Males	40.9	11.7	47.4	100.0	403

- The choice of row/column percentage is determined by the nature of hypothesis
- Are women more likely to be democrats? Row%
- Are democrats mostly women? Column%
- Conditional probability of being in each of the categories of the dependent variable (e.g. party identification) given that an individual is in a particular category of independent variable (gender)

Cross-tabulations: technical points

- Always include the percentage totals
- Always include the number of cases on which the percentages are based (the denominators)
- Sometimes useful to add a total column
- Control variable should be put on the outside of the tabulation, so that it changes most slowly (e.g. education)
- Some categories are combined to improve clarity and to avoid zero cells

Three Dimensional Tables

Table 2.5. Per Cent Militant by Religiosity and Educational Attainment, Urban Negroes in the U.S., 1964.

	Grammar School			High School			College		
Militancy	V	S	N	V	S	N	V	S	N
Militant	17%	22%	32%	34%	32%	47%	38%	48%	68%
Non-militant	83	78	68	66	68	53	62	52	32
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%
N	(108)	(201)	(44)	(96)	(270)	(138)	(26)	(61)	(49)

Source: Adapted from Marx, 1967a: Table 6.
*V=very religious; S=somewhat religious; N=not very religious or not at all religious.

2. Visual Representation (Graph)

- Pie chart
- Bar chart
- Histogram
- Frequency curve
- Box plot

- When to use what kind of graph is the most effective/appropriate?

Group Activity#3 Graph

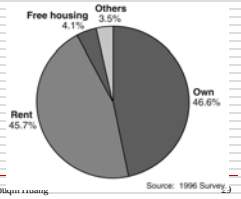
- For categorical variables (nominal, ordinal), which of the following is the most appropriate?
  - A) bar chart
  - B) pie chart
  - C) histogram
  - D) line chart
  - E) box plot

Group Activity#3 Graph

- For quantitative variables (ratio), which of the following is most appropriate?
  - A) bar chart
  - B) pie chart
  - C) histogram
  - D) line chart
  - E) box plot

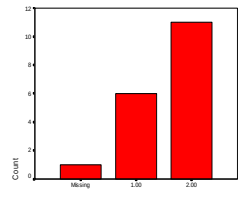
Pie Chart

- For nominal and ordinal variables
- Size of the slice represent the share of cases in a specific category (sum=100%)



Bar Chart

- For Nominal & Ordinal Variables Only
- Height of bars represent number of cases, or share of cases



Clustered Bar Chart

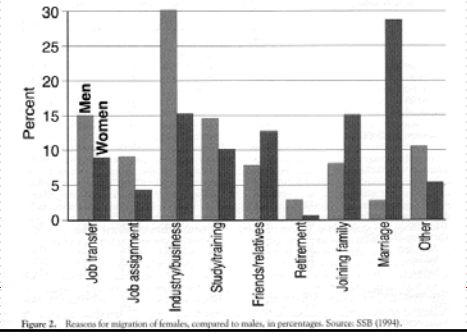
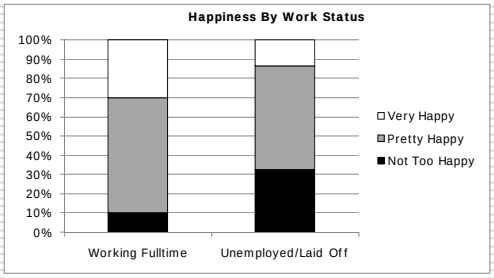


Figure 2. Reasons for migration of females, compared to males, in percentages. Source: SSB (1994).

A Clustered Bar Graph

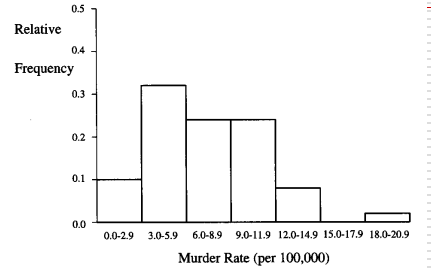
A Stacked Bar Graph



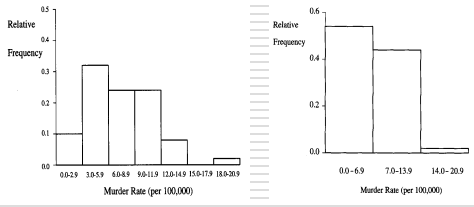
Histogram

- For continuous measures, ratio
- Height of bar represents number of cases or % **within a given range of values**
- Data needs to be "grouped", some info is lost

Histogram (Example)



Group Activity: how do you interpret these histograms?



Histogram

- Interval width (bin): Histograms look very different depending on how wide you set the value for intervals.
  - Choose width carefully
  - Try multiple widths
  - Different bins may lead to different interpretations

Frequency Curve

- Connecting the mid-points of intervals
- Use a single line, reduce the visual emphasis on boundaries

Frequency Curve

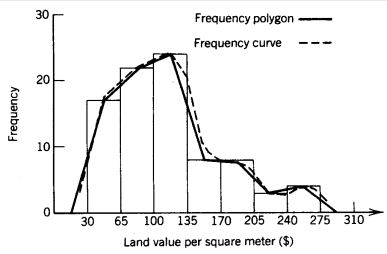


Figure 2.7 Frequency polygon and frequency curve for land value per square meter. *Source:* Urban Land Value Survey.

Sample vs. Population Distribution

- For a continuous varaible

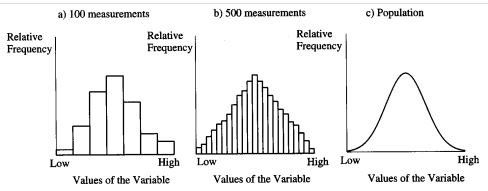


Figure 3.6 Histograms for a Continuous Variable

Frequency Curve

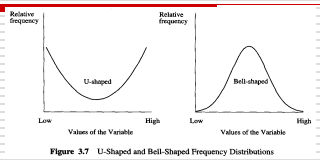


Figure 3.7 U-Shaped and Bell-Shaped Frequency Distributions

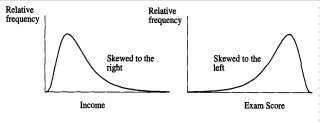


Figure 3.8 Skewed Frequency Distributions

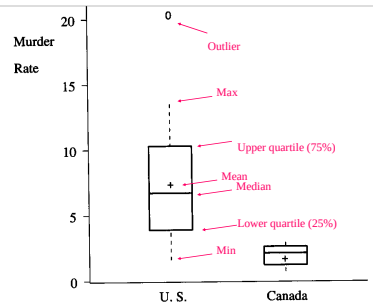
Box Plot

- Measure both central tendency and "spread"

Box Plot

Outlier: > 1.5 IQR above upper quarter or below lower quartile

How do you interpret this box plot?



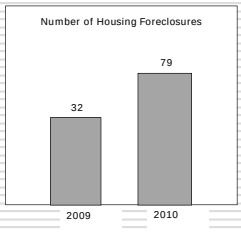
3. Mapping

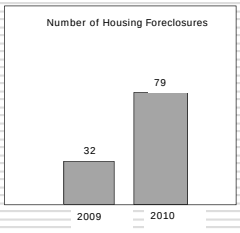
- Spatial, 3-dimensional
- Use point, line or areal symbols
- SKIP

Individual activity: Table and Graph in SPSS

- Dataset: housing sale
  - Create a frequency table for bedrooms
  - Create a frequency table for dateblt
    - Recode the variable first!
  - Create a cross-tab for dprice by garage
    - Request appropriate %
- Create charts
  - Pie: bedrooms
  - Bar: bedroom
  - Boxplot: unempl * district
  - Histogram: price

Interpret the bar chart





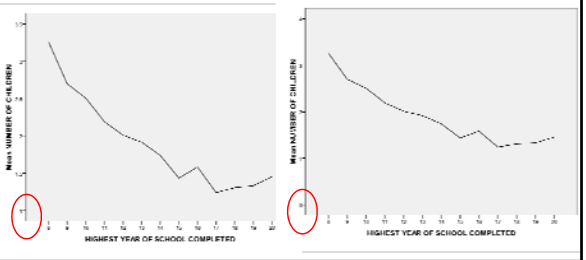
A Graph that Lies: Under-representing the change in the data

A Graph that does not Lie:

Lie Factor

- change shown in graphic/change shown in data
- 1: truthful representation of the data
- >1: over-representing, exaggerating
- <1: under representing
- Bar Chart 1:
  - Change in the data: (79-32)/32=147%
  - Change shown in graphic: height of the bars, (4.5-3)/3=50%
  - Lie factor=50%/147%=0.34
  - The chart under-representing the increase in number of foreclosure in 2010

An SPSS Graph that Somewhat Lies



Y-axis did not start at zero, lie factor=1.38, over-representing the effect of education on number of children

Recap: Ways to Summarize Data

- Tabulation
- Graph
- Mapping
- Summary statistics
  - Measure central tendency
  - Measure variability
  - Measures of skewness, kurtosis
  - Measures of relative position
  - Geographic data

Summary Measures for Frequency Curve

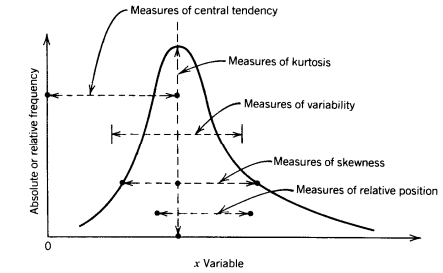


Figure 3.3 Summary measures for frequency distributions.

Measuring the Central Tendency

- The "center" of a distribution, "typical" case
  - Mean, median, mode

Variables

- Each column of a dataset is considered a variable, generally referred as "Y", or "X"

Person	# Guns owned
1	0
2	3
3	0
4	1
5	1

The variable "Y"

Central Tendency: Mean

- Arithmetic Mean, or "average", "Y-bar"
  - Sum of the Y for all cases divided by the number of subjects
  - Most frequently used measure

$$\bar{Y} = (Y_1 + Y_2 + \ldots + Y_n)/n$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Calculating the Mean

- Y_i represents "i^{th}" case of variable Y
- i goes from 1 to n
- Y_1 = value of Y for first case in spreadsheet
- Y_2 = value for second case, etc.
- Y_n = value for last case

Person	# Guns owned (Y)
1	$Y_1 = 0$
2	$Y_2 = 3$
3	$Y_3 = 0$
4	$Y_4 = 1$
5	$Y_5 = 1$

Calculating the Mean

$$\sum_{i=1}^{5} Y_i = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$$

$$= 0 + 3 + 0 + 1 + 1 = 5$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{5} \times 5 = 1$$

Person	# Guns owned (Y)
1	$Y_1 = 0$
2	$Y_2 = 3$
3	$Y_3 = 0$
4	$Y_4 = 1$
5	$Y_5 = 1$

TABLE 3.3 Work Table for Calculation of Mean and Variance for Land Value pe Square Meter

Class (j)	Frequency (f_j)
30–64	17
65–99	22
100–134	24
135–169	8
170–204	8
205–239	3
240–274	4
TOTAL	86

What is the mean land value?

TABLE 3.3 Work Table for Calculation of Mean and Variance for Land Value pe Square Meter

Class (j)	Frequency (f_j)
30–64	17
65–99	22
100–134	24
135–169	8
170–204	8
205–239	3
240–274	4
TOTAL	86

$$\bar{Y} = \frac{\sum_{j=1}^{k} M_j f_j}{n} = \frac{9817}{86} = 114.15$$

What if the last class is open-ended?

Mean of Groups

- Mean of groups is the weighted mean (average) of group means.
- Two groups of size n_1, n_2

$$\bar{Y} = (n_1\bar{Y}_1 + n_2\bar{Y}_2)/(n_1 + n_2)$$

- More generally,

$$\bar{Y} = \frac{\sum_{j=1}^{k} M_j f_j}{n}$$

Team Activity (graded): Mean

- Which of the following variables has a meaningful mean?

A) household income ($)

B) housing value ($)

C) education in highest degree (1: <hs; 2: hs; 3: college; 4: graduate)

D) age

E) gender (1: female; 0: male)

Properties of the Mean

- Pros:
  - Gives a sense of "typical" case
  - Useful for continuous data
  - Easy to calculate
  - Center of gravity

$$\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \langle \sum_{i=1}^{n}(Y_i - A)^2$$

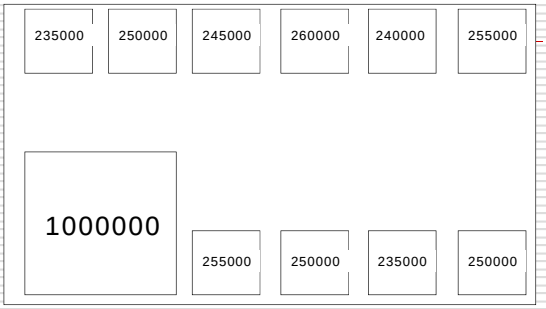
Properties of the Mean

- Cons:
  - Every case influences outcome
  - Extreme cases (outliers) affect results a lot. (e.g. Mean income is often not very meaningful)
  - Doesn't give you a full sense of the distribution
  - Appropriate only for quantitative data

The Mean and Extreme Values

Case	Num CD's	Num CD's2
1	20	20
2	40	40
3	0	0
4	70	1000
Mean	32.5	265

Hypothetical Block After One Heck of a Remodel:
Mean housing price/value is not very meaningful

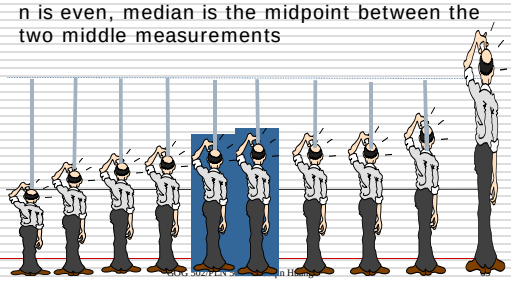


Central Tendency: Median

- The middle measurement of a ranked sample: (n+1)/2.

Central Tendency: Median

- If n is odd, median is a single measurement; if n is even, median is the midpoint between the two middle measurements



Group Activity:
What is the median education?

TABLE 3.7 *Highest Degree Completed, for a Sample of Americans*

	Highest Degree	Frequency	Percentage
A)	Not a high school graduate	38,012	21.4%
B)	High school only	65,291	36.8%
C)	Some college, no degree	33,191	18.7%
D)	Associate's degree	7,570	4.3%
E)	Bachelor's degree	22,845	12.9%
	Master's degree	7,599	4.3%
	Doctorate or professional	3,110	1.7%

Central Tendency: Median

- Median: Appropriate for both ratio and ordinal data, but not for nominal data
- Same as the mean for symmetric distributions; for skewed distribution, median lies toward the shorter tail related to the mean

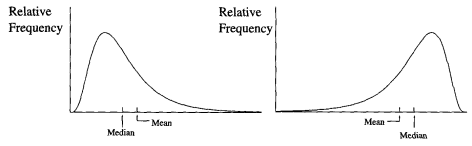


Figure 3.10 The Mean and the Median for Skewed Distributions

Team Activity: Median

- Median is not appropriate for_____?

A) household income ($)

B) housing value ($)

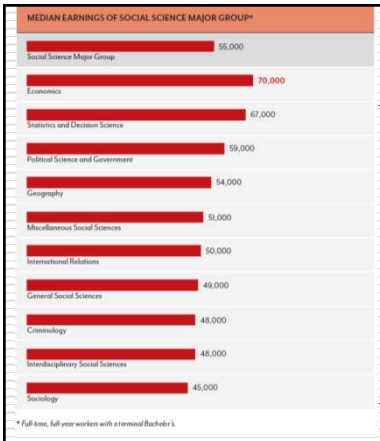
C) education in highest degree (1: <hs; 2: hs; 3: college; 4: graduate)

D) age

E) gender (1: female; 0: male)

Central Tendency: Median

- Pros:
  - Unaffected by outliers (appropriate for variables such as income, housing price)
- Cons:
  - Insensitive to the distances of the measurements from the middle.
    - 8, 9, 10, 11, 12
    - 1, 2, 10, 100, 500



How much can you earn with your degree? (by Georgetown Univ.)

Central Tendency: Mode

- The value that occurs most frequently -- the "Modal" value
- Appropriate for all types of data.
  - Commonly used for categorical (nominal, ordinal) data
  - Only useful for continuous (interval/ratio) variables if you have **grouped** data
    - Otherwise, all values may very likely be unique
- Modes = Peaks
  - Uni-modal distribution: One peak
  - Bi-modal distribution: Two peaks
  - Multi-modal distribution: Multiple peaks (usually more than two).

Central Tendency: Mode

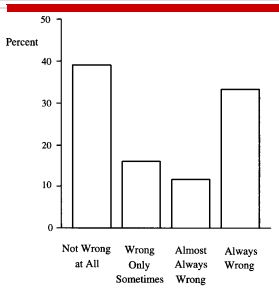


Figure 3.12 Bimodal Distribution for Opinion about Abortion

What is the mode? Why is the distribution bimodal?

Reasons for Multi-Modal Distributions

- The sample is heterogeneous (i.e., made up of more than one group)
  - Height forms a bell-shaped distribution for men and for women, but the peaks are different. A combined sample has two peaks

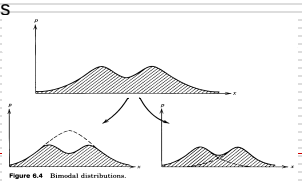


Figure 6.4 Bimodal distributions.

Reasons for Multi-Modal Distributions

- The sample is heterogeneous (i.e., made up of more than one group)
  - Height forms a bell-shaped distribution for men and for women, but the peaks are different. A combined sample has two peaks
- The sample reflects some exogenous structural ordering process
  - Years of education completed is peaked at 12 (high school), 16 (college)

Mode

- Pro: Easy, useful
- Con:
  - Do not necessarily close to the center
  - Not very helpful (even misleading) in certain circumstances, e.g. if there are many peaks, or a single unusual one; if the variable is distributed quite evenly

Comparing Mean, Median, Mode

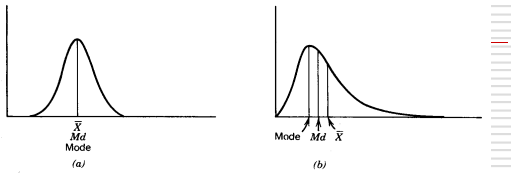


Figure 3.4 Locations of mean, median, and mode in (*a*) symmetric and (*b*) skewed unimodal distributions.

- For a symmetrical unimodal distribution, the three are identical
- For a smoothed unimodal frequency curve, the mode defines the peak value; the median divides the area under the curve into two equal parts; the mean divides the curve into two equally balanced parts through the center of gravity

Comparing Mean, Median, Mode

- Both mean and median can be easily calculated for grouped or ungrouped data. Mode is usually used for grouped data
- Unequal class intervals in grouped data do not hinder the calculation of mean, median, but severely limit the calculation of mode
- The presence of an open-ended class do not affect the median or mode, but severely limit the calculation of mean.

Group Activity: Central Tendency

- In Canada, based on the 2001 census, for religious affiliation (Catholic, Protestant, other Christian, Muslim, Jewish, None, others), the relative frequencies were 42%, 28% 4%, 2%, 1%, 16%, 7%.

A) the mean religion is Protestant

B) the mode religion is Catholic

C) The median religion is Protestant

D) only 2.5% of the subjects fall within one standard deviation of the mean

E) The Jewish response is an outlier

Levels of Measurement and Measures of the Centre

	Nominal	*Ordinal*	*Ratio*
Mode	YES	YES	YES
Median	NO	YES	YES
Mean	NO	NO	YES

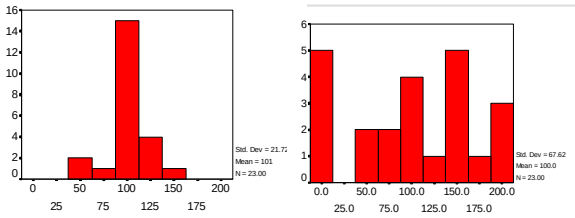
If appropriate, report all three. The differences between them tell something important about the distribution

Summary statistics

- Measures of central tendency
  - Mean, median, mode
- Measures of variability
  - Describing how "spread out" a distribution is around its center

Variability

- Very different groups can have the same means:



Number of CDs (Group 1) Number of CDs (Group 2)

Variability

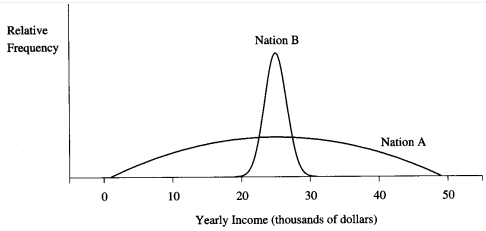


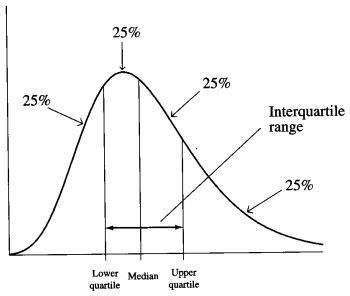
Figure 3.13 Distributions with the Same Mean but Different Variability

Which country would you prefer to live in?

Measures of Variation

- Range ($Y_{max} - Y_{min}$)
  - Doesn't tell you much about the middle cases
  - Influenced by extreme values… may not be representative
- Interpercentile (usually interquartile) range
  - Percentile: p% scores below it, (100-p)% above it
  - Lower quartile (P_{25}), upper quartile (P_{75})
  - IQR =P_{75} - P_{25}
  - Not sensitive to extreme value
  - Outlier: >1.5 IQR above the upper quartile, or 1.5 IQR below the lower quartile

Quartile and Interquartile Range



Measures of Variation

- Deviation $d_i = Y_i - \bar{Y}$
- Variance (s_Y^2) $s_Y^2 = \frac{\sum_{i=1}^{n} d_i^2}{n-1} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$
- Standard deviation (s_Y) $s_Y = \sqrt{s_Y^2} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}$

Example

Case	Num CD's
1	3
2	5
3	1
4	7

Variance? Standard Deviation?

Example

Case	Num CD's	Mean (Ybar)
1	3	4
2	5	4
3	1	4
4	7	4

Variance? Standard Deviation?

Example

Case	Num CD's	Mean (Ybar)	Deviation (Yi-Ybar)
1	3	4	-1
2	5	4	1
3	1	4	-3
4	7	4	3

Variance? Standard Deviation?

Case	Num CD's	Mean (Ybar)	Deviation (Yi-Ybar)	Square of deviation
1	3	4	-1	1
2	5	4	1	1
3	1	4	-3	9
4	7	4	3	9
sum			0	20

$s_Y^2 = \frac{\sum_{i=1}^{n} d_i^2}{n-1} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$

Variance=20/(4-1)=6.67
St.Dev=sqrt(6.67)=2.58

Properties of Variance (s_Y^2)

- $s_Y^2 >= 0$
  - Zero if all points cluster exactly on the mean
  - Larger for more "spread" distributions
- Pros:
  - Comparable across samples of different size
- Cons:
  - Values get fairly large, due to "squaring"

Properties of Standard Deviation

- s >= 0
- s=0 when all observations have the same value, grows larger if points are spread further from the mean
- s is the average distance of an observation from the mean
- Most commonly used measure of dispersion
- Comparable across different sample sizes
- The Empirical Rule

Team Activity (graded)

- Data: 1,1,1,2,2,3,3,5,1,1
- By hand, find the mean, media, mode, variance, and standard deviation

Empirical Rule

If the histogram of the data is approximately bell-shaped, then

1. About 68% of the data fall between $\bar{Y} - s$ and $\bar{Y} + s$.
2. About 95% of the data fall between $\bar{Y} - 2s$ and $\bar{Y} + 2s$.
3. All or nearly all the data fall between $\bar{Y} - 3s$ and $\bar{Y} + 3s$.

The rule is called the Empirical Rule because many distributions encountered in practice (that is, *empirically*) are approximately bell-shaped. Figure 3.15 is a graphical portrayal of the rule.

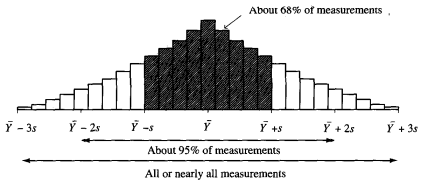


Figure 3.15 Empirical Rule: Interpretation of the Standard Deviation for a Bell-Shaped Distribution

Team Activity: Standard Deviation

- SAT math score is approximately bell shaped, with a mean 500 and standard deviation 100. How many people scored more than 700?
- A) 0.5%
- B) 1%
- C) 2.5%
- D) 4%
- E) 5%

The Alternative

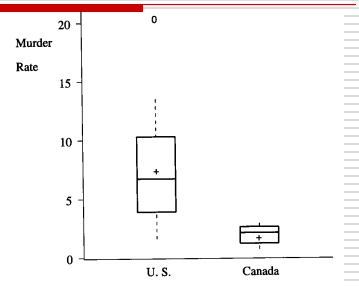
- Average Absolute Deviation (AAD)
  - Very intuitive interpretation
  - Has non-ideal statistical properties

$$AAD = \frac{\sum_{i=1}^{N} |d_i|}{N} = \frac{\sum_{i=1}^{N} |Y_i - \bar{Y}|}{N}$$

Measure Central tendency and Variation

- Box plots

How to interpret this box plot?



Measuring Skewness

- Is the distribution symmetrical?
- Skewness measuring the degree of asymmetry around a measure of central tendency
- Zero = perfectly symmetrical
- Higher number = increasingly skew

Measureing Skewness

- A "tail" is referred to as "skewness"
  - Tail on left = skewed to left = negative skew
  - Tail on right = skewed to right = positive skew
- Pearson's Coefficient of Skewness
  - Based on distance from Mean to Median
  - Mean moves more if there are extreme cases, as when there is a "tail"

$$\text{skew} = \frac{3(\bar{Y} - \text{Mdn})}{s_Y}$$

Measuring Skewness

- Pearson's Coefficient of Skewness
- Quartile skewness
  - Measures distance between median and lower & upper quartiles
  - Extreme values move lower/upper quartiles further out, resulting in larger skewness

$$\text{skew} = \frac{P_{25} + P_{75} - 2Mdn}{2}$$

Skewness in SPSS

$$\text{skew} = \frac{n}{(n-1)(n-2)} * \sum\left(\frac{(X_i - \bar{X})}{s}\right)^3$$

Group Activity

At a business, we find that the male employees' salary distribution is positively skewed, whereas the female employees' salary distribution is negatively skewed. Which of the following is true about this business?

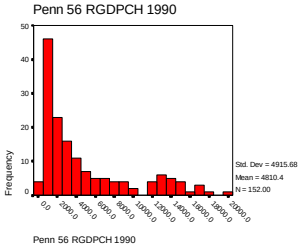
A) there are a small number of women have very low salary, and a small number of men have very high salary

B) there are a small number of women have very high salary, and a small number of men have very low salary

C) there are a large number of women have very low salary, and a large number of men have very high salary

D) there are a large number of women have very high salary, and a large number of men have very low salary

Interpreting Skewness



Which way is it skewed?
What is the social interpretation?
What would be the interpretation if it were skewed in the opposite direction?

- Skewness provides information about inequality
  - Example: Economic wealth of nations

Interpreting Skewness

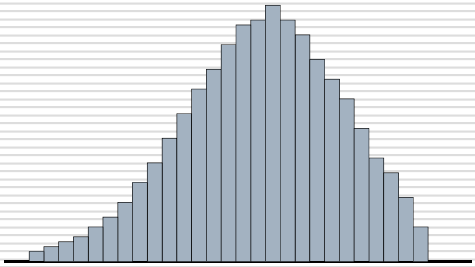
- Skewness may reflect "floor" or "ceiling" effects
  - Example: Number of crimes committed by individuals in a sample. Lower bound is zero. Mode is very low. A few cases are high.
  - Example: National secondary school enrollment ratio. Cannot exceed 100%

Notes on Skewness

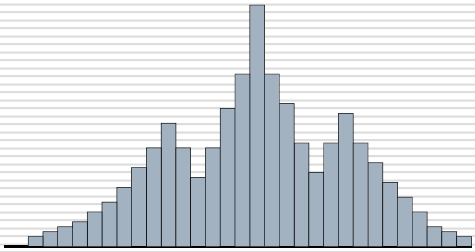
- More often assessed informally "by eye" than calculated as a value.
  - Look at a histogram to identify skewness
- Some statistical techniques work properly **only** on variables that are **not skewed** (e.g. empirical rule).
  - It can be very important to identify highly skewed variables.
- Note: mode, skew sound like "jargon", but are actually quite helpful in communicating descriptive information about your variables

Example:

- How would you describe this variable?



Example: How would you describe this variable?



Measuring Relative Position

- Rank (R_i)
  - Sort the data, the position of score
- Cumulative frequency list/curve
  - Number of cases (percentage of cases) falling in or below a given interval

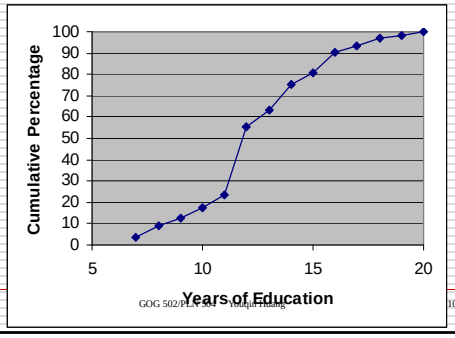
Cumulative Frequency List

Years of Education (N=2904)

Value	Frequency	Percent	Cumulat %
7 or less	21	1.4	3.9
8	82	5.3	9.3
9	51	3.3	12.6
10	70	4.6	17.2
11	95	6.2	23.4
12	489	31.8	55.4
13	125	8.1	63.5
14	184	12.0	75.6
15	76	4.9	80.5
16	152	9.9	90.5
17	40	2.6	93.1
18	61	4.0	97.1
19	18	1.2	98.2
20	27	1.8	100.0

Indicates that 55% of students have 12 years of education or less

Cumulative % Graphs



Measuring Relative Position

- Rank (R_i)
- Cumulative frequency list/curve
- Quantile
  - Percentiles, quartiles, deciles, etc…
  - General term = quantile
  - Dividing cases up into fixed number of equal "chunks"
    - 100 chunks = percentiles (1% each)
    - 10 chunks = deciles (10% each)
    - 5 = quintiles (20% each)
    - 4 = quartiles (25% each)

Is History Siding With Obama's Economic Plan? (NY Times)

Family Income Growth

Annual average for 1948-2005, by income level. *Adjusted for inflation.*

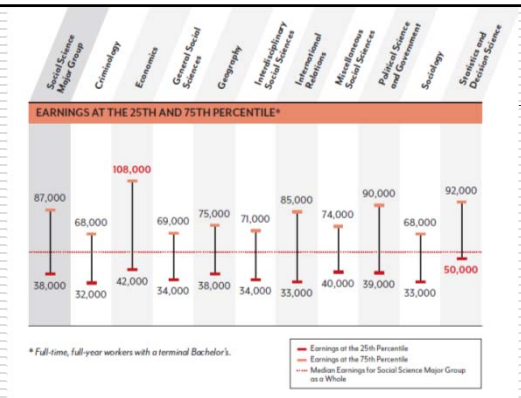
Percentile	Under Democratic presidents *26 years*	Under Republican presidents *32 years*
20th	+2.64%	+0.43%
40th	+2.46	+0.80
60th	+2.47	+1.13
80th	+2.38	+1.39
95th	+2.12	+1.90

Sources: Census Bureau; Larry M. Bartels

THE NEW YORK TIMES

Benefit of Quantiles

- Quantiles allow you to identify cases (or groups of cases) in relation to the larger group
  - Who is "high", who is "low"
  - Regardless of the unit of measurement
- Advantage: Allows comparison between variables with different scales (or with different means)
  - Example: Reading test scored 1-100, Math test is scored 1-25. How do you know which you scored better on? Answer: percentile



How much can you earn with your degree? (by Georgetown Univ.)

Measuring Relative Position: Ratio

- Ratio
  - The position of an individual score in relation to some other score (e.g. mean, max, min…)
  - Standardized score (Z-score): deviation from mean divided by standard deviation

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

Z-Score Example

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

- Number of CD's: Mean = 32.5, s = 29.8

Case	Num CD's (Y)	Mean (Y bar)	Deviation (d)	Z-score (d_i/s)
1	20	32.5	-12.5	-.42
2	40	32.5	7.5	+.25
3	0	32.5	-32.5	-1.1
4	70	32.5	37.5	+1.3

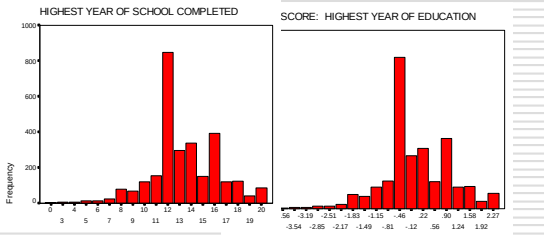
Z-Score (Standardized Score)

- Unit of Z-scores is "standard deviation"
  - A Z-score of -1.1 indicates a case is nearly one standard deviation below the mean
  - Z=0.5 → 0.5 St. Dev above the mean
- You can convert any or all values of a variable to a common scale
  - mean = 0
  - negative = below mean
  - positive = above mean
  - range approximately from –3 to +3. WHY?

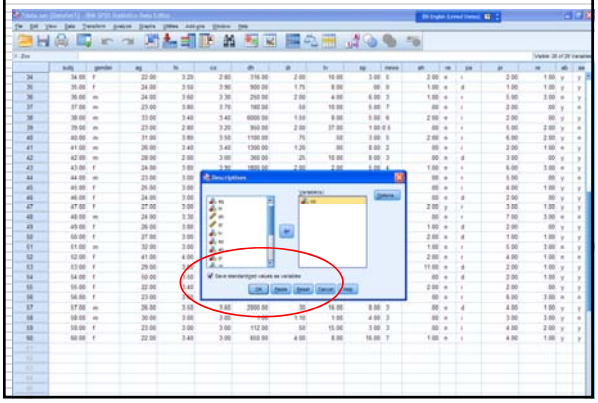
Z-Scores

- Z-scores can be compared across variables with different units or means
  - Examples: height and weight; a person is -0.3 on math, but 1.2 on income
  - Simple deviations can't be compared if units of measurement are different
- Convert an entire variable (all cases) to Z-scores, creating a new variable with useful properties
  - preserves the shape of the distribution, but unit is changed
  - Mean = zero, because it is based on deviations
  - Standard Deviation (s_y) = 1
  - Easier to compare different variables

Converting Variables to Z-scores GSS Data, N=2904



Create Z-score in SPSS



Team Activity: Z-score

Here is a frequency distribution of number of household members less than 6 years old for respondents aged 20 to 29:

Number of children (<6 years old)	Frequency
0	216
1	59
2	36
3	6
4	2

For variable "number of children", mean=0.49, s=0.82.

Use z-score to describe the "uniqueness" of someone who has four young children.

Summary statistics

- Measures of central tendency
  - Mean, median, mode
- Measures of variability
  - Range, IQR, variance, s.d., AAD
- Measures of skewness
- Measures of relative position
  - Rank, quantile, Z-score

Special case: Dichotomous Variables

- Mean, variance, and S.D. are generally NOT too useful for nominal variables
- Exception: Mean of dichotomous variables
  - Dichotomous variable = nominal, w/ 2 categories, often called "dummy" variables
  - E.g.: Do you approve of gun control (yes/no)?
  - People saying "yes" assigned 1, no = 0

Dichotomous (Dummy) Variables:

1 = Presence of something, 0 = absence of it

Person	View On Gun Control	Support? (Dummy)
1	Favor	1
2	Oppose	0
3	Favor	1
4	Favor	1
5	Oppose	0

1 = Presence of support for gun control

0 = Absence of support for gun control

Dichotomous Variables

- Interpretation:
  - Mean = proportion indicating yes
- Example: "Do you approve of gun control?"
  - 14 yes, 24 no. (37% yes, 63% no)
  - Mean of variable = .37

Dichotomous Variables

- The Standard Deviation for dichotomous variables

$$s_Y = \sqrt{s_Y^2} = \sqrt{(p_0)(p_1)}$$

- p_0 = the proportion of cases scoring 0
- p_1 = the proportion of cases scoring 1
- Q: What is s_Y if the sample is half 0, half 1?
- Answer:
  - p_0 =0.5, p_1 =0.5
  - square root of 0.25, = 0.5

Geographic Data

- Statistical and spatial distribution
- Summary statistics:
  - Important to accessibility and dispersion
  - Centrality: mean center, median center
  - Dispersion: standard distance, relative distance

Geographic Data: Mean Center

- Minimize the sum of squared distance
  - Point data:

$$(\bar{X}, \bar{Y})$$

  - Areal data:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}; \bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

TABLE 3.8 Coordinate Locations for California Metropolitan Areas

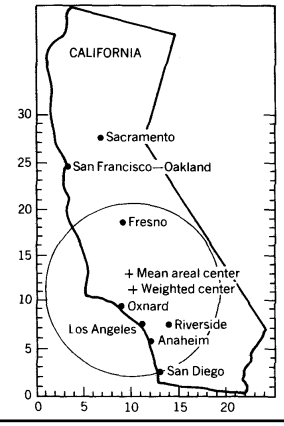
	x Coordinate	y Coordinate	Population (Millions)	Weighted x	Weighted y
Sacramento	7	27	1.0	7.0	27.0
San Francisco/ Oakland	3	24	3.3	9.9	79.2
Fresno	9	18	0.5	4.5	9.0
Oxnard/Ventura	9	9	0.5	4.5	4.5
Los Angeles	11	7	7.5	82.5	52.5
Riverside/ San Bernardino	14	7	1.6	22.4	11.2
Anaheim	12	5	1.9	22.8	9.5
San Diego	13	2	1.9	24.7	3.8
TOTAL	78	99	18.2	178.3	196.7

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}; \bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

- Mean of X=78/8=9.8
- Mean of Y=99/8=12.4
- Weighted mean of X=178.3/18.2=9.8
- Weighted mean of Y=196.7/18.2=10.8

Mean Center

Why are they different?



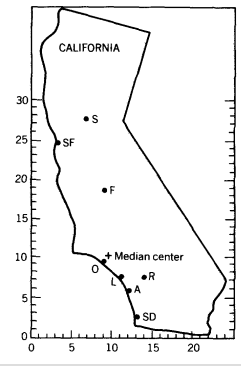
Geographic Data: Median Center

- **NOT** the point defined by the medians of x and y
- Minimum aggregated distance to all points

$$\sqrt{\sum_{i=1}^{n}[(X_i - X_0)^2 + (Y_i - Y_0)^2]}$$

- Application in location theory

Median Center



Geographic Data: Standard Distance

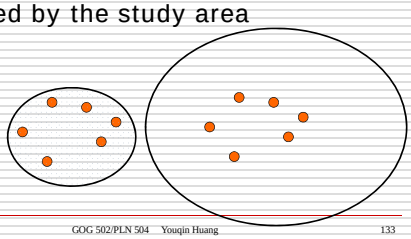
- Spatial equivalent to standard deviation
- Average distance of observations to mean center, or radius around mean center

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}}$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n} d_{ic}^2}{n}} \qquad SD = \sqrt{s_x^2 + s_y^2}$$

Geographic Data: Standard Distance

- affected by the unit which distance is measured
- Affected by the study area



Geographic Data: Relative Distance

- Dividing SD by the radius of a circle with area equal to the size of the study area

$$RD = \frac{SD}{r}$$

Summary

- Measures of central tendency
  - Mean, median, mode
- Measures of variability, skewness, kurtosis
  - Variance, standard deviation, range, IQR
  - Pearson's coefficient, Quartile skewness, kurtosis
- Measures of relative position
  - Rank, quantile, Z-score
- Dummy variables
- Measures for geographic data
  - Mean center, median center, standard distance, relative distance

Individual activity: SPSS application

- Dataset: crimedata
- Variable: crime rate; poverty rate
- Find mean, median, mode; variance, standard deviation; quartiles, IQR
- Interpret the results, and compare these two variables.
- Create a z-score for crime rate; what can you say about Washington DC?

SPSS: Descriptive Statistics