

# Inferential Statistics

Pam Perlich  
URBPL 5/6010:  
Urban Research  
University of Utah

## Uses of Inferential Statistics

- Determine the nature and strength of relationships between variables.
- Compare populations to determine similarities and differences
- Generalize to the population from sample data
- Evaluate uncertainty

## Variable Types

- Univariate = Single Variable
- Quantitative variables = numbers
- Qualitative = categorical

## Quantitative Variable

- Quantitative variables = numbers
  - Discrete: defined list of numbers
    - Six-sided die: 1,2,3,4,5,6
  - Continuous: wide range of possible values
    - Weight = 124.63 pounds

## Qualitative Variable

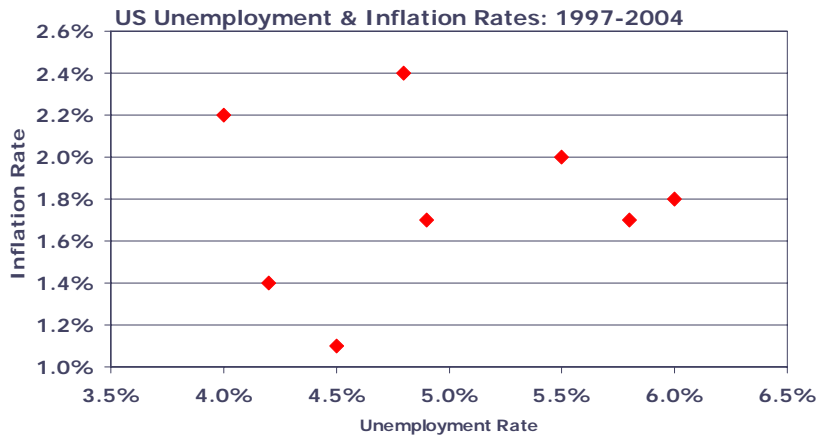
- Qualitative variables = categories
  - Ordinal: order indicates ranking
    - Example: How likely are you to vote in the presidential election?
      - Very unlikely
      - Somewhat unlikely
      - Possibility
      - Somewhat likely
      - Very likely
  - Nominal: categories cannot be put into any meaningful order
    - Examples: Race, state of residence, gender

## Time Series Plot



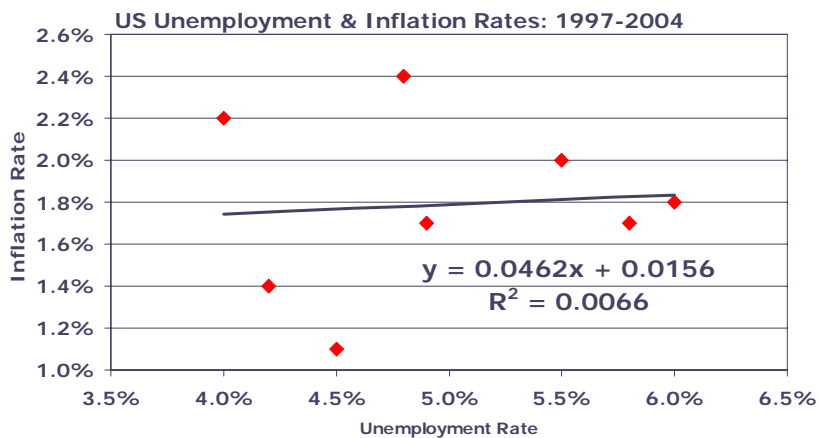
Note: Changes in the GDP deflator measure inflation.  
National data from Global Insight.

## Scatter Plot



Note: Changes in the GDP deflator measure inflation.  
National data from Global Insight.

## Scatter Plot: Add Trendline



Note: Changes in the GDP deflator measure inflation.  
National data from Global Insight.

## Probability

Probability of an event ranges from 0 to 1

Theoretical Probability of an event =

$$\frac{\text{(Number of possible ways of obtaining the event)}}{\text{(Total number of equally likely possible outcomes)}}$$

Probability of flipping a coin and getting a head =

$\frac{1}{2} = 50\%$

## Probability Distribution

- Pattern of probabilities for a set of events
- Probability ranges from 0 to 1
- Sum of the probabilities across all events = 1
  - Something will happen
- Go to [ProbabilityDistribution.xls](#)

## Discrete Probability Distribution

- Probabilities associated with a set of discrete events
- Example: Poisson Distribution
  - Use when the outcome event involves counts within a specified time period

$$P_{(x)} = \left( \frac{\lambda^x}{x!} \right) \times e^{-\lambda}$$

$$P_{(x)} = \left( \frac{\lambda^x}{x!} \right) \times e^{-\lambda}$$

Normally there are 3 accidents in an intersection in a year. What is the probability of 2 in a year?

$\lambda$  Number of events in a specified time period

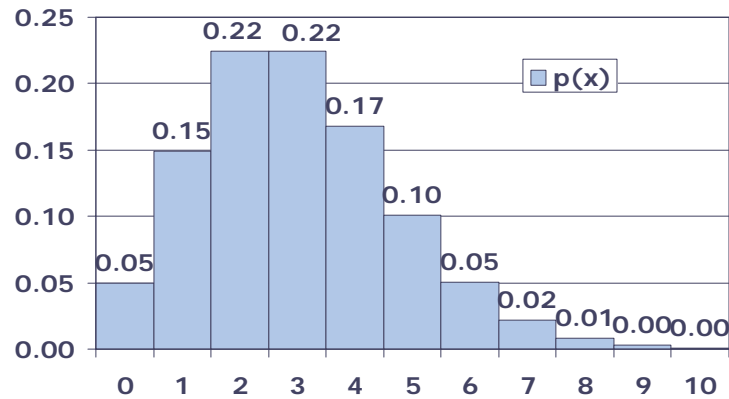
$x$  Number of times the event occurs  
(Calculating the probability of this)

$$P_{(x)} = \left( \frac{3^2}{2!} \right) \times e^{-3}$$

$$p_{(x)} = \left( \frac{3^2}{2!} \right) \times e^{-3}$$

**Poisson Probability Distribution**

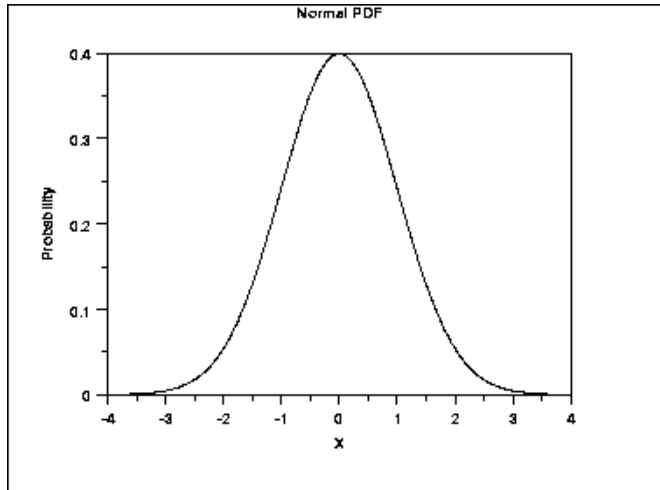
$$p_{(x)} = \left( \frac{9}{2 \times 1} \right) \times \left( \frac{1}{2.718282^3} \right) = 0.224$$



## Continuous Probability Density Function

- Probabilities are assigned to ranges
- Probability at a point is zero
- Probability is calculated as area under the curve between two values
- Example – normal distribution or bell curve

$$\text{probability density} = \exp(-\frac{1}{2}(x - \mu)^2 / \sigma^2) / (\sigma \sqrt{2\pi})$$



<http://psych.colorado.edu/~mcclella/java/zcalc.html>

<http://www-stat.stanford.edu/~naras/jsm/FindProbability.html>

## Excel Function

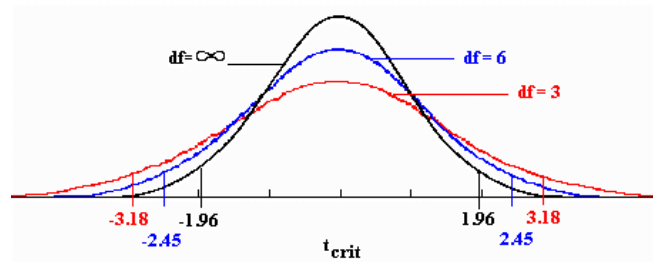
- **NORMDIST(x,mean,standard\_dev,cumulative)**
- X is the value for which you want the distribution.
- Mean is the arithmetic mean of the distribution.
- Standard\_dev is the standard deviation of the distribution.
- Cumulative is a logical value that determines the form of the function. If cumulative is TRUE, NORMDIST returns the cumulative distribution function; if FALSE, it returns the probability mass function.

## t-distribution

- Use when the standard deviation is not known
- Use with sample data
- Fatter tails than normal distribution
- Has additional parameter: degrees of freedom
- Sample size  $\rightarrow$  Degrees of freedom

## t-distribution

- Smaller degrees of freedom  $\rightarrow$  flatter distribution
- When  $df \rightarrow$  infinity  $\rightarrow t \rightarrow$  normal

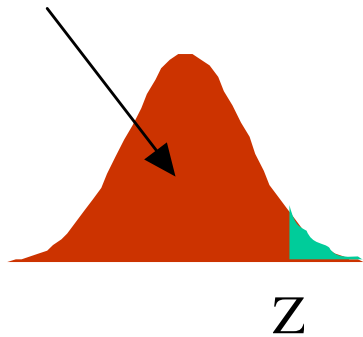


## Calculating Probabilities

- Given a normal distribution
- Convert values to standard normal
  - Mean 0, SD=1
- Set up one or two tailed test
- Calculate areas

## One & Two Tailed Tests

Total - infinity to  $Z$

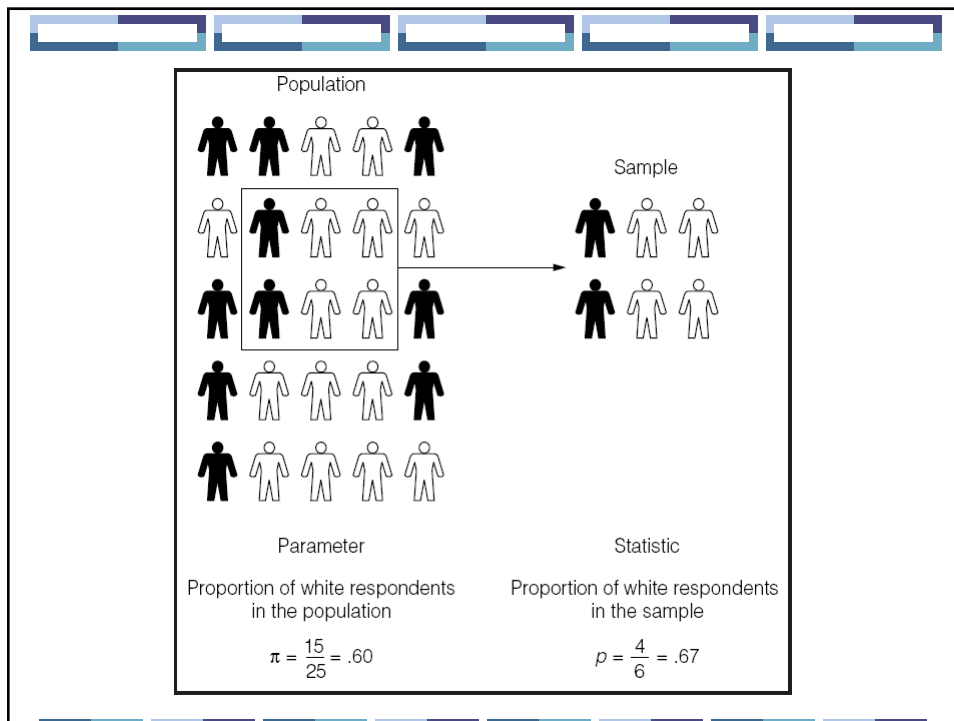


-infinity to  $-Z$   
plus  
 $+Z$  to + infinity



## Quality Sample Data is Critical

- If sample data is not created in an appropriate way, it is useless
- No statistical technique can correct for bad data
- No model can overcome bad data



## Simple Random

### Requirements / Definition

- Every member of the population has an equal chance of being chosen
- Every combination of N members has an equal chance of being chosen.

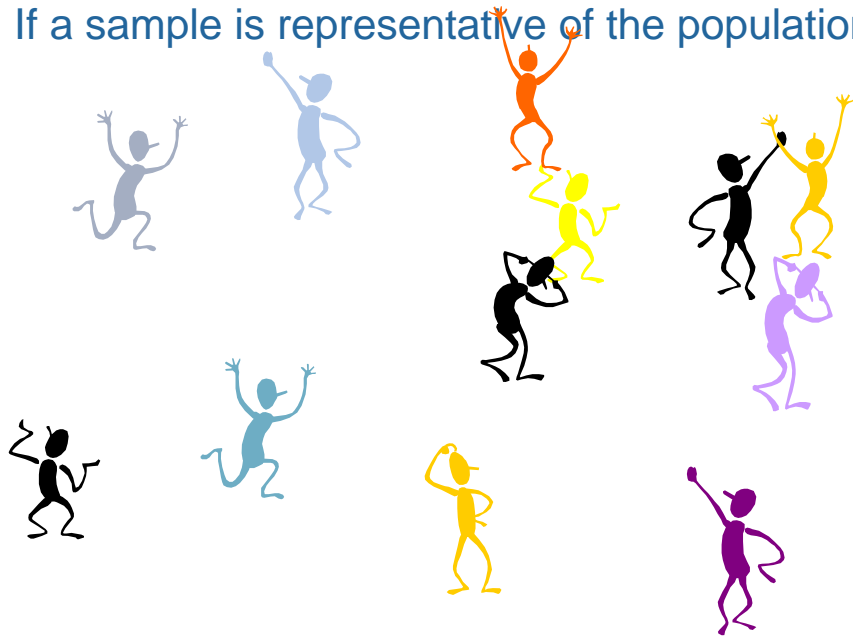
### Variety of methods

- Computer
- Calculator
- Table of random numbers

Population inferences can be made...



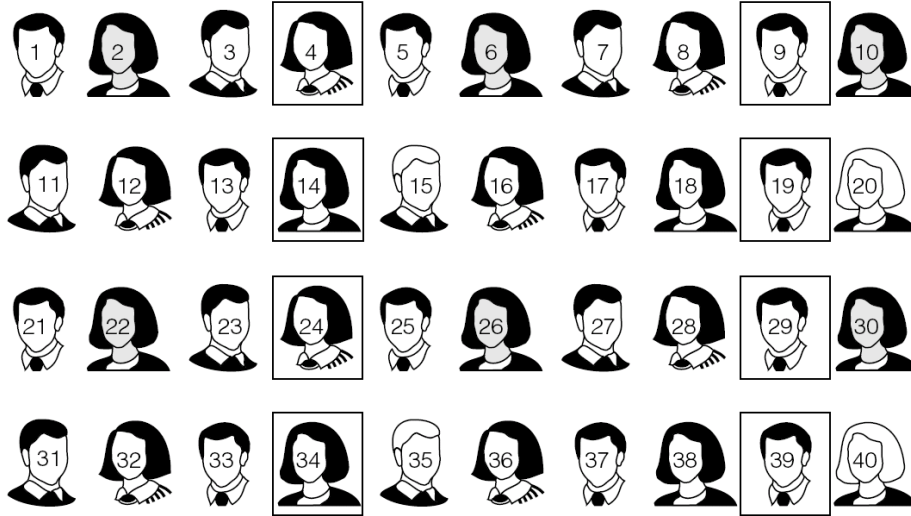
If a sample is representative of the population



## Systematic Random Sampling

- Every  $K$ th member in the total population is chosen for inclusion in the sample
- $K = \text{population size} \div \text{desired sample size}$
- The first selection is random
- All subsequent selections are  $K$  items later

From a population of 40 students, let's select a systematic random sample of 8 students. Our skip interval will be 5 ( $40 \div 8 = 5$ ). Using a random number table, we choose a number between 1 and 5. Let's say we choose 4. We then start with student 4 and pick every 5th student:



Our trip to the random number table could have just as easily given us a 1 or a 5, so all the students do have a chance to end up in our sample.

## Stratified Random Sample

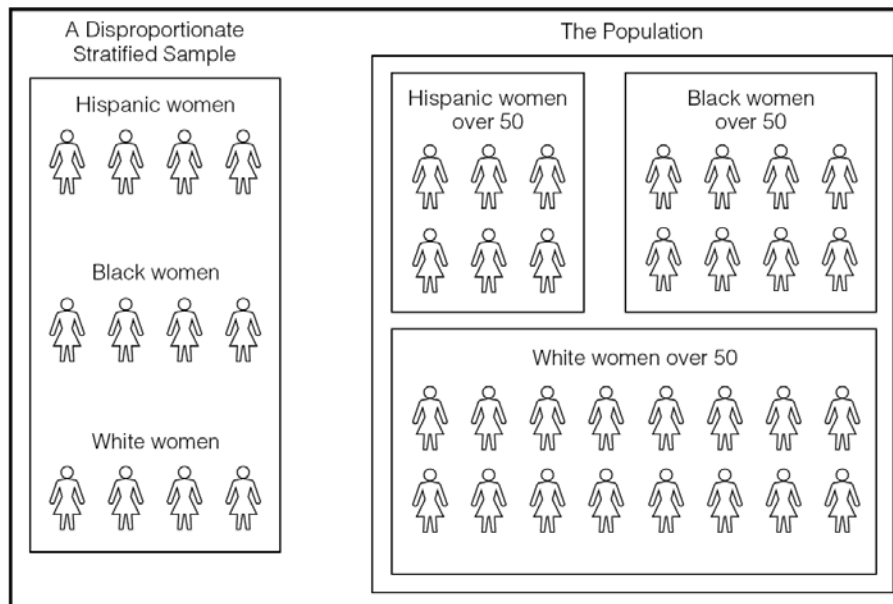
### ● Procedure

- Dividing the population into subgroups based on one or more relevant variables central
- Draw a simple random sample from each of the subgroups

## Stratified Sample Types

- **Proportionate** – sample size from each subgroup is proportional to the size of that subgroup in the entire population
- **Disproportionate** – sample size from each subgroup is disproportional to the size of that subgroup in the population

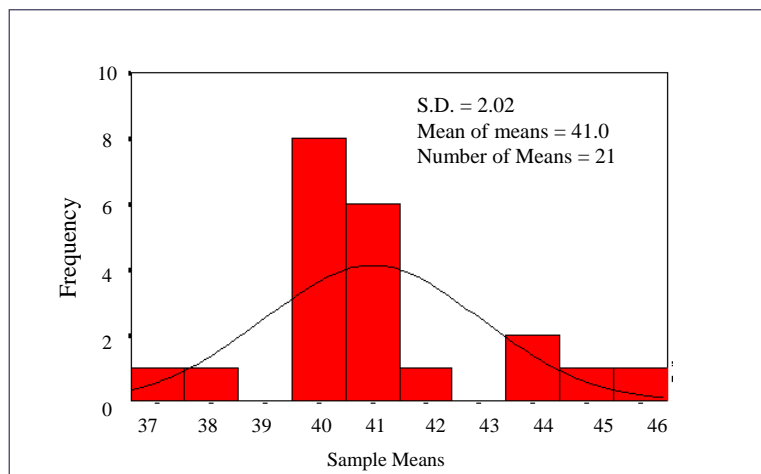
### A Random Sample Stratified by Race/Ethnicity



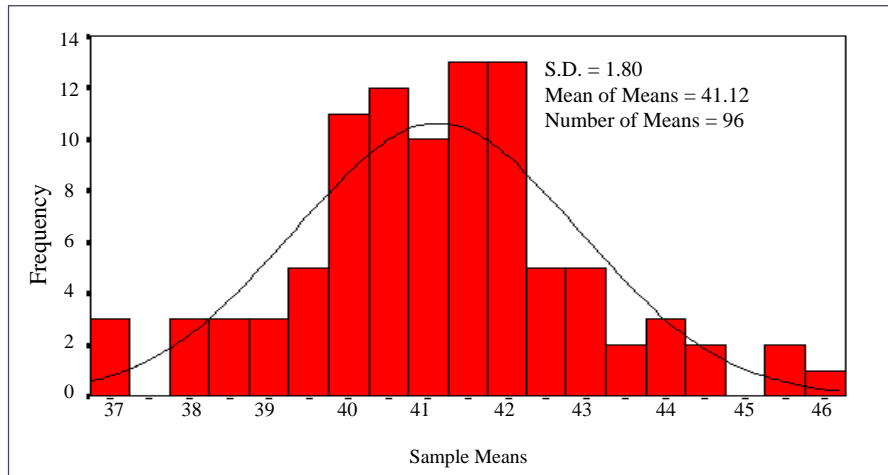
## Sampling Distributions

- **Sampling error** – Discrepancy between a sample estimate of a population parameter and the real population parameter
- **Sampling distribution** – Theoretical distribution of all possible sample values for the statistic we are estimating
  - Take larger number of samples → mean converges to population mean

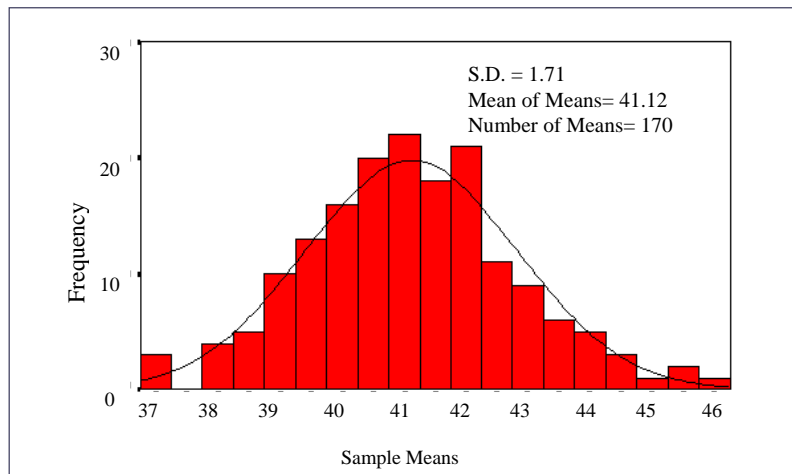
## Distribution of Sample Means with 21 Samples



## Distribution of Sample Means with 96 Samples



## Distribution of Sample Means with 170 Samples



## Estimation

### ● Process

- Select a random sample from a population
- Utilize the sample statistic to estimate a population parameter

### ● Types

- Point
- Interval – confidence interval

## Point and Interval Estimation

- **Point Estimate** – sample statistic estimates the exact value of a population parameter
- **Confidence interval (*interval estimate*)** – range of values defined by the confidence level within which the population parameter is estimated to fall within
- **Confidence Level** – the probability (expressed as a percentage) that a specified interval will contain the population parameter

## Inferential Statistics → 3 Distributions

- **Population** - variation in the larger group that we want we are attempting to discover (estimate)
- **Sample of observations** - variation in our sample (We can observe this).
- **Sampling distribution** – Normal distribution whose mean and standard deviation are unbiased estimates of the parameters of the “true” population

## The Central Limit Theorem

- No matter what the distribution of the population, the sampling distribution of the mean is normally distributed
- Larger sample size →
  - the mean of the sampling distribution becomes equal to the population mean
  - the standard error of the mean decreases in size
  - the variability in the sample estimates from sample to sample decreases

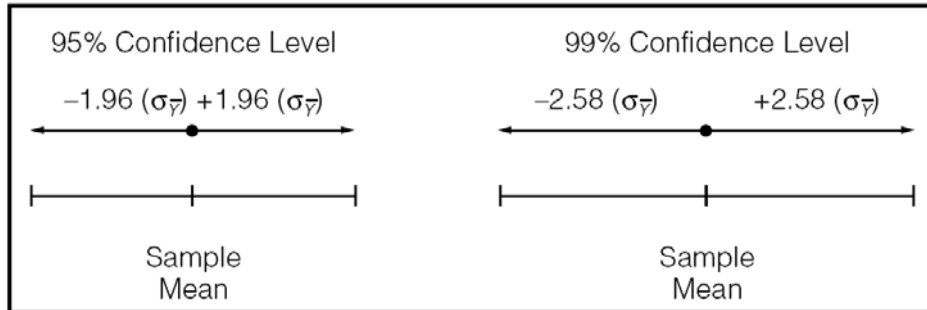
## Practical Note

- Researchers do not typically conduct repeated samples of the same population
- Rather, they use the knowledge of theoretical sampling distributions to construct confidence intervals around estimates

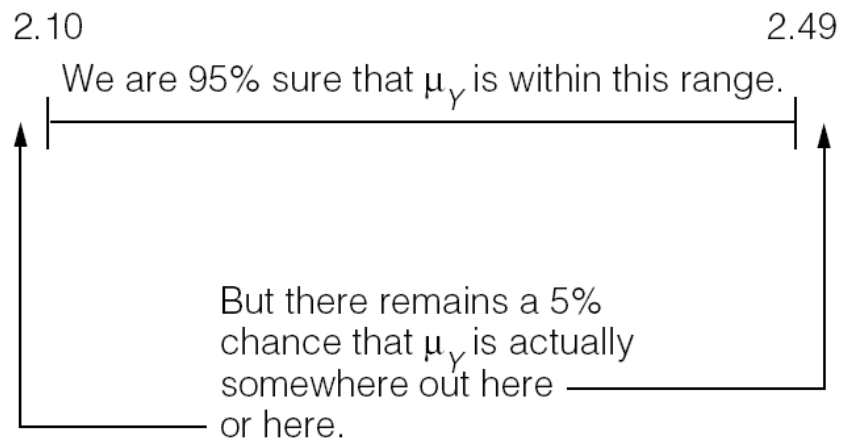
## Confidence Levels

- Probability that a specified interval will contain the population parameter
  - **95% confidence level**
    - There is a .95 probability that a specified interval **DOES** contain the population mean
    - There are 5 chances out of 100 (or 1 chance out of 20) that the interval **DOES NOT** contain the population mean.
  - **99% confidence level**
    - There is 1 chance out of 100 that the interval **DOES NOT** contain the population mean

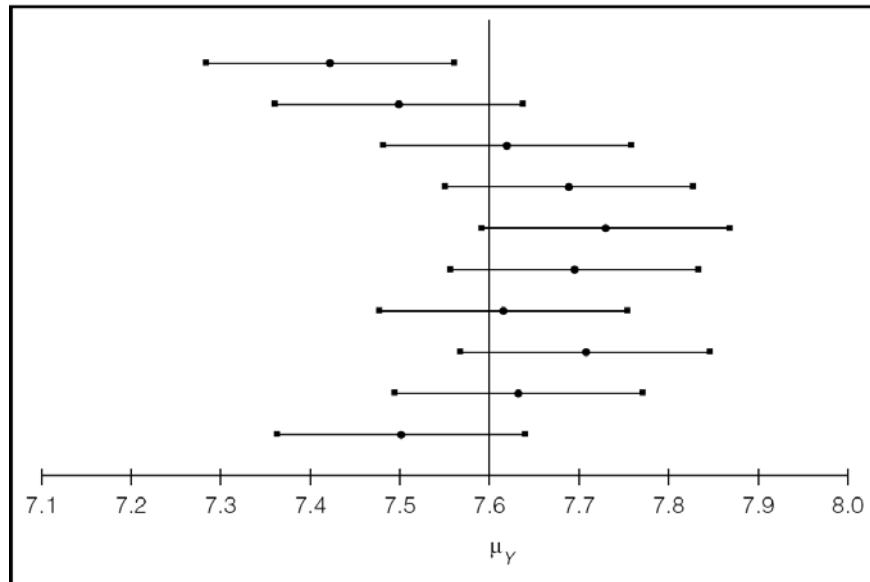
## Confidence Interval Width



### A 95% confidence interval



### 95 Percent Confidence Intervals for Ten Samples



## Confidence Interval Width

$$\bar{Y} \pm Z \left( \frac{s_Y}{\sqrt{N}} \right)$$

### • Larger Sample Size →

- Smaller standard errors
- Sampling distributions that are more clustered around the population mean
- confidence intervals will be narrower and more precise

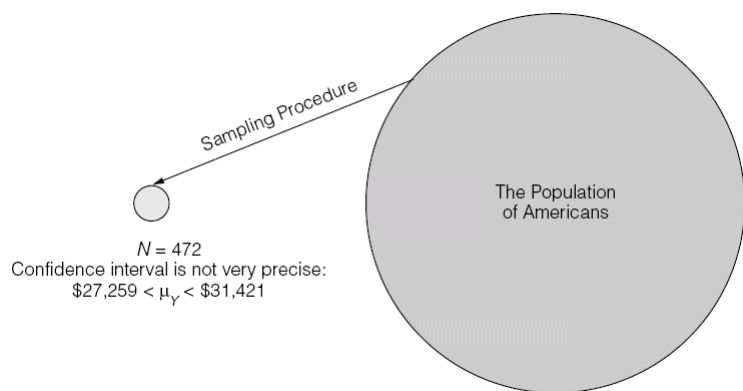
## Confidence Interval Width

$$\bar{Y} \pm Z \left( \frac{s_Y}{\sqrt{N}} \right)$$

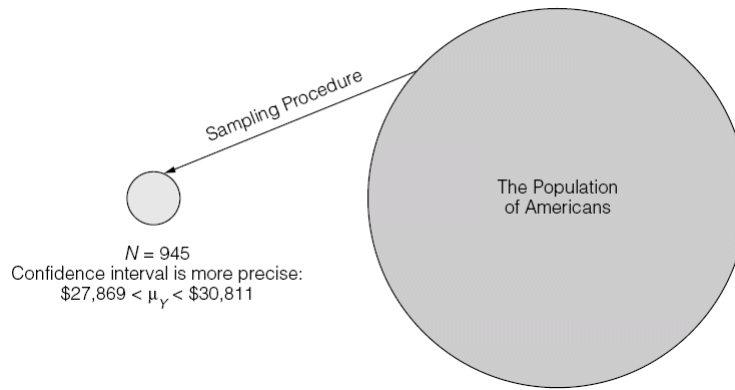
**Standard Deviation** – Smaller sample standard deviations → smaller, more precise confidence intervals

Note: *Unlike sample size and confidence level, the researcher plays no role in determining the standard deviation of a sample.*

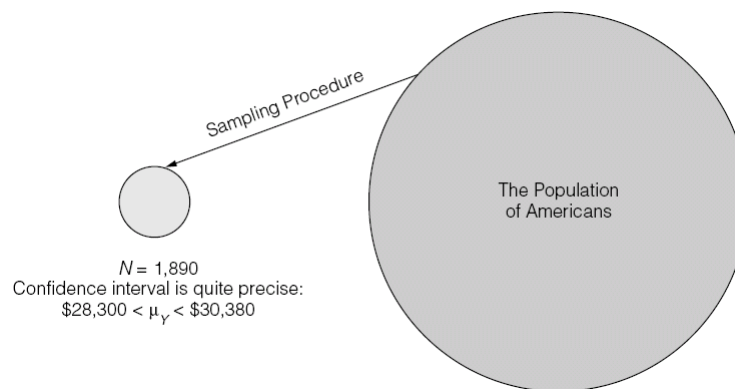
## Sample Size and Confidence Intervals

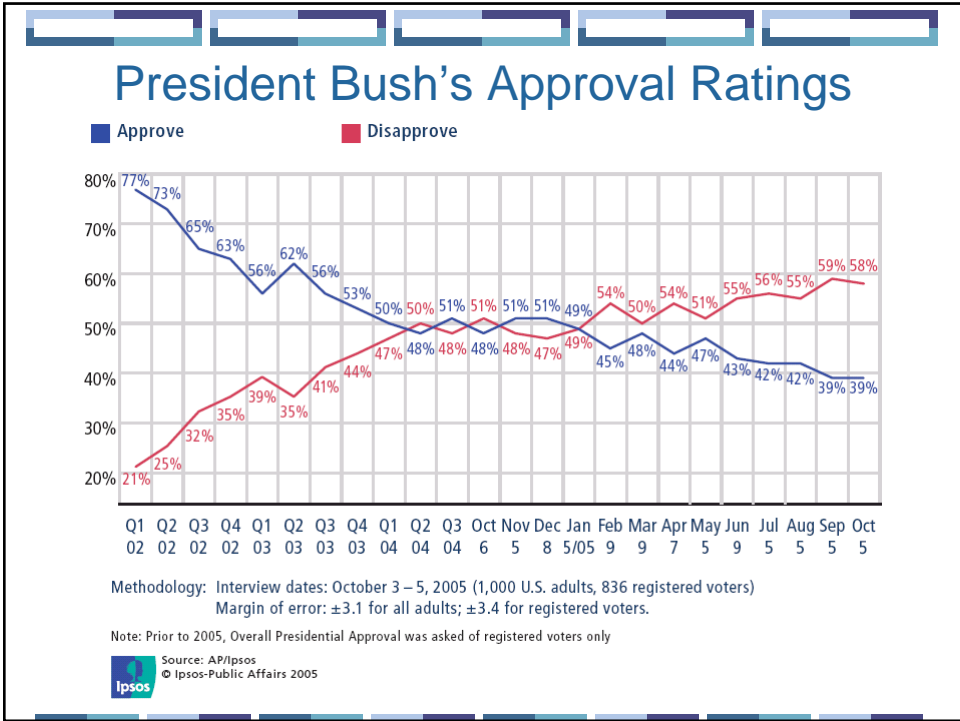


## Sample Size and Confidence Intervals



## Sample Size and Confidence Intervals





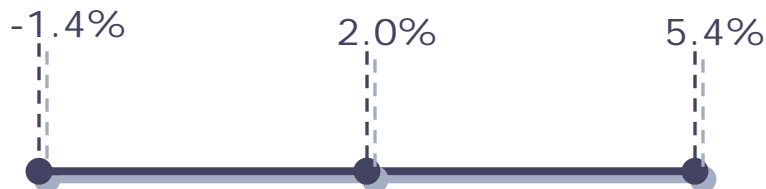
## 2.0% Approval Rating

### 3.4 % Margin of Error for Results

- “In what may turn out to be one of the biggest free-falls in the history of presidential polling, President Bush's job-approval rating among African Americans has dropped to 2 percent, according to a new NBC/Wall Street Journal poll.”
- 3.4% margin of error reported for poll

“A Polling Free-Fall Among Blacks,” By Dan Froomkin, Special to washingtonpost.com, Thursday, October 13, 2005, 3:09 PM

## President Bush's Approval Rating Among Blacks



**2.0% Approval Rating**  
**3.4 % Margin of Error for Results**

"A Polling Free-Fall Among Blacks," By Dan Froomkin, Special to  
washingtonpost.com, Thursday, October 13, 2005, 3:09 PM

## Conclusion

- Data quality is absolutely essential to produce valid statistical work
  - Sampling procedures must be correct (e.g., random, "internet polls," telephone polls, representative, etc.)
- Samples → Population
  - Larger sample size → more reliable and precise results
  - Large samples that are not representative are invalid
- Sampling error can be reduced but never eliminated
- Confidence intervals are essential to interpreting survey results