

HANDOUT 1 STATISTICS AND VARIABLES

- > Descriptive and inferential statistics
- > Population and sample
- > Dependent and independent variable
- > Level of measurement

Statistics

Refer to the numbers that summarize information quantitatively.

Refer to the methods used to calculate numbers and to generalize them and refer to ways we make information more manageable.

A Examples of statistics are average age at marriage in Eau Claire, Income level in WI, enrollment rate in UWEC, Birth rate in Chippewa Falls, Average SAT in class 4, Percent of urban population, Percent of girl students in SOC 331, average height of students in UWEC. etc.

Data

Are records of observations and the unsummarized raw information.

Data come from surveys, experiments, or from systematic observations of any kind.

Data Set

Data in an organized form or in some systematic way are called data set.

Data Files

Data Sets stored so that they can be read by a computer are called data files. Some data files include descriptions of the data or are stored on the computer diskette such as the General Social Survey (GSS).

Unit of Analysis

It is an entity that is studied. It is the person, object, or event that a researcher is studying such as individuals, students, villagers, cities, countries, schools, families, etc.

Descriptive and Inferential Statistics

Descriptive Statistics refer to methods for summarizing information so that the information is more useful and can be communicated more effectively. Procedures used for organizing and summarizing data.

Inferential Statistics refer to procedures used to generalize from a sample to the larger population. Draw conclusions about a population from a sample drawn from that population.

Populations and Samples

Populations pertain to all or almost all cases to which a researcher wants to generalize. Whenever information is from all or nearly all cases that a researcher wants to describe, the data is population data.

A sample refers to a subset of cases or elements selected from a population or refers to a smaller number of individuals taken from a population (for the purpose of generalizing to the entire population from which it was taken).

Parameters and Statistics

A parameter is a characteristic of a population.

Such as the average age, percent of married, and average of income of *all Americans* are examples of parameters since they pertain to an entire population. Other examples are the average at marriage for American, Women's labor force participation rate in USA, % of the world's nations that are highly industrialized, etc.

A statistic is a characteristic of a sample.

Such as the average of age, percent of married, and average of income of *a sample of Americans* are statistics. Other examples are percentage of GSS respondents who voted in election, the percentage of adults watching TV in Eau Claire, Car accident rate of a sample of 500 Eau Claire residents. Etc.

Variables and Constants

A variable is any characteristic that varies. It is some property that differs in value from one case to another.

Your family income, Educational attainment in years, Height, Weight, President's popularity, SAT and ACT scores, Population size, Crime rates in Eau Claire, Attitudes toward abortion, Support for legalization of marijuana, etc.

A constant is something that never varies. It is opposite to a variable

The United States, Eau Claire, UWEC, Nixon's age when he died, Number of historical sights in Madison, etc.

Independent and Dependent Variables

An independent variable has a causal role in relation to the dependent variable. It is a cause of the dependent variable.

Education predicts Income, Income predicts Prestige, Dad & Mom's Education predicts Son's Education, Sun rays predict plant growth, Rain predicts river flooding, Fires predict property damages, SAT predicts Scholarships, ACT scores predict awards, etc.

A dependent variable has a consequent or affected role in relation to the independent variable. It is a result of an independent variable.

A There are numerous dependent variables such as Height, Health, Longevities, Good SAT scores, Marital quality, Family life satisfaction, Reputations, Income, Occupation, etc.

Continuous and Discrete Variables

A continuous variable may in principle take on any value in its range of possible values.

A For example, age (years, months, days, hours, minutes, seconds, etc.), time (hours, minutes, seconds), attitudes toward premarital sex - always wrong, usually wrong, sometimes wrong, never wrong, Percent of WI population, Per capita income of WI residents, etc.

A discrete variable has only certain values within its range.

A For example, family size, # of friends / courses / CDs, Dwelling types - single house, mobile, apartment, etc. Sex ratio = # of males per 100 females, etc.

Levels of Measurement

Measurement is the assignment of numbers or labels to units of analysis to represent variable categories. The various meanings of these numbers are what are meant by the levels of measurement.

Nominal - the lowest level of measurement. It is a system in which cases are classified into two or more categories on some variable such as gender, race, religious preference, etc. These categories are mutually exclusive and exhaustive. Numerals are merely labels and quantitatively meaningless.

For example, a marital status / Religious preference may be defined and measured as

- | | | | |
|----|----------|----|------------|
| 1. | married | 1) | Protestant |
| 2. | single | 2) | Catholic |
| 3. | divorced | 3) | Jewish |
| 4. | widowed | 4) | Others |
| 5. | other | | |

Dichotomous and non-dichotomous nominal variables

A dichotomous variable has exactly two values. Dichotomous variables are common in social sciences.

A For example, geographic location: rural and urban; place of birth: native or foreign; Sex/Gender: male and female; etc.

But there are many non-dichotomous nominal variables.

A Such as region (urban, rural, towns, cities, costal areas, mountain areas, etc.); ethnicity (Americans, Africans, Laos, etc.); academic major (Sociology, Social work, Business, Accounting, etc.); Dwelling types (single house, apartment, trailer, dormitory, etc.); Favorite TV shows/channels (Discover, NSC news, Local, Animal world), etc.

Ordinal- the process of ordering or ranking cases in terms of the degree to which they have any given characteristic. Numbers indicate the rank order of cases on some variable. The categories can be rank ordered. For example, an attitude toward abortion/ Opinion on work performance can be measured as:

- | | | | |
|----|-------------------|----|----------------------|
| 1. | strongly disagree | 1) | Very satisfactory |
| 2. | disagree | 2) | Satisfactory |
| 3. | no opinion | 3) | Dissatisfactory |
| 4. | agree | 4) | Very dissatisfactory |
| 5. | strongly agree | | |

Interval- the process of assigning a score to cases so that the magnitude of differences between them is known and meaningful. Can be rank ordered. Interval variable is measured using some fixed unit of measurement.

A Such as the dollar, the year, the pound, or the inch. Numerals represent mathematical equivalences on the variables being measured. Temperature: 0, 10, 20,90, 100, is another example.

Ratio- like an interval variable, it has a standard unit of measurement; unlike an interval variable, a ratio variable has a nonarbitrary zero point. That is, the zero point represents the absence of the characteristic being measured. Ratio variables convey the most information and thus have the highest level of measurement. As there are not many truly interval variables in social sciences, many of the interval and ratio variables are often used interchangeably.

A Such as Crime rate, percent of urban/rural, poverty rate, employment rate, marriage rate, birth rate, etc.

Table 1. Level of Measurement of Variables

Level of measurement	Is there an intrinsic order to values	Is there a standard unit of measurement
Nominal	No	No
Ordinal	Yes	No
Interval	Yes	Yes
Ratio	Yes	Yes

Table 2. Information Provided by the Four Levels of Measurement

Information Provided	Nominal	Ordinal	Interval	Ratio
Classification	X	X	X	X
Rank Order		X	X	X
Equal Interval			X	X
Absolute Zero				X

Aggregate Data

Individual scores or individual respondents are combined into larger groupings. Data of this kind in which cases are larger units of analysis are called aggregate data. Data of these kinds are birth rates, death rates, population density, urbanization rate, divorce rate, marriage rate, suicide rate, unemployment rates, etc.

Ecological Variables

If the larger units are spatial or geographic areas like states, provinces, or countries, the aggregate data are ecological data and their variables are ecological variables.

Examples are murder rate, percent urban, average income, and percent Hispanic in the 50 states, average age at marriage in Eau Claire, # of newspapers in WI, Percentage voting for Democrats, etc.

Ecological Fallacy

Inferring individual characteristics from analysis of aggregate data (1), or inferring aggregate characteristics from data for individuals (2), entails a logical error, called Ecological Fallacy.

For instance, Japan has a high suicide rate for elderly over 80s. Since Sakio is a Japanese aged 80 and above, she may commit a suicide (1).

It is found from an individual survey of 100 people who are poor are likely to steal cars. Thus it is argued that states with higher proportion of poor people are likely having more auto theft (2)

Mutually Exclusive and Collectively Exhaustive

Variables are most useful if their values are both mutually exclusive and collectively exhaustive.

Mutually Exclusive means that the values do not overlap. Each case has only one value. For example, a person can be male or female, but not both male and female. Likewise, a person can be Protestant or Catholic, but not both Protestant and Catholic.

Collectively Exhaustive means that the set of values include all cases. Every case falls into some category. For example, the categories such as Protestant, Catholic, Jewish, none, and other include all possible answers to a question about religious preference.

Dr. Ji
SOC 331

HANDOUT 2

FREQUENCY AND PERCENTAGE DISTRIBUTION

Frequency Distribution

The summarization of the pattern of variation of a variable is a frequency distribution. Frequency is helpful in summarizing information.

EXAMPLES 1: Pattern of Variation of Grades in Class One

Table 1: Grades in Social Statistics (Raw Data)

Case #	Grades
1	A
2	A
3	A
4	B
5	B
6	B
7	B
8	C
9	C
10	C
11	D
12	D
13	D
14	F
15	A
16	B
17	B
18	B
19	C
20	C
21	C
22	B
23	D
24	D
25	B

Table 2: Tally and Categorize Grades (Numbers)

Grades Level	Tally	Frequency		
A	////	4	499	49999
B	////////	9	999	99999
C	/////	6	699	69999
D	////	5	599	59999
F	/	1	99	999
Total (N=size)		25	2895	280995

Disadvantages of Frequency

Although good in summarizing information, frequencies of different variables are difficult to interpret, especially in larger numbers. Frequency distributions are hard to compare with bigger cases such as US census. Percentage distribution can solve the problem.

Table 3: Frequency Table (percentage / %)

Grades Level	Frequency (f)	Percent % (f/N × 100)		Rounded %
A	4 (4/25k)=	16	49999	18 (17.7935)
B	9 (9/25k)=	36	99999	36
C	6 (6/25k)=	24	69999	25
D	5 (5/25k)=	20	59999	21
F	1 (1/25k)=	4	999	0 (0.0035)
Total (N)	25	100	280995	100

Percentage Distributions

Percentages are what frequencies would be if there were a total of 100 cases. In other words, percentages are the proportion taken by the frequency from a total of 100. Percentage distribution is the standardized summary distribution of the pattern of variation of a variable. It reduces the magnitudes of large frequencies to manageable numbers (percentages) that range from 0 to 100 and percentages can be easily compared.

When to percentage? When N is or over 100. Some argue $N > 30$ or $N > 50$ is not reliable. So let us keep N over 100 and it is safe to have percentage.

Percentage Formula

f = frequency, N = total number of cases

Percent = frequency divided by total number of cases

$$= f/N \times 100$$

Table 4: Grades Distribution in Percentages

Grades Level	Frequency (f)	Percent % $f/N \times 100$
A	4	$4/25 \times 100 = 16$
B	9	36
C	6	24
D	5	20
F	1	4
Total (N)	25	100% (16+36+24+20+4=100)

Proportion and Percentage

Percentage = $f/N \times 100$, the value ranges from 0 to 100. So the percentage of Grades A is 16 percent and F 4 percent.

Proportion = f/N , the value ranges from 0 to 1. So the proportion of A grades is .16 while that of F is .04.

Cumulative Distribution

Cumulative percentage is the percentage of all scores that have a given value or less. To find a cumulative percentage, one needs to 1) add all frequency for the given value and all lesser values; 2) divide that sum by the total number of cases; 3) multiply that result by 100.

Cumulative percent = $F/N (100)$, where F (the capital letter F) is the cumulative frequency and N is the total number of cases.

Table 5: Grades Cumulative Distribution

Grades Level	Frequency f	Percent %	Cumulative %
A	4	16	16
B	9	36	52
C	6	24	76
D	5	20	96
F	1	4	100
Total (N)	25	100	100

Collapsing Variables

To collapse value categories into a more manageable number for frequency or percentage tables. For example, age grouping, educational levels, etc.

To collapse in a way that makes a sense to your research objective:

- Establish categories that are collectively exhaustive and mutually exclusive;

- Don't obscure important patterns of the distribution;

- Keep categories homogeneous;

- Keep the same width;

- Follow cultural conventions (\$0 – 19,999, \$20,000 – 39,999, etc);

- Establish categories with about equal numbers of cases.

Table 6: Collapsing Ages

0	0-4	0-9
1		...
2		...
3		...
4		...
...	5-9	...
...	10-14	10-19
...	15-19	20-29
...
100	95-99	90-99

Table 7: Collapsing Education (Years)

0	0-6
1	7-8
2	9-11/12
3	...
4	...
...	13-15
...	16
...	17-20
...	...
21	21

Missing Data

Clear up data that are missing so that one can work with the VALID DATA. Missing data are usually recorded as DK=don't know; NA and N/A=not available/no answer/not applicable (NAP). Some are written as No Answer, No Opinion, Can't Answer, Not Ascertain, Refused, inapplicable, etc. Or coded as 8, 9, 999, etc.

Missing data is included or excluded makes a difference in result.

For example, to answer the question:

“Is US likely in World War in 10 years?” Yes or No.

The following answer of percentages both increased when missing data were excluded.

Table 7: Missing Data Included or Excluded

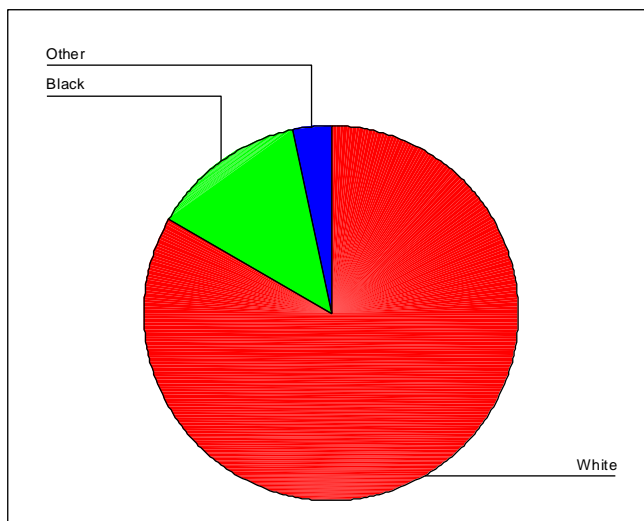
Makes a Difference in Results
(in percentages)

Answer	Missing Data	Missing Data
--------	--------------	--------------

	Included	Excluded
Yes	38.6	41.1
No	55.4	58.9
Don't Know	4.5	----
No Answer	1.6	----
Total	100.0	100.0
(N)	1960	1841

Pie Charts

- 1) One of the most common graphing technique for univariate analysis is pie chart; pie chart is quicker for human to grasp information;
- 2) Display either frequency or percentages;
- 3) The greater the percentage of the cases, the larger is the slice of the pie;
- 4) Due to limit, pie charts are usually used for variables with less than 8-9 values;
- 5) Pie charts are usually for nominal variables such as religious preference, marital status and dichotomous variables such as Male versus Female, Yes versus No, Urban versus Rural, etc.



White=83.3%

Black=13.4%

Other=3.2%

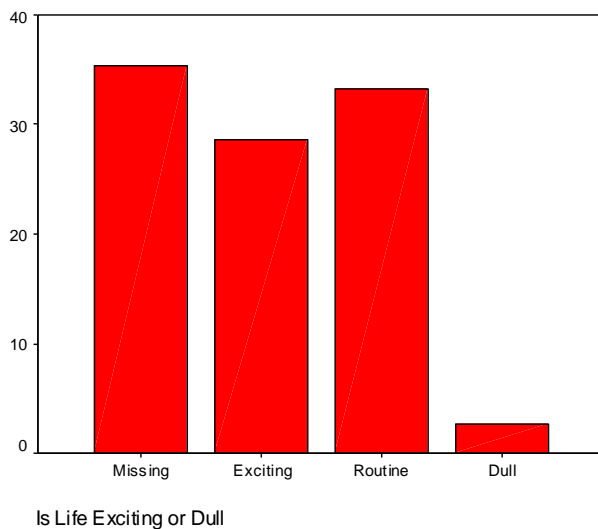
*1991 US GSS Data. Variable= race of respondent.

Bar Graphs

- 1) One of the most common graphing technique for univariate analysis;

- 2) Bar graphs can allow variables to use more than 10 or more values;
- 3) Display for either frequency or percentages;
- 4) The more cases with a given value, the higher the bar;
- 5) Bar graphs are usually separate while histogram touch to each other;
- 6) Usually for multi-valued ordinal and interval/ratio variables;
- 7) histograms are usually with continuous ordinal or interval/ratio variables such as hours watching TV or years of educational attainment (treated as discrete also).

See the following example of bar graph representing the question” Is life exciting or dull?



Exciting=28.6%; Routine=33.3%; Dull=2.7%; Missing=34.5%

*1991 US GSS Data. Variable= is life exciting or dull?

Conventions that should follow in creating bar graphs:

- 1) keep the scales of graphs consistent;
- 2) bars should be of equal width;
- 3) should be spaces between bars;
- 4) label the vertical and horizontal axis (with frequency/percent or names);
- 5) title the graphs.

Outliers

Scores on variables that stand alone as unusually high or unusually low, and that are isolated at the end of a distribution, remote from most other scores. They are also called extreme scores or numbers. For instance, based on US GSS survey, the average hours for watching TV a day ranges from 0 to 24 and the majority rank from 1 to 4 a day. But some reported 8, 10, 15, 18, 22, and 24 hours a day? Therefore,

- 1) Outliers may be exaggerated report;
- 2) Outliers may be mistakes in data collection or data management;

- 3) Do they make sense?
- 4) Determine the meaning or theoretical implications;
- 5) Decide to delete or to keep;
- 6) Can affect analysis results.

Mapping Ecological Variables

Ecological data are often continuous interval/ratio variables such as rates, averages, or percentages. Spatial distributions over geographical areas are better displayed on maps. For example, fertility rate among 50 countries can be best presented by area maps (see textbook p.51) and 50 most populous nations can be seen by spot map (see book p.51).

FERT.RT -- THE AVERAGE NUMBER OF CHILDREN THAT WOULD BE BORN PER WOMAN
WOMEN LIVED TO THE END OF THEIR CHILD BEARING YEARS AND BORE CHILDREN /



Area map gives us an overall picture of the areas of high fertility countries.

FERT.RT -- THE AVERAGE NUMBER OF CHILDREN THAT WOULD BE BORN PER WOMAN
WOMEN LIVED TO THE END OF THEIR CHILD BEARING YEARS AND BORE CHILDREN /



Spot map provides a clearer picture of where the high fertility is concentrated.

Subset of Cases

A subset of cases is one selected for analysis based on their scores on some particular variable. For example, respondents over age 65, or African-American respondents, or students from Eau Claire.

Conventions for Tables

When constructing a table, one should pay attention to the following:

Title of the table

Values are mutually exclusive

Values are collectively exhaustive

Consistent decimal places

Total percentage should be 100

Present total percent and total number of cases

HANDOUT 3 AVERAGES

Averages

An average is a typical value for set of scores. In statistical jargon, it is called “measure of central tendency.” There are many kinds of averages of which three are addressed in this class: Mode, Median, and Mean.

Mode (Mo)

Mode is the most frequent, most typical, or most common value in a distribution. In other words, we find the mode simply by identifying the value that occurs most often among scores.

Example 1,

In the distribution, “1, 2, 3, 1, 1, 6, 5, 4, 1, 4, 3,”

Mo=1, because it is the number that occurs more than any other score in the set.

Example 2,

In the distribution, “6, 6, 7, 2, 6, 1, 2, 3, 2, 4,”

Mode= 6 and 2, because 6 and 2 are the two points of maximum frequency, suggesting two humps on a camel’s back.

Example 1 is a Unimodal – where there is only one score that is most common or the general distribution shows only one hump. That is, only one dominant score or a pronounced hump is in a distribution.

Example 2 is a Bimodal – where there are two humps in a distribution or the distribution shows a camel-shaped distribution with two humps (see p-63).

Modes can be found at any level of measurement-nominal, ordinal, or interval/ratio.

The mode is the only kind of average that can be used for nominal variables.

Religious Affiliation

- | | | |
|---|------------|---------|
| 1 | Protestant | ←← Mode |
| 2 | Catholic | |
| 3 | Jewish | |
| 4 | Protestant | ←← Mode |
| 5 | Buddhist | |

Modes can be identified with the help of a bar graph, a frequency, or a percentage table.

Median (Md)

The median is the value that divides an ordered set of scores in half. It is the midpoint or center of the ordered scores. It is the point at which half the scores are lower and half the scores are higher.

Remember, scores must be rank-ordered before one can find a median; otherwise the median would be meaningless!!!

Formula used to find the **position** of Medina:

$$\text{Position of Median} = \frac{N + 1}{2}, \text{ where } N \text{ is the number of cases.}$$

In an **odd** number of scores, the middle score is the median:

For example, in a distribution of “11, 12, 13, **16**, 17, 20, 25,”

1 2 3 4 5 6 7

The middle score is 16 and 16 is the Median.

Based on the formula, the position of Md = $(N+1)/2 = (7+1)/2 = 4$.

In an **even** number of scores, the median is the sum of two middle scores divided by 2:

For example, in a distribution of “ 11, 12, 13, **16, 17**, 20, 25, 26,”

-----|-----
1 2 3 4 | 5 6 7 8
4.5

The middle two scores are 16 and 17. The median is $(16+17)/2 = 16.5$, or using the formula: $Md = (8 + 1)/2 = 4.5$ which falls midway between the fourth 16 and the fifth cases 17.

It doesn't matter if there is more than one the same score in the middle of a distribution:

4, 6, 8, 8, 8, 8, 12, where the Md is still 8;

Nor does it matter if an **ordinal variable** has alphabetical rather than numeric values. Its scores can be rank-ordered and we can find the median.

Rank-ordered
scores for social class

Lower
Working
Working
Middle ← ← ← ← ← Median
Middle

Middle
Upper

Outliers or Extreme low or high scores do not affect the median; actually this is the advantages of the Median.

Set A	Set B	Set C
10	1	10
20	2	20
30	30	30
50 ← ←	50 ← ←	50 ← ← Median
60	60	60
70	65	69
80	75	9,000

There is no Median for nominal variables!!! For example,

Marital Status	or	Party Affiliation	
Single		Democrats	
Married		Republicans	
Separated		Independents	NO MEDIAN!!!
Divorced		Greens	
Widowed		None	

Formula used to find the **REFINED** median for variables that are conceptually continuous but measured using discrete and integer scores.

For Example,
“Do you support legal abortion?”

Value	f	F
Agree strongly	170	170
Agree	446	616
Neither agree nor disagree	299	915 ← ← ← Median
Disagree	301	1216
Disagree strongly	65	1281

(N) 1281

The median should be the value of the $1281 \div 2 = 641^{\text{st}}$ score. That is, the Median should lie between the values 616 and 915 but not exactly the 3, because the variable is conceptually continuous and the scores on this variable may range from the lower limit of the interval containing the median 2.5, 2.6, 2.7, up to the upper limit of the interval 3.5. Therefore, we use the formula:

$$\text{Md} = L + \left(\frac{N/2 - F}{f} \right) (i) = 2.5 + \left(\frac{1281/2 - 616}{299} \right) 1 = 2.5 + \frac{24.5}{299} = \mathbf{2.6}$$

Where,

L = lower limit of the interval containing the median

N = total number of scores

F = cumulative frequency of scores less than the interval containing the median

f = number of scores in the interval containing the median

i = width of the interval containing the median (the interval's upper limit minus its lower limit)

While 3 is the median, the “refined or the true median” should be 2.6.

Mean (0)

Mean is the arithmetical average obtained by dividing the sum of all scores by the number of scores. That is, add all scores and divide by the number of scores.

Formula:

$$\text{Mean} = \frac{\text{Sum of all scores}}{\text{Number of scores}}$$

$$\bar{0} = \frac{\sum x}{N}, \text{ where } \Sigma = \text{sign of summary}$$

x = raw score in set of scores
N = number of cases
0 is pronounced “X-bar.”

For example,

Table 3.1

Respondent	X(IQ)
Gene	125
Steve	92
Bob	72
Michael	126
Joan	120
Jim	99
Jane	130
Mary	100
$\Sigma = 864$	

The mean is, $\bar{X} = \Sigma X/N = 864/8 = 108$

An Illustration of Mode, Median, and Mean

- 1 The following is a set of scores as donations (in dollars) in a neighborhood received by Bob for a local charity. Find the mode, median and mean.

5 10 25 15 18 2 5

Arrange the scores from the highest to lowest

25
18
15
10
5
5
2

Mo = \$5

Md = \$ 10

Mean: the sum = $\Sigma X = 25 + 18 + 15 + 10 + 5 + 5 + 2 = \80
 $N = 7$
Mean = $\Sigma X/N = \$80/7 = \11.43

Implication:

The Mode, Median, and Mean provide a very different picture of the average level of donations. The mode suggests that the donations are typically small (\$5), whereas the Median and Mean suggest greater generosity (\$10 and \$11.43).

2 Another Example;

4
8
10
10
12
600

Median = 10

Mean = $644/6 = 107$

Implication:

With outliers, either extremely low or high, in a distribution, mean may not be a typical representative of most scores. The median works better as an average.

3 The sum of deviations of scores from the mean is 0

X	$X_i - 0$	$(X_i - 0)^2$
4	$4 - 8 = -4$	16
8	$8 - 8 = 0$	0
10	$10 - 8 = 2$	4
11	$11 - 8 = 3$	9
9	$9 - 8 = 1$	1
6	$6 - 8 = -2$	4
Sum	0 (Sum of Squares)	34
0 = 8		

Implications:

If we subtract the mean from each score and add all these differences, the sum is always 0.

$\Sigma(X_i - 0) = 0$ for a Sample Data.

$\Sigma(X_i - \mu) = \mu$ for a Population Data.

$\Sigma(X_i - 0)^2$ produce a lower sum of squares for a Sample Data.

$\Sigma(X_i - \mu)^2$ produce a lower sum of squares for a Population Data.

Please note that the Sum of Squares $[(X_i - 0)^2]$ is 34. That is the Sum of Squared deviations from the mean. This is a minimum. No other number could be substituted for the mean (8) that would produce a lower sum of squares.

Guidelines to use Mode, Median, and Mean

- 1 Means are preferred for interval and ratio variables.
- 2 Means take advantage of more information than mode and median.
- 3 Median is preferred for ordinal variables because generally we do not compute the mean for such variables.
- 4 Mode is preferred for nominal variable because we do not have much choice for nominal variables that are not sensible for either median or mean.
- 5 Median is far more resistant to extreme scores, whereas mean obscures the differences among extremes.
- 6 If outliers are present, consider either excluding them out or using the median.
- 7 Remember:
The MODE is the value of the most frequently occurring score;
The MEDIAN divides a distribution in 2-half: One half higher and one lower;
The MEAN is what we usually refer to as the "AVERAGE."

Relationships between Mode, Median and Mean

(See Textbook p. 76-77).

In a **symmetric distribution** - the right half is the mirror image of the left half, the median and mean are the same (see p-76).

In a symmetric distribution, the distribution is unimodal, the mode, median and mean are the same.

In a **non-symmetric distribution** – the distribution of a variable has more cases in one direction than the other; we have the **skewness** (p-77).

Positively Skewed – if a distribution is skewed to the right, the mean is larger than the median, because the high score pull the means in their direction (p-77).

Negatively Skewed – if a distribution is skewed to the left, the mean is smaller than the median, because the low scores pull the mean in their direction (p-77).

HANDOUT 4 MEASURES OF VARIATION

- ⇒ Define and calculate variances and standard deviations for samples and populations
- ⇒ Explain what a normal distribution/curve is
- ⇒ Calculate and interpret Z-scores/standard scores
- ⇒ Calculate and interpret confidence intervals

Measures of Variation

Measures of Variation are also called Measures of Dispersion or Measures of Variability.

They are the measures that summarize how close together or spreading out scores are distributed around the mean.

Variance and Standard Deviation

The variance and the standard deviation are two very closely related measures of variation that summarize how narrowly or widely scores are distributed around the mean.
We use the mean as a point of reference from which to measure deviations.

Variance (s²)

A variance is a measure of the variation of a distribution. It measures how scores distributed around the mean.

To get the variance, we divide the Sum of Squares by the number of cases (N) for a population but by number of cases minus one (N-1) for a sample.

By dividing the sum of squares by N, the variance is the average squared deviation of scores from the mean. That is, the variance is the average squared deviations from the mean.

Formula for a population:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

Where σ^2 = variance for population data (Greek lowercase σ^2 pronounced “sigma squared”)
 X_i = score of the i^{th} case
 μ = mean of the population (mju:)
 N = total number of cases in the population

Since the average of all sample variances always underestimate the population variance, statisticians figured out a formula to estimate population variance based on sample data.

Formula for a sample:

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{N - 1}$$

Where s^2 = variance for sample data

X_i = score of the i^{th} case

\bar{X} = mean of the sample

N = total number of cases in the sample

$N - 1$ = degree of freedom (df)

(Note that “N-1” in the denominator only matters when N is small. In larger samples like GSS, it makes little difference whether we divide by 1501 or by 1502).

Degree of Freedom (df)

The denominator $N-1$ is called the degree of freedom (df) of the variance. The term so called has to do with the property of the mean.

- 1) the sum of deviations from the mean is 0
 $\sum (X_i - \bar{X}) = 0$
- 2) if we know all scores and all the deviations but one, we can calculate the last score and deviation.
- 3) The last score is the number that makes the sum of the deviations equal to 0
- 4) The deviations from the mean are restricted a little, only $N-1$ are free to vary

Example.

$\sum (X_i - \bar{X})$	
4	$4-5 = -1$
5	$5-5 = 0$
X	$X-5 = ?$
0 = 5	0

$$\begin{aligned}\sum (X_i - \bar{X}) &= (4-5) + (5-5) + (X-5) = 0 \\ &= -1 + 0 + X-5 = 0 \\ &= X-1-5 = 0 \\ X &= 6\end{aligned}$$

Example,

N	X	$X_i - 0$	$(X_i - 0)^2$
1	64	$64-68 = -4$	16
2	68	$68-68 = 0$	0
3	70	$70-68 = 2$	4
4	71	$71-68 = 3$	9
5	69	$69-68 = 1$	1
6	66	$66-68 = -2$	4
Total		$\Sigma (X_i - 0) = 0$	$\Sigma (X_i - 0)^2 = 34$
<hr/>			
N = 6			
$0 = \Sigma X_i / N = (64+68+70+71+69+66)/6 = 68$			

Therefore, the Variance is

$$\begin{aligned}
 s^2 &= \frac{\Sigma (X_i - 0)^2}{N - 1} \\
 &= \frac{34}{6 - 1} \\
 &= 6.80
 \end{aligned}$$

The variance is 6.80 - the average squared deviation of scores from the mean. It measures how scores distributed around the mean of 68.

What is the use of variance? (Social Statistics p 218)

- 1) To compare the distribution of one variable with the distribution of another.
- 2) To compare groups like males and females or groups like Whites and Blacks.
- 3) To find the standard deviation (s) - the square root of the variance- which is very powerful and useful in analyzing distributions.

Steps to get the variance (s^2):

- 1 Calculate the mean 0
- 2 Subtract the mean from each score $X_i - 0$
- 3 Square each difference $(X_i - 0)^2$
- 4 Add all the squared differences $\Sigma (X_i - 0)^2$
- 5 Divide that sum by the total number of cases

$$\frac{\sum (X_i - \bar{X})^2}{N - 1}$$

Standard Deviation (σ)

A standard deviation is a measure of dispersion for numerical variables. It measures how scores distributed around the mean. It is the square root of the variance.

Formula for the standard deviation for sample data:

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N - 1}}$$

Formula for the standard deviation for population data:

$$\sigma = \sqrt{\text{Variance}} = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Example,

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N - 1}}$$

$$s = \sqrt{6.80} = 2.61$$

The standard deviation is 2.61.

Importance of the standard deviations:

- 1) It is a measure of dispersion.
- 2) It gives us an indication of the dispersion of responses to a variable.
- 3) It uses the mean as a point of comparison.
- 4) It helps us to assess variability by using what is “standard” about standard deviation.
- 5) If the scores are widely distributed about the mean, then the deviations from the mean are large, the sum of squares is large, the variance is large, and the standard deviation is large, there is more variation among scores.
- 6) If the sum of square is smaller, the variance is smaller, and the standard deviation is smaller, then there is less variation among scores.

Example:

There are three sets of data that distribute differently from each other. So the variances, standard deviations, and standard errors are also different.

	Data A	Data B	Data C
	64	44	34
	68	63	58
	70	80	90
	71	91	101
	69	74	79
	66	56	46
Mean	68	68	68
Variance s^2	6.8	290.80	686.80
Sta. Dev. s	2.61	17.05	26.21
S. E. σ_0	1.06	6.96	10.69

Implications

The lesser the variation among scores

The smaller the sum of squares

The smaller the variance and standard deviation

Thumbs of Rules

In reporting statistical analysis, we usually report the standard deviation rather than variances because the sta. dev. makes more cognitive sense to our minds;

The variance and std. dev. are calculated appropriately only for variables measured at the interval/ratio level.

⇒ To understand what is “standard” about a standard deviation, we need to learn normal distribution

Normal Distribution

- 1 It is a symmetrical and unimodal (one hump) and bell-shaped distribution.
- 2 Normal distribution has one thing in common: all normal distributions are symmetric; the mean is always at the center of a distribution and always equals the median and the mode. That is, the mean, median, and mode are the same.
- 3 When a distribution is normal, the area under the curve is divided into two equal portions at the mean. The mean divides the area under the curve in half – half of the area under the curve falls below the mean and half falls above the mean.

- 4 The “standard” in the standard deviation is the mathematical “given” that 34.13% of the area under a normal curve always falls between the mean and one standard deviation above the mean, while another 34.13% of the area falls between the mean and one standard deviation below it.
- 5 That is, 1 standard deviation from the mean includes about 68 percent of scores. 2 standard deviations include a little over 95 percent and 3 standard deviations from the mean include 99.7 percent of scores (Remember: 68-95-99.7).
- 6 Normal distributions may differ in different standard deviations - the larger the standard deviation, the more flat the bell-shaped curve is; the smaller the standard distribution, the narrower the curve is.

- 7 In a normal distribution, 95 percent lie within 1.96 standard deviations of the mean, and 99 percent lie within 2.58 standard deviations. These particular distances from the mean will be used in confidence intervals.
- 8 A normal distribution with a mean of μ and a standard deviation of σ is denoted $N(\mu, \sigma)$.

Shapes of Distributions

Distributions come in all kinds of forms. Three of these forms are typical.

Leptokurtic Distribution is tall and thin.

Platykurtic Distributions are low and flat.

Mesokurtic distributions are moderate in between.

Kurtosis, refers to the peakedness of distributions of interval/ratio variables.

Skewness refers to a non-symmetric distribution where the distribution of a variable has more cases in one direction than the other. As a result, distributions may skew to the right or left.

Measure of Skewness

Skewness is obtained by the difference between mean and the median, multiplied by 3 and divided by the standard deviation. It measures the extent to which a distribution skews.

The more skewed a distribution, the greater the difference between mean and median. Skewness is positive for distribution skewed to the right and negative for distribution skewed to the left. The skewness is 0 when the mean equals median, the numerator is also 0, and the distribution is symmetric.

$$\text{Skewness} = \frac{3(\bar{X} - Md)}{S}$$

Example for the years of education of males and females in GSS.

Statistic	Males	females
Mean	13.56	13.21
Median	13.12	12.66
Sta. Dev.	2.95	2.91
Skewness	.45	.57

Conclusion: The skewness coefficients show that the education distribution of both males and females are positively skewed, with distribution of females skewed more than that of males, .57 versus .45.

Standard Scores = Z-Scores

Describes how many standard deviations from the mean a score is located.

Formula

$$Z_i = \frac{X_i - \bar{X}}{s}$$

Where Z_i = standard score of the i^{th} case

X_i = score of the i^{th} case

\bar{X} = mean

s = standard deviation

Example,

Bob has a test score of 87 and his class test mean is 81 with a standard deviation of 6. Mary has a test score of 83 and her class mean is 76 with a standard deviation of 4. Tom has a test score of 76 and his class mean is 86 with a standard deviation of 10. What are the Z-scores for the three students?

For Bob,

$$Z_i = \frac{X_i - \bar{X}}{s} = \frac{87 - 83}{6} = 1.00$$

For Mary,

$$Z_i = \frac{X_i - \bar{X}}{s} = \frac{83 - 76}{4} = 1.75$$

For Tom,

$$Z_i = \frac{X_i - \bar{X}}{s} = \frac{76 - 86}{10} = -1.00$$

Answer:

Bob's score is better than his class average, 1.00 standard deviation above the class mean. Mary's score is even more impressive, 1.75 standard deviation above the class mean. But Tom's score is bad, 1.00 standard deviation below the class mean.

Note:

- 1 Z-scores may be positive or negative. A positive sign indicates that the score is above the mean while the negative sign suggests the score below the mean.
- 2 Z represents the deviation from the mean in standard-deviation units.
- 3 Z-scores have a mean of 0 and a standard deviation of 1.00. Z-scores represent standardized variables whose scores have been all converted to standard score. Every distribution of Z-scores has a mean of 0 and standard deviation of 1.00.
- 4 Z-scores are used when we compare scores from distributions that have different means and standard deviations.
- 5 Z-scores are meaningful only for interval/ratio variables.

Sampling Distributions

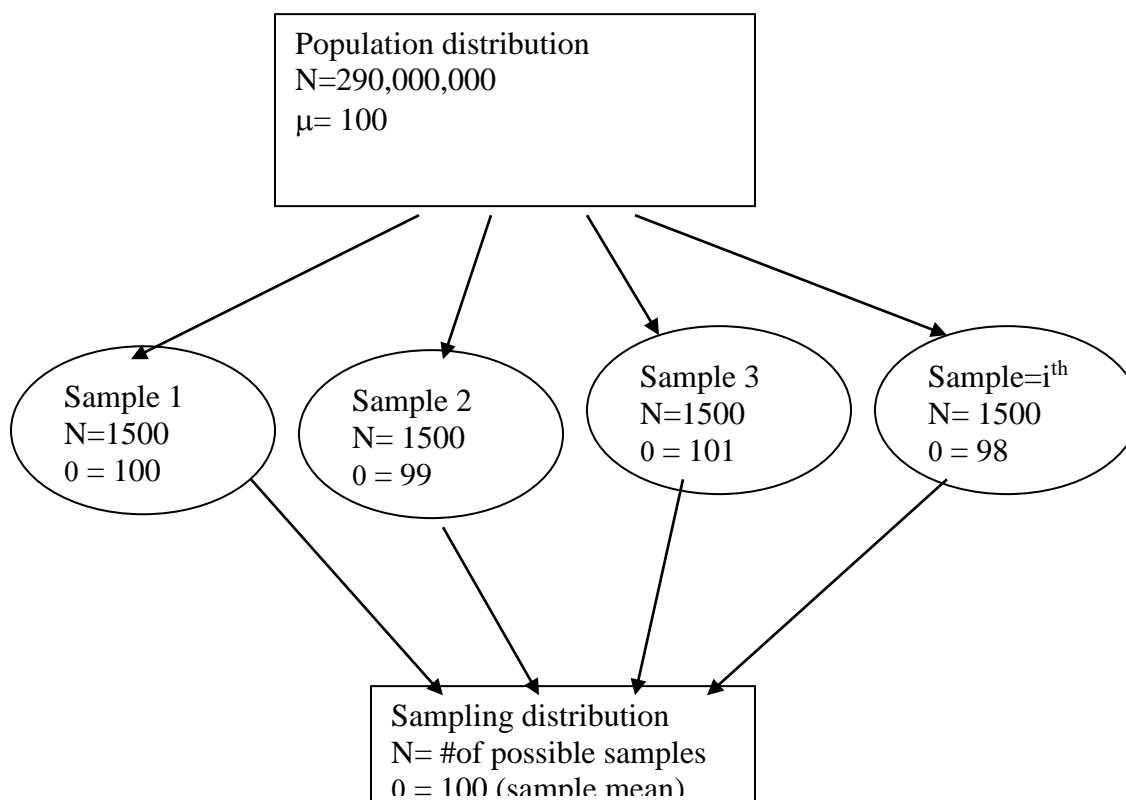
The distribution of some statistic (e.g., the mean) in all possible samples of a given size. The theoretical distribution of sampling statistics from all possible samples of a given size that could be drawn from the same population.

Population distribution - The distribution of scores in a population.

Distribution of a sample - The distribution of scores in a sample of a given size n.

Sampling Distribution - The distribution of some statistic (e.g., the mean) in all possible samples of a given size.

Example,
Population Distribution, Distribution of a Sample, and Sampling Distribution.



Central Limit Theorem

As the sample size N increases, the sampling distribution of the mean more and more closely resembles a normal distribution with a mean equal to the population mean and a standard deviation of σ/\sqrt{N} . This distribution is described symbolically as $N(\mu, \sigma/\sqrt{N})$. Statistics refer to this tendency as the Central Limit Theorem.

Most possible samples will have means somewhere around the population mean.
Most of possible samples we could draw will be pretty much like the population

Standard Error of the Mean (σ_0)

The standard deviation of the sampling distribution of the means (from all possible samples of a given size drawn randomly from a population).

Formula,

$$\sigma_0 = \sigma / \sqrt{N}$$

Where σ_0 = standard error

σ = standard deviation

N = sample size

Example, the variable daily hours of TV watching has a standard deviation of 2.14 on 1940 cases in the GSS. Find the standard error.

$$\sigma_0 = \sigma / \sqrt{N} = 2.14 / \sqrt{1940} = 2.14 / 44.045 = .049$$

Therefore, the standard error is .049.

- ⇒ The standard error of the mean is used to estimate the standard deviation of the sampling distribution of sample means.
- ⇒ Used to establish confidence interval.
- ⇒ Used to estimate population mean
- ⇒ The smaller the standard error, the lesser the variation in the sampling distribution of sample means, the more confident that our predictions about populations based on the sample will be accurate. The more homogeneous / lesser heterogeneous the characteristics of the sampling distribution.

Confidence Interval

We use a sample statistic to estimate a population parameter. Population mean may be greater or less than the sample mean. To estimate a population mean, we must establish a range around the sample mean within which we think the population mean lies. This range is called confidence interval.

Two commonly used formula to find confidence intervals: 95 and 99 percent.

95 percent confidence interval = $\bar{x} \pm 1.96\sigma_0$

The chances are 95 out of 100 that the population mean lies within this range.

99 percent confidence interval = $\bar{x} \pm 2.58\sigma_0$

The chances are 99 out of 100 that the population mean lies within this range.

Steps to find the 95/99 percent confidence interval:

- 1 Subtract 1.96/2.59 standard errors from the sample mean to find the lower limit of the confidence interval

- 2 Add 1.96/2.58 standard errors from the sample mean to find the upper limit of the confidence interval

Example,

The mean for watching TV daily is 2.90 with a standard deviation of $\sigma = 2.14$, a sample size of $N=1940$, and a standard error of $\sigma_0 = .049$. Therefore,

[Still remember the formula? $\sigma_0 = \sigma/\sqrt{N} = 2.14/\sqrt{1940} = 2.14/44.045 = .049$]

$$\begin{aligned} 95 \text{ percent confidence interval} &= \bar{x} \pm 1.96\sigma_0 \\ &= 2.90 \pm 1.96 (.049) \\ &= 2.90 \pm .096 \\ &= 2.80 \text{ to } 3.00 \end{aligned}$$

Thus the 95 percent confidence interval is between 2.80 and 3.00. The chances are 95 out of 100 that the population mean lies within this range.

$$\begin{aligned} 99 \text{ percent confidence interval} &= \bar{x} \pm 2.58\sigma_0 \\ &= 2.90 \pm 2.58 (.049) \\ &= 2.90 \pm .126 \\ &= 2.77 \text{ to } 3.03 \end{aligned}$$

Thus the 99 percent confidence interval is between 2.77 and 3.03. The chances are 99 out of 100 that the population mean lies within this range.

CROSS-TABULATION

- A Is there a relationship between two variables?
 A How to interpret the strength of a relationship?
 A What is the direction and shape of that relationship?

Cross-Tabulation

Cross-tabulation, also called cross-tab, cross-classification or tabular analysis, examines association between two variables by comparing percentage distributions. It is also used to examine a causal relationship in which we hypothesize that an independent variable affects a dependent variable. Cross-tab is one of several methods to analyze bivariate relationships. Other methods include difference of means, analysis of variance, regression, multivariate analysis, etc.

Research Hypothesis

An expectation that how variables are related: positive or negative; causal or non-causal.
 An expected but unconfirmed relationship among two or more variables.

Independent and Dependent Variable

An independent variable is the presumed cause of the dependent variable while the dependent variable is the presumed effect of the independent variable.

Bivariate Frequency Distribution

Frequency distribution for two variables in one table.

Example.

Table 5.1 Civil Disobedience by Education (in frequencies)

	Education			Total
	<high school	high school	College	
Follow one's Conscience	4	13	12	29
Obey Law	6	11	4	21
Total	10	24	16	50

Marginal Distributions/Marginals

The row totals and column totals are called marginal distribution or marginals. In Table 5.1, the numbers like 10, 24, 16, and 29, 21, are marginals.

Grand Total

The total number of cases presented in the lower, right-handed cell of the frequency table is called grand total. In Table 5.1, the grand total is 50.

*****As a rule, Independent variable is put in columns and dependent variable put in rows.**

Table 5.1 tells us:

6 cases with less than high school education believe in always obeying the law, compared with 11 high school graduates. But the problem is there are fewer cases with less than high school graduates (10) than there are high school graduates (24). It is hard to tell who are more likely to obey law, less than high school or high school graduates?

Then we need standardize the frequency by percentaging the frequency.

Table 5.2 Civil Disobedience by Education (in percentages)

	Education			Total
	<high school	high school	College	
Follow one's Conscience	40	54	75	58
Obey Law	60	46	25	42
Total (N)	100 (10)	100 (24)	100 (16)	100 (50)

Percentaging

Percentaging is a way to standardize distributions regardless the number of cases. Percentages tell us how many cases there are in a cell if there are 100 cases with the value of the independent variable.

The rule is to compute percentages within categories of the independent variable.

The sum within each category of the independent variable is 100.

The percentage of each cell is obtained by dividing each cell frequency by the total for the column, and then multiply by 100.

Table 5.2 tells us:

40 percent of the less educated believe in following their conscience, compared with 54% of high school graduates and 75% of college graduates. It is easy to draw conclusion that Among 50 cases, respondents with more education are more likely to believe that people should follow conscience and they are less likely to believe that people should always obey the law.

Table 5.3 Civil Disobedience by Education (Standard Cross-Tab)

Education

		<hr/>			
Civil Disobedience		<high school	high school	College	Total
<hr/>					
Conscience		4	13	12	29
		0.40	0.54	0.75	0.58
Obey Law		6	11	4	21
		0.60	0.46	0.25	0.42
<hr/>					
Total	count	10	24	16	50
	%	100.00%	100.00%	100.00%	100.00%
<hr/>					

How to describe tables

The rule is “r by c” Table, where r refers to the number of rows and c refers to the number of columns. Tables with 2 rows and 2 columns are described as 2 by 2 table; 2 rows and 3 columns as 2 by 3 table; 3 rows and 4 columns as 3 by 4 tables. Thus Table 5.1, 5.2 and 5.3 are described as 2 by 3 table.

How to interpret percentage tables

- 1 Compare percentages across categories of the independent variable.
- 2 The smaller the differences between percentages across categories of the independent variable, the weaker the relationship. The larger such differences, the stronger the relationship.
- 3 Magnitude of Strength-Rule of Thumb:

Small difference:	less than 10 %	→	the weaker relationship
Moderate difference:	10 to 30 %	→	the moderate relationship
Larger difference:	more than 30 %	→	the stronger relationship

Positive Relationships

A relationship in which higher scores on one variable are associated with higher scores on the other variable. For instance, when education levels go up, financial situation gets better.

Table 5.4 Financial Situation by Education (in percentages)

<hr/>		Education Level			
-------	--	-----------------	--	--	--

Financial	< HS	HS	College	BA	MA
Better	23	39	43	47	49
The same	52	39	38	35	34
Worse	25	22	19	18	17
Total	100	100	100	100	100
N	445	1652	186	470	223

***In a positive relationship, cases tend to cluster along the minor diagonal of the table running from the lower left to the upper right like a forward slash “/”.

Negative Relationships

A relationship in which higher scores on one variable are associated with lower scores on the other variable. For instance, when education level goes up, daily watching TV hours get less.

Table 5.5 TV Watching Hours by Education (in percentages)

TV Hours	Education Level				
	< HS	HS	College	BA	MA
5 +	29	17	11	7	4
3 - 4	35	34	28	25	20
2	22	29	37	30	34
0 – 1	14	20	24	38	42
Total	100	100	100	100	100
N	293	1068	117	314	154

***In a negative relationship, cases tend to cluster along the major diagonal from the upper left to the lower right like a backward slash “\”.

Curvilinear Relationships

Curvilinear relationships take a variety of forms. The simplest forms are the relationships in which cases with low and high scores on the independent variables are similar in their scores on the dependent variables. The typical curvilinear relationships have patterns shaped roughly like the letter “V” or an upside-down “V”.

For instance, both low and high educated respondents have greatest closeness to neighborhood.

Table 5.6. Feelings of Closeness to Neighborhood by Education (in percentages)

Closeness to Neighborhood	Education Level				
	< HS	HS	College	BA	MA
Feel Close	62	55	49	60	59
Not Close	38	45	51	40	41
Total	100	100	100	100	100
N	183	740	81	220	101

***In a curvilinear relationship, the typical form is shaped like either a V or an upside-down V.

***See patterns of these relationships on textbook page 124.

Format Conventions for Bivariate Tables

- 1 Conform to format conventions of American Sociological Association
- 2 Number tables with Arabic numerals
- 3 Title form: dependent by independent variable (row by column)
- 4 Label the values of dependent and independent variables
- 5 Include a Total row and footnote if necessary (for rounding errors)
- 6 Including an N row presenting the number of cases
- 7 Retain only significant digits in percentages
- 8 Be consistent in decimal places
- 9 Do not put % signs after cell entries
- 10 Do not draw vertical lines in a table
- 11 Be neat. Keep cell entries lined up and horizontal lines the same length
- 12 For interval/ratio variables, values of independent variable are listed from the lowest at left to highest at the right. Values of dependent variable are ranged from the highest at the top to the lowest at the bottom.
- 13 For nominal variables, there are no specific orderings. But it is preferred that nominal categories are listed from the most frequently to the least frequently occurring.

Stacked Bar Graphics

Stacked bar graphs offer an efficient way to visually describe bivariate relationships. While tables provide us more detailed information about the relationship between variables, stacked bars give us a faster and more vivid overall impression of the relationship (See page 128)

Rule of Thumbs for Ns

- 1 In generally, the larger the total frequencies (N) in the independent variable's categories, the more stable and reliable the percentages, and the more confidence we have in them. When N in the column total is small, the shift of just few scores from one dependent variable value to another may radically change the percentage distribution.
- 2 Collapsing values in the dependent variable may help identify and interpret relationships. For example, age group 100 + may be collapsed with group 90 –99 to increase the N.
- 3 As rule of thumb, percentages should be based on at least,
 $N \geq 30$ or ≥ 50 or ≥ 100 or more.

Relationship between Association and Causation

Think of these assumptions:

- 1 The more firefighters at a fire, the greater the property loss.
- 2 The bigger the kids' shoes, the better they spell.
- 3 When the Ganges River floods/overflows its banks, the street crime increases in New York.
- 4 The taller (height) the students, the better grades they get.

Are these relationships associated or causal?

They are ridiculous!

Association

Association refers to two variables that are associated. Association does not suggest directionality. Association simply means two variables are related but not necessarily causal.

Causal Relationship

A relationship said to be causal implies that changes made in one variable will cause changes in another variable. Causally related variables must be related whereas variables that are associated are not necessarily causally related.

CAUSATION:

1. Association: if X causes Y, then X and Y must co-vary;
2. Time order: if X is a cause of Y, then the occurrence of X must precede the occurrence of Y.
3. Non-spurious: if X is a cause of Y, then the relationship between X and Y can't be explained away by a third factor.
4. Theory: Theory is needed to explain why X is a cause of Y.

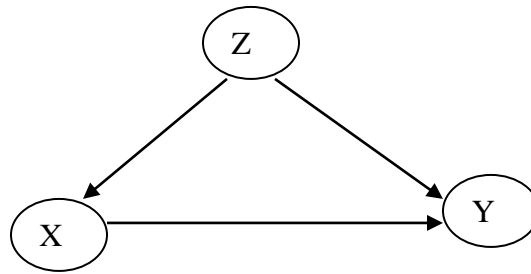
Spurious Relationship

When statistical associations that are not really causal relationships are called spurious relationships. Alternatively, if a relationship between two variables is explained away by a third variable, that relationship is called spurious.

For example,

- 1 Changes in X bring changes in Y;
- 2 Changes in Z bring changes both in X and Y. Then

3 The relationship between X and Y is spurious.



Dr. Ji
SOC331

HANDOUT 6

THE CHI-SQUARE TEST OF STATISTICAL SIGNIFICANCE

A What is a null hypothesis?

A How to interpret chi-square tests on cross-tabulation?

A How to interpret chi-square values in terms of statistical significance?

Statistical Significance (Sig.)

- 1 Ways to decide whether a relationship found in sample data is also likely to be found in the population from which the sample was drawn.
- 2 The relationship between the variables under study is not due to chances alone.
- 3 Statistical significant is a demonstration that the probability of the null hypothesis being true is very small and that the probability of the research hypothesis being true is very big.

Probability (p)

A probability multiplied by 100 is the number of times an event is likely to occur out of 100 trials. A probability of 0 means that the chance of something (sth) happening is non-existent. A probability of 1.00 means that sth is likely to occur 100 out of 100 times (sth is completely certain – a sure bet). Thus the smaller the probability of sth, the less likely it is to occur; the higher the probability, the more likely it is to occur.

Similarly, a probability of 0.01 indicates that the chance of sth happening is 1 out of 100 ($p \leq 0.01$); a probability of 0.05 means that the chance of sth happening is 1 out of 20, or alternatively, 5 out of 100, ($p \leq 0.05$). By the same token a probability of 0.001 means that the chance of sth happening is 1 out of 1000, ($p \leq 0.001$).

The $p \leq 0.05$, $p \leq 0.01$, and $p \leq 0.001$, are the three **Levels of Significance** and the $p \leq 0.05$ level is a widely used cut-off for statistical significance in sociological world.

Null Hypothesis (H_0)

Null hypothesis refers to the assumption that there is no relationship in the population.

Therefore, we reject the null hypothesis of no population relationship. However, if there is a relationship in the population, say the probability we find in a relationship is less than 1-in-20 or 5-in-100 = $p \leq .05$, we say that there is a relationship in the population and that the relationship is statistically significant.

Type I Error and Type II Error

When we reject a null hypothesis that is really true, we commit a Type I error, also called alpha error.

If we fail to reject a null hypothesis that is really false, we commit a Type II Error, also called beta error.

The Type I and Type II errors are inversely related. Reducing the chance of Type I error increases the chance of a Type II error, and vice versa.

The Chi-square (χ^2)

Symbolized with the Greek χ^2 (pronounced “ki square”), Chi-square is a statistic that compares the actual frequencies in a bivariate table with the frequencies expected if there is no relationship between the two variables. It is used for tests of statistical significance and for measures of association between nominal variables in cross-tabulation.

Chi square Test

If observed frequencies are similar to the expected frequencies, then we can not reject the H_0 hypothesis. We conclude that there is no relationship in the population from which the sample was drawn. (Thus we risk a Type II error – fail to reject a hypothesis when it is really false).

On the other hand, if the observed frequencies are different from the expected frequencies, then we reject the null hypothesis. We conclude that there is a relationship between the variables. (Thus we may risk a Type I error – rejecting a hypothesis when it is true).

Chi square Formula***

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where Σ (sigma) = to add everything that comes after it

f_o = observed frequency in each cell

f_e = expected frequency in each cell

*** The formula indicates that the larger the difference between the observed f_o and expected frequencies f_e , the larger the value of Chi-square χ^2 , the stronger the relationship at a certain significant level.

Expected Frequency Formula

$$f_e = \left(\frac{\text{Row Marginal}}{N} \right) \text{Column Marginal} = \frac{\text{RM}}{N} \text{CM}$$

Steps to Calculate Chi square

- 1 Calculate expected frequency first
- 2 Subtract the expected frequency from the observed frequency for each cell
- 3 Square each difference
- 4 Divide each squared difference by the expected frequency for that cell
- 5 Add all these numbers for all cells.

Example 1

Table 6.1. Role of Government by Race (in frequencies)

	Race		Total
	White	Black	
Role of government			

Should do more	334	100	434
	<u>375.5</u>	<u>58.5</u>	
About right	581	103	684
	<u>591.9</u>	<u>92.1</u>	
Should do less	582	30	612
	<u>529.6</u>	<u>82.4</u>	
<hr/>			
Total	1497	233	1730
<hr/>			

Null Hypothesis (H₀)

Opinion about role of government is not related to race of respondents. Or
 Opinion about role of government is the same between Whites and Blacks.

Research Hypothesis

Opinion about role of government is related to race of respondents.
 Opinion about role of government is NOT the same between Whites and Blacks.

$$f_e = \left(\frac{\text{Row Marginal}}{\text{N}} \right) \text{Column Marginal} = \frac{434}{1730} \cdot 1497 = 375.5$$

$$f_e = \frac{684}{1730} \cdot 1497 = 591.9$$

$$f_e = \frac{612}{1730} \cdot 1497 = 529.6$$

$$f_e = \frac{434}{1730} \cdot 233 = 58.5$$

$$f_e = \frac{684}{1730} \cdot 233 = 92.1$$

$$f_e = \frac{\text{-----}}{1730} 233 = 82.4$$

Table 6.2. Calculate the Chi Square for Table 6.1

f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$(f_0 - f_e)^2 / f_e$
334	375.5	-41.5	1722.25	4.586
581	591.9	-10.9	118.81	.201
582	529.6	52.4	2745.76	5.184
100	58.5	41.5	1722.25	29.440
103	92.1	10.9	118.81	1.290
30	82.4	-52.4	2745.76	33.322
Sum			$\chi^2 = 74.023$	

$\chi^2 = 74.023$; $df = 2$; $p \neq .001$; (Reject the H_0).

Degree of Freedom (df)

Degrees of freedom for the Chi square test are equal to $(r-1)(c-1)$, where r and c are the numbers of rows and columns in the table.

$$df = (r-1)(c-1) = (3-1)(2-1) = 2 \times 1 = 2$$

Decision about H_0

- 1 Obtain χ^2 and df .
- 2 Find the probability ($p < ?$).
- 3 Decide to reject H_0 or Not to reject H_0 .

Table 6.3. The Chi-Square Distribution (Probability)

df	.05	.02	.01	.001	← significance levels
1	3.841	5.412	6.635	10.827	← critical values
2	5.991	7.824	9.210	13.815	
3	7.815	9.837	11.345	16.266	

4	9.488	11.668	13.277	18.467
5	11.070	13.388	15.086	20.515
6	12.592	15.033	16.812	22.457
...
...
29	42.557	46.693	49.588	58.302
30	43.773	47.962	50.892	59.703

Note: If the calculated value is bigger than the critical value at a required DF and a required significance level, Reject the Null Hypothesis and accept the Research Hypothesis.

$$\chi^2 = 74.023,$$

$$df = 2$$

$$\chi^2 = 74.023 > 13.815 \text{ at the } p \# .001 \text{ significance level}$$

Conclusion:

Since Chi-square is larger than the table value ($74.023 > 13.815$) at the $p < .001$ significant level, we reject the null hypothesis that opinion about role of government is not related to race of respondents. The results support our research hypothesis that opinion about role of government differs from Whites to Blacks. This suggests that opinion about role of government is related to race of respondents.

Example 2

Table 6.4. Adjustment to College by Size of Home Town
(in frequencies)

Adjustment To College	Size of Home Town		Total
	Small	Large	
Good	7 (8)	13 (12)	20
Poor	3 (2)	2 (3)	5
Total	10	15	25

Table 6.5. Calculate the Chi Square for Table 6.4

f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$(f_0 - f_e)^2 / f_e$
7	8	-1	1	.125

3	2	1	1	.500
13	12	1	1	.083
2	3	-1	1	.333
Sum				$\chi^2 = 1.041$

$\chi^2 = 1.041$; df = 1; p # n.s. (Do not reject the H_0).

Null Hypothesis (H_0)

Adjustment to college is not related to size of home town. Or

Adjustment to college is the same between small and large size of home town.

Research Hypothesis

Adjustment to college is related to size of home town. Or

Adjustment to college is NOT the same between small and large size of home town.

Be Cautious About Small Expected Frequencies

When 20% or more of the cells have expected frequencies less than 5.0, we report the significance as not applicable (n.a.). Because we violate the assumption of expected frequencies at least 5.0 and we are on weak grounds. In this case, we may switch to use Fisher's exact test which is good for small expected frequencies. We may also collapse values so that the expected frequencies become 5 or more, or consider excluding categories responsible for the small expected frequencies if the variable is nominal.

Statistical Significance and Sample Size (N)

Statistical significance means that a relationship found in a sample data can be generalized to a larger population. But remember, statistical significance both depends on the strength of a relationship and the number of cases in a sample. With few cases, even a larger difference cannot be generalized with confidence; with enough cases, even a very small difference of no substantive significance can be generalized to the population. Statistically significance does not necessarily mean substantive significance. Chi-square is a function of sample size. It is sensitive to N. Therefore, students must be caution concerning the chi-square test, while drawing conclusions.

The example below illustrates the relationships between statistical significance, sample size, and substantive significance (see textbook p. 146)

Table 6.6-a. Being a Duck by Walking & Quacking Like a Duck

	Walks & Quacks Like a Duck?		
Is a Duck?	Yes	No	Total
Yes	15	10	25

No	10	15	25
N	25	25	50
$\chi^2 = 2.00$. $p = \text{n.s.}$ $df=1$ (Do not reject H_0)			

Table 6.6-b. Being a Duck by Walking & Quacking Like a Duck

	Walks & Quacks Like a Duck?		
Is a Duck?	Yes	No	Total
Yes	30	20	50
No	20	30	50
N	50	50	100
$\chi^2 = 4.00$; $p < .05$; $df=1$; (Reject H_0).			

Table 6.6-c. Being a Duck by Walking & Quacking Like a Duck

	Walks & Quacks Like a Duck?		
Is a Duck?	Yes	No	Total
Yes	150	100	250
No	100	150	250
N	250	250	500
$\chi^2 = 20.00$; $p < .001$; $df=1$; (Reject H_0).			

Significance and Population Data

- 1 With population data, we do not need to test significance since we are already 100 percent certain ($p = 1.00$) that the relationship found in the data occurs in the population.
- 2 Chi square test is a major application of **inferential statistics** in which we carry out test of statistical significance to decide if we can generalize relationship in the sample data to larger population.

- 3 Nevertheless, significant tests are often carried out in order to assess the likelihood that a relationship found in the population is due to chance or random process of any kind. It is our job to explain why that relationship exists.

Relationships between Degree of freedom, Chi square, and Probability

- 1 Degrees of freedom for the Chi square test are equal to $(r-1)(c-1)$, where r and c are the numbers of rows and columns in the table. [$df = (r-1)(c-1)$]
- 2 Degrees of freedom depend only on the number of rows and columns, not the ordering of categories (change in category orderings has no effect on Chi square)
- 3 Degrees of freedom depend on the number of rows and columns but not the number of cases in table cells
- 4 Chi-square depends on the differences between observed and expected frequencies
- 5 Chi square depends on the number of cases. It is sensitive to sample size N
- 6 Increasing the case to frequencies increases the chi square but decreases the p (the significant levels)

- A Basics of the measures of association
- A Major advantages and disadvantages of C, V, ϕ , and λ
- A Major advantages and disadvantages of gamma, Somers'D, tau-b, and tau-c

MEASURES OF ASSOCIATION FOR CROSS-TABULATIONS

Overview of Measures of Association

Range

From 0 to 1.00 for nominal variables
 From - 1.00 to 0 to 1.00 for ordinal, interval, and ratio variables

Magnitudes/Strength

0 indicates no relationship; 1.00 reflects a perfect relationship.
 The larger the magnitude of a measure of association between variables is, the stronger the relationship they will be.

Directionality

The sign of “+” or “-” of a measure of association indicates the direction of the relationship. The sign of “+” indicates Positive Relationship while the sign of “-” suggests a Negative Relationship.

Criterion for choosing a measure

Choosing one measure over another is primarily based on the level of measurement of the variables used.

Chi Square-Based Measures for Nominal Variables - C, V, and ϕ

Advantages of C and V over Chi-Square.

(C= Pearson's Coefficient of Contingency; V = Cramer's V).

Chi-square is sensitive to number of cases (sample size N). To eliminate the effect of the number of cases and take the table size into account to adjust for a maximum value in order to obtain a measure of the strength of association, we use C and V- the improved chi-square-based measures.

Pearson's Coefficient of Contingency (C)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}},$$

Where, C = contingency coefficient
 χ^2 = chi square
 N = total number of cases

Example:

Are Blacks' disadvantages (income, education, housing, etc.) due to discrimination?

Table 7.1. Opinion about Cause of Black Disadvantage
by Race (in percentage)

Discrimination	Race		
	White	Black	Other
Yes	34.6	64.2	53.1
No	65.4	35.8	46.9
Total (N = 1851)	100.0 (1493)	100.0 (260)	100.0 (98)

$\chi^2 = 89.127$; df = 2; $p \leq .001$.
C = 0.21

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{89.127}{89.127 + 1851}} = \sqrt{\frac{89.127}{1940.127}} = \sqrt{0.0459} = 0.21$$

Conclusion:

The C of 0.21 suggests that there is a moderately strong relationship between race and opinion about discrimination as the main source of Blacks' disadvantages.

*** Serious problems with C:

- 1 Its upper limit depends on the number of rows and columns. The upper limit increases as the minimum number of rows and columns increases;
- 2 It can never reach the value of 1.00;
- 3 Cramer's V adjusts for the numbers of rows and columns so that the value of V can reach 1.00.

Cramer's V, [Compared with Pearson's Coefficient of Contingency (C)]

$$V = \sqrt{\frac{\chi^2}{(N) \text{Min}(r-1, c-1)}}, \quad [C = \sqrt{\frac{\chi^2}{\chi^2 + N}}]$$

Where, V = Cramer's V

χ^2 = Chi-square

N = total number of case

r = number of rows

c = number of columns

$\text{Min}(r-1, c-1) = \text{either } r-1 \text{ or } c-1, \text{ whichever is less.}$

Applied to the Example of Table 7.1, we have

$$\begin{aligned}
 V &= \sqrt{\frac{\chi^2}{(N) \text{Min}(r-1, c-1)}} = \sqrt{\frac{89.127}{(1851) \text{Min}(2-1, 3-1)}} = \sqrt{\frac{89.127}{1851}} \\
 &= \sqrt{.0482} \\
 &= .22
 \end{aligned}$$

Conclusion:

The V of 0.22 suggests that there is a moderately strong relationship between race and opinion about discrimination as the main source of Blacks' disadvantages.

Mean Square Contingency ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad \left[= \sqrt{\frac{89.127}{1851}} = .22, \text{ a special case of } V: \text{ a } 2 \times 2 \text{ table} \right]$$

Property of ϕ

- 1 Phi (ϕ) is only used for tables with 2 rows and 2 columns. ϕ works well with tables or 2 by 2. It is the special case of V.
- 2 ϕ is always positive but never negative. It is good because relationships between nominal variables have no direction.
- 3 ϕ has an unfortunate property, that is, its upper limit may be greater than 1.00 for tables with more than 2 rows and 2 columns.
- 4 When reporting phi result, we can either report as $\phi = .22$ or $\phi^2 = .05$.

Relationships between χ^2 , C, V, and ϕ

- 1 C, V, and ϕ are symmetric measures of association: their values do not depend on which variable is dependent or which variable is independent.
- 2 They are based on Chi-square which makes no distinction between dependent or independent.

- 3 If the chi square for the table is statistically significant, so too is the chi square-based measure of association (C, V, and ϕ); if the chi square is not significant, neither is the measure of association (C, V, and ϕ).
- 4 The assumption of random sampling necessary for chi square test also applies to the significance test for C, V, and ϕ .

Lambda (λ) [Guttman's coefficient of predictability]

- 1 Another measure of association for nominal variables (Or one is nominal and other is ordinal).
- 2 Lambda is NOT based on Chi square.
- 3 Unlike C and V, Lambda is asymmetric – the value of lambda depends on which is dependent and which is independent.
- 4 Lambda measures the strength of a relationship by calculating the proportion by which we reduce errors predicting a dependent variable score if we know the independent variable score for each case.
- 5 Lambda is calculated from bivariate frequencies not percentages.

$$\lambda = \frac{E_1 - E_2}{E_1}$$

Where, λ = lambda

E_1 = number of errors we would make guessing the dependent variable if we did not know the independent variable.

E_2 = number of errors we would make guessing the dependent variable if we knew the categories of the independent variable.

E_1 = subtract the largest row marginal total from N.

E_2 = add up the highest frequencies of each category of the independent variable and subtract the sum from N.

Example.

Table 7.2. Opinion about Cause of Black Disadvantage
by Race (in frequencies)

Discrimination	Race			Total
	White	Black	Other	
Yes	516	167	52	735
No	977	93	46	1116
Total (N)	1493	260	78	1851

- E_1 = subtract the largest row marginal total from N
 = $1851 - 1116$
 = 735
 E_2 = add up the highest frequencies of each category of the independent variable and subtract the sum from N
 = $1851 - (977 + 167 + 52)$
 = $1851 - 1196$
 = 655

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{735 - 655}{735} = \frac{80}{735} = .11$$

Note: The numerator tells us how much we reduce errors if we know independent variable scores. The denominator is the total number of errors.

Answer:

We reduce our errors in predicting discrimination opinion about 11 percent if we know each respondent's race.

Properties of Lambda:

- 1 Lambda is a **Proportional Reduction in Error (PRE)** measure of association. It tells us the proportion by which we reduce errors in predicting dependent variable if we know the score of each on the independent variable.
- 2 By convention, we set up table with independent variable in column and the dependent in rows.
- 3 Lambda's range in value is from 0 to 1.0. The closer the value is to 0, the weaker is the association; the closer the value is to 1.0, the stronger is the association.
- 4 Lambda has no direction.
- 5 Lambda may produce a value of 0 even when there is really a relationship between variables.
- 6 Lambda is a poor measure of association to use with a skewed dependent variable.
- 7 Lambda is rarely used for PRE interpretation.

Choosing A Measure: χ^2 , C, V, ϕ or λ

- 1 χ^2 is sensitive to sample size.
- 2 C has an upper limit of less than 1.00, making it hard to compare.
- 3 V and ϕ are preferable to C especially with small table (2 by 2).
- 4 χ^2 , C, V, and ϕ are symmetric measures/no distinction b/w dep or ind variable.
- 5 λ is asymmetric measure. It tends to understate a relationship for a skewed dependent variable (skewed = the distribution of a variable has more scores in one direction than the other).
- 6 V or C is preferred with dependent variable skewed; If not, go with λ .

Example of comparisons:

Table 7.3 Beliefs in afterlife by gender (in percentage)

After life	Gender	
	Male	Female
Yes	82	83
No	18	17
Total	100	100
N	759	978

$\chi^2 = .471$; df=1; n.s.
 $C = .02$
 $V = \phi = .02$
 $\lambda = .00$

Table 7.4 Gun ownership by gender (in percentage)

Own gun	Gender	
	Male	Female
Yes	48	35
No	52	65
Total	100	100
N	848	1066

$\chi^2 = 32.7$; df=1; $p < .001$
 $C = .13$
 $V = \phi = .13$
 $\lambda = .00$

Table 7.5 fear of walking in neighborhood by gender (in percentage)

Fear walk	Gender	
	Male	Female
Yes	26	55
No	74	45
Total	100	100
N	846	1057

$\chi^2 = 169.6$; df=1; $p < .001$
 $C = .29$
 $V = \phi = .30$
 $\lambda = .14$

Implications from the above tables:

- 1 Table 7.3 - The mode of category is “Yes” for both male and female and seriously skewed. Not significant; Lambda failed.
- 2 Table 7.4 – The modal category is both “No” even the skewedness is moderate.
- 3 Table 7.5 – Lambda picked up relationship because the modal categories are found in both Yes and No.
- 4 Choosing what measures to use is dependent on
 - 1) purpose of researcher.
 - 2) based on data.
 - 3) characteristics of or strength and weakness of the various measures of association.
 - 4) levels of measurement – nominal, ordinal, interval, or ratio.
 - 5) degrees of skewedness.

Gamma (G)

- 1 Gamma is a measure of association for both ordinal variables.
- 2 The magnitude of G is from -1.00 to 1.00 . While -1.00 suggests a perfect negative relationship, 1.00 indicates a perfect positive relationship.
- 3 G will be positive if there are more pairs in the same direction. G will be negative if there are more pairs in the opposite direction.
- 4 It is a symmetric and PRE measure.
- 5 How Gamma Works: For every pair of cases in a bivariate table, we predict that the rank-order of their scores on the dependent variable will be the same as the pair's rank-ordering of scores on the independent variable. If positively related, cases with higher scores on the independent will be associated with higher scores on the

dependent variable. This is what it means that variables are positively related. If negatively related, cases with higher scores on the independent will be associated with lower scores on the dependent variable. This is what it means that variables are negatively related.

- 6 Gamma is the proportion of the difference between the positive direction pairs and negative direction pairs to the total pairs of scores. In the same ordered direction of pairs of cases, higher scores in the independent variable are associated with higher scores in the dependent variable; in the opposite ordered direction of pairs of scores, higher scores in the independent variable are associated with lower scores in the dependent variable.

Table 7.6. Civil Disobedience by Education (frequency)

Civil Disobedience	Education		
	<i>Low Value</i> <HS	HS	<i>High Value</i> College
<i>High Value</i>			
Conscience	4	13	12
Obey Law	6	11	4
<i>Low Value</i>			

$$\begin{aligned}
 \# \text{ of same-ranking-pairs} &= [(6 \times 13) + (6 \times 12)] + (11 \times 12) \\
 &= [78 + 72] + 132 \\
 &= 282
 \end{aligned}$$

This same-ranking-pairs means that 282 pairs are ordered in the same direction on both variables. For each of these 282 pairs, the case with a higher Education score also has a higher Civil Disobedience score. In other words, for each of 282 pairs, the case with a higher score in the independent variable also has a higher score in the dependent variable.

$$\begin{aligned}
 \# \text{ of opposite-ranking-pairs} &= [(4 \times 11) + (4 \times 4)] + (13 \times 4) \\
 &= [44 + 16] + 52 \\
 &= 112
 \end{aligned}$$

This opposite-ranking-pairs means that 112 pairs are ordered in the opposite direction on both variables. For each of these 112 pairs, the case with a higher Education score has a lower Civil Disobedience score. In other words, for each of 112 pairs, the case with a higher score in the independent variable has a lower score in the dependent variable.

$$\begin{aligned}
 G &= \frac{\text{\# of same-ranking-pairs} - \text{\# of opposite-ranking-pairs}}{\text{\# of same-ranking-pairs} + \text{\# of opposite-ranking-pairs}} \\
 &= \frac{282 - 112}{282 + 112} \\
 &= \frac{170}{394} \\
 &= .43
 \end{aligned}$$

Answer:

Gamma of .43 indicates a fairly strong relationship between education and civil disobedience. It suggests that as education goes up, civil disobedience tends to follow conscience.

*The above calculation suggests that Gamma is the proportion of the difference between the positive direction pairs scores and negative direction pairs scores to the total pairs of scores.

Somers' D_{YX}

- 1 Named for Robert H. Somers, D_{YX} treats the row variable as the dependent variable called the Y and the column variable as the independent variable called X.
- 2 Like Gamma, D_{YX} is based on the number of pairs in the same and opposite directions.
- 3 Unlike Gamma, D_{YX} takes into account the number of pairs tied on the row-Y variable T_Y (sum of horizontal multiplication between cells).
- 4 D_{XY} takes also into account the number of pairs tied on the column -X variable T_X (sum of vertical multiplication between cells).

Table 7.7. Civil Disobedience by Education (frequency)

Civil Disobedience	Education		
	<HS	HS	College
Conscience	4	13	12
Obey Law	6	11	4

$$\begin{aligned}
 T_Y &= \text{sum of horizontal multiplication between cells} \\
 &= [(4 \times 13 + 4 \times 12 + 13 \times 12) + (6 \times 11 + 6 \times 4 + 11 \times 4)] \\
 &= 52 + 48 + 156 + 66 + 24 + 44 \\
 &= 390
 \end{aligned}$$

$$\begin{aligned}
 D_{YX} &= \frac{\# \text{ of same-ranking-pairs} - \# \text{ of opposite-ranking-pairs}}{\# \text{ of same-ranking-pairs} + \# \text{ of opposite-ranking-pairs} + T_Y} \\
 &= \frac{282 - 112}{282 + 112 + 390} \\
 &= \frac{170}{784} \\
 &= .22
 \end{aligned}$$

$$\begin{aligned}
 T_X &= \text{sum of vertical multiplication between cells} \\
 &= 4 \times 6 + 13 \times 11 + 12 \times 4 \\
 &= 24 + 143 + 48 \\
 &= 215
 \end{aligned}$$

$$\begin{aligned}
 D_{XY} &= \frac{\# \text{ of same-ranking-pairs} - \# \text{ of opposite-ranking-pairs}}{\# \text{ of same-ranking-pairs} + \# \text{ of opposite-ranking-pairs} + T_X} \\
 &= \frac{282 - 112}{282 + 112 + 215} \\
 &= \frac{170}{609} \\
 &= .28
 \end{aligned}$$

In comparison, D_{YX} is more used than D_{XY} .
All signs of interpretations of D_{YX} are same as Gamma.

Tau-b

Called Kendall's tau-b, tau-b shares characteristics of both gamma and the D_{YX} . Like Gamma, tau-b is symmetric and is PRE. Like D_{YX} , tau-b takes ties into account in predicting the rank-ordering of pairs.

$$\text{Tau-b} = \sqrt{D_{YX} D_{XY}}$$

Or,

$$\text{Tau-b} = \frac{\text{Same} - \text{Opposite}}{\sqrt{(\text{same} + \text{Opposite} + T_Y)(\text{Same} + \text{Opposite} + T_X)}}$$

Example.

$$\begin{aligned}\text{Tau-b} &= \sqrt{D_{YX} D_{XY}} \\ &= \sqrt{(.22)(.28)} \\ &= \sqrt{.0616} \\ &= .25\end{aligned}$$

Or,

$$\begin{aligned}\text{Tau-b} &= \frac{\text{Same} - \text{Opposite}}{\sqrt{(\text{same} + \text{Opposite} + T_Y)(\text{Same} + \text{Opposite} + T_X)}} \\ &= \frac{282 - 112}{\sqrt{(282 + 112 + 390)(282 + 112 + 215)}} \\ &= \frac{170}{\sqrt{477,456}} \\ &= .25\end{aligned}$$

Tau-c

Tau-c is not PRE though its interpretation is similar to tau-b.

$$\text{Tau-c} = \frac{2 \text{Min (r, c) (Same – Opposite)}}{N^2 \text{Min (r-1, c-1)}}$$

Where, N = total number of cases
 r = number of rows
 c = number of columns

Example,

$$\begin{aligned} \text{Tau-c} &= \frac{2 \text{Min (r, c) (Same – Opposite)}}{N^2 \text{Min (r-1, c-1)}} \\ &= \frac{2 (2) (282 - 112)}{(50)^2 (1)} \\ &= \frac{680}{2500} \\ &= .27 \end{aligned}$$

Note: Read carefully “Measures of Association: An Overview” on textbook page 172.

CHAPTER 7 SUPPLEMENTARY

Example 1 Civil Disobedience by Education (frequency)

Education	
Civil	
	<i>L</i> <i>H</i>

Disobedience	<HS	HS	College
<i>H</i>			
Conscience	4	13	12
Obey Law	6	11	4
<i>L</i>			

$$\begin{aligned}
 \# \text{ of same-ranking-pairs} &= [(6 \times 13) + (6 \times 12)] + (11 \times 12) \\
 &= [78 + 72] + 132 \\
 &= 282
 \end{aligned}$$

These same-ranking-pairs mean that 282 pairs are ordered in the same direction on both variables. For each of these 282 pairs, the case with a higher Education score also has a higher Civil Disobedience score. In other words, for each of 282 pairs, the case with a higher score in the independent variable also has a higher score in the dependent variable, because we assume they are positively related.

$$\begin{aligned}
 \# \text{ of opposite-ranking-pairs} &= [(4 \times 11) + (4 \times 4)] + (13 \times 4) \\
 &= [44 + 16] + 52 \\
 &= 112
 \end{aligned}$$

These opposite-ranking-pairs mean that 112 pairs are ordered in the opposite direction on both variables. For each of these 112 pairs, the case with a higher Education score has a lower Civil Disobedience score. In other words, for each of 112 pairs, the case with a higher score in the independent variable has a lower score in the dependent variable.

$$\begin{aligned}
 G &= \frac{\# \text{ of same-ranking-pairs} - \# \text{ of opposite-ranking-pairs}}{\# \text{ of same-ranking-pairs} + \# \text{ of opposite-ranking-pairs}} \\
 &= \frac{282 - 112}{282 + 112} \\
 &= \frac{170}{394} \\
 &= .43
 \end{aligned}$$

Gamma is the proportion of the difference between the positive direction pair scores and negative direction pair scores to the total pairs of scores.

Answer:

Gamma of .43 indicates a fairly strong relationship between education and civil disobedience. It suggests that as education goes up, civil disobedience tends to follow conscience.

Somers' D_{YX}

D_{YX} treats the row variable as the dependent variable called the Y and the column variable as the independent variable called X- an asymmetric measure.

Like Gamma, D_{YX} is based on the number of pairs in the same and opposite directions.

Unlike Gamma, D_{YX} takes into account the number of pairs tied on the row-Y variable T_Y (sum of horizontal multiplication between cells).

$$D_{YX} = \frac{\text{\# of same-ranking-pairs} - \text{\# of opposite-ranking-pairs}}{\text{\# of same-ranking-pairs} + \text{\# of opposite-ranking-pairs} + T_Y}$$

Unlike Gamma, D_{XY} takes also into account the number of pairs tied on the column -X variable T_X (sum of vertical multiplication between cells).

$$D_{XY} = \frac{\text{\# of same-ranking-pairs} - \text{\# of opposite-ranking-pairs}}{\text{\# of same-ranking-pairs} + \text{\# of opposite-ranking-pairs} + T_X}$$

Tau-b

Tau-b shares characteristics of both gamma and the D_{YX} .

Like Gamma, tau-b is symmetric and is PRE.

Like D_{YX} , tau-b takes ties into account in predicting the rank-ordering of pairs.

$$\text{Tau-b} = \frac{\text{Same} - \text{Opposite}}{\sqrt{(\text{same} + \text{Opposite} + T_Y)(\text{Same} + \text{Opposite} + T_X)}}$$

Tau-c

Tau-c is not PRE.

Its interpretation is similar to tau-b.

$$2\text{Min}(r, c)(\text{Same} - \text{Opposite})$$

$$\text{Tau-c} = \frac{\text{N}^2 \text{ Min (r-1, c-1)}}{\text{N}^2 \text{ Min (r-1, c-1)}}$$

Example 2 Liking of Popular and Rock Music by Education (in frequencies)
 (#6 Chapter 7 – an illustration of how to do the assignment)

Liking of Pop/Rock	Education			Total
	<HS	HS	College	
Like It	104	462	284	850
Mixed	37	170	89	296
Dislike It	129	184	84	397
Total	270	816	457	1543

How many pairs of cases are ordered in the same direction?

$$= 37 \times 462 + 37 \times 284 + 170 \times 284 + 129 \times 170 + 129 \times 89 + 184 \times 89 + \\ 129 \times 462 + 129 \times 284 + 184 \times 284 \\ = 274159$$

How many pairs of cases are ordered in the opposite direction?

$$= 104 \times 170 + 104 \times 89 + 462 \times 89 + 37 \times 184 + 37 \times 84 + 170 \times 84 + \\ 104 \times 184 + 104 \times 84 + 462 \times 84 \\ = 158930$$

How many pairs of cases are tied on the dependent variable?

$$T_y = 104 \times 462 + 104 \times 284 + 462 \times 282 + 37 \times 170 + 37 \times 89 + 170 \times 89 + \\ 129 \times 184 + 129 \times 84 + 184 \times 84 \\ = 283533$$

How many pairs of cases are tied on the independent variable?

$$T_x = 104 \times 37 + 104 \times 129 + 37 \times 129 + 462 \times 170 + 462 \times 184 + 170 \times 184 + \\ 284 \times 89 + 284 \times 84 + 89 \times 84 \\ = 273473$$

$$G = 274159 - 158930 / 274159 + 158930 = .27$$

$$D_{yx} = 274159 - 158930 / 274159 + 158930 + 283533 = .16$$

Same - Opposite

$$\text{Tau-b} = \frac{\text{Same - Opposite}}{\sqrt{(\text{same} + \text{Opposite} + T_y) (\text{Same} + \text{Opposite} + T_x)}}$$

$$= \frac{274159 - 158930}{\sqrt{(274159 + 158930 + 283533)(274159 + 158930 + 273473)}} = .16$$

$$\text{Tau-c} = \frac{2 \text{Min}(r, c) (\text{Same} - \text{Opposite})}{N^2 \text{Min}(r-1, c-1)}$$

$$= \frac{2 \times 3 \times (274159 - 158930)}{1543^2 \times 2} = .14$$

Interpretation:

Results show that liking of popular and rock music is associated with the level of education. The directionality of the relationship is positive. The strength of the relationship is moderate although values vary from one measure to another ($G = .27$; $Dyx = .16$; $\text{tau-b} = .16$; and $\text{tau-c} = .14$). In comparison, Dyx is relatively appropriate for this is an asymmetric relationship. $Dyx = .16$ suggests a moderate relationship b/w education and liking of popular music.

Dr. Ji
Soc331

HANDOUT 8

COMPARISON OF MEANS AND T TEST

A Compare means with interval/ratio dependent variable and dichotomous independent variable such as gender (male and female), region (urban and rural), and race (whites and blacks).

Box-and-Whisker Diagram and Differences between Means

- 1 Box-and whisker diagram is also called BOXPLOT. It offers a visual display of means and the differences between means. It is used to describe and compares means and mean differences between two groups of the independent variable.
- 2 The vertical axis represents the dependent variable and the horizontal axis represents independent variable.
- 3 The vertical arrays of dots show the distribution of dependent variable scores within each category of the independent variable listed along the horizontal axis.
- 4 Within the box, two lines of vertical dots are displayed representing scores for the two groups of data. Horizontal bars are displayed within the boxes showing the means of that group data.
- 5 Each elongated box extends one standard deviation above and one standard deviation below the mean dependent variable score for that value of the independent variable. Data points extend out from the standard deviation boxes which are called whiskers of the boxplot.
- 6 We calculate means within categories of the independent variable but we compare means across independent variable categories.

Example: (Go-MacroCase-GSS-#14V – t-test for SEX-X axis and TV-Y axis)

Using GSS data to compare mean difference between male and female TV watching hours (#14 Sex and #86 Watching TV), the mean hours of watching TV per day are 2.78 for men and 3.11 for women. So the mean difference is $2.78 - 3.11 = -.33$, indicating women watch TV about .33 (60/100x.33=20 minutes) hour a day more than men. The mean of watching TV for both men and women is 2.96 hour a day.

It can be displayed in a boxplot (Textbook page 183).

A Review of Variance, Standard Deviation, and Standard Error

The variance (s^2 / σ^2) – is the average squared deviation of scores from the mean

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

The standard deviation (s / σ) - is the square root of the variance

$$s = \sqrt{\frac{(X_i - \bar{X})^2}{N - 1}}$$

The standard error of the mean ($\sigma_{\bar{X}}$)

The standard deviation of the sampling distribution of the means (from all possible samples of a given size drawn randomly from a population).

$$\sigma_0 = \sigma / N \text{ (for population data),} \quad S_0 = S / N \text{ (for sample data)}$$

Where σ_0 = standard error;

σ = standard deviation;

N = sample size

S_0 = standard error;

s = standard deviation;

T-Test for the Difference between Means

- 1 T-test is also called student's t.
- 2 While Chi Square is used to test the significance that a relationship found in a sample can be found in a population, t-test is used to test the null hypothesis that there is no difference between means of a dichotomous independent variable in the population:

$$H_0: \Phi_1 = \Phi_2, \quad \text{or} \quad \Phi_1 - \Phi_2 = 0.$$

H_0 : the mean of Population One is equal to the mean of Population Two. or,
 H_0 : the difference of the two population means is 0.
- 3 If there is a difference between means, we reject the null hypothesis and conclude that the variables are indeed related; if there is no difference between means, we can not reject the null hypothesis and conclude that the variables are not related in the population.
- 4 Thus, t-test is used to generalize a difference between means of a dichotomous independent variable in sample data to the population. It is the way through which we can confidently generalize the difference between means to the larger population from which the sample was drawn.
- 5 The conclusion of t-test is based on sample size N , degree of freedom DF , and the significant level of P (p #.05, .01, .001).

T-Test Formulas:

$$t = \frac{(O_1 - O_2)}{S_{O_1 - O_2}}$$

The numerator is the difference b/w the dependent variable means
The denominator is the standard error of the difference b/w means
t tells us how many standard errors our difference is from the 0

t expresses difference in terms of standard error units.

Where, $t = t\text{-statistic}$
 $0_1 - 0_2 = \text{the difference between means}$
 $S0_1 - 0_2 = \text{standard error (see formula below).}$

$$S0_1 - 0_2 = \theta \left(\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2} \right) \left(\frac{N_1 + N_2}{N_1 N_2} \right)$$

Where, $S_1^2 = \text{dependent variable variance for category 1 of the independent variable;}$

$S_2^2 = \text{dependent variable variance for category 2 of the independent variable;}$

$N_1 = \text{number of cases in category 1 of the independent variable;}$

$N_2 = \text{number of cases in category 2 of the independent variable.}$

$$df = N_1 + N_2 - 2$$

$N_1 = \text{number of cases in category 1 of the independent variable;}$

$N_2 = \text{number of cases in category 2 of the independent variable.}$

Example:

A researcher studies employee turnover at eight randomly selected McDee restaurants of similar size. 4 of the restaurants hold regular employee meetings to discuss the operation of the restaurants; the other 4 do not hold meetings. Table 8.1 presents the number of employee resignations in a 12-week period at McDees.

1) present a boxplot for the data and indicate the means of each independent variable category;

2) carry out a difference of means test to find means, difference of means, t statistic, df, and p.

Table 8.1. Employee resignations by employee meetings
at McDees Restaurant (exercise p.159 #5)

Employee	Number of
----------	-----------

Restaurant	Meetings	Resignations
------------	----------	--------------

McDee 01	No	12
McDee 02	No	8
McDee 03	No	11
McDee 04	No	9
McDee 05	Yes	9
McDee 06	Yes	10
McDee 07	Yes	5
McDee 08	Yes	8

Dependent V	employee resignations
Independent V	employee meeting: Yes/No

Null hypothesis:

$H_0: \Phi_1 = \Phi_2$, or $\Phi_1 - \Phi_2 = 0$.

The mean score of number of resignations between McDee Restaurants who have meetings and those who do not are equal.

Step 1: Calculate means for 0₁ and 0₂.

$$0_1 = \frac{\sum 3x_i}{N} = \frac{12+8+11+9}{4} = \frac{40}{4} = 10$$

$$0_2 = \frac{\sum 3x_i}{N} = \frac{9+10+5+8}{4} = \frac{32}{4} = 8$$

Step 2: Calculate dependent variable variance for categories 1 and 2 of the independent variable.

$$S_1^2 = \frac{\sum 3(x - 0)^2}{N - 1} = \frac{(12-10)^2 + (8-10)^2 + (11-10)^2 + (9-10)^2}{4 - 1} = \frac{10}{3} = 3.333$$

$$S_2^2 = \frac{3(x - 0)^2}{N - 1} = \frac{(9-8)^2 + (10-8)^2 + (5-8)^2 + (8-8)^2}{4 - 1} = \frac{14}{3} = 4.667$$

Step 3: Calculate the standard error.

$$S_{0_1 - 0_2} = \theta \left(\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2} \right) \left(\frac{N_1 + N_2}{N_1 N_2} \right)$$

$$= \theta \left(\frac{4(3.333) + 4(4.667)}{4 + 4 - 2} \right) \left(\frac{4 + 4}{(4)(4)} \right)$$

$$= \theta \left(\frac{32}{6} \right) \left(\frac{8}{16} \right) = 02.667 = 1.663$$

Step 4: Calculate t.

$$t = \frac{(0_1 - 0_2)}{S_{0_1 - 0_2}} = \frac{10 - 8}{1.663} = 1.225$$

Step 5: Calculate the degree of freedom (df):

$$df = N_1 + N_2 - 2 = 4 + 4 - 2 = 6$$

Step 6: Go to the t distribution table on page 304 to find the p <.05, or p <.01, or p < .001, to decide the significant level if any.

- 1 Find df in left column.
- 2 Find the right significant level.

- 3 If the result value (t) is equal or larger than the table value, we reject the null hypothesis that there is no difference between means. Then we accept the research hypothesis that there is a difference between the means.

Results:

$$O_1 = 10$$

$$O_2 = 8$$

$$\text{Difference between means} = 10 - 8 = 2$$

$$t = 1.225$$

$$df = 6$$

$$p < \text{n.s.}$$

Conclusion:

The analysis shows that the difference between mean resignations of McDee restaurants with meetings or without meetings is not statistically significant ($t = 1.225$ [which is smaller than the table value of 1.943], $df=6$, $p \# \text{ n.s.}$). Therefore we can not reject the null hypothesis. This indicates that having employee meetings or not makes no difference regarding employees' resignation in all McDee Restaurants.

Last: Boxplot of Resignations By Employee Meetings (See Text exercise p. 160)

One-tailed and two-tailed test

- 1 In a normal distribution, there are two tails: left tail \square 2.5% and right tail \square 2.5%. The middle part is the sample means taking about \square 95%. All together under the normal curve, we have a value of $1.00 = 2.5\% + 2.5\% + 95\%$.
- 2 When the sampling mean falls far away from the mean of the sampling distribution of sample means - in one of these tails of the normal distribution, we conclude that the sampling mean was less likely to be representative of the population. When the sampling mean was included in the 95% of sampling means that fall near the mean of the sampling distribution, we conclude that the sampling mean was more likely to be representative of the population.
- 3 If a hypothesis is directional - which tells us which category is likely to score higher than the other one, we apply one-tail test. If the hypothesis is non-directional - which does not say any thing about the nature of the association between variables, we use two-tailed test of significance.
- 4 Keep in mind that we are always testing the null hypothesis - the hypothesis of no association. Keep in mind that we are always aiming to test the research hypothesis - the hypothesis of having association between variables.

- 5 The t statistic can be negative or positive. Either a larger negative t or a larger positive t would lead us to reject the null hypothesis of no population difference.
- 6 Two-tailed test has become customary - the default, unless there is a good reason for a one-tailed test.
- 7 Each cell entry in t table is the minimum value that t needs to reject the null hypothesis for a given degree of freedom.

Confidence intervals for differences between means

In chapter 4, we discussed confidence interval - where we use a sample statistic to estimate a population parameter. Population mean may be greater or less than the sample mean. To estimate a population mean, it is useful to establish a range around the sample mean within which we think the population mean lies. This range is called confidence interval.

The confidence interval for difference between means

is analogous to interpretation of the confidence interval around a mean.

General formula:

$$(1 - \alpha) \text{ confidence interval} = (\bar{O}_1 - \bar{O}_2) \pm t_{\alpha} S_{\bar{O}_1 - \bar{O}_2}, \quad \text{where,}$$

$\bar{O}_1 - \bar{O}_2$ = difference between means;

α = the alpha level corresponding to the confidence level

If $\alpha = .05$, then $(1 - \alpha) = 95\%$; If $\alpha = .01$, then $(1 - \alpha) = 99\%$.

t_{α} = the value of t associated with the α level for given df;

$S_{\bar{O}_1 - \bar{O}_2}$ = the standard error of the difference between means.

Two commonly used formula to find confidence intervals around the difference between means: 95 and 99 percent.

- 1) 95 percent confidence interval $= (\bar{O}_1 - \bar{O}_2) \pm t_{.05} S_{\bar{O}_1 - \bar{O}_2}$
The chances are 95 out of 100 that the population difference between means lies within a 95 percent confidence interval.
- 2) 99 percent confidence interval $= (\bar{O}_1 - \bar{O}_2) \pm t_{.01} S_{\bar{O}_1 - \bar{O}_2}$
The chances are 99 out of 100 that the population difference between means lies within a 99 percent confidence interval.

Steps to find confidence interval around the difference between means:

- 1) Find means of the 2 groups \bar{O}_1 and \bar{O}_2 .
- 2) Find t value associated with the given degree of freedom.
- 3) Find the standard error $S_{\bar{O}_1 - \bar{O}_2}$.

Example,

The mean for watching TV daily is $\mu_1 = 2.75$ for men and $\mu_2 = 3.01$ for women, the standard error is $S_{\mu_1 - \mu_2} = .098$, and $t_{.01} = 2.576$ ($t = 2.653$ at $df=1938$).

Thus,

99 percent confidence interval around the difference between means

$$\begin{aligned} &= (\mu_1 - \mu_2) \pm t_{.01} S_{\mu_1 - \mu_2} \\ &= (2.75 - 3.01) \pm (2.576) (.098) \\ &= -.26 \pm .25 \\ &= -.51 \text{ to } -.01 \end{aligned}$$

Thus the chances are 99 out of 100 that difference between means in the population lies between $-.51$ to $-.01$. In other words, women on average have $(60/100 \times .01 \text{ and } 60/100 \times .51) .006 - .306 =$ less one minute to half an hour more TV hours a day.

Distinguish Confidence Intervals for Means and for Differences between MeansConcepts differ:

Confidence Interval for Means is to establish a range around the sample mean within which the population mean lies (to estimate a population mean).

Confidence Interval for Differences between Means is to find out the difference between means within which the population difference between means lies. It is to test for the significance of a difference between means.

The formulas differ:

The first CI is for mean and the second CI is for difference between means:

$$1). 95 \text{ percent confidence interval} = \bar{x} \pm 1.96\sigma_{\bar{x}}$$

The chances are 95 out of 100 that the population mean lies within this range.

We need only mean = \bar{x} , and standard error = $\sigma_{\bar{x}}$ for calculation.

$$2). 95 \text{ percent confidence interval} = (\mu_1 - \mu_2) \pm t_{.05} S_{\mu_1 - \mu_2}$$

The chances are 95 out of 100 that the population difference between means lies within a 95 percent confidence interval.

We need two means μ_1 & μ_2 , t statistic at .05 / .01 significant level (t .05/ t .01), and standard error of the difference between means = $S_{\mu_1 - \mu_2}$ for calculation.

**Means, Standard Deviations and Number of Cases of Dependent Var:
WATCH TV**

by Categories of Independent Var: SEX

Difference of means across groups is statistically significant (prob. 0.002)

	N	Mean	Std.Dev.
MALE	856	2.776	2.133
FEMALE	1091	3.105	2.556

Analysis Of Variance

Dependent Variable: WATCH TV

Independent Variable: SEX

N: 1947

Missing: 957

ETA Square = 0.005

Source	Sum of Square	DF	Mean Square	t	Prob.
Between	52.142	1	52.142	3.035	0.002
Within	11009.813	1945	5.661		

TOTAL 11061.955 1946

TABLE 2: THE T DISTRIBUTION /student t (page 304)

Level of Significance for One-tailed Test						.10
.05	.025	.01	.005	.0005		
Level of Significance for One-tailed Test						.20
.10	.05	.02	.01	.001		
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.0965	9.925	31.598
3	1.638	2.353	3.182	4.451	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	-	-	-	-	-	-
8	-	-	-	-	-	-
9	-	-	-	-	-	-
10	-	-	-	-	-	-
11	1.363	1.796	2.201	2.718	3.106	4.437
12	-	-	-	-	-	-
13	-	-	-	-	-	-
14	-	-	-	-	-	-
15	-	-	-	-	-	-
20	1.325	1.725	2.086	2.528	2.845	3.850
	-	-	-	-	-	-
	-	-	-	-	-	-
	-	-	-	-	-	-
30	1.310	1.697	2.042	2.457	2.750	3.646
	-	-	-	-	-	-
	-	-	-	-	-	-
	-	-	-	-	-	-
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
4	1.282	1.645	1.960	2.326	2.576	3.291

Note: 1) Find Degree of Freedom; 2) Find the corresponding significance level; 3) If the result value is greater than the table t-value, reject the H_0 ; otherwise don't reject H_0 .

Dr. Ji
Soc. 331

HANDOUT 9

ANALYSIS OF VARIANCE

A Unlike the dichotomous independent variable in Chapter 8, this Chapter focuses on ANOVA with a non-dichotomous independent variable measured at any level and an interval/ratio dependent variable.

Analysis of variance - ANOVA

Analysis of variance has an acronym: ANOVA. It is a statistical technique that makes full use of the interval/ratio level of dependent variables.

Box-and-Whisker Diagrams/Differences Among Means

- 1 Vertical axis is dependent variable - interval or ratio, whereas horizontal axis is independent variable - interval, ratio, ordinal, or nominal;
- 2 The elongated boxes extend one standard deviation above and one standard deviation below mean dependent variable scores for values of the independent variable;
- 3 Lines can be connected between midpoint of each box showing the direction of the relationship between variables - positive, negative, or curvilinear;
- 4 If independent variable is nominal, the midpoint in each box can not be connected because they are meaningless. With independent variable as interval, ratio, or ordinal, lines can be attached.

Example:

Table 9.1. TV Hours by Education Levels

N	case #	Education	TV Hour / Day	
1	1	<HS	3	
2	2	<HS	4	N=2
				$0 = \frac{3+4}{2} = 3.5$
1	3	HS	2	
2	4	HS	2	
3	5	HS	3	
4	6	HS	4	$2+2+3+4+4$

5	7	HS	4	N=5	$0 = \frac{\text{-----}}{5} = 3$
1	8	Cel	1		
2	9	Cel	2		$1+2+3$
3	10	Cel	3	N=3	$0 = \frac{\text{-----}}{3} = 2$

Total number of cases = N= 10

Number of categories/groups of the independent variable = K = 3

Figure 9.1. Boxplot for TV Hours by Education

T
V

H
O
U
R
S

Education Level

Compare Differences Among Means

Table 9.2. Mean Daily TV Hours by Edu

TV H	Education			Total
	<HS	HS	Cel	
Mean Hour	3.5	3.0	2.0	2.8
Sta. Dev.	0.7	1.0	1.0	1.1

Rule of Comparing Means

- 1 Like percentage tables in cross-tabulation, we calculate means within categories of the independent variable and compare means across categories of the independent variable.

- 2 The logic is exactly the same as the comparison of two means as shown in Chapter 8. The only difference is that we have more means.
- 3 To compare means to determine if TV watching is related to education.
- 4 The relation is clear: lower the education level, higher the mean hours TV watching; higher the education level, lower the mean hours TV watching. The relationship between education and TV watching is negatively related.

Purpose and Assumptions of ANOVA

- 1 Do differences we find in means in a sample happen by chance? How confident we are that the differences found in a sample exist in the population? Can we generalize the differences from the sample to the population from which the sample was drawn?
- 2 Required assumptions in Analysis of variance -ANOVA
 - 1) Dependent variable must be interval or ratio. Ordinary variable is also often used by some researchers.
 - 2) Random sampling. The sample is drawn randomly from a population.
 - 3) Independence. The category means must be independent of one another. The scores of the sub-categories in the independent variable must be sampled independently. For example, the chance of a student A being selected has nothing to do with the chance student B is selected.
 - 4) Normal Distribution. The dependent variable is assumed to be normally distributed in the population.
 - 5) Homoscedasticity. It is assumed that the distributions of the dependent variable within independent variable categories have equal variances in the population. This condition is called homoscedasticity.

Logic of ANOVA

- 1 Break the total variation in dependent variable scores into two parts:
 - 1) the variation that occurs within each independent variable group;
 - 2) the variation that occurs between independent variable groups.
- 2 If variables are associated, the independent variable groups will differ quite a bit from one another.
- 3 The variation between groups is greater than the variation within groups.
- 4 **To get variation**, we need to calculate sum of squares (the sum of the squared deviations from the mean - the numerator) and the degree of freedom (the denominator).

Recall the formula for calculating the variance

Sum of squares	A	$(\sum x_i - 0)^2$
Variance	A	$s^2 = \frac{\text{Sum of squares}}{\text{Degree of Freedom}}$
Degree of Freedom	A	$N - 1$

Formula for Sum of Squares

$$\sum (\bar{x}_i - 0_T)^2 = \sum (\bar{x}_i - 0_G)^2 + \sum N_G (0_G - 0_T)^2$$

Total sum of squares = Within-group sum of squares + Between-group sum of squares

$$\sum (\bar{x}_i - 0_T)^2 = \text{Total sum of squares}$$

$$\sum (\bar{x}_i - 0_G)^2 = \text{Within-group sum of squares}$$

$$\sum N_G (0_G - 0_T)^2 = \text{Between-group sum of squares}$$

\bar{x}_i = dependent variable score of the i^{th} case

0_T = total mean

0_G = mean of the G^{th} independent variable group that the i^{th} case is in

N_G = number of cases in each group of the independent variable

Example

Total sum of square

$$\begin{aligned} \sum (\bar{x}_i - 0_T)^2 &= (3-2.8)^2 + (4-2.8)^2 + (2-2.8)^2 + (2-2.8)^2 + (3-2.8)^2 + (4-2.8)^2 + (4-2.8)^2 + \\ &\quad (1-2.8)^2 + (2-2.8)^2 + (3-2.8)^2 \\ &= (.2)^2 + (1.2)^2 + (-.8)^2 + (-.8)^2 + (.2)^2 + (1.2)^2 + (1.2)^2 + (-1.8)^2 + (-1.8)^2 + (.2)^2 \\ &= .04 + 1.44 + .64 + .64 + .04 + 1.44 + 1.44 + 3.24 + .64 + .04 \\ &= 9.60 \end{aligned}$$

Within-group sum of square

$$\begin{aligned} \sum (\bar{x}_i - 0_G)^2 &= (3-3.5)^2 + (4-3.5)^2 + (2-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (4-3)^2 + \\ &\quad (1-2)^2 + (2-2)^2 + (3-2)^2 \\ &= (-.5)^2 + (.5)^2 + (-1)^2 + (-1)^2 + (0)^2 + (1)^2 + (1)^2 + (-1)^2 + (0)^2 + (1)^2 \\ &= .25 + .25 + 1 + 1 + 0 + 1 + 1 + 1 + 0 + 1 \\ &= 6.50 \end{aligned}$$

Between-group sum of square

$$\begin{aligned} \sum N_G (0_G - 0_T)^2 &= 2(3.5 - 2.8)^2 + 5(3.0 - 2.8)^2 + 3(2.0 - 2.8)^2 \\ &= 2(.7)^2 + 5(.2)^2 + 3(-.8)^2 \\ &= .98 + .20 + 1.92 \\ &= 3.10 \end{aligned}$$

Therefore,

Total sum of square = Within-group sum of square + Between-group sum of square

$$\sum (\bar{x}_i - 0_T)^2 = \sum (\bar{x}_i - 0_G)^2 + \sum N_G (0_G - 0_T)^2$$

$$9.6 = 6.50 + 3.10$$

Formula for Variance

$$\text{Total Variance} = \frac{\sum (\bar{x}_i - \bar{0}_T)^2}{\text{Total Degree of Freedom } A} = \frac{9.6}{N - 1} = \frac{9.6}{10 - 1} = 1.067$$

$$\text{Within-Groups Variance} = \frac{\sum (\bar{x}_i - \bar{0}_G)^2}{\text{Within-Group Degree of Freedom } A} = \frac{6.50}{N - k} = \frac{6.50}{10 - 3} = .928$$

$$\text{Between-Groups Variance} = \frac{\sum N_G (\bar{0}_G - \bar{0}_T)^2}{\text{Between-Group Degree of Freedom } A} = \frac{3.10}{k - 1} = \frac{3.10}{3 - 1} = 1.551$$

Where,

<u>Total Degree of Freedom</u>	df = N - 1 = 10 - 1 = 9
	N = total number of cases

<u>Within-Group Degree of Freedom</u>	df = N - k = 10 - 3 = 7 (N ₂)
	k = number of categories of the independent variable

<u>Between-Group Degree of Freedom</u>	df = k - 1 = 3 - 1 = 2 (N ₁)
----------------------------------------	------------------------------------------

***These estimated variances are called Mean Sum of Squares or Mean Squares. Statisticians have proven that if we divide each sum of squares by these degrees of freedoms, we can get good estimates of variance.

F Ratio and Uses of F Ratio

$$F_{N_1, N_2} = \frac{\text{Between-groups variance}}{\text{Within-groups variance}} = \frac{1.550}{.928} = 1.67$$

- 1 F ratio is a ratio of between-groups variance to within-groups variance. It assesses the overall strength of a relationship. The stronger the relationship, the larger the ratio; the weaker the relationship, the smaller the ratio.
- 2 When Between-groups variance is larger, we have big ratio; when Between-groups variance is small, we have little ratio; when Between-groups variance is moderate, the ratio is also moderate.
- 3 F ratio is used to establish statistical significance at various levels. The minimum F ratios required for significance is at the .05, .01, and .001 levels.
- 4 F ratios and associated probabilities are based on the assumptions underlying ANOVA - random sampling, independent means, interval/ratio variable that is normally distributed (DV), and homoscedasticity (equal variance).
- 5 The significance of F ratios depends on two kinds of degrees of freedom: Within-Group $df = N - k$ and between-groups $df = k - 1$.
- 6 In F distribution table, it shows the F ratio required for significance at various levels. If the calculated F ratio is larger than the F value in the table, it means that the relationship is statistically significant. The F test tells us that the variables under study are really related in the larger population.
- 7 For dichotomous independent variable such as Gender (male/female) and Region (urban/rural), t test and F test are identical. That is, $t = \sqrt{F}$. T test is preferred.
- 8 For independent variable with three or more values, ANOVAs/F test are preferred.
- 9 ANOVA table and Interpretation.

Table 9.3. Analysis of Variance of Daily Hours of TV Watching by Education

Source	Sum of Squares	df	Mean Sum of Squares	F	p
Between Groups	3.10	2	1.550	1.67	n.s.
Within Groups	6.50	7	.928		
Total	9.60	9	1.067		

$F_{2,7} = 1.67$; $p < n.s.$;

***See more examples about ANOVA via SPSS, employee data.sav - ANOVA - means plot b/w mean of current salary and employment category.

Conclusion:

The results of ANOVA show that the relationship between TV watching and education levels is not statistically significant ($F_{2,7} = 1.67$; $p < n.s.$).

The Correlation Ratio (E^2)

Eta squared (E^2) is a measure of association. It describes how strongly the dependent variable is related to the independent variable. It is a PRE measure of association. Another equivalent way to describe Eta is that E^2 is the proportion of variation in the dependent variable explained by the independent variable, which is the same interpretation of R square as in the case of regression.

Formula

$$E^2 = \frac{\text{Between-Groups Sum of Squares}}{\text{Total Sum of Squares}} = \frac{3.10}{9.60} = .32$$

Interpretation:

As indicated by Eta, education explains about 32 percent of the variation in hours of TV watching.

One-Way Analysis of Variance

Analysis of variance involving a single independent variable is called One-Way Analysis of Variance.

Two-Way Analysis of Variance and Beyond

Analysis of variance involving two independent variables and their interaction effects is called Two-Way Analysis of Variance.

Multivariate ANOVA (MANOVA)

Multivariate ANOVA technique handles two or more dependent variables that are related to one another (beyond the scope of this text).

Cautions about F Ratios

- 1 F test is largely based on sample size. Statistical significance is easier to achieve with larger samples. Therefore, statistical significance does not imply substantive significance. Be sure to inspect means to see whether differences among them are substantively important.
- 2 A statistically significant F does not necessarily mean that all the means are different from one another. ANOVA tests for an overall difference among means, not for differences between particular means.
- 3 ANOVA does not necessarily imply that the two variables are causally related. Many pairs of variables are associated that are not causally related.

Example

Describe the relationship between RACE and EDUCATION.

Data

GSS

Task ANOVA
 Dependent #9 Education
 Independent #15 Race
 View Means, ANOVA, and Boxplot

Means, Standard Deviations and Number of Cases of Dependent Var: EDUCATION
 by Categories of Independent Var: RACE
 Difference of means across groups is statistically significant (prob. 0.000)

	N	Mean	Std.Dev.
WHITE	2344	13.492	2.871
BLACK	400	12.505	2.907
OTHER	151	13.669	3.469

Analysis Of Variance
 Dependent Variable: EDUCATION
 Independent Variable: RACE N: 2895 Missing: 9

ETA Square = 0.014

TEST FOR NON-LINEARITY:
 R Square = 0.003 F = 32.044 Prob = 0.000

Source	Sum of Squares	DF	Mean Square	F	Prob.
Between	347.527	2	173.763	20.522	0.000
Within	24487.279	2892	8.467		
TOTAL	24834.807	2894			

is

Data from GSS
(Chapter 10
Exercise #10)
indicates that there
a relationship
between race and
years of education.
The mean
education is 12.5
years for Blacks,
13.5 for Whites,

and 13.7 for other races. The $F_{2, 2892} = 20.522$, is statistically significant ($p < .001$). The Eta squared is .014, indicating that race only explains about 1% of the variation in years of education. The relationship is very weak.

Dr. Ji
SOC331

HANDOUT 10

REGRESSION AND CORRELATION

- A Analyze relationships b/w two interval/ratio variables (Dep./Ind)
- A Calculate and interpret the slope and intercept of a regression line
- A Calculate and explain a correlation coefficient
- A Interpret as proportion of variation on dependent variable explained by the model

Scatterplots

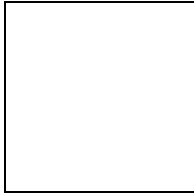
- 1 Definition. Also called scattergram, scatterplots are graphs that help visualize relationships b/w interval/ratio variables. By convention, the independent variable is arrayed along the horizontal X-axis and the dependent variable is arrayed along the vertical Y-axis. Scatterplots are often set up so that the X-and Y-axes cross at the origin -

the zero point for each variable. A box-and whisker diagram is a typical type of scatterplot.

- 2 Direction. Scatterplot can identify the direction of a relationship immediately. If the relationship is positive, scatterplot points form a pattern from the lower left to the upper right (Graph 1). If the relationship is negative, the pattern of points sweeps from the upper left to the lower right (Graph 2).

Graph 1. Scatterplot about Fertility by Infant Mortality

Graph 2. Scatterplot about Infant Mortality by Women's Literacy

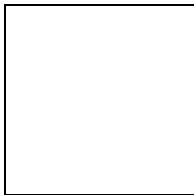


The above is an example about scatterplot- SPSS- world95.sav. It illustrates the relationship b/w fertility and infant mortality, infant mortality and females' literacy.

- 3 Strength (See text pages 230-231)
 - 1) if a relationship is perfect and positive, the data points form a straight line (/).

- 2) if a relationship is strong, the scatter plot points cluster tightly around a straight line.
- 3) if a moderate relationship, points are more dispersed with the direction still evident.
- 4) if a weak relationship, scatterplot points are widely dispersed with direction even hard to identify.
- 5) if no relationship, all points are randomly distributed about the scatterplot. Graph 3 is an example of no relationship b/w climate and women's literacy (Graph 3).

Graph 3. Climate by Women's Literacy



- 4 Scatterplot shows both the direction and strength. But it also has limitations.
- 1) It is most useful when N is small ($N < 100$). Otherwise, scatterplots will pile up on top of one another, making visualization difficult.
- 2) It is only two-dimensional space of a scatterplot, making it difficult to display data for a larger number of cases. Even a triangle might stand for 5 cases, a square for 6 to 10 cases, and so on.
- 3) It is very useful with aggregate data like those of the world, American states, or Canadian provinces, etc.

Regression Line/Least-Square Line

Regression line is also called least-squares line. This single straight line summarizes the relationship more precisely and best describes the relationship. This line runs through the data points and best represents the general pattern of the scatterplot dots quite closely.

- 1 Using a straight line to describe a relationship is called linear regression (see graph on the right)
- 2 It is called least-squares line because the line minimizes the sum of squared distances to dependent variable scores and best predicts a case's score on the dependent variable if we know the case's independent variable score.
- 3 Regression line does not go through each and every data point but most cases lie somewhere off the regression line.
- 4 If you measure the length of each residual, square those lengths, add up the squares, and the sum would be the line that is less than any other line you could draw (see graph on the right).

Regression Equation

$$Y = a + bX$$

where,

Y = score on the dependent variable

a = Y-intercept
the value of Y when the line crosses the Y-axis when X = 0;

b = slope
the change in Y for every one-unit change in X

the steepness of the line

X = scores on the
independent variable

- 1 Wording. We speak of “dependent variable Y’s regression on the independent variable X.”
- 2 The Y-intercept (a) is also called the CONSTANT.
- 3 The slope b is also called the REGRESSION COEFFICIENT. It tells us how much change in the dependent variable is associated with one unit increase in the independent variable.

Example

$$Y = 5.720 - .041X$$

This equation indicates that if urbanism goes up one percentage point, the average of number of children born to each woman will drop .041.

$$\begin{aligned} Y &= 5.720 - .041 (45) \\ &= 5.720 - 1.845 \\ &= 3.875 \end{aligned}$$

This equation means that if a country’s urbanism is 45 Percent, then the fertility in that country will be 3.875, namely, on average, a woman will bear 3.875 children at the urbanization level of 45%.

$$\begin{aligned} Y &= 5.720 - .041 (0) \\ &= 5.720 - 0 \\ &= 5.720 \end{aligned}$$

This equation suggests that if a country’s urbanization level is 0 (hypothetical), then the fertility rate will be 5.720, namely, on average, a woman will bear 5.72 children. Then one will not use urbanism to predict fertility because urbanization is non-existent.

Calculating Regression Coefficients

Formula for Regression Coefficient (b)

$$b = \frac{3(X - 0)(Y - y)}{3(X - 0)^2} !!$$

$$Y = a + bX$$

b is the ratio of the sum of cross-products to the sum of squares of the independent variable. This coefficient tells us how much change in the dependent variable is associated with one unit increase in the independent variable.

Example

(page 343: “Elementary Statistics in Social Research” by Levin and Fox)

Table 10.1 Sentence Length by
Prior Convictions for 10 Defendants

Priors (X)	Sentence in Years (Y)
0	12
3	13
1	15
0	19
6	26
5	27
3	29
4	31
10	40
8	48

We assume prior convictions are independent and years of sentences are dependent variable.

Table 10.2. Regression Calculations for Data in Table 10.1

X	Y	X - 0	Y - y	(X - 0)(Y - y) SP	(X - 0) ² SSx	(Y - y) ² SSy	Zx	Zy	ZxZy
0	12	-4	-14	56	16	196	-1.201	-1.187	1.426
3	13	-1	-13	13	1	169	-.300	-1.103	.331
1	15	-3	-11	33	9	121	-.901	-.933	.841
0	19	-4	-7	28	16	49	-1.201	-.594	.713
6	26	2	0	0	4	0	.601	0	0
5	27	1	1	1	1	1	.300	.085	.025
3	29	-1	3	-3	1	9	-.300	.254	-.076
4	31	0	5	0	0	25	0	.424	0
10	40	6	14	84	36	196	1.802	1.187	2.140
8	48	4	22	88	16	484	1.201	1.866	2.241
3X=40 3Y=26		3SP =300		3SSx=100		3SSy=1,250		ZxZy = 7.641	

$$0 = \frac{3X}{N} = \frac{40}{10} = 4$$

$$y = \frac{3X}{N} = \frac{260}{10} = 26$$

$$N = 10$$

SP =sum of products

SSx =sum of squares for IV

SSy =sum of squares for DV

$$\text{IV-Variance } s^2 = 3(X - 0)^2/N - 1 = 100/9 = 11.11 \quad \text{St.D } s = \sqrt{11.11} = 3.33$$

$$\text{DV-Variance } s^2 = 3(Y - y)^2/N - 1 = 1250/9 = 138.89 \quad \text{St.D } s = \sqrt{138.89} = 11.79$$

Z-score (standardized score): $Z_x = X - 0 / s$

$$\text{Correlation Coefficient: } r = 3Z_x Z_y / N - 1 \quad r = 7.641/10 - 1 = .85$$

$$b = \frac{3(X - 0)(Y - y)}{3(X - 0)^2} = \frac{3SP}{3SS_x} = \frac{300}{100} = 3$$

Because $Y = a + bX$,
thus $a = y - bX$

$$a = 26 - (3)(4) = 26 - 12 = 14$$

$$b = 3$$

$$y' = a \text{ (when } X \text{ is } 0)$$

$$\begin{aligned} y' &= a + bX \\ &= 14 + 3(10) \\ &= 14 + 30 \\ &= 44 \end{aligned}$$

Note: One should select a larger value of X so that the point is far from the Y-intercept and plug it into the equation. In this example, the larger X is 10.

y' is pronounced “Y” prime to indicate the predicted value of the sentence as opposed to the actual value of sentence Y. Because prediction always has an error which is denoted by e in the equation. Since this error is very small, we usually omit the error.

$$\begin{aligned} \text{Based on } y' &= a + bX \\ &= 14 + 3(10) \\ &= 14 + 30 \\ &= 44 \end{aligned}$$

We can draw a regression line to illustrate the relationship between Sentence length (Y) and number of prior convictions (X). (Draw line in Class- page 343 from “Elementary Statistics in Social Research” by Levin and Fox).

Correlation Coefficient (r)

Correlation coefficient is a shortened form for Pearson’s product-moment correlation coefficient. Its role is to assess the strength of a relationship. It is a summary measure of how tightly cases are clustered around the regression line. It is the mostly widely used measure of association in the social science.

- 1 If cases bunch very closely along the regression line, r is large in magnitude, indicating a strong relationship; if cases are widely dispersed about the regression line, then r is small in magnitude, indicating a weak relationship; if $r = 0$, indicating two variables are statistically un-associated and there is no relationship.
- 2 The value of r ranges between -1.00 to + 1.00 with the + and - signs indicating the direction of the relationship. $R = +1.00$ or -1.00 indicating perfect relationship with the signs denoting direction.
- 3 Formula for the correlation coefficient (r)

$$r = \frac{\sum Z_x Z_y}{N}$$

where, r = Pearson product-moment correlation coefficient / Correlation Coefficient

Z_x = standardized score on the independent variable X

Z_y = standardized score on the dependent variable Y

N = number of cases (for population data, we use N while for sample data, we use N-1. But that makes little difference)

$$\text{The formula for Z-score} = \frac{X - \bar{X}}{s}$$

\bar{X} ≈ subtract mean for a particular score
 s ≈ standard deviation

- 4 All Z-scores are standardized scores and not affected by the unit of measurement. After standardized, both DV and IV have means of 0 with standard deviation of 1.00.
- 5 r is a symmetric measure of association. It does not matter which of the two variables is dependent or independent.
- 6 Thumb of rules to assess r

	Negative relationship					No Relationship	Positive Relationship				
R =	<u>-1.00</u>	<u>-.80</u>	<u>-.60</u>	<u>-.40</u>	<u>-.20</u>	<u>.00</u>	<u>.20</u>	<u>.40</u>	<u>.60</u>	<u>.80</u>	<u>1.00</u>
	Perfect	strong	moderate	weak	None		weak	Moderate	Strong	Perfect	

Example

The correlation coefficient b/w Bridges and Trolls (p.203) is .56, a moderate relationship. The correlation coefficient b/w son's education and Dad's education is (p.206) is .39, a moderate rel.

The correlation coefficient b/w Dropout and Murder (p.210) is .80, a strong relationship. The correlation coefficient b/w birth rate and Infant mortality (spssworld95) is .865, a strong relationship.

- 7 Correlation Matrix
 - 1) Correlation matrix is a format that presents correlation coefficients describing interrelationships among three or more variables.
 - 2) Correlations in the major diagonal are always 1.00 because variables are perfectly correlated with themselves.
 - 3) Asterisks are used to indicate significant levels ($p < .05$, .01, or .001).
 - 4) When doing correlation matrix, we will have two deletions: listwise and pairwise. Listwise deletion is to exclude any cases that is missing information on any variable in the entire analysis. Pairwise deletion is to exclude missing information on particular variable. We prefer listwise option (same N) to pairwise (different N).

Example: Correlation among income, age, political view, and religion.

Correlation Coefficients
 N: 2437 Missing: 467
 Cronbach's alpha: Not calculated--negative correlations

LISTWISE deletion (2-tailed test) Significance Levels: ** =.01, * =.05

	INCOME	AGE	POL. VIEW	RELIGION
INCOME	1.000	-0.037	0.051 *	-0.006
AGE	-0.037	1.000	0.060 **	-0.184 **
POL. VIEW	0.051 *	0.060 **	1.000	-0.171 **
RELIGION	-0.006	-0.184 **	-0.171 **	1.000

- 8 If r^2 squared, it becomes R square. Then the r^2 has a special meaning in regression analysis

r^2 as Proportion of Variation Explained (r^2 is pronounced as R-square)

- 1 In linear regression analysis, the r^2 is called the coefficient of determination. It is ratio of explained variation out of the total variation. As formula indicated below:

$$r^2 = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Example: Income's regression on education (use education to predict income)

Analysis of Variance
 Dependent Variable: INCOME
 N: 2557 Missing: 347
 Multiple R-Square = 0.145 Y-Intercept = 6.111
 LISTWISE deletion (1-tailed test) Significance Levels: **=.01, *=.05

Source	Sum of Squares	DF	Mean Square	F	Prob.
REGRESSION	9368.014	1	9368.014	432.334	0.000
RESIDUAL	55362.969	2555	21.668		
TOTAL	64730.980	2556			

	Unstand.b	Stand.Beta	Std.Err.b	t
EDUCATION	0.660	0.380	0.032	20.793 **

- 2 Interpretation: The R-square is .145, indicating that 14.5 percent of variance on dependent variable of income is explained by the independent variable Education. In other words, changes in education will bring increase in income. The regression coefficient is $b = .660$, indicating that one unit increase in education will be associated with .660 increase in income. The relationship is statistically significant ($F_{1, 2555} = 432.334$, $p < .001$).
- 3 R-square = .145. If un-squared, it becomes Pearson's r which is $r = .380$.

That is, $r \times r = R\text{-square}$ ($.38 \times .38 = .145$) or ($.145 = .38$).

- 4 Since R-square is the proportion explained by the independent variable, then $1 - r^2$ will be the proportion that is not explained by the independent variable. It is called the coefficient of alienation. That is to say, education only explains about 14.5% percent of variance on income. There are 85.5 percent of variance on income unexplained. Many independent variables are needed to identify.

Test of Significance (F)

- 1 This test of significance is to estimate the probability that a given correlation coefficient occurs only by chance if there is no population relationship. In other words, this F test is to test if there is a statistical relationship b/w variables in the population if they are correlated in a sample.
- 2 3 assumptions:
 1) a linear relationship in the population;
 2) random sample of the population;
 3) two variables are normally distributed.

- 3 F Ratio Formula

$$F = \frac{r^2 (N-2)}{1 - r^2}$$

Where, r = correlation coefficient
 N = number of cases

Example: $r^2 = -.56$, $N = 50$. ($df = 1$, and $N - 2 = 50 - 2 = 48$) Is this correlation significant?

$$F = \frac{r^2 (N-2)}{1 - r^2} = \frac{(-.56)^2 (50 - 2)}{1 - (-.56)^2} = \frac{(.314) (48)}{1 - .314} = \frac{15.07}{.686} = 21.968$$

Interpretation:

Appendix Table 3 shows that $F_{1, 48} = 21.968$ is larger than the table value of 12.61 with degree of freedom at 1 and 48. It indicates that there is less than one chance in a thousand that we would find a correlation this large if there is really no relationship. In other words, there are more than 999 chances in a thousand that we would find this correlation this large if there is a relationship. Therefore, we reject the null hypothesis and conclude that there is a relationship b/w urbanism and fertility.

Linear and Nonlinear Relationships

- 1 Regression line is a straight line because straight lines are easiest way to summarize relationship both conceptually and mathematically.
- 2 Not all relationships are linear, however. Linear regression and correlation are inadequate to handle such patterns as “U-curve,” upside down “U-curve,” or sleeping shape of “S-curve.” (See text page 246). They are nonlinear relationships.
- 3 Therefore, one is suggested to run scatterplot to examine the distribution of individual scores before doing regression analysis.
- 4 If relationship appears nonlinear, then linear regression and correlation technique should not be used. Other skills should be considered such as logarithm of scores that is beyond the scope of this text.

Dr. Ji
Soc331

Handout 11

CHAPTER 11

MULTIVARIATE CROSS-TABULATION

FOCUS

- A Conditions to meet a causal relationship
- A Distinguish causal and spurious relationships
- A The logic of control variables
- A Create and describe multivariate tables

The logic of causal relationships

Causality deals with a cause-and-effect relationship. A causal relationship involves 4 conditions that must be satisfied to meet causal relationship: association, time order, non-spurious, and rationale.

- 1 Association:

Certain values of the dependent variable must be found occurring with certain values of the independent variable more often than we expect by chance. For example, Cancer rates must be higher among smokers (smoking causes Ψ cancer), crime rates must be higher among the poorer areas (poverty causes Ψ crime), etc. Their associations are with more evidences.

$X \quad) \quad Y \quad \text{or} \quad Y \quad) \quad X$

If X causes Y, then X and Y must co-vary.

If X changes, Y also changes.

X must be associated with Y, or Y must be associated with X. The two variables are associated.

2 Time order/Temporal condition:

Time order deals with the sequence of occurrence of variables between the independent and the dependent. If X is a cause of Y, X must happen before Y, not the reverse.

$X \quad \Psi \quad Y$

X occurs prior to Y but not Y occurs prior to X.

Y will not occur without the prior occurrence of X.

X causes Y, or X is a cause of Y, or X is a cause and Y is an effect.

3 Non-spurious explanation:

If X is a cause of Y, then the relationship between X and Y can't be explained away by a third variable/factor.

That is,

If X causes Y, there should be no antecedent variables that occur prior to the independent variable, nor prior to the dependent variable

If X causes Y, then the relationship between X and Y can not be explained away by Z.

$X \quad \begin{matrix} Z \\ \beta \\ \Psi \end{matrix} \quad Y$

If the relationship between X and Y can be explained as the above, then this is an example of a spurious relationship.

4 Rationale:

The theoretical explanation that explains why X is a cause of Y. Rationale provides theoretical or conceptual sense of understanding. We need a theory to guide our establishment of the causal relationship.

Examples of Spurious Relationships

Example 1:

Hypothesis: Storks cause high/low Birth Rate?

Storks Ψ

Birth Rate

Table 11.1 Birth Rate by Number of Storks (%)

Ψ

Zero order

Birth rate	Number of Storks	
	Few	Many
High	44	62
Low	56	38
Total	100	100
N	100	100

$$\chi^2 = 6.50; p < .05; G = .35$$

Table 11.1 provides a ridiculous relationship between birth rate and number of birds/storks. But surprisingly enough, they are statistically significant! High birth rate is related to high number of birds. Low birth rate is related to low number of birds. Can we conclude that number of birds is causal to the birth rate in a region? It is SPURIOUS!

Hint: Do not just reply upon statistics. Thinking/theory is more important!!!

What are the causes of birth rate?

A Is there still a relationship by controlling for location (rural and urban)?

Table 11.2 Birth Rate by Number of Storks
Controlling for Location (%)

Birth rate	Location			
	Rural		Urban	
	Number of Storks		Number of Storks	
	Few	Many	Few	Many
High	80	80	20	20
Low	20	20	80	80
Total	100	100	100	100

N	(40)	(70)	(60)	(30)
$\chi^2 = 0.00$; $p < \text{n.s.}$ $G = 0$; $\chi^2 = 0.00$; $p < \text{n.s.}$ $G = 0$				

Table 11.2 shows that after controlling for location, the relationship between storks and birth is gone!!!

- 1 80% with few storks and 80% with many storks have high birth rates - No Difference there - among rural area the birth rate is NOT related to prevalence of storks;
- 2 20% with few storks have high birth rates and 20% with many storks have high birth rates - No Difference - No relationship - among urban area the birth rate is NOT related to prevalence of storks;

Questions? If no relationship between birth rate and number of storks controlling for location, is there a relationship between storks and location, and between birth rate and location? In other words, is location influencing both birth rates and number of storks?

Table 11.3 Number of Storks by Location (%)

Number of Storks	Location	
	Rural	Urban
Many	64	33
Few	36	67
Total	100	100
N	(110)	(90)
$\chi^2 = 18.18$; $p < .001$; $G = -.56$		

Table 11.4 Birth Rate by Location (%)

Birth Rate	Location	
	Rural	Urban

High	80	20
Low	20	80
Total	100	100
N	(110)	(90)

$$\chi^2 = 71.54; p < .001; G = -.88$$

Table 11.3 and Table 11.4 demonstrate that

- 1 Rural areas have more storks than urban areas
- 2 Rural areas have higher birth rates than urban areas
- 3 The relationships between number of storks and location, and birth rate and location, are strong by Chi square and Gamma.

Conclusion:

- 1 The relationship between storks and birth rates of the first example is SPURIOUS
- 2 Storks and birth rates have no relationship at all
- 3 Location is really associated with both birth rate and storks
- 4 Location affect both birth rate and storks

Storks ; Birth rate <spurious>

Location < Storks
: Birth Rate

Example 2

Hypothesis: The more firefighters fighting a fire, the greater the property loss.

Firefighters ; property loss

- 1) Do firefighters cause property damage? ; ridiculous?
- 2) The bigger the fire, the greater the damage? ; possible?

Size of fire ; property loss

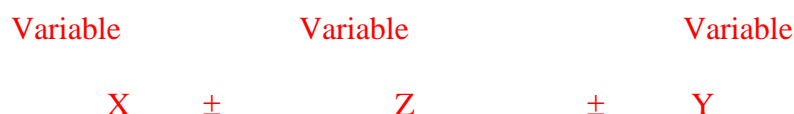
Size of Fire < More firefighters
: Greater property loss

Yes. The antecedent variable should be the SIZE OF FIRE.

Some Terminology

<u>Control Variable</u>	A variable that we hold constant while examining a bivariate relationship is called control variable. The <u>location</u> in Table 11.2 is a control variable.
<u>Spurious</u>	If a bivariate relationship largely disappears when we control for an antecedent variable, that bivariate relationship is spurious. The bivariate relationship between storks and birth rate in Table 11.1 is spurious.
<u>Explanation</u>	The description of the results of an analysis or the description of the conclusion from a research analysis.
<u>Zero-order</u>	Zero order means there is <u>no control variable involved in a bivariate relationship</u> . We use zero-order to distinguish bivariate relationships from those with control variables.
<u>First-order</u>	If having <u>only one control variable in a table</u> , we call the table as first order partial table. Table 11.2 is first order partial table.
<u>Second-order</u>	if <u>two control variables are introduced simultaneously into the table</u> , the table is called second-order partial table.
<u>Replication</u>	Replication is to re-do the same research analysis but with newly introducing control variables in order to recheck or verify the results. <u>If the resultant bivariate relationship reveals the same patterns or the similar patterns of the original relationship</u> , then this research is a replication.
<u>Suppressor Variables</u>	<p>When two independent variables are highly correlated with one another and each is strongly correlated with the dependent variable but in the opposite direction.</p> <p>In this case, each of the independent variables may suppress the effect of the other. As a result, two independent variables may appear NOT to be correlated with the dependent variable when in fact they are. This is called a suppressor variable.</p> <p>Suppressor variables can explain, weaken, or specify an original relationship.</p>
<u>Intervening Variables</u>	Intervening variable occurs after the independent variable but before the dependent variable in a causal chain. It causally links an independent variable to a dependent variable. The schematic illustration of the relationships is as follow:

Independent \pm Intervening \pm Dependent



Multivariate Analysis and Experimental Design

- 1 Researchers use randomization to assign subjects to experimental and control groups entirely on the basis of chance; whereas in multivariate analysis, researchers cannot assign subjects randomly with people (say use people for lab examination).
- 2 Experimental and control groups are entirely under control by the researchers whereas in multivariate they are not.
- 3 Experimental design provides true or genuine relationship especially the cause-effect whereas multivariate analysis can never be sure the results are truly causal.
- 4 While experimental design of research is best for causal relationship research, multivariate analysis can provide us with as many as possible the antecedent variables to be included in our data and do the best what we can.

Chapter 11: Examples of Spurious Relationships

Hypothesis: Storks cause high/low Birth Rate?

Storks Ψ Birth Rate

Table 11.1 Birth Rate by Number of Storks (%)

(Zero order)

Birth rate	Number of Storks	
	Few	Many
High	44	62
Low	56	38
Total	100	100
N	100	100

$$\chi^2 = 6.50; p < .05; G = .35$$

Table 11.1 provides a ridiculous relationship but hey are statistically significant!

A Is there still a relationship by controlling for location (rural and urban)?

Table 11.2 Birth Rate by Number of Storks
Controlling for Location (%) (First order partial table)

	Location			
	Rural		Urban	
	Number of Storks		Number of Storks	
	Few	Many	Few	Many
High	80	80	20	20
Low	20	20	80	80
Total	100	100	100	100
N	(40)	(70)	(60)	(30)

$$\chi^2 = 0.00; p < \text{n.s. } G = 0; \quad \chi^2 = 0.00; p < \text{n.s. } G = 0$$

Table 11.2 shows that after controlling for location, the relationship between storks and birth is gone!

Are there a relationship between storks and location and between birth rate and location?

Table 11.3 Number of Storks by Location (%)

Number of Storks	Location	
	Rural	Urban
Many	64	33
Few	36	67
Total	100	100
N	(110)	(90)

$$\chi^2 = 18.18; p < .001; G = -.56$$

Table 11.4 Birth Rate by Location (%)

Birth Rate	Location	
	Rural	Urban
High	80	20

Low	20	80
Total	100	100
N	(110)	(90)

$$\chi^2 = 71.54; p < .001; G = -.88$$

Table 11.3 and Table 11.4 demonstrate that

- 1 Rural areas have more storks than urban areas
- 2 Rural areas have higher birth rates than urban areas
- 3 The relationships between number of storks and location, and birth rate and location, are strong by Chi square and Gamma.

Conclusion:

- 1 The relationship between storks and birth rates of the first example is SPURIOUS
- 2 Storks and birth rates have no relationship at all
- 3 Location is really associated with both birth rate and storks
- 4 Location affect both birth rate and storks

Storks ; Birth rate <spurious>

Location
β
Storks ; Birth Rate

Dr. Ji
Soc331

HANDOUT 12

MULTIPLE REGRESSION AND CORRELATION

- A Extend regression analysis to two or more independent variables
- A Interpret multiple correlation coefficients
- A Interpret significance tests for multiple correlations

Simple Linear Regression

Simple linear regression deals with a relationship between two variables - a dependent and an independent, nothing more.

Equation

$$Y = a + bX$$

where,

Y = score on the dependent variable

a = Y-intercept, the value of Y when the line crosses the Y-axis and when $X = 0$; it is also called constant because it remains unchanged.

b = slope, the change in Y for every one-unit change in X. It is the steepness of the line.
 X = scores on the independent variable.

Multiple Regression

Multiple regression deals with a relationship between one dependent and two more independent variables.

Equation

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 \dots$$

where,

Y = score on the dependent variable

a = Y-intercept, the value of Y when the line crosses the Y-axis and when $X = 0$; it is also called constant because it remains unchanged.

A Note:

b_1 = slope, the change in Y for every one-unit change in independent variable X_1 .

X_1 = scores on the independent variable X_1 .

b_2 = slope, the change in Y for every one-unit change in independent variable X_2 .

X_2 = scores on the independent variable X_2 .

b_3 = slope, the change in Y for every one-unit change in independent variable X_3 .

X_3 = scores on the independent variable X_3 . (And so on).

Unstandardized regression coefficients

b_1 , b_2 , b_3 are called **unstandardized regression coefficients**, or simply **regression coefficients**, in distinguishing them from **standardized regression coefficients** or **Beta Coefficients denoted by β** . Unstandardized regression coefficients are **partial slopes or partial regression coefficients** because each of them only explains a part of the total variance on the dependent variable. They describe the change in the dependent variable Y associated with an increase of one unit in the independent variable X, controlling for the other independent variables.

Example:

Y' = fertility rate

$a = 5.59$

$b_1 = -.032$

$b_2 = -.010$

X_1 = percent urban

X_2 = radios per 100 persons

Equation $Y' = 5.59 - .032X_1 - 0.10X_2$

Interpretation

With radios (X_2) controlled, one percentage point increase in urbanism will decrease fertility rate by .032 child per woman.

With urbanism (X_1) controlled, one radio increase per 100 persons is associated with a reduction in fertility rate by .010 child per woman.

Prediction by Regression Equation

$$Y = a + b_1X_1 + b_2X_2 \dots$$

Suppose

Y' = fertility rate in Egypt

$$a = 5.59$$

$$b_1 = -.032$$

$$b_2 = -.010$$

X_1 = percent urban is 45%

X_2 = radios per 100 persons is 25

Then

$$Y = 5.59 - .032 (45) - .010 (25) = 5.59 - 1.44 - .25 = 3.90$$

Interpretation

Our knowledge of Egypt's urbanism (45%) and radio availability give us a prediction of Egypt's fertility level is 3.90, namely, on average a woman has almost 4 children.

Actually, Egypt has a fertility of 4.35. Our equation mis-predicted and less predicted about $4.35 - 3.90 = .45$ child per woman. This is no surprise because we only used available two independent variables - urbanism and radios. We may use more other independent variable such as education level, income, women's labor force participation, and so on to be included in the model. Then the prediction would be different.

Nevertheless, multiple linear regression provides us a very useful tool to work on social scientific research.

Assumptions

- 1 The independent variables are linearly related to the dependent variable. If non-linear relationship occurs between the two variables, the model may poorly describe the relationship.
- 2 The effects of the independent variables on the dependent variable are additive - with no statistical interaction between them. That is, no interaction effect or "extra effect" on the dependent variable because of the combination of the independent variables. For instance, although both urbanism and radio are related to fertility, there is no combination of urbanism and radio that has an "extra effect."
- 3 Independent variables in the model are not correlated with each other.
- 4 Interval / ratio levels of measurement.

Interpretation of Multiple Regression Results

Analysis of Variance

Dependent Variable: # CHILDREN

N: 2544 Missing: 360

Multiple **R-Square** = 0.055

LISTWISE deletion (1-tailed test)

Y-Intercept = 3.361

Significance Levels: **=.01, *=.05

Source	Sum of Squares	DF	Mean Square	F	Prob.
REGRESSION	382.359	2	191.180	73.757	0.000
RESIDUAL	6586.326	2541	2.592		
TOTAL	6968.685	2543			

	Unstand.b	Stand.Beta	Std.Err.b		t
EDUCATION	-0.144	-0.252	0.012	-12.107	**
INCOME	0.025	0.077	0.007	3.712	**

$$R^2 = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}} = \frac{382.350}{6968.685} = 0.055$$

$$F = \frac{\text{Mean Square}}{\text{Residual}} = \frac{191.180}{2.592} = 73.757$$

$p < 0.001$ — significant or not for the whole model, regardless partial regression coefficients significant or not.

Unstandardized coefficient (b) describes the change in the dependent variable associated with an increase of one unit in the independent variable, controlling for the other independent variables in the model.

t describes whether or not the relationship between the dependent variable and a particular independent variable is statistically significant or not, expressed by ($t = -12.107$, $p < .01$).

Beta Coefficient (β)

- 1 Unstandardized coefficients are of different units of measurement. For instance, radios are expressed as a radio per 100 people; urbanism as expressed as percentage of urban; education as expressed as years of schooling; income as expressed as dollars, etc. This makes it difficult to make comparisons.
- 2 Standardized regression coefficient (also called as beta or beta-weight) describes the effect of an independent variable on the dependent variable in standard

deviation units. Beta reports the standard deviation change in the dependent variable for a one standard deviation increase in the independent variable.

- 3 Beta coefficient can be greater than 1.00 since a one standard deviation change in an independent variable may produce more than one standard deviation change in the dependent variable.
- 4 Interpretation. β describes the change in the dependent variable associated with an increase of one unit in the independent variable in terms of standard deviations, controlling for the other independent variables. For example, with one standard deviation increase in education, there will be .252 standard deviation decrease in the # of children born to women.

Significance Test for Multiple Coefficients (F)

In multiple linear regression, we use F distribution for the significance test for R-square. F for R^2 is given by the ratio of the mean explained sum of squares to the mean unexplained sum of squares. It is also expressed as a ratio of regression sum of squares to the residual sum of squares.

F Formula for significant test for R^2

$$F = \frac{\text{mean explained sum of squares}}{\text{mean unexplained sum of squares}}, \quad \text{or} \quad = \frac{\text{mean square}}{\text{residual}}$$

Another way to calculate F

$$F = \left(\frac{R^2}{1 - R^2} \right) \left(\frac{N - k - 1}{k} \right)$$

Where,

N = number of cases

k = number of independent variables

R^2 = Value of the R-square

Dr. Ji
Soc331

HANDOUT 13

Dummy Variables and Applications

Dummy Variables

- 1 Multiple regression can include dichotomous independent variables such as Sex (male/female), Region (urban/rural), or other nominal variables such as Religion (Catholic, Jewish, Protestant, none, or other).
- 2 Dummy variables only have values of 0 and 1, with 0 indicating the absence of an attribute and 1 indicating its presence.
- 3 For dichotomous dummy variable, we usually treat modal category as “reference group,” or “reference value,” which will be coded as “0.”
- 4 Similarly, for dummy variables with five categories, the modal category is treated as reference value.
- 5 The value of the constant a or Y intercept, is the mean score for “reference group,” and the means for all other dummy variables are equal to “The intercept + that dummy variables’ unstandardized regression coefficient.” These means are the effects on the dependent variable from each dummy variable.
- 6 When reporting dummy variables’ regression coefficients, we compare the mean difference between the category coded as 1 and the category coded as 0 = “reference group.”

Example: Religion with 5 values:

- 1 Protestant
- 2 Catholic
- 3 Jewish
- 4 None
- 5 Other

Steps:

- 1 Pick up the modal category as reference group. As Protestants are the most, we treat Protestant as Modal Category or “Reference Group” or “Reference Value” and coded as 0.

2 Recode these dummy variables

Dummy Variables	Codes
R-Cath	1 if Cath 0 otherwise
R-Jewish	1 if Jewish 0 otherwise
R-None	1 if None 0 otherwise
R-Other	1 if Other 0 otherwise

3 Leave Reference Group as 0

4 Run regression

5 The Y intercept/constant is the mean of the reference group, which is equal to the effect on the dependent variable by the dummy independent variable. Here is the Protestant.

6 The effect of other dummy variables on the dependent variable are equal to: “The Y intercept + that dummy’s unstandardized regression coefficient.”

5 R^2 is the total proportion of variance on the dependent variable by all dummy independent variables.

Table 13.1 Regression of Years Education on Religion Dummy Variable

$R^2 = .02$ $F(3, 2887) = 17.309$ $p < .001$			
Variable	b	Beta	Mean
Constant	13.100		13.100
R-Cath	.410	.059	13.510
R-Jewish	2.268	.117	15.368
R-Other	1.361	.100	14.461
R-None	.419	.046	13.519

1 The constant 13.100 is the average for Protestant. It is its effect on the education.

2 Catholic’s effect on education is “constant + b” = $13.100 + .410 = 13.510$.

3 Similarly, the effects on education by R-Jewish, R-Other, and R-None are 15.368, 14.461, and 13.519 respectively.

4 Overall, religion only explains about 2% of the variance on the dependent variable.

5 The model is statistically significant ($F_{3, 2887} = 17.309, p < .001$).

Interpret and Calculate Multiple Linear Regression

Analysis of Variance
 Dependent Variable: # CHILDREN
 N: 1913 Missing: 991
 Multiple R-Square = 0.218
 Y-Intercept = 1.432
 LISTWISE deletion (1-tailed test)

Significance Levels: **=.01, *=.05

Source	Sum of Squares	DF	Mean Square	F	Prob.
REGRESSION	1066.382	4	266.596	133.159	0.000
RESIDUAL	3819.974	1908	2.002		
TOTAL	4886.356	1912			

	Unstand.b	Stand.Beta	Std.Err.b	t	
AGE	0.040	0.387	0.002	17.270	**
EDUCATION	-0.070	-0.120	0.013	-5.366	**
DAD EDUC.	-0.030	-0.076	0.011	-2.822	**
MOM EDUC.	-0.010	-0.020	0.013	-0.753	

Important statistics calculation:

R-Square = regression ss/total ss = 1066.382/4886.356 = 0.218
 F = mean squares/ residual = 266.596/2.002 = 133.159
 p < .001

a=1.432

b age = .040 B = .380 p < **
 b education = -.070 B = -.120 p < **
 b Dad edu = -.030 B = -.076 p < **
 b Mom edc = -.010 B = -.020 p < n.s.

***Interpretation of multiple regression coefficients.

Bivariate relationship between age and fertility:

- 1) The model explains about 21.8 % of the variance in the fertility by all the independent variables ($R^2 = .218$). The relationship between the dependent and all the independent variable are statistically significant except the variable "MOM education" (F = 133.159; $p < .001$).
- 2) The partial regression coefficients show that, with one unit increase in age, there will be " b= .040" increase in fertility. The bivariate relationship is positive.
- 3) Based on the standardized regression coefficient, with one standard deviation increase in age, there will be B= .380 standard deviation increase in fertility.

- 4) Beta provides an indication of the relative moderate strength of the effect of age on fertility.
- 5) The relationship between age and fertility is statistically significant ($t = 17.270$; $p < .001$).
- 6) Therefore, I conclude that the independent variable “age” has a statistically significant bivariate effect on the dependent variable “fertility.”

Bivariate relationship between education and fertility:

- 1) The partial regression coefficients show that, with one unit increase in education, there will be “ $b = -.070$ ” decrease in fertility. The bivariate relationship is negative.
- 2) Based on the standardized regression coefficient, with one standard deviation increase in education, there will be $B = -.120$ standard deviation decrease in fertility.
- 3) Beta provides an indication of the relative weak strength of the effect of education on fertility.
- 4) The relationship between education and fertility is statistically significant ($t = -5.366$; $p < .001$).
- 5) Therefore, I conclude that the independent variable “education” has a statistically significant bivariate effect on the dependent variable “fertility” although the effect is negative.