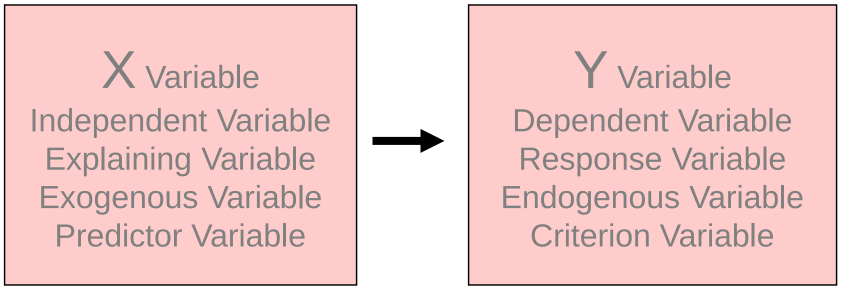
CRP 272

Introduction To Regression Analysis

Relationships Among Two Variables: Interpretations

One variable is used to “explain” another variable



X Variable
Independent Variable
Explaining Variable
Exogenous Variable
Predictor Variable

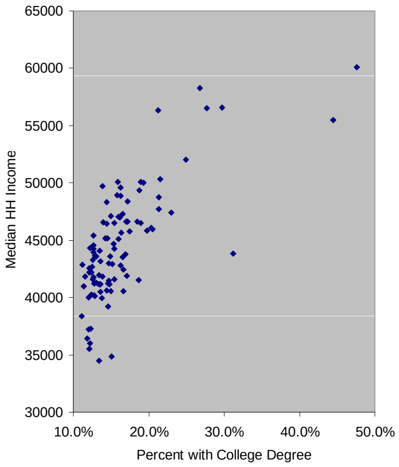
→

Y Variable
Dependent Variable
Response Variable
Endogenous Variable
Criterion Variable

Today's Material

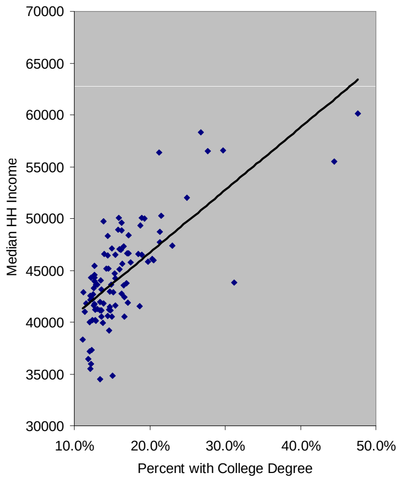
- Introduction to regression analysis, a commonly used statistical technique
  - Regression analysis is used to understand relationships between two interval or ratio variables where the relationship is thought to be linear
  - One variable is the dependent variable (Y) and the other is the independent variable (X)
  - Multiple regression analysis is a more complex technique where there is one dependent variable and more than one independent variables

Example Scatter Plot



- This is a linear relationship
- It is also a positive relationship.
- As percent of population with a college degree increases so does median HH income
- Is this a causal relationship?

Regression Line



- The regression line is the best-fitting straight line description of the plotted points and use can use it to describe the association between the variables.
- If all the points fall exactly on the line then the squared variation from the line is 0 and you have a perfect relationship.

Things to Remember

- Regression analysis is focused on association, not causation – causation involves theory and hypothesis testing.
- Association is a necessary prerequisite for inferring causation, but also:

1. The independent variable must precede the dependent variable in time or sequence
2. The two variables must be plausibly linked by a theory,
3. Competing independent variables must be eliminated as potential causes.

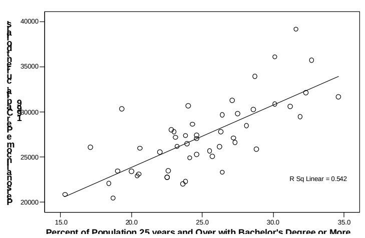
Regression Coefficients

- For the equation Y = a + bX
- The intercept or constant (a) tells you the value of Y if X = 0.
- The regression coefficient (b) is the slope of the regression line and tells you what the nature of the relationship between the variables is (positive or minus).
- The regression coefficient indicates how much change in the independent variables is associated with how much change in the dependent variable.
- The larger the regression coefficient the more change.

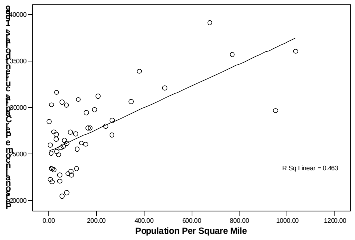
Regression Basics

- The regression coefficient (slope of the line) is not a good indicator for the strength of the relationship.
- Two scatter plots with very different dispersions could produce the same regression line.

Percent of Population with Bachelor's Degree by Personal Income Per Capita



Percent of Population with Bachelor's Degree by Personal Income Per Capita



Pearson’s r

- To determine relationship strength, look at how closely the dots are clustered around the regression line. The more tightly the cases are clustered, the stronger the relationship, while the more distant, the weaker.
- The first statistical indicator of the strength and direction of a relationship is Pearson’s r.
- Pearson’s r has a possible range of -1 to + 1 with 0 being no linear relationship at all.
  - The sign (+ or -) indicates whether the relationship is positive or negative or the slope of the line is positive or minus

The linear equation

Y = a + bX

Y = Median HH income (what we are explaining – **the dependent variable)**

X = Percent with College Degree (what is doing the explaining – **the independent variable)**

Also described as “regressing Y on X” or “regressing income on education”

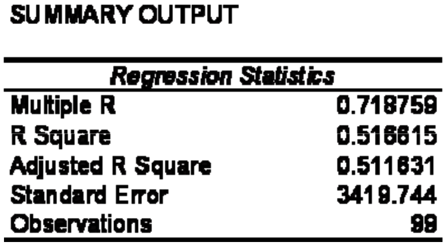
Reading the Tables

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.718759
R Square	0.516615
Adjusted R Square	0.511631
Standard Error	3419.744
Observations	99

- When you run regression analysis in Excel or Minitab you get several tables. Each tells you something about the relationship. This table is from Excel
- One will be a model summary.
- Typically, you will find the following indicators: R, R Square, Adjusted R Square, and the Standard Error of Estimate.
- Note: Multiple R is the same as Pearson’s r

R-Square

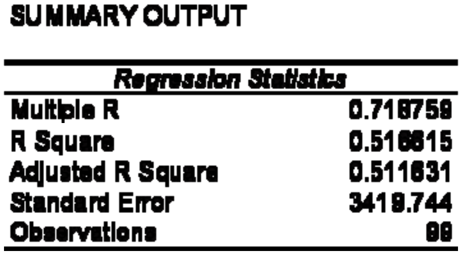


SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.718759**
R Square	**0.516615**
Adjusted R Square	**0.511631**
Standard Error	**3419.744**
Observations	**99**

- R-Square is the proportion of variance in the dependent variable (**income per capita**) which can be predicted or explained from the independent variable (**level of education**).
- This value indicates that 51.7% of the variance in income can be predicted from the variable **education**.
- Note that this is an overall measure of the strength of association, and does not reflect the extent to which any particular independent variable is associated with the dependent variable if there is more than one (as there is in multiple regression).
- R-Square is also called the **coefficient of determination**.

Adjusted R-square



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.718759
R Square	0.516615
Adjusted R Square	0.511631
Standard Error	3419.744
Observations	99

- As more variables are added to the model, the degrees of freedom are reduced.
- The adjusted R-square attempts to yield a more honest value to estimate the R-squared for the population by accounting for the degrees of freedom in the model. The value of R-square was .517, while the value of Adjusted R-square was .512. There isn't much difference because we are dealing with only one variable.
- When the number of observations is small and the number of predictors is large, there will be a much greater difference between R-square and adjusted R-square.
- By contrast, when the number of observations is very large compared to the number of predictors, the value of R-square and adjusted R-square will be much closer.

ANOVA

ANOVA

	df	*SS*	*F*	*Significance F*
Regression	1	1212361499	103.6680407	0.00
Residual	97	1134381094	11694650.46	
Total	98	2346742594		

- **ANOVA** – Analysis **o**f **Va**riance
- These values may be used to answer the question "Do the independent variables reliably predict the dependent variable?".
- It gives us the sum of squares (SS) due to regression and the sum of squares (SS) due to the residual or the error.
- The F statistic is only used when we are analyzing sample data. An F value of greater than 4 usually means that we have a low probability that these findings occurred by chance.
- Here the significance or the “p” of F is very low.

Coefficients

Table of Coefficients

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	34,607.14	1,030.15	33.59	0.00
College Degree Percentage	60,488.59	5,940.88	10.18	0.00

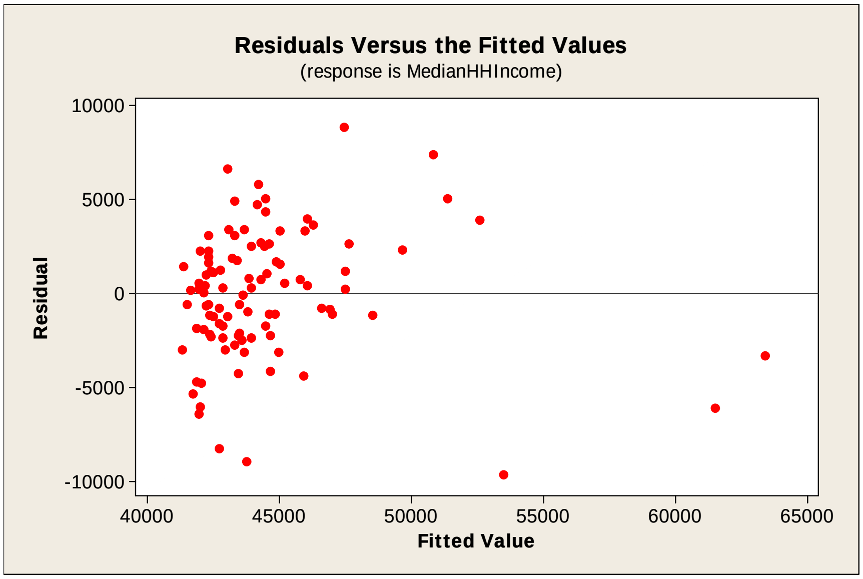
- In the equation Y = a + bX,

  a = 34,607.14 and b = 60,488.59. So,

  Y = 34607.14 + 60488.59X
- The *t* statistic is the coefficient value divided by the standard error – the lower the standard error, the higher the *t*.
- The p-value applies only to sample data – it tells us the probability that our coefficients occurred by chance.
- In population analysis we ignore the *t* statistics

Part of the Regression Equation

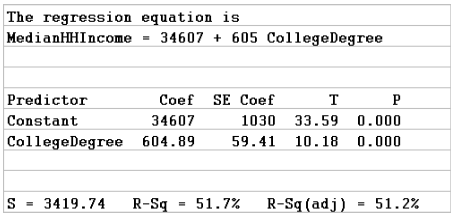
- *b* represents the slope of the line
  - It is calculated by dividing the change in the dependent variable by the change in the independent variable.
  - The difference between the actual value of Y and the calculated (or predicted) amount is called the **residual**.
  - We analyze our residuals to look for patterns
  - This represents how much error there is in the prediction of the regression equation for the y value of any individual case as a function of X.



Multiple Regression: Incorporating Two Or More Independent Variables

- Multiple regression analysis is useful for comparing two variables to see whether controlling for other independent variable affects your model.
- For the first independent variable, education, the argument is that a more educated populace will have higher-paying jobs, producing a higher level of per capita income in the state.
- The second independent variable is included because we expect to find better-paying jobs, and therefore more opportunity for state residents to obtain them, in urban rather than rural areas.

Bivariate Regression

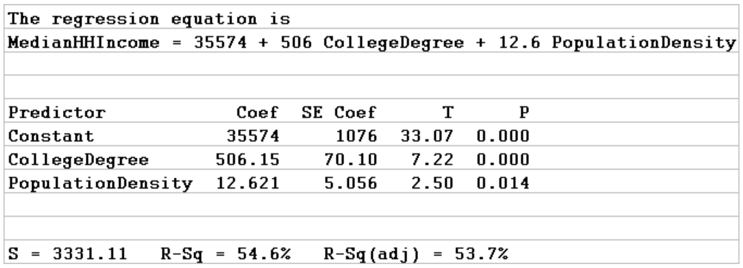


```
The regression equation is
MedianHHIncome = 34607 + 605 CollegeDegree


Predictor        Coef  SE Coef      T      P
Constant        34607     1030  33.59  0.000
CollegeDegree  604.89    59.41  10.18  0.000


S = 3419.74   R-Sq = 51.7%   R-Sq(adj) = 51.2%
```

Multiple Regression



```
The regression equation is
MedianHHIncome = 35574 + 506 CollegeDegree + 12.6 PopulationDensity


Predictor             Coef  SE Coef      T      P
Constant             35574     1076  33.07  0.000
CollegeDegree       506.15    70.10   7.22  0.000
PopulationDensity   12.621    5.056   2.50  0.014


S = 3331.11   R-Sq = 54.6%   R-Sq(adj) = 53.7%
```

Last Points

- When we do multiple regression, a fundamental assumption is that the independent variables are not highly correlated with each-other.

- Graduation rate = f(poverty, welfare)

  Y = HS graduation rate

  X_1 = child poverty rate

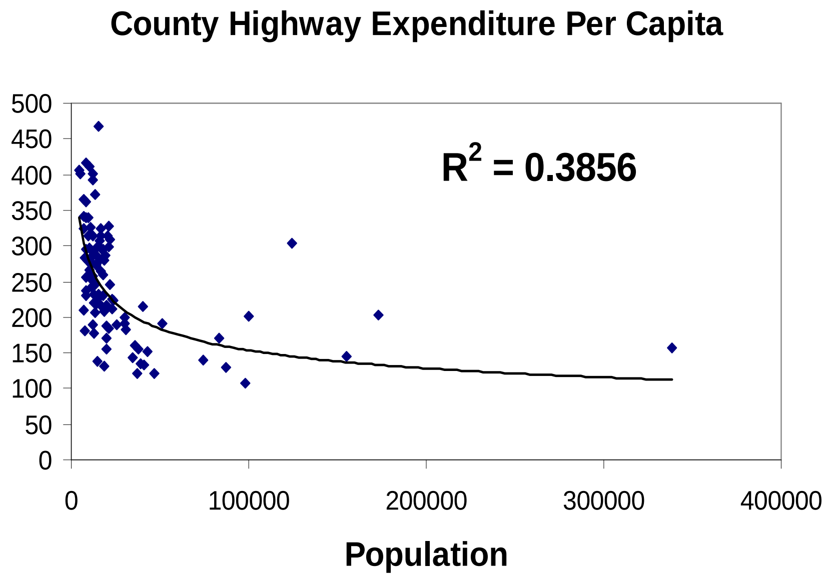
  X_2 = children receiving public assistance

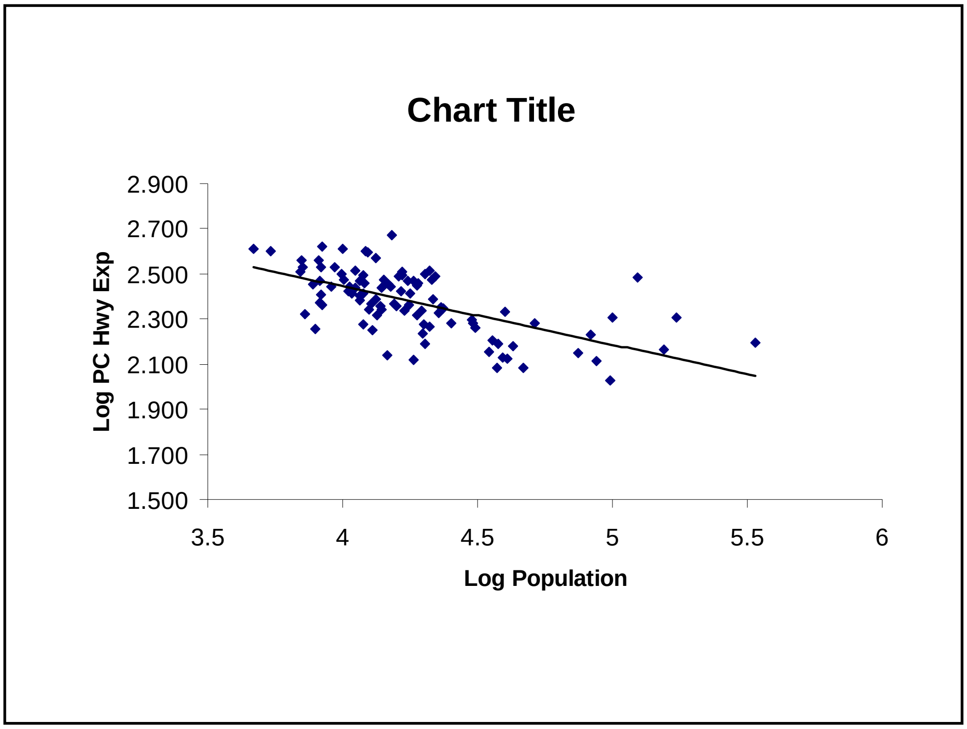
Last Points

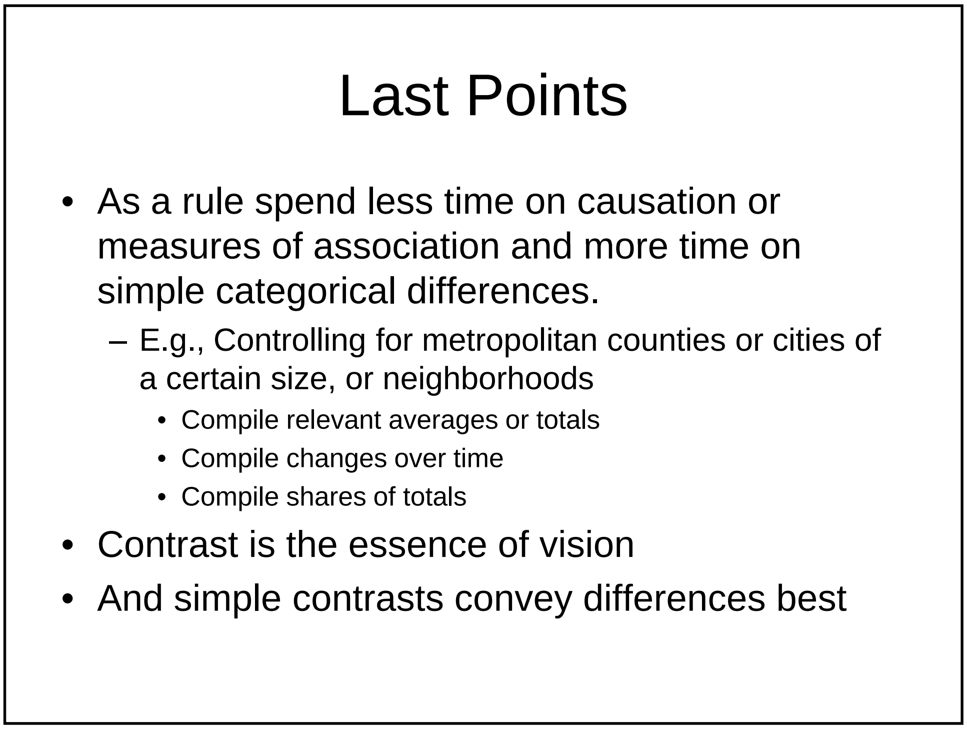
- If your scatter plot produces a curve of data, you can
  - Transform the data so that they are linear
    - Log transformations
    - If numbers are very large, you can take their square root.
  - Introduce a term into your equation to account for the curve:

$Y = a + bX + cX^2$

$Y = ab^X$

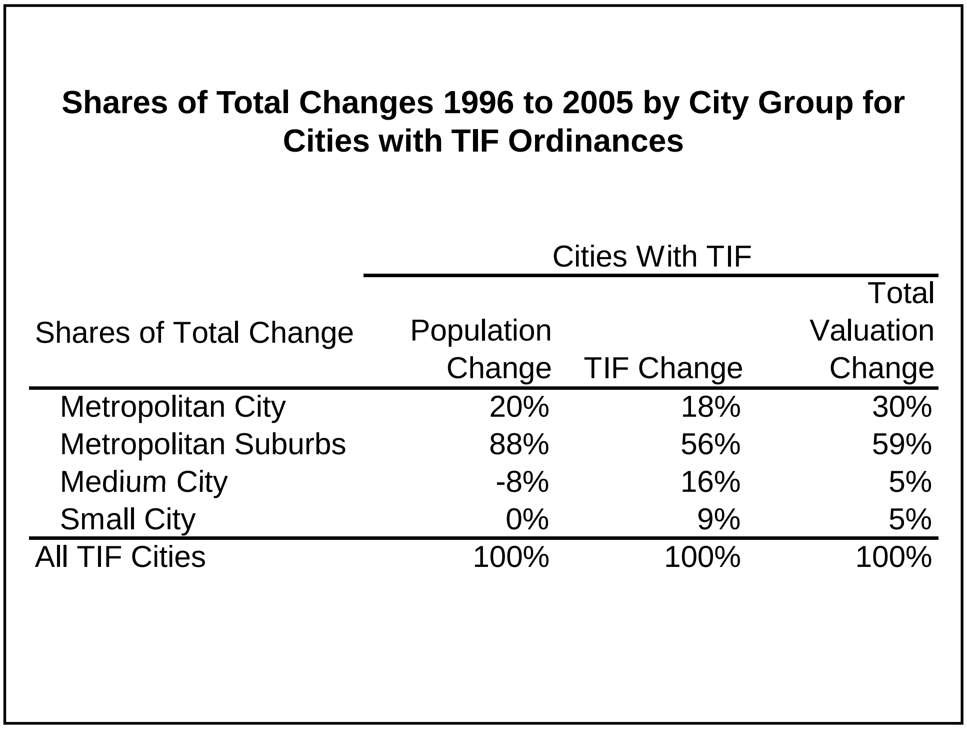






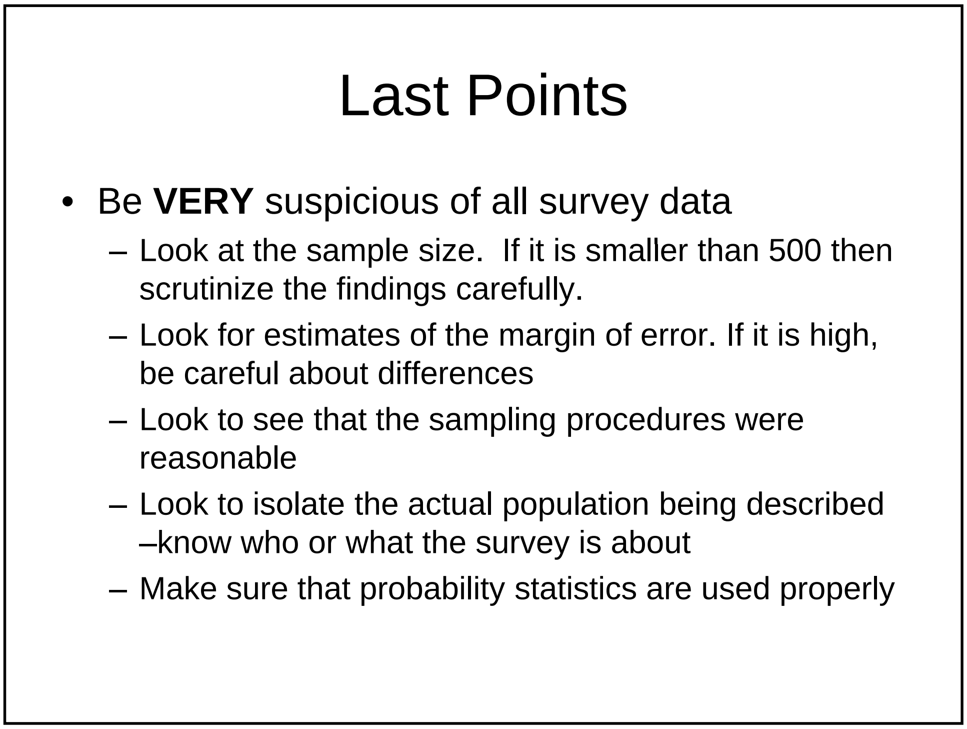
Last Points

- As a rule spend less time on causation or measures of association and more time on simple categorical differences.
  - E.g., Controlling for metropolitan counties or cities of a certain size, or neighborhoods
    - Compile relevant averages or totals
    - Compile changes over time
    - Compile shares of totals
- Contrast is the essence of vision
- And simple contrasts convey differences best



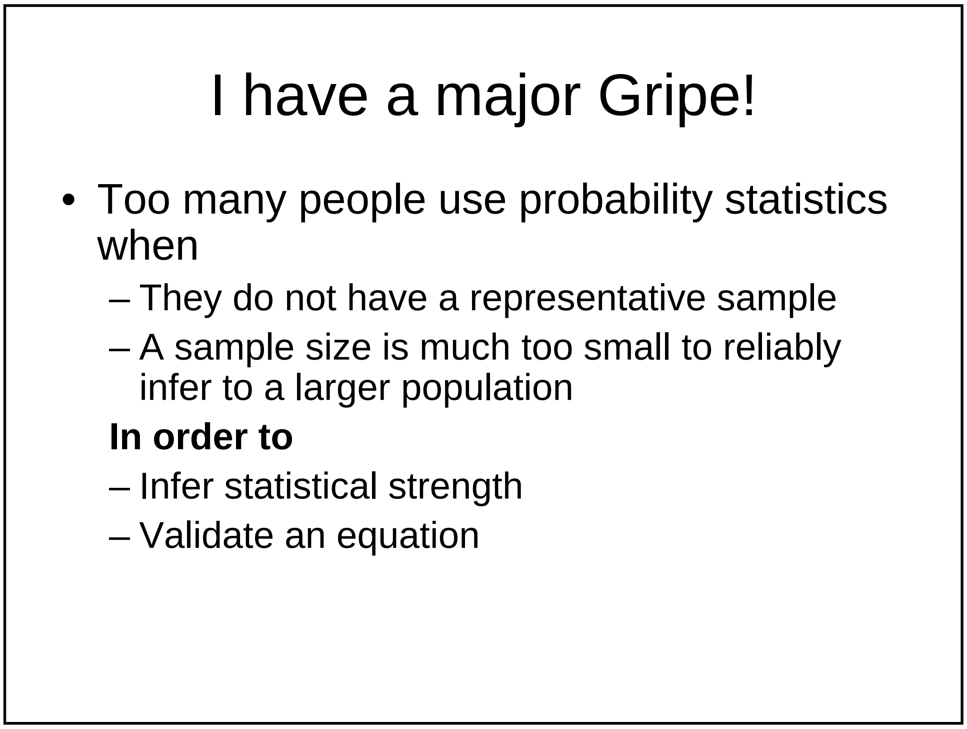
Shares of Total Changes 1996 to 2005 by City Group for Cities with TIF Ordinances

	Cities With TIF		
Shares of Total Change	Population Change	TIF Change	Total Valuation Change
Metropolitan City	20%	18%	30%
Metropolitan Suburbs	88%	56%	59%
Medium City	-8%	16%	5%
Small City	0%	9%	5%
All TIF Cities	100%	100%	100%



Last Points

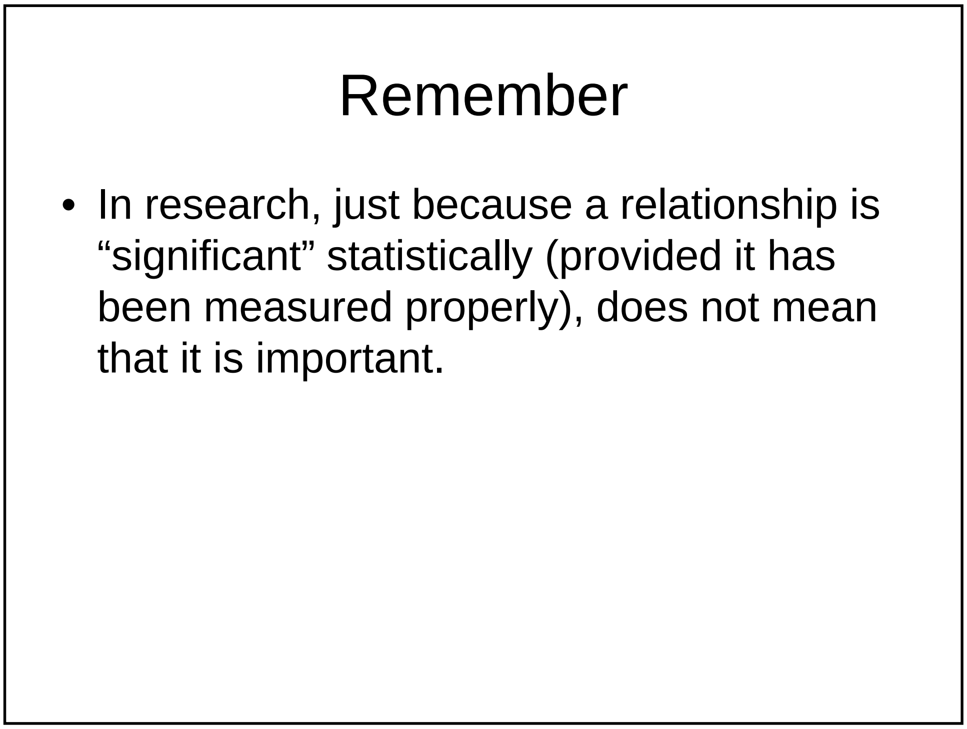
- Be **VERY** suspicious of all survey data
  - Look at the sample size. If it is smaller than 500 then scrutinize the findings carefully.
  - Look for estimates of the margin of error. If it is high, be careful about differences
  - Look to see that the sampling procedures were reasonable
  - Look to isolate the actual population being described –know who or what the survey is about
  - Make sure that probability statistics are used properly



I have a major Gripe!

- Too many people use probability statistics when
  - They do not have a representative sample
  - A sample size is much too small to reliably infer to a larger population

  In order to
  - Infer statistical strength
  - Validate an equation



Remember

- In research, just because a relationship is “significant” statistically (provided it has been measured properly), does not mean that it is important.