Entering and recoding variables

<u>To enter:</u>
You create a New data file
Define the variables on Variable View
Enter the values on Data View

<u>To create the dichotomies</u>:

Transform -> Recode into Different Variable [Name Output Variable]
-> Old and New Values
Then you specify the Old Value then the New Value
then -> Add
So if the 1st and 2nd groups (Old Value = 1,2 or Range 1 thru 2)
should be collapsed into one (the 1st group or New Value=1) you will see
in the right lower window 1 thru 2 --> 1. Etc

SOC 103M

Looking for Patterns

1	0	1	1	1	1	1	1	2	0	0	0	1
	1	0	0	1	0	0	0	0	0	0	1	0
	0	1	0	0	0	0	1	2	2	0	0	2
	2	0	0	0	2	3	0	0	0	0	1	0
	0	0	1	0	1	2	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	2	0	0	0	0	0	2	0	1	0	1
	0	0	0	0	0	0	1	0	0	0	0	0
	0	0	0	2	2	0	0	2	0	0	0	0
	1	3	0	2	1	1	2	1	1	1	1	1
	0	0	0	1	0	1	0	0	2	0	0	0
	0	2	0	0	0	0	0	1	1	2	0	0
	1	2	2	0	0	2	0	1	0	1	0	2
	1	2	0	0	1	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	2	0	0	0
	0	0	0	0	0	1	1	1	0	0	0	0
	2	0	1	0	0	0	1	2	0	0	0	0
	0	0	0	0	2	0	0	1	1	0	2	0
	0	0	1	0	2	0	1	0	1	0	0	0
	1	0	1	1	1	0	0	2	1	0	1	0
	0	0	0	2	3	2	1	1	1	1	0	0
	2	0	0	0	2	1	0	0	0	0	0	0
	0	0	0	2	1	1	1	0	1	0	0	3
	1	1	1	1	0	1	1	0	0	1	2	0
	0	1	2	1	0	1	0	2	1	2	0	0
	0	1	0	1	1	0	0	0	0	1	2	0
	2	0	1	0	1	0	2	0	0	2	0	0
	2	0	0	0	1	2	1	0	9	1	0	1
	2	1	0	1	1	1	0	0	0	0	0	0
	0	2	1	3	1	1	1	2	0	0	0	1
	0	2	2	1	0	3	1	1	1	1	1	1
	2	2	1	0	3	2	2	2	1	0	0	0
	2	2	0	1	0	0	3	3	0	0	0	0
	1	1	0	1	1	2	0	2	1	2	0	1
	0	0	2	1	1	0	2	0	2	0	0	0
	1	2	2	0	2	2	1	0	2	0	0	0
	1	0	0	1	0	0	0	0	0	0	1	0
	2	0	1	1	0	1	1	0	0	0	0	2
	1	0	1	2	0	0	1	0	1	0	1	1
	2	1	1	3	0	1	0	0	0	0	0	0
	0	0	2	1	2	2	0	2	0	0	2	0
	0	1	0	1	0	1	0	0	0	0	0	0
	0	2	0	1	0	2	1	1	0	0	1	1
	0	1	0	1	1	1	2	1	0	0	2	2

The first 12*44+1=529 cases.
There are another 2,338 cases I could not fit on this slide.

0= IAP
1= Very happy
2= Pretty happy
3=Not too happy

Frequencies
Relative frequencies (percentages)

SPSS:

FREQUENCIES VARIABLES=HAPMAR
/ORDER=ANALYSIS.

OR:

Analyze → Descriptive Statistics → Frequencies

Frequencies

Notes

Output Created		07-JAN-2019 16:21:51
Comments		
Input	Data	C:\Users\aronatas\Documents\My Documents\Class\Soc103M\GSS2016.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	2867
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on all cases with valid data.
Syntax		FREQUENCIES VARIABLES=HAPMAR /ORDER=ANALYSIS.
Resources	Processor Time	00:00:00.09
	Elapsed Time	00:00:00.10

Statistics

Happiness of marriage

N	Valid	1204
	Missing	1663

Happiness of marriage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	726	25.3	60.3	60.3
	PRETTY HAPPY	430	15.0	35.7	96.0
	NOT TOO HAPPY	48	1.7	4.0	100.0
	Total	1204	42.0	100.0	
Missing	IAP	1654	57.7		
	DK	1	.0		
	NA	8	.3		
	Total	1663	58.0		
Total		2867	100.0		

Describing ONE Variable

- What is the typical value?

- **Central Tendency Measures**

  Mode

  Median

  Mean

  How Typical is the typical value?

  Measures of Variation

  Range

  InterQuartile Range IQR

  Variance/Standard Deviation

Describing Relationships Between TWO Variables

- Tables
  - **Independent Variable Column/Dependent Variable Row**
- Percentage Difference
  - For dichotomies difference of two column percentages in the same row
- Cramer's V
  - For nominal variables $0 \leq V \leq 1$
- Gamma
  - For ordinal variables $-1 \leq \gamma \leq +1$

Central Tendency Measures

- What is the typical value?
-
- **Mode**
- most frequent value
- **Median**
  - 50^{th} percentile

- **Mean (Average)**
- $\Sigma X_i/N$

Examples

- Number of children people have:
- 0,0,0,0,0,1,1,2,2,2,3,3,4,5,7
- 0 5
- 1 2
- 2 3
- 3 2 ← Frequency Distribution
- 4 1
- 5 1
- 7 1
- N= 15

  Mode

  0

  Median

  2

  Mean
- 0+0+0+0+0+1+1+2+2+2+3+3+4+5+7=30
- 30/15=***2***

Which central tendency measure to use when?

	Mode	Median	Mean
Nominal	**Yes**	**No**	**No**
Ordinal	**Yes**	**Yes**	**No**
Interval and Ratio	**Yes**	**Yes**	**Yes**

Measures of Variability

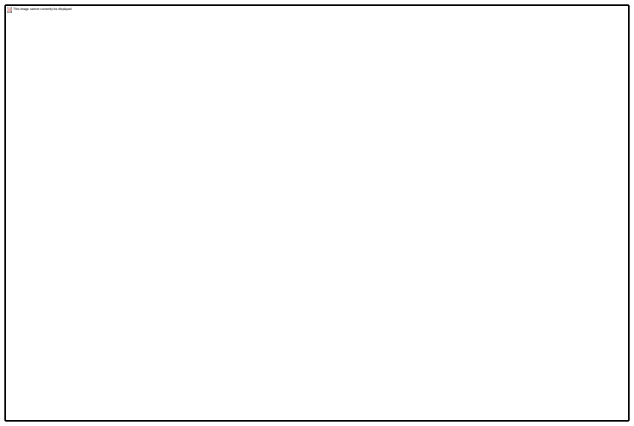
- How typical is the typical value?
-
- **Range**
  - Maximum-Minimum NOTE: There is NO plus 1 (+1) in the formula!
- **Interquartile Range**
  - Difference between the 25^{th} and 75^{th} percentile
-
- **Variance**
  - Average Squared Deviation from the Mean
- $\Sigma[Xi\text{-}Mean(Xi)]^2/N$
- Corrected variance

  $\Sigma[Xi\text{-}Mean(Xi)]^2/(N\text{-}1)$

Measures of Variability (cont.)

- **Standard Deviation**
  - **Square root of variance**

$$s = \sqrt{\frac{\sum (Xi - \sum Xi / N)^2}{N - 1}}$$



Example
N=15 Mean=2

# kids(X_i)	[X_i-Mean(X_i)]	[X_i-Mean(X_i)]^2
0	0-2=-2	$(-2)^2=4$
0	0-2=-2	$(-2)^2=4$
0	0-2=-2	$(-2)^2=4$
0	0-2=-2	$(-2)^2=4$
0	0-2=-2	$(-2)^2=4$
1	1-2=-1	$(-1)^2=1$
1	1-2=-1	$(-1)^2=1$
2	2-2=0	$(0)^2=0$
2	2-2=0	$(0)^2=0$
2	2-2=0	$(0)^2=0$
3	3-2=+1	$(1)^2=1$
3	3-2=+1	$(1)^2=1$
4	4-2=+2	$(2)^2=4$
5	5-2=+3	$(3)^2=9$
7	7-2=+5	$(5)^2=25$
Total		62

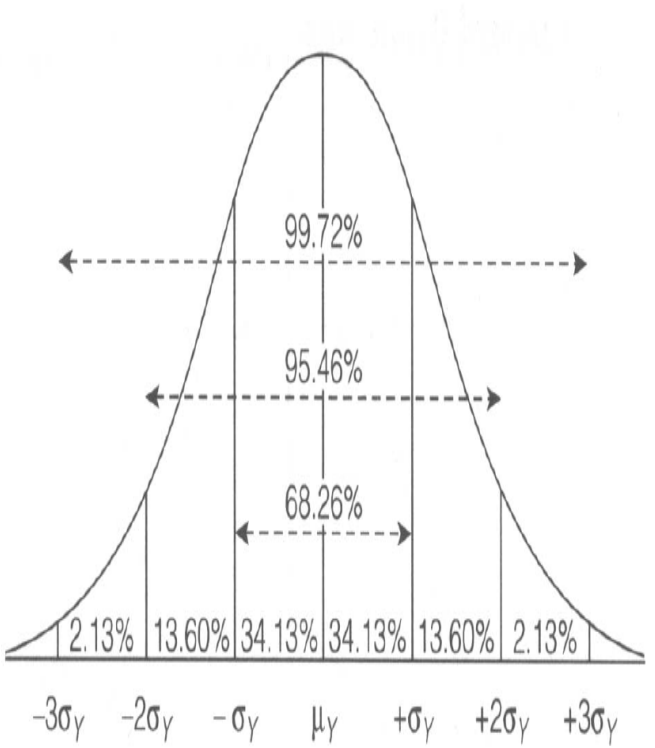
- Variance:
- 62/15=4.1333

- Corrected Variance
- 62/14=4.4286

Measures of Variability (cont.)

- **Standard Deviation σ**
  - **Square root of variance**
- Z-score or Standard Score
  - **Z=(Score-Mean)/Standard Deviation**

Tells you how many standard deviations away your score is from the mean.



Z for a childless family

Score=0 Mean=2 Standard Deviation=SQRT(4.1333)

Z(0)=(0-2)/2.033= -.984

Z for a family with 7 kids

Score=7 Mean=2 Standard Deviation=SQRT(4.1333)

Z(7)=(7-2)/2.033= +2.46

Which variability measure to use when?

	Range	Interquartile Range	Variance/ Stand.Dev.
Nominal	**No**	**No**	**No**
Ordinal	**Yes**	**Yes**	**No**
Interval/ Ratio	**Yes**	**Yes**	**Yes**

Two ways:

SPSS:

DESCRIPTIVES VARIABLES=HAPMAR
/STATISTICS=MEAN STDDEV VARIANCE RANGE MIN MAX.

OR:

Analyze → Descriptive Statistics → Descriptives
then Options and choose the statistics

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Happiness of marriage	1204	2	1	3	1.44	.571	.326
Valid N (listwise)	1204						

OR

SPSS:

FREQUENCIES VARIABLES=HAPMAR
/STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM MEAN MEDIAN MODE
/ORDER=ANALYSIS.

OR:

Analyze → Descriptive Statistics → Frequencies
then Statistics and choose the statistics

Statistics

Happiness of marriage

N	Valid	1204
	Missing	1663
Mean		1.44
Median		1.00
Mode		1
Std. Deviation		.571
Variance		.326
Range		2
Minimum		1
Maximum		3

Strictly speaking the Mean, Std. Deviation and Variance are inappropriate as the variable HAPMAR is ordinal

How much TV people watch?

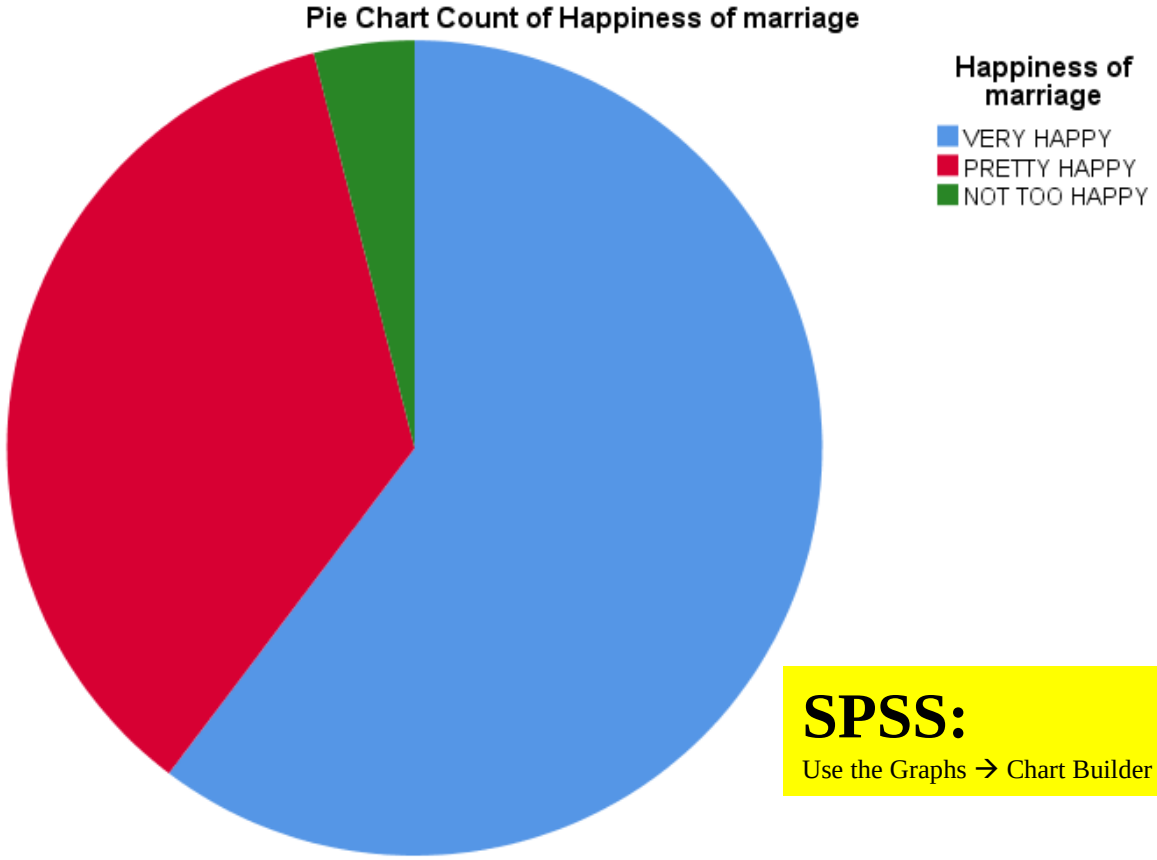
Statistics

Hours per day watching TV

N	Valid	1883
	Missing	984
Mean		3.03
Median		2.00
Mode		2
Std. Deviation		2.811
Variance		7.900
Range		24
Minimum		0
Maximum		24

TV Hour is ratio variable

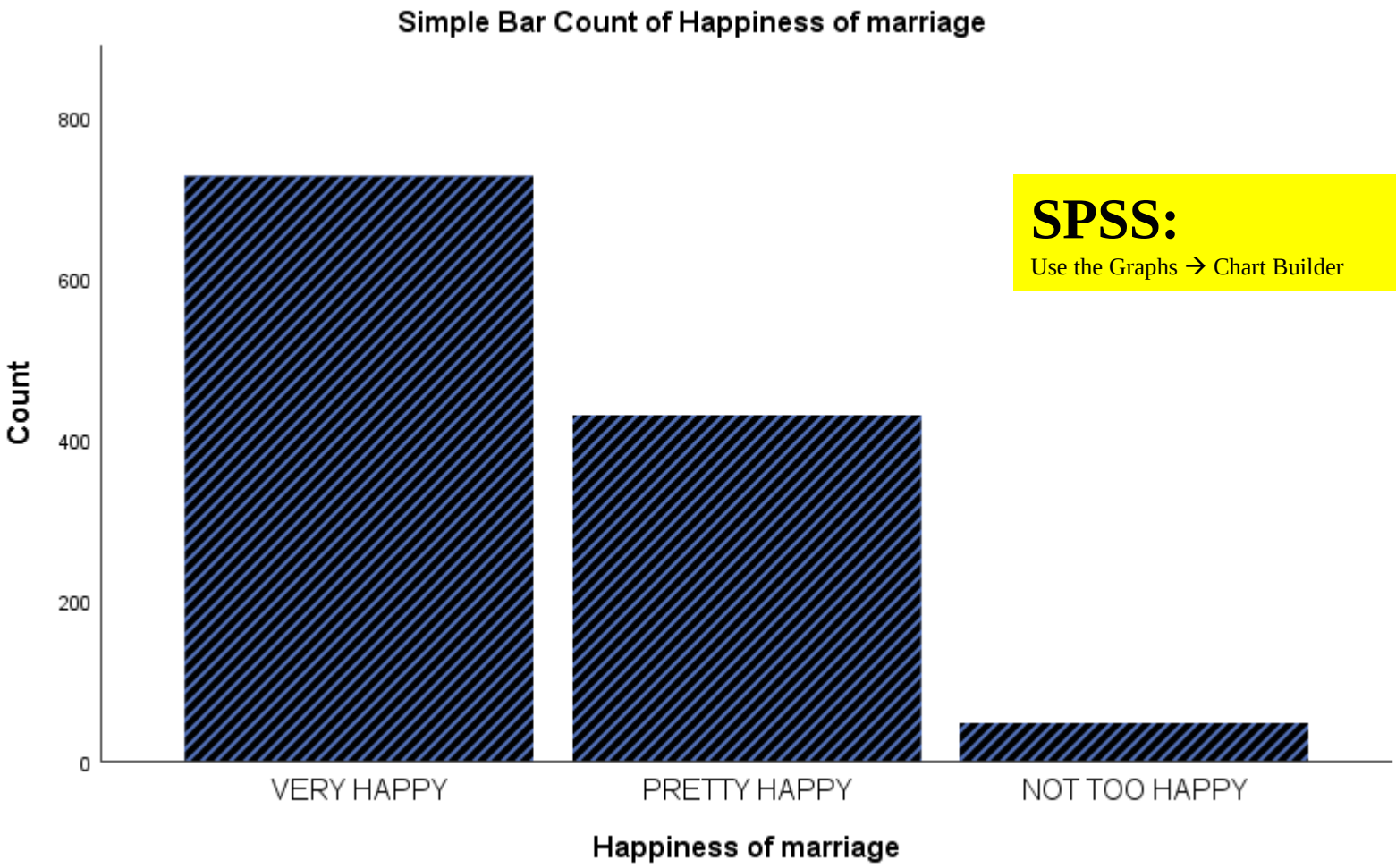
Visualization
Pie Chart (Mostly for nominal variables)



SPSS:

Use the Graphs → Chart Builder

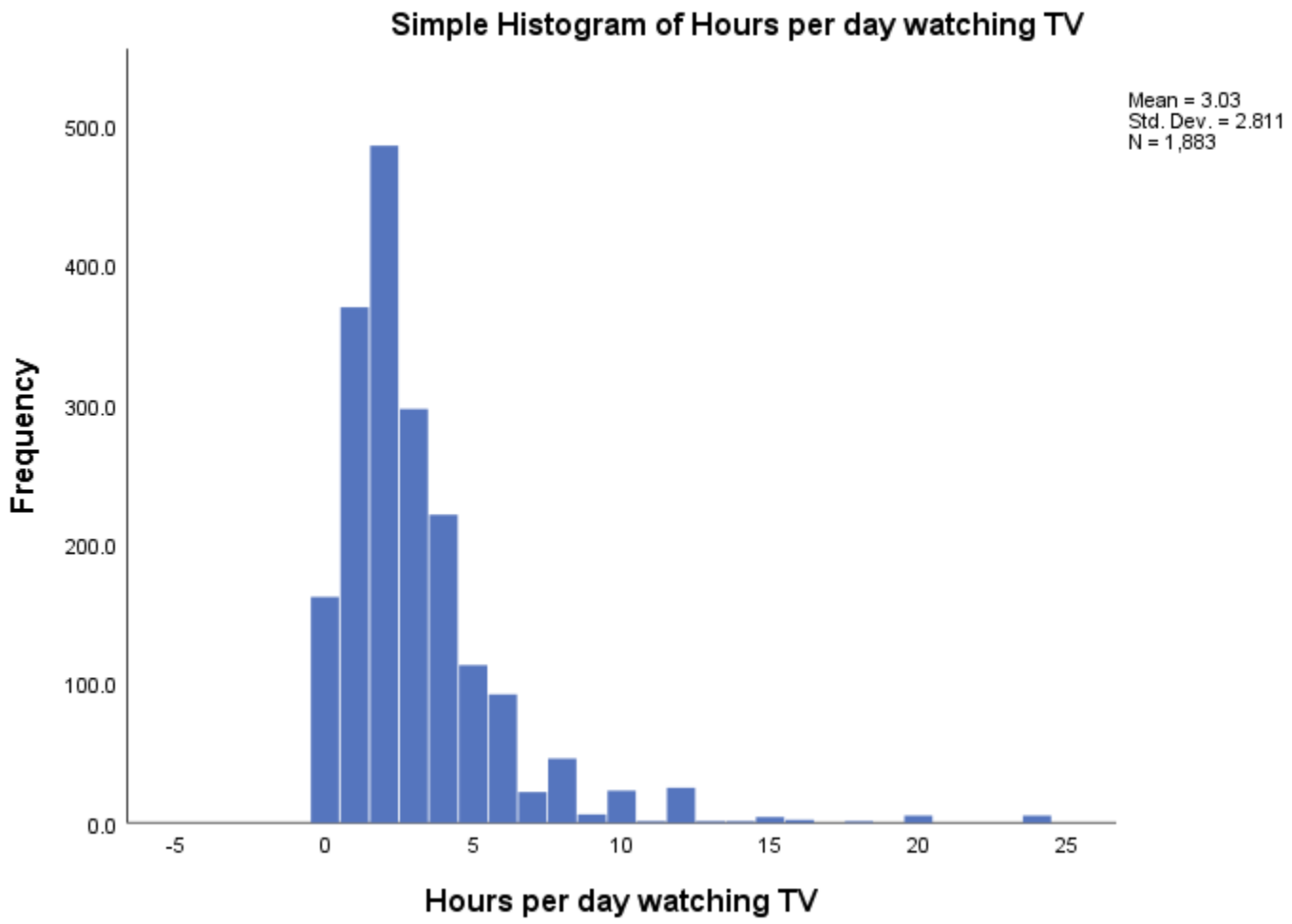
Visualization
Bar Chart



SPSS:

Use the Graphs → Chart Builder

Visualization
Histogram



SOC 103M

Quantitative Analysis

II.

Analyzing Relationships

John Stuart Mill's 3 Main Criteria of Causation

- ***Empirical Association***
- ***Appropriate Time Order***
- ***Non-Spuriousness (Excluding other Forms of Causation)***

Analyzing Relationships

	MEN	WOMEN	
DEMOCRAT	100	120	220
REPUBLICAN	100	80	180
TOTAL	200	200	400

- **Crosstabulating Variables**
  - Bivariate distributions
  - Marginal distributions
  - Row and column marginal
  - Grand total
-
-

Percentage Table

	MEN	WOMEN	TOTAL
DEMOCRAT	50%	60%	55%
REPUBLICAN	50%	40%	45%
TOTAL	100% N=200	100% N=200	100% N=400

Describing Associations

- Strength
  - Percentage difference
    - |50%-60%|=10%

Observed vs. Expected Tables

	MEN	WOMEN	TOTAL		MEN	WOMEN	TOTAL
DEM	**100** (50%)	**120** (60%)	220 (55%)		**110** (55%)	**110** (55%)	220
REPUBLICAN	**100** (50%)	**80** (40%)	180 (45%)		**90** (45%)	**90** (45%)	180
TOTAL	200	200	400		200	200	400

Chi -Square

- **$(100-110)^2/110+(120-110)^2/110+$**
- **$(100-90)^2/90 \quad +(80-90)^2/90=$**

- **100/110+100/110+100/90+100/90=.909+.909+1.111+1.111=**
- **4.04**

- **$SUM[Fo_{ij}-Fe_{ij}]^2/Fe_{ij}$=Chi-Square**

Cramer's V

- Cramer's V=SQRT[Chi-Square/(N*Min(c-1,r-1)]

- Cramer's V is between 0 (no relationship)
- and 1 (perfect relationship)

- V=SQRT[4.04/400*1]=.1005

Trust and TV Watching

CAN PEOPLE BE TRUSTED * Watching television Crosstabulation

			Watching television		
			0 to 2 hours per day	3 or more hours per day	Total
CAN PEOPLE BE TRUSTED	CAN TRUST	Count	95	72	167
		% within Watching television	38.3%	29.6%	34.0%
	CANNOT TRUST	Count	153	171	324
		% within Watching television	61.7%	70.4%	66.0%
Total		Count	248	243	491
		% within Watching television	100.0%	100.0%	100.0%

Strength and Statistical Significance of the Relationship

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.092	.042
	Cramer's V	.092	.042
N of Valid Cases		491	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Trust and Newspaper Reading

CAN PEOPLE BE TRUSTED * HOW OFTEN DOES R READ NEWSPAPER Crosstabulation

			HOW OFTEN DOES R READ NEWSPAPER		
			EVERYDAY	Less than everyday	Total
CAN PEOPLE BE TRUSTED	CAN TRUST	Count	90	77	167
		% within HOW OFTEN DOES R READ NEWSPAPER	40.9%	28.4%	34.0%
	CANNOT TRUST	Count	130	194	324
		% within HOW OFTEN DOES R READ NEWSPAPER	59.1%	71.6%	66.0%
Total		Count	220	271	491
		% within HOW OFTEN DOES R READ NEWSPAPER	100.0%	100.0%	100.0%

Strength and Statistical Significance of the Relationship

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.131	.004
	Cramer's V	.131	.004
N of Valid Cases		491	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Does Education Make You Happy?

GENERAL HAPPINESS * Degree recoded Crosstabulation

			Degree recoded			
			LT HIGH SCHOOL	HIGH SCHOOL	AT LEAST SOME COLLEGE	Total
GENERAL HAPPINESS	VERY HAPPY	Count	54	249	148	451
		% within Degree recoded	25.4%	30.2%	32.9%	30.3%
	PRETTY HAPPY	Count	121	482	258	861
		% within Degree recoded	56.8%	58.4%	57.3%	57.9%
	NOT TOO HAPPY	Count	38	94	44	176
		% within Degree recoded	17.8%	11.4%	9.8%	11.8%
Total		Count	213	825	450	1488
		% within Degree recoded	100.0%	100.0%	100.0%	100.0%

Strength and Statistical Significance of the Relationship

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.086			.026
	Cramer's V	.061			.026
Ordinal by Ordinal	Gamma	-.111	.042	-2.641	.008
N of Valid Cases		1488			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Gamma

- For two nominal variables: Cramer's V
- For two ordinal variables: Gamma

- Gamma is between
- –1(perfect negative relationship) and
- +1 (perfect positive relationship)
- 0 means no relationship

Does Money Buy Happiness?

GENERAL HAPPINESS * TOTAL FAMILY INCOME Crosstabulation

			TOTAL FAMILY INCOME			
			LOWER	MIDDLE	HIGHER	Total
GENERAL HAPPINESS	VERY HAPPY	Count	108	135	145	388
		% within TOTAL FAMILY INCOME	23.1%	31.2%	36.3%	29.8%
	PRETTY HAPPY	Count	273	255	228	756
		% within TOTAL FAMILY INCOME	58.5%	58.9%	57.0%	58.2%
	NOT TOO HAPPY	Count	86	43	27	156
		% within TOTAL FAMILY INCOME	18.4%	9.9%	6.8%	12.0%
Total		Count	467	433	400	1300
		% within TOTAL FAMILY INCOME	100.0%	100.0%	100.0%	100.0%

Strength and Statistical Significance of the Relationship

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.175			.000
	Cramer's V	.124			.000
Ordinal by Ordinal	Gamma	-.238	.040	-5.864	.000
N of Valid Cases		1300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Evaluating Relationships

- Existence
- Strength,
- Direction
- Pattern
- **Statistical Significance**:
  - Can we generalize from our sample to the population?
  - The values show the probability of making a mistake if we did.

  More precisely: The probability of getting a relationship this strong or stronger from a population where that relationship does not exist, just by sampling error.