# It is not yet time to panic?


**After this week's class -> START ASSIGNMENT 5!!**
**SHOULD BE ABLE TO COMPLETE QUESTIONS 1-4, 5a of Assignment 5**
***Part 5b – 5e (after next weeks lecture)***

***I highly recommend that you get started on the final "problem solving assignment" TODAY (and avoid a last minute panic)***

***Assignment 5 due Friday, after the last class: DECEMBER 6th, my office LH208 5:00 p.m. sharp***

***FINAL EXAM IS SCHEDULED FOR:***
***SATURDAY DECEMBER 14th, 2:00 p.m. – 5:00 p.m. ROOM SA 150***

**YOU CAN PICK UP MARKED ASSIGNMENT 5 (marked) on Wednesday, DECEMBER 11th 4:00-5:00 p.m.**

Last week:

Examining associations.
- Is the association significant? (chi square test)
- Strength of the association (with at least one variable **nominal**)
- maximum difference approach
- phi / Cramer's v/ lambda ( 3 alternative measures of association)
- Nature of the relationship -> column percentages

14-2

# This week you will learn about:

- Gamma as PRE Measure Measures of Association for **Ordinal**-Level Variables
- Determining the Direction of Relationships
  Introduced last week:
  "ordinal/interval ratio" variables .. NOT NOMINAL
- Limitations of Gamma
- Testing Gamma for Statistical Significance
- INTRODUCTION TO ASSOCIATION WITH "**INTERVAL/RATIO** VARIABLES!!

14-3

- 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.

14-4

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**).

14-7

# 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:

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

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.

14-10

# 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 14.1:

FORMULA 14.1

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

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 14.1:

**$G$ = (300-50)/(300+50) = +250/350 = +.71**

## Example *(continued)*

- THE FOLLOWING Table 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 the text is intended as a general guideline only:

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 0.0000 and 0.0999 | weak |
| Between 0.1000 and 0.2999 | moderate |
| Greater than 0.3000 | 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

**FORMULA 14.1**

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

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

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

**FORMULA 14.1**

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

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

Using  our Table, 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

| Volunteerism | Education: 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

| Volunteerism | Education: 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 |

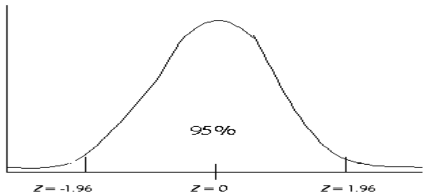
# Testing Statistical Significance of Gamma

- 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

## 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

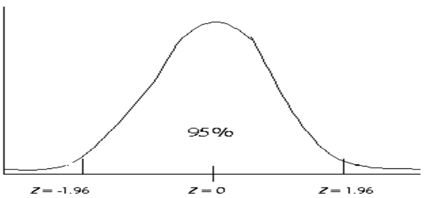
- Step 2: Null and Alternate hypotheses

$H_o$: γ=0,
$H_1$: γ≠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

14-27

## Testing Gamma for Significance (cont.)

- Part 4: Calculating Test Statistic:

- Formula :

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

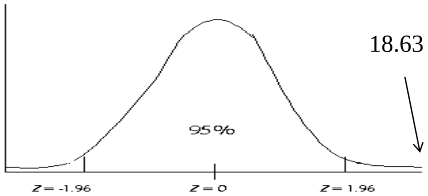
In previous example (education and volunteerism)

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(12+3)+10(3)= 724

Nd = 9(15+10+8+12)+8(15+8)+18(8+12)+10(8)= 1029

**FORMULA 14.1**

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

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

*Ns = 724; Nd=1029*

**FORMULA 14.1**

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

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

Consult Table

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 0.0000 and 0.0999 | weak |
| Between 0.1000 and 0.2999 | moderate |
| Greater than 0.3000 | strong |

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

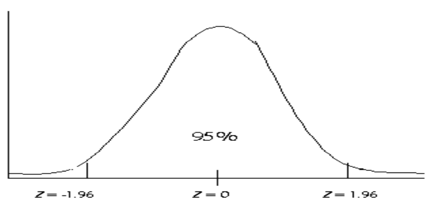
Step 1: Assumptions
Random sample, ordinal,
Sampling Dist. is normal

Step 2: Null and Alternate hypotheses

$H_o$: γ=0,
$H_1$: γ≠0, where γ is the population value of G

14-33

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.

14-34

# Testing Gamma for Significance (cont.)

- Part 4: Calculating Test Statistic:

- Formula :

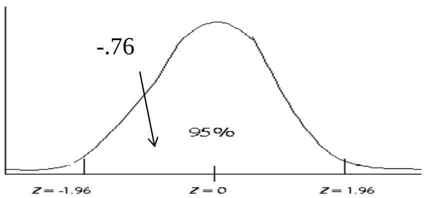
$$z = G\sqrt{\frac{n_s + n_d}{N\left(1 - G^2\right)}}$$

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…

## Association Between Variables Measured at the Interval-Ratio Level: Bivariate Correlation and Regression

Last couple of classes:

Measures of Association:
Phi, Cramer's V and Lambda (nominal level of measurement)
Gamma (ordinal level of measurement)

**Today: what if: interval/ratio level of measurement**

## Introduction:

- Interval/ratio level of measurement
- Scores are actual numbers and have a true zero point and equal intervals between scores

- E.g. Age (in years);
- Income (in dollars);
- Education (in years)
- Weight (in pounds)
- Hours worked (hours)
- etc.

## Introduction:

- Interval/ratio level of measurement

| Age | Family Size |
|---|---|
| 5 | 6 |
| 23 | 3 |
| 5 | 4 |
| 26 | 2 |
| 33 | 1 |
| 84 | 4 |
| 33 | 8 |
| 44 | 9 |
| 62 | 6 |
| 69 | 3 |
| 76 | 4 |
| 82 | 2 |
| 89 | 1 |
| 23 | 6 |
| 5 | 3 |
| 26 | 4 |
| 33 | 2 |
| 84 | 1 |
| 33 | 4 |
| 44 | 8 |
| 62 | 9 |
| 69 | 6 |
| 64 | 3 |
| 70 | 4 |
| 77 | 2 |
| 11 | 1 |

Example:

Two variables,
Age (0-100 years) &
Family Size (1-10)

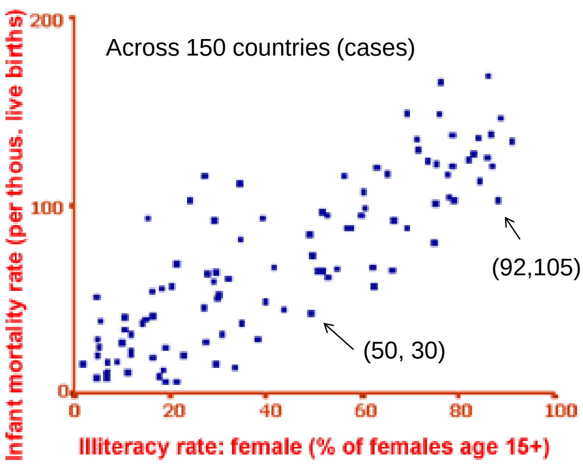
Sometimes bivariate tables are impractical with interval/ratio variables (10 rows X 100 columns)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Age | | | | | | | | | |
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Family Size | | | | | | | | | | |
| | | | | | | | | | | |
| 1 | | | | | | | | | | |
| 2 | | | | | | | | | | |
| 3 | | | | | | | | | | |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |
| 8 | | | | | | | | | | |

Rather than working with Bivariate Tables, we work with "scattergrams"..

100 columns

## Example of a Hypothetical Scattergram Showing the Relationship Between X and Y



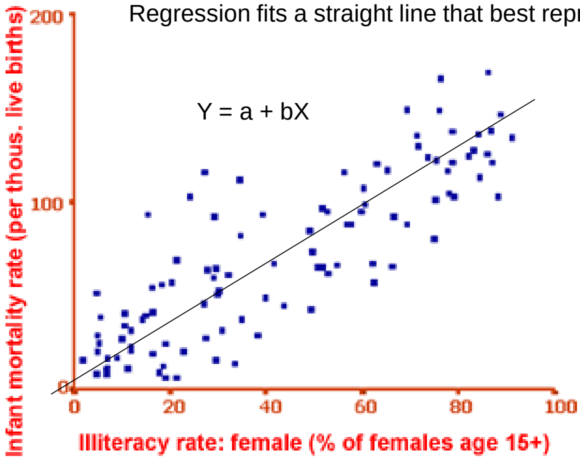

| | X | Y |
|---|---|---|
| | | |
| Case 1 | 2 | 8 |
| Case 2 | 20 | 4 |
| Case 3 | 25 | 10 |
| Case 4 | 4 | 11 |
| | .. | .. |
| | ... | ... |
| Case 150 | 92 | 120 |

Retaining as much information as possible

Regression uses directly all of this detailed information!!

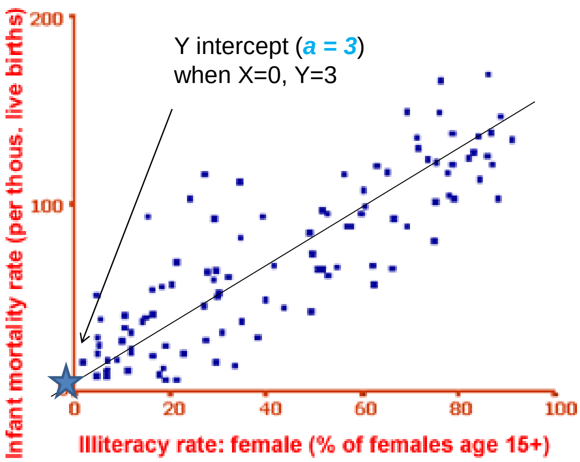
# Regression is all about representing a relationship linearly..

Regression fits a straight line that best represents the data




Where:
a is the y intercept
b is the slope

Assume our regression line (Y= a + bX) is: ***Y = 3 + 1.4 X***



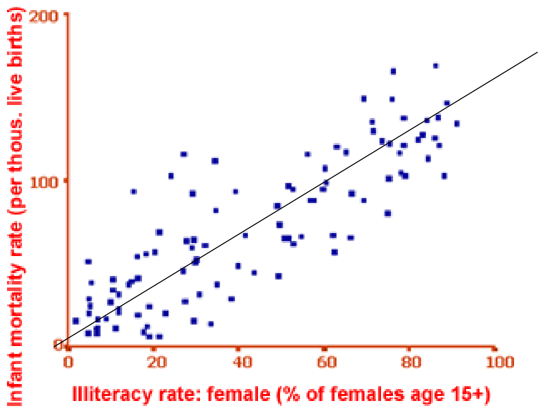

Slope (***b = 1.4***)

For every 1 unit increase in the illiteracy rate, we can expect a 1.4 unit increase in the dependent

Positive and negative associations are possible

Positive associations are represented by "positive slopes"

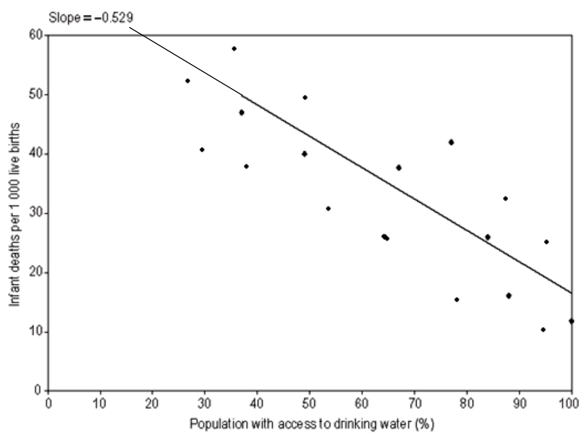
In this case, the higher a society scores in terms of the illiteracy rate, the higher we would predict the infant mortality rate…



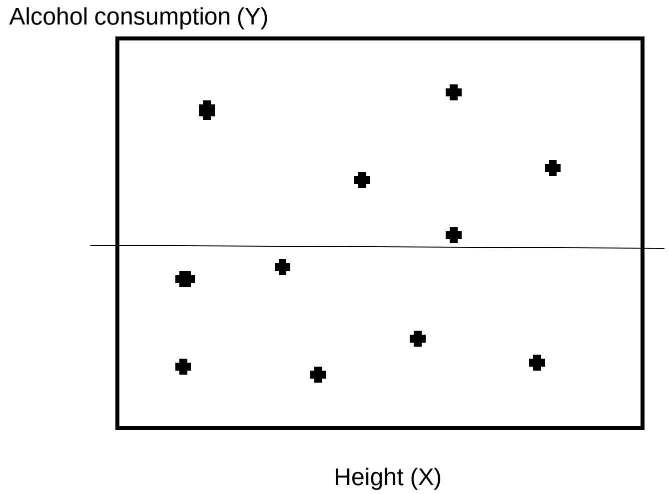

Negative associations are possible -> negative slope

In this case, the "higher" the percentage with access to drinking water the "lower" the observed infant mortality rate

FIGURE 2. Infant mortality and population access to safe drinking water 1997

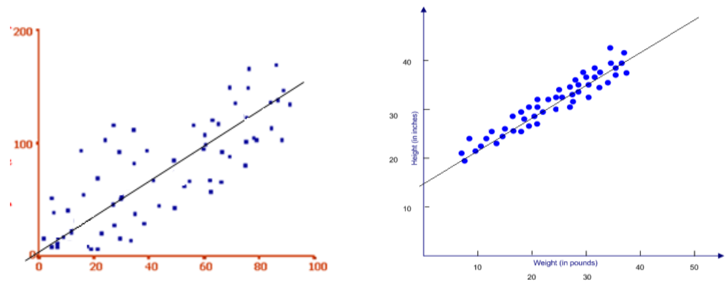


An absence of an association has a slope of zero




In addition to the direction (positive or negative),
we are also interested in both the “strength” and “significance” of relationships..

Linear relationships vary in terms of the strength of the associations involved:



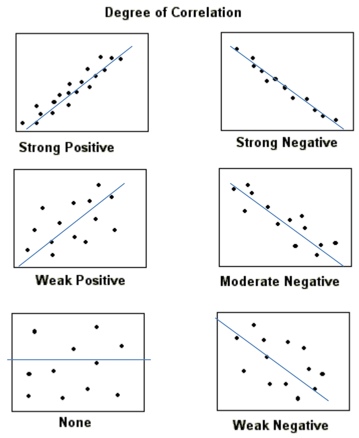

Example: the right graph portrays a much stronger association

The greater the cases are clustered around the regression line, the stronger the relationship.

Based on the regression slope, we can calculate an additional statistic:
***Pearson’s R*** (also called the correlation coefficient) which serves as a “measure of association” for interval variables (details forthcoming)

Like Gamma, ranges form -1.0 thru +1.0

# Regression is all about representing a relationship linearly..




NEXT CLASS:

How do we obtain the Regression Line:
$y = a + bx$ ?

How do we obtain Pearson's *R?*

*What about "statistical significance"?*

**GET STARTED ON ASSIGNMENT # 5!!**
**You can now do: Q1 – Q5a.,, remainder after next class.**