Last week:

Examining associations.

- Is the association significant? (chi square test)
- Strength of the association (with at least one variable nominal)
- maximum difference approach
- chi/cramer's v/lambda
- Nature of the relationship

    -> column percentages

This week:

Association Between Variables Measured at the Ordinal Level

In this presentation you will learn about:

- Gamma as PRE Measure Measures of Association for Ordinal-Level Variables
- Determining the Direction of Relationships
- Limitations of Gamma
- Testing Gamma for Statistical Significance

- Gamma is measure of association for two ordinal-level variables that have been arrayed in a bivariate table.

- Recall: Ordinal
- Can rank order cases but "without precision"
- Example -> attitudinal or crude measurement ; likert scales

- **Happiness**
- 1. very unhappy
- 2. unhappy
- 3. happy
- 4. very happy

Level of education

1. Low
2. Medium
3. high

Level of satisfaction

1. Very dissatisfied
2. Dissatisfied
3. Satisfied
4. Very satisfied

  - Gamma measures both strength and direction of relationship
  - Note: can't measure direction with "nominal variables", so if one of your variables is nominal, don't use Gamma -> use measures reviewed last week
  - Gamma is a symmetrical; that is, the value of gamma will be the same regardless of which variable is taken as independent.

Gamma can answer the questions (beginning with a bivariate table:

1. Is there an association? (note: significance test is available)
2. How strong is the association?
3. What direction (because level is ordinal) is it?

Gamma's significance test involves a corresponding sampling distribution of "gammas"
If N>100, this sampling distribution is "normal"
A Z test is possible to see if the association (relationship) between two ordinal level variables is significant

In this case, you would use the 5 step method similar to previous "tests of significance" reviewed in previous chapters

- Gamma is a PRE (Proportional Reduction in Error) measure of association
- In other words, it tells us how much our error in predicting ***y*** is reduced when we take ***x*** into account.

- This statistic is based on the logic of the “order of pairs of cases.”
- i.e. it involves predicting the order of *pairs of cases* (predict whether one case will have a higher or lower score than another) on a given variable..

- To compute Gamma, two quantities must be found:
- N_s # of pairs with same ranking
- N_d # of pairs with different ranking

- N_s is the total number of pairs of cases ranked in the same order on both variables.
  - For example, Dick and Jane are among 50 respondents to a survey investigating the relationship between education (independent variable) and income (the dependent variable).

	Education	**Income**
Dick	(High)	(High)
Jane	(Low)	(Low)

  - For this "pair" of cases, Dick reports a *higher* level of education than Jane **and** Dick also reports a *higher* level of income than Jane
  - Thus this "pair" of cases is said to be **similar** (**same**).

Gamma *(continued)*

- N_d is the total number of pairs of cases ranked in different order on the variables.
  - For example, Peter and Susan are also among the 50 respondents.

	Education	**Income**
Peter	(High)	(Low)
Susan	(Low)	(High)

  - For this "pair" of cases: Peter reports a *higher* level of education than Susan **but** Peter has a *lower* level of income than Susan
  - This "pair" of cases is said to be **dissimilar** (**different**).

Thinking about it: Which of these two scenarios would be more likely given what we know about the relationship between income and education? (i.e. a + relationship)

	Education	Income
Dick	(High)	(High)
Jane	(Low)	(Low)

an example of N_s pairs with same ranking

	Education	Income
Peter	(High)	(Low)
Susan	(Low)	(High)

an example of N_d pairs With dissimilar ranking

The first is more likely given what we know about education and income,.. If Dick has a higher education than Jane, we would predict that he also have the higher income??

The second is less likely… Susan with greater income despite less education

If in a sample: Ns predominates, we would expect a positive relationship

If Nd predominate, we would expect a negative relationship

If neither Nd or Ns predominate, we would expect neither: no relationship

Gamma is calculated by finding the ratio of cases that are ranked the same on both variables minus the cases that are not ranked the same (N_s – N_d) to the total number of cases (N_s + N_d).

- Formula for Gamma (Formula 12.1):

$$G = \frac{N_s - N_d}{N_s + N_d}$$

This ratio can vary from:

+1.00 for a perfect positive relationship to

-1.00 for a perfect negative relationship.

Gamma = 0.00 means no association between two variables.

Note that when N_s is greater than N_d, the ratio with be positive, and when N_s is less than N_d the ratio will be negative.

Gamma *(continued)*

- If N = 50, the **overall** number of pairs of cases will be 1,225.
  - We can calculate the overall number of pairs of cases with this simple formula:
    - $(N * (N-1))/2 = (50*49)/2 = 1225$
  - The pairs “Dick and Jane” and “Peter and Susan” are just **2** out of **1225** possible pairs of cases when N=50.

- Gamma uses only the total number of **similar** pairs, N_s, and total number of **dissimilar** pairs, N_d, and ignores all **tied** pairs of cases.

  - Examples of tied pairs:

	Education	Income
Joseph	(High)	(High)
Steve	(High)	(High)

	Education	Income
Joseph	(High)	(High)
Steve	(High)	(Low)

	Education	Income
Joseph	(High)	(High)
Steve	(low)	(High)

- Gamma ignores all types of tied pairs:
- Pairs **tied** on both the independent and dependent variable;
- Pairs **tied** on the independent variable;
- and pairs **tied** on the dependent variable.

Gamma: An Example

- Let's now consider the survey on education and income for *all* 50 respondents.

- However, rather than looking at **each pair individually** to determine if it is similar or dissimilar (as we did above for the pairs Peter and Susan), we can use the bivariate table

- We can conveniently compute the **total number** of pairs of cases ranked in the same order on both variables (N_s) and the **total number** of pairs of cases ranked in different order on both variables (N_d).

Example

- To compute **Ns**, start with the **Low-Low** cell (upper left) and multiply the cell frequency by the cell frequency **below and to the right**.

	Education		
Income	Low	High	Totals
Low	15	10	25
High	5	20	25
Totals	20	30	50

- For this 2x2 table: $\boldsymbol{N_s}$**:15 x 20 = 300**
- There are 300 pairs whereby one case scores low/low and the other scores higher on both variables

Example *(continued)*

- For N_d, start with the **High-Low** cell (upper right) and multiply each cell frequency by the cell frequency **below and to the left**.

	Education		
Income	Low	High	Totals
Low	15	10	25
High	5	20	25
Totals	20	30	50

N_d: 5 x 10 = 50

There are 50 pairs here that scored Low/High on one case and high/low on the other (i.e. the opposite ranking)

Example *(continued)*

- Gamma is computed with Formula 12.1:

$$G = \frac{N_s - N_d}{N_s + N_d}$$

Since there is a substantial *preponderance* of similar pairs (300) relative to dissimilar pairs (50), we know the value of Gamma will large (closer to 1) and positive.

Using Formula 12.1:

$$\boldsymbol{G = (300-50)/(300+50) = +250/350 = +.71}$$

Example *(continued)*

- Table 12.2 provides a guide to interpret the strength of gamma.
  - As before, the relationship between the values and the descriptive terms is arbitrary, so the scale in Table 12.2 is intended as a general guideline only:

TABLE 12.2 THE RELATIONSHIP BETWEEN THE VALUE OF GAMMA AND THE STRENGTH OF THE RELATIONSHIP

Value	Strength
If the value is	*The strength of the relationship is*
between 0.00 and 0.30	weak
between 0.31 and 0.60	moderate
greater than .60	strong

Example *(continued)*

- The computed value of gamma of +.71 suggests:
- this relationship appears to be strong and positive: as education increases, income increases.

- PRE interpretation:
- predicting the order of pairs of cases on the dependent variable (income)
- we would make 71% fewer errors by taking the independent variable (education) into account.

- NOTE: we haven't determined whether this is "significant or not".. Our total sample N=50.., so it might not be!!

Prior to addressing the issue of “STATISTICAL SIGNIFICANCE”..

ANOTHER EXAMPLE!!

Another example: examining the nature and strength of the association Between “level of education” and “volunteerism”..

Volunteerism by Education

	Education				
Volunteerism	**Less than HS**	**HS**	**Some PS**	**University Grad**	**Total**
Low	1719	1330	2833	1010	6892
Moderate	852	958	1416	1212	4438
High	799	1022	3144	2310	7275
Totals	3370	3310	7393	4532	18605

To compute Gamma, two quantities must be found: N_s and N_d

Ns = 1719 (958+1416+1212+1022+3144+2310) + 852 (1022+3144+2310) + 1330 (1416+1212+3144+2310) + 958 (3144+2310) + 2833 (1212+2310) + 1416 (2310) = 52,036,908

Nd = 1010 (852+958+1416+799+1022+3144) + 1212 (799+1022+3144) + 2833 (852+958+799+1022) + 1416 (799+1022) + 1330 (852+799) + 958 (799) = 30,116,921

$$G = \frac{N_s - N_d}{N_s + N_d}$$

To compute Gamma, two quantities must be found: N_s and N_d

Ns = 52,036,908; Nd=30,116,921

$$G = \frac{N_s - N_d}{N_s + N_d}$$

G = (52,036,908 - 30,116,921)/(52,036,908+30,116,921) = .27

Using Table 12.2, we conclude that we have a relatively “weak” positive association between the two variables.. By positive, we mean “as level of education increases, so too does “level of volunteerism”..

The computed value of gamma of +.27 means that, when predicting the order of pairs of cases on the dependent variable (volunteerism), we would make 27% fewer errors by taking the independent variable (level of education) into account. as education increases, income increases.

Limitations of Gamma

- When variables are not coded from low to high (e.g., **high** education=**1**; **low** education =**2**), we must exercise caution in using the sign (+ or -) of gamma to determine actual direction of the relationship

- Gamma ignores all tied pairs of cases, which can potentially "distort" the real strength of association.

- Alternatively, other ordinal measures that take ties into account, such as Somer's *d* and Kendall's *tau-b,* may be used instead of gamma (not covered in this class)

- Also -> examine column % of bivariate table., are the results consistent???

Volunteerism by Education

	Education				
Volunteerism	**Less than HS**	**HS**	**Some PS**	**University Grad**	**Total**
Low	1719	1330	2833	1010	6892
Moderate	852	958	1416	1212	4438
High	799	1022	3144	2310	7275
Totals	3370	3310	7393	4532	18605

↓

Volunteerism by Education

	Education								
Volunteerism	**Less than HS**		**HS**		**Some PS**		**University Grad**		**Total**
Low	1719	51.01%	1330	40.18%	2833	38.32%	1010	22.29%	6892
Moderate	852	25.28%	958	28.94%	1416	19.15%	1212	26.74%	4438
High	799	23.71%	1022	30.88%	3144	42.53%	2310	50.97%	7275
Totals	3370	100.00%	3310	100.00%	7393	100.00%	4532	100.00%	18605

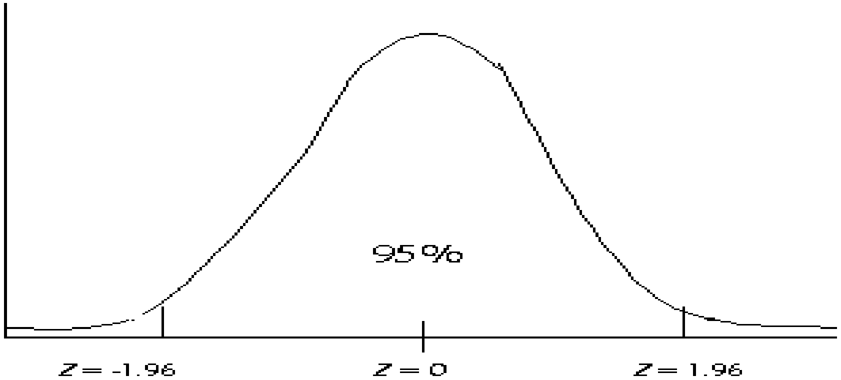
Testing Statistical Significance of Gamma

- In testing gamma for statistical significance, the null hypothesis states that there is no association between the variables in the population.
- To test the significance of gamma, the familiar five-step model should be used to organize the hypothesis testing procedures.
- *Z* is used to test of the significance of gamma
- Section 12.7 provides details on testing the null hypothesis of "no association" with gamma.

Testing Gamma for Significance

- The test for significance of Gamma is a hypothesis test, and the 5 step model should be used.

- Step 1: Assumptions
  - Random sample, ordinal,
  - Sampling Dist. is normal



What is the sampling distribution here?

Assume that in the population there is no relationship between two variables (independence).

If we repeatedly sampled the population with samples of size N, and repeatedly calculated gamma, the gammas would take on a normal distribution with a mean of zero

- Step 2: Null and Alternate hypotheses

H_o: $\gamma=0$,

H_1: $\gamma \neq 0$, where γ is the population value of G

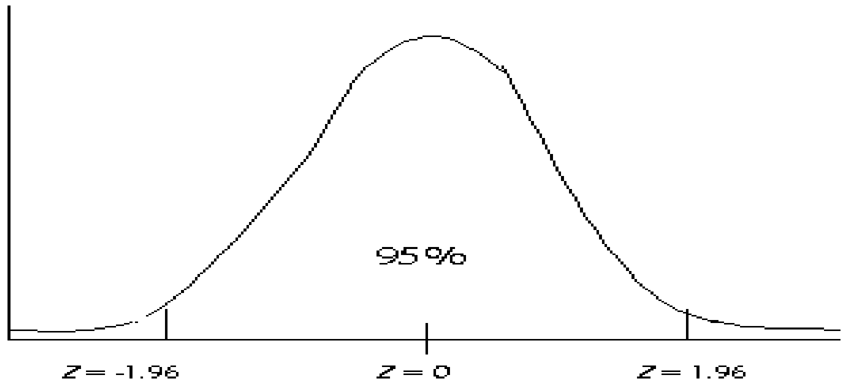
In other words:

our null hypothesis is that there is no association between the variables in our population

our research hypothesis is that gamma is significantly different from zero in our population..

Step 3: Sampling Distribution and Critical Region

Z-distribution, $\alpha = .05$, $z = +/-1.96$



Assuming our gamma in the population is zero (our null hypothesis), and knowing that our sampling distribution is normal,
we would expect only a 5% chance of obtaining a gamma from a sample that is more than 1.96 standard deviations (standard errors) above the mean or 1.96 standard deviations (standard errors) below.

After converting “standard errors” into Z scores:
If more than or less than 1.96 Z scores away from the mean, we reject our null hypothesis and accept our research hypothesis

Testing Gamma for Significance (cont.)

- Part 4: Calculating Test Statistic:

- Formula :

$$z(obtained) = G\sqrt{\frac{N_s + N_d}{N\left(1 - G^2\right)}}$$

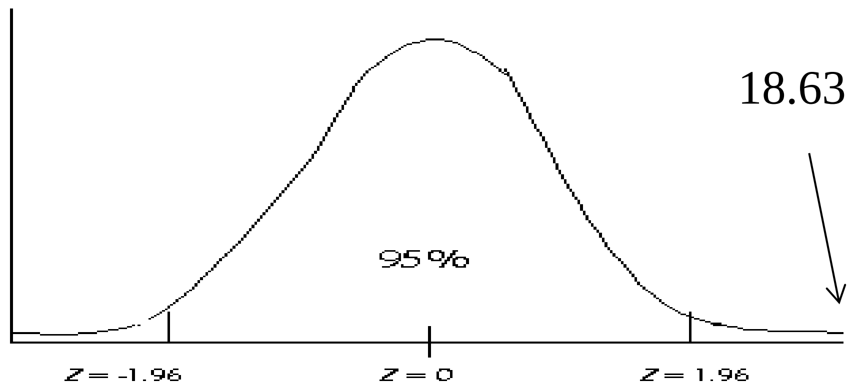
In previous example (education and volunteerism)

G = .27
Ns = 52,036,908
Nd = 30,116,921
N = 18,605 (note: VERY LARGE SAMPLE< SO LIKELY SIGNIFICANT)

- Calculate:

$$z = .27\sqrt{\frac{52{,}036{,}908 + 30{,}116{,}921}{18{,}605\left(1 - .27^2\right)}} = .27(69.01) = 18.63$$

- Step 5: Make Decision and Interpret



- Z_{obt}=18.63 falls in our critical region…
- Reject H_o
- The association between two variables is clearly significant.

- In other words, if there was no relationship in the population, it is extremely unlikely that we would come up with a gamma that is fully 18 standard errors away from the population gamma of 0, so we reject the null hypothesis

- Let’s try one more example
- A sample of children has been observed and rated for symptoms of depression (few, some, many). Their parents have been rated for authoritarianism (low, moderate, high).
- What’s the level of measurement involved?
- Ordinal in both cases
- What the likely dependent variable?
- Childhood depression
- What is the nature of the relationship between the two variables given the following data?

Gamma:
Is there an association?
How strong is the association?
What direction?

Symptoms of Depression

	Authoritarianism			
	Low	Moderate	High	Totals
Few	7	8	9	24
Some	15	10	18	43
Many	8	12	3	23
Totals	30	30	30	90

Example: examining the nature and strength of the association Between “level of education” and “volunteerism”..

Symptoms of Depression

	Authoritarianism			
	Low	**Moderate**	**High**	**Totals**
Few	**7**	**8**	**9**	**24**
Some	**15**	**10**	**18**	**43**
Many	**8**	**12**	**3**	**23**
Totals	**30**	**30**	**30**	**90**

To compute Gamma, two quantities must be found: N_s and N_d

Ns = 7(10+12+18+3)+8(18+3)+15(13+3)+10(3)= <u>724</u>

Nd = 9(15+10+8+12)+8(15+8)+18(8+12)+10(8)= <u>1029</u>

$$G = \frac{N_s - N_d}{N_s + N_d}$$

To compute Gamma, two quantities must be found: N_s and N_d

$Ns = 724; \quad Nd=1029$

$$G = \frac{N_s - N_d}{N_s + N_d}$$

Consult Table 12.2:

$$= \frac{724 - 1029}{724 + 1029} = \frac{-304}{1753} = -0.17$$

TABLE 14.2 THE RELATIONSHIP BETWEEN THE VALUE OF GAMMA AND THE STRENGTH OF THE RELATIONSHIP

Value	Strength
If the value is	*The strength of the relationship is*
between 0.00 and 0.30	weak
between 0.31 and 0.60	moderate
greater than .60	strong

We have evidence of a relatively "weak" negative association between the two variables.. By negative, we mean "as authoritarianism increases, "symptoms of depression go down"..

BUT??? POTENTIAL PROBLEM HERE!!!!

Recall: Gamma
Is there an association?
How strong is the association?
What direction?

We have evidence of a weak negative association, but is it significant?

Step 1: Assumptions

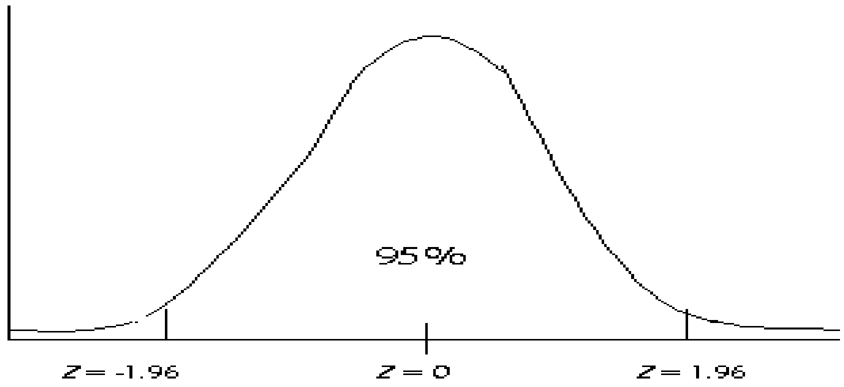
Random sample, ordinal,
Sampling Dist. is normal

Step 2: Null and Alternate hypotheses

H_o: $\gamma=0$,
H_1: $\gamma\neq0$, where γ is the population value of G

Step 3: Sampling Distribution and Critical Region

Z-distribution, $\alpha = .05$, $z = +/-1.96$



Assuming our gamma in the population is zero (our null hypothesis), and knowing that our sampling distribution is normal,
we would expect only a 5% chance of obtaining a gamma from a sample that is more than 1.96 standard deviations (standard errors) above the mean or 1.96 standard deviations (standard errors) below.

Testing Gamma for Significance (cont.)

- Part 4: Calculating Test Statistic:

- Formula :

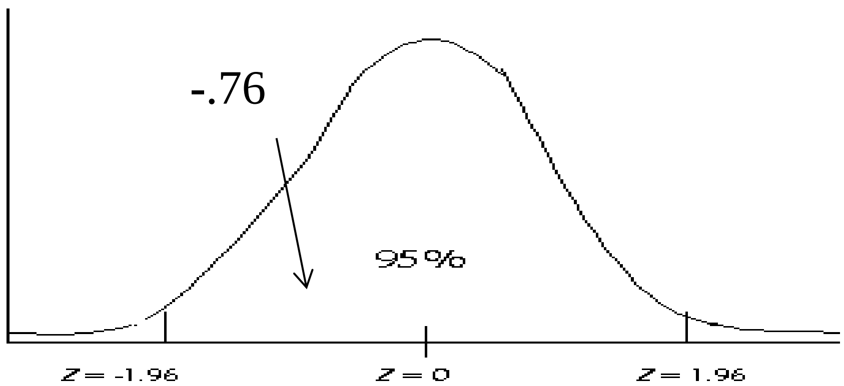
$$z = G\sqrt{\frac{n_s + n_d}{N\left(1 - G^2\right)}}$$

G = =-.17
Ns = 724
Nd = 1029
N = 90 (note: SMALL SAMPLE< SO POSSIBLY NOT SIGNIFICANT)

- Calculate:

$$z = -.17\sqrt{\frac{724 + 1029}{90\left(1 - (-.17^2)\right)}} = -.17\sqrt{\frac{1753}{87.399}} = -.76$$

- Step 5: Make Decision and Interpret



- Z_{obt}=-.76 does not fall in our critical region…
- Can not reject H_o
- The association between two variables is not significant.

- We have no real way of knowing whether an association exists, because our sample size is too small…