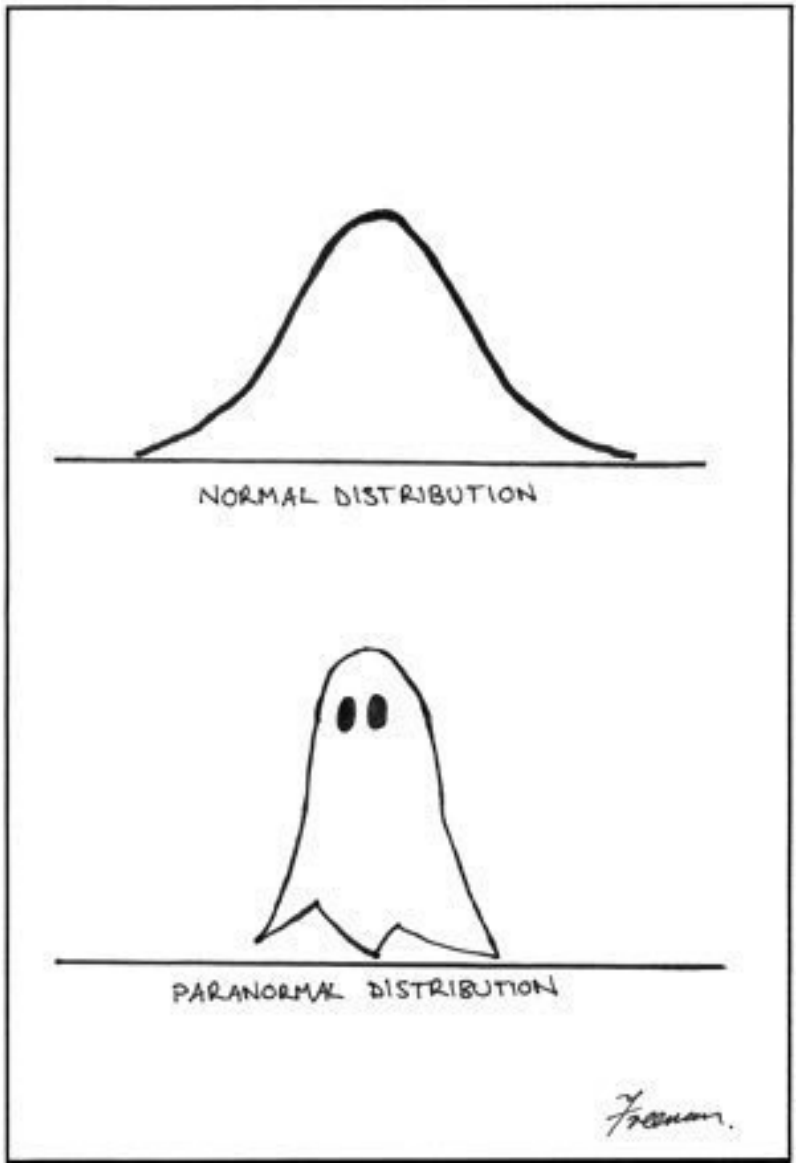
# Association of Variables

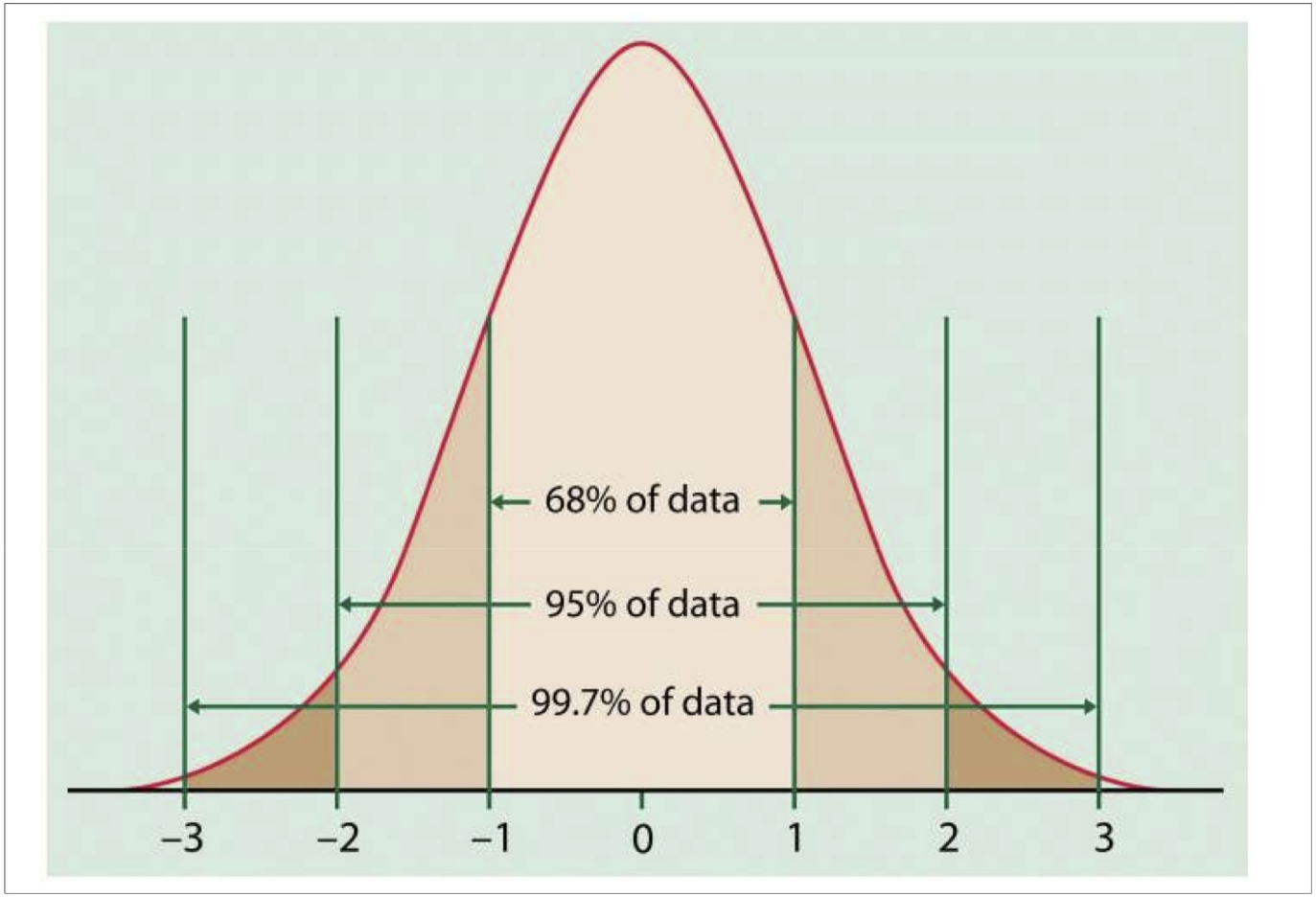
POLI 205 Doing Research in Politics

Fall 2015

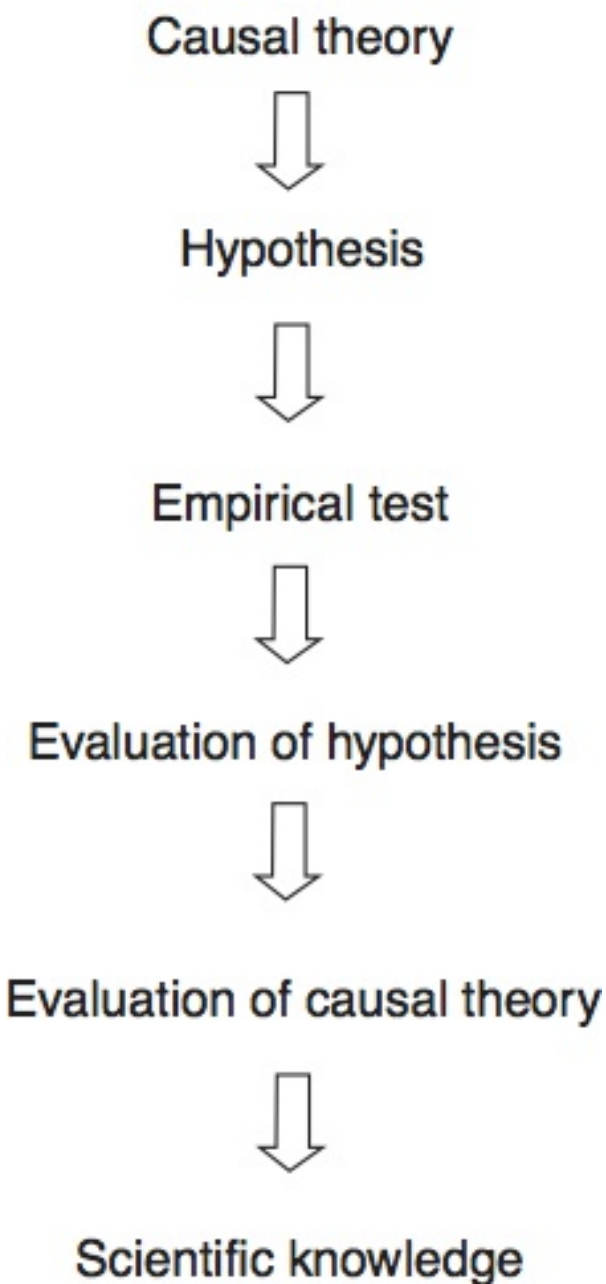
# Normal Distribution



# Normal Distribution



# Road To Scientific Knowledge



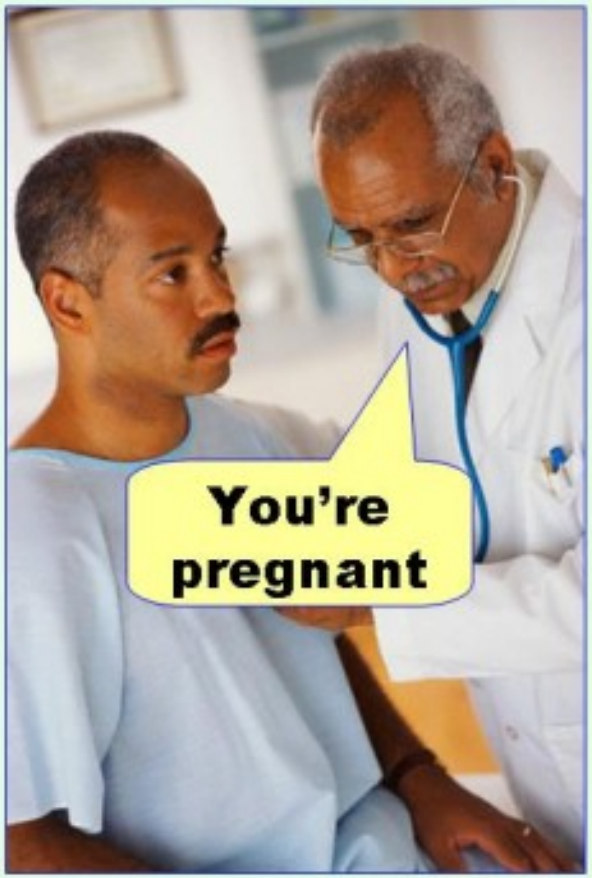
# Bivariate Hypothesis Testing

- These techniques are good at helping us answer the question, "*Are X and Y related?*"
- Bivariate: *two variables*
  - Can not help with controlling for $Z$ variables
- Hypothesis testing:
  - Null vs. working hypothesis
  - Type I and Type II errors
- Variable type and measurement metric

# Type I and Type II errors

**Type I error**

(false positive)




**Type II error**

(false negative)

# Variable Types and Statistical Tests

| | | Independent variable type | |
|---|---|---|---|
| | | Categorical | Continuous |
| Dependent variable type | Categorical | *Tabular analysis* | Probit/logit (Ch.11) |
| | Continuous | *Difference of means* | *Correlation coefficient*; bivariate regression model (Ch.8) |

*Note*: Tests in italics are discussed in this chapter.

# $p$ -values

- $p$ -value:
  - $p$ means probability
  - Ranges from 0 to 1
- Based on the assumption that you are drawing a perfectly random sample from the underlying population
- "Is there a *relationship* between $X$ and $Y$?"

# Logic of $p$ -values

- We compare the actual relationship between $X$ and $Y$ in sample data with what we would expect to find if $X$ and $Y$ *were not* related in the underlying population
- The *more different* the empirically observed relationship is from what we would expect to find if there were *not* a relationship, the more confidence we have that X and Y are related in the population
- **Critical value**: a pre-determined standard based on the level of statistical significance and *degrees of freedom*
- **Degrees of freedom**: number of independent pieces of information (data) that are used to estimate the parameter. Usually $N$ - number of parameters to be estimated.

# Logic of $p$ -values

- $p$ -value:
  - the probability that we would see the relationship that we are finding because of random chance
  - the probability that we would see the observed relationship between the two variables in our sample data if there were truly no relationship between them in the unobserved population
- The *lower* the $p$ -value, the *greater* confidence we have that there is a systematic relationship between the two variables
- Level of confidence with which we can reject the null hypothesis

# Statistical Significance

- **Statistical significance**: The *lower* the $p$ -value, the *greater* confidence we have that there is a systematic relationship between the two variables
- Depends on:
  - "Statistical significance" is achieved only to the extent that the assumptions underlying the calculation of the $p$ -value hold
  - There are a variety of different standards for what is a statistically significant $p$ -value. Most social scientists use the standard of a $p$ -value of *.05*

# Hypothesis Test: Tabular Analysis

- When would we use tabular analysis?
- Reading a table
  - What variables define the rows and columns?
  - What do the cells represent?
  - What general patterns are there?
- How do we test hypothesis with tabular analysis?
  - $\chi^2$ test–calculate critical value and degrees of freedom
  - $\chi^2 = \frac{(O-E)^2}{E}$
  - $df = (r - 1)(c - 1)$

# Hypothesis Test: Difference of Means

- When do we use a difference of means test?
- Are the means of the DV different across the values of the IV?
- How do we test hypotheses using difference of means?
  - $t$ -test–calculate critical value and degrees of freedom
  - $t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)}$
  - $se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2})} X \sqrt{(\frac{1}{n_1+\frac{1}{n_2}})}$
  - $df = n_1 + n_2 - 2$

# Hypothesis Test: Correlation Coefficient

- When do we use correlation coefficients?
- Is there covariation between the DV and the IV?

# Hypothesis Test: Correlation Coefficient

- How do we test hypothesis using correlation coefficients?
- Pearson's $r$ and $t$ -test for $r$
  - $r = \frac{cov_{XY}}{\sqrt{var_X var_Y}}$
  - $t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
- $r$ ranges from -1 to 1
- Perfect positive correlation $r = 1$
- Perfect negative correlation $r =$ -1
- No correlation $r = 0$

# Covariance

- **Covariance**: A statistical way of summarizing the general pattern of association between two continuous variables
- $cov_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$
- Positive vs. negative relationship

# Covariance

- In a covariance table, the cells across the main diagonal (from upper-left to lower-right) are cells for which the column and the row reference the same variable
- In this case the cell entry is the variance for the referenced variable
- Each of the cells off of the main diagonal displays the covariance for a pair of variables

# Covariance

- If the difference between $X$ and the mean of $X$ multiplied by the difference between $Y$ and the mean of $Y$ are both $>$ 0, that case's contribution to the numerator in the covariance equation will be positive

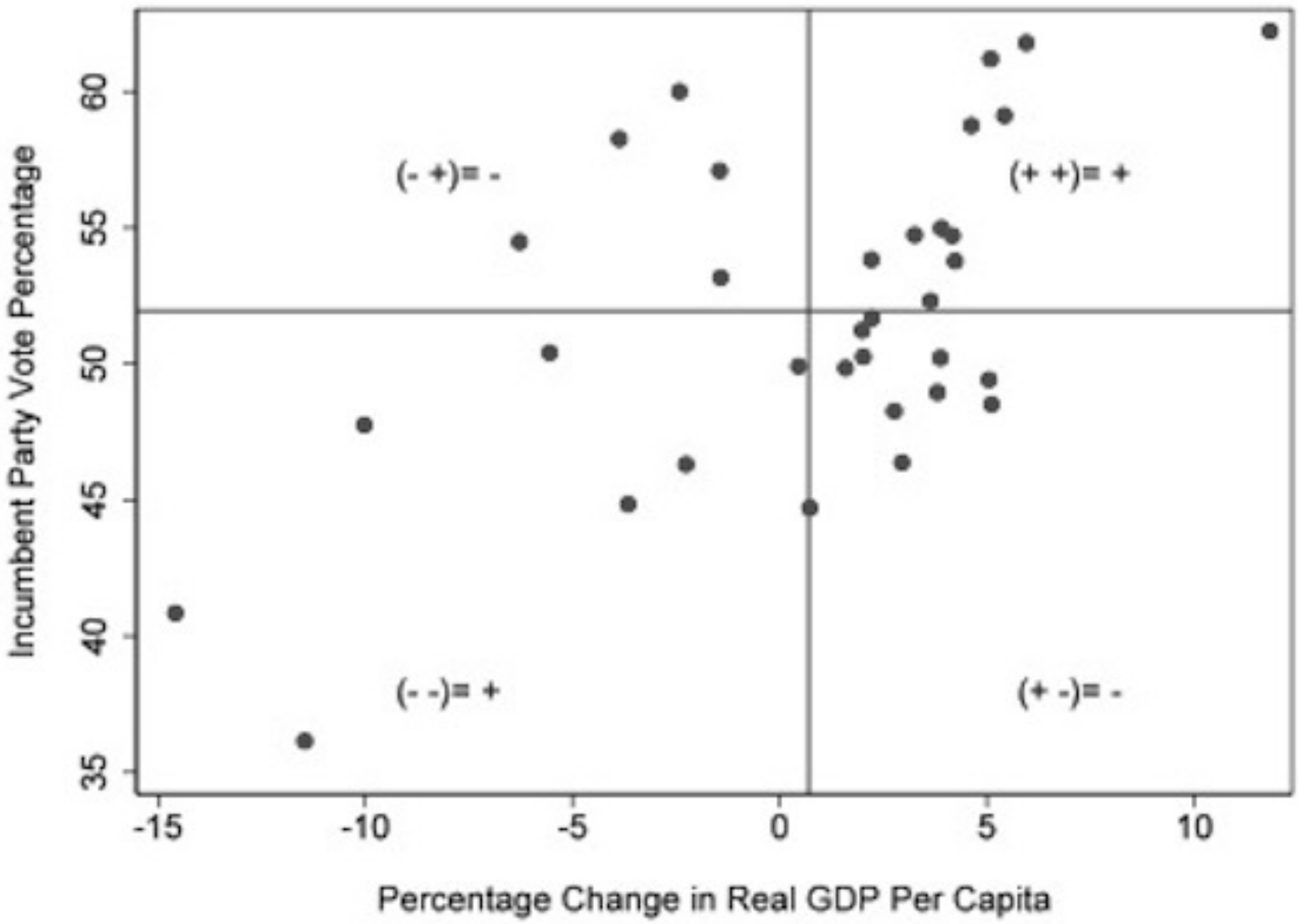
# Covariance

- If the difference between $X$ and the mean of $X$ multiplied by the difference between $Y$ and the mean of $Y$ are both $<$ 0, that case's contribution to the numerator in the covariance equation will also be positive, because multiplying two negative numbers yields a positive product

# Covariance

- If a case has a combination of one value greater than the mean and one value less than the mean, its contribution to the numerator in the covariance equation will be negative because multiplying a positive number by a negative number yields a negative product

# Covariance


(- +)= -
(+ +)= +
(- -)= +
(+ -)= -
Incumbent Party Vote Percentage
35
40
45
50
55
60
-15
-10
-5
0
5
10
Percentage Change in Real GDP Per Capita