Stats 95

Two Branches Of Statistics

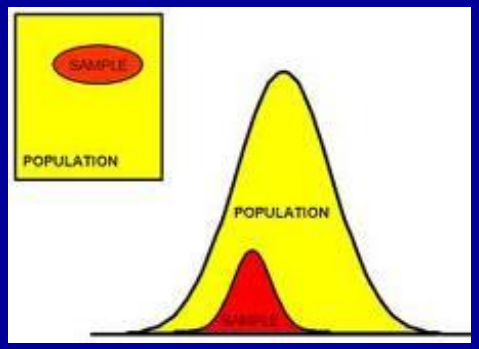
Descriptive

- Organize
- Summarize
- Communicate

  … a body of observed data
- Describe a **Population** or a **Sample**

Inferential

- Using sample data to make estimates of the rest of the population
- Can infer only from a **Sample to the Population**



Populations & Samples

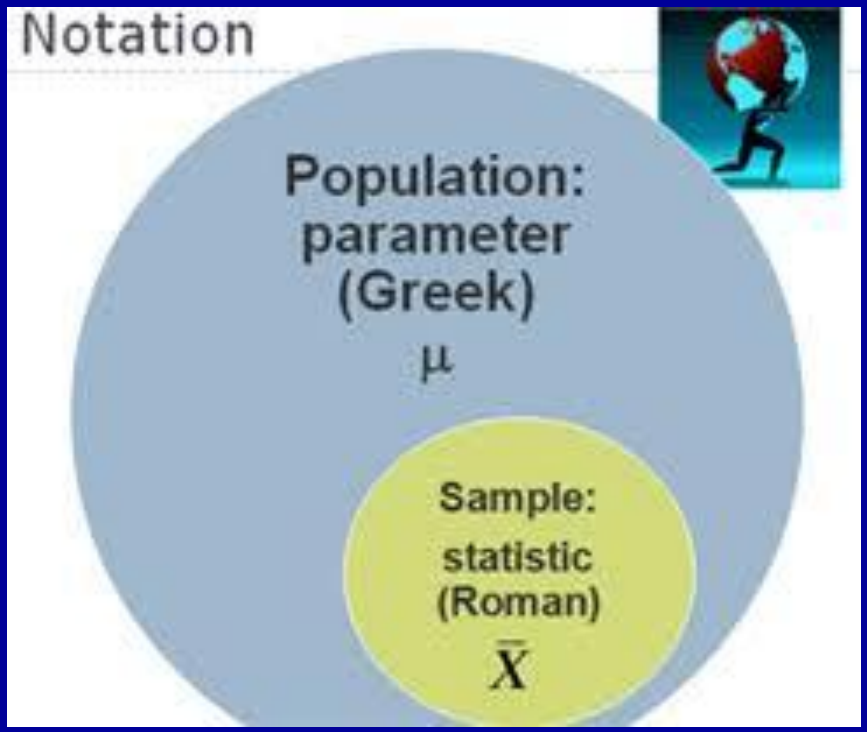
or Why Stats Is All Greek To Me

Population

- Includes ALL POSSIBLE OBSERVATIONS
- Greek Letters

Sample

- A set of observations from a population
- Roman Letters



Data & Experimental Variables

- Data Variables: Types of Data
- Experimental Variables
  - Independent Variable
  - Dependent Variable
  - Extraneous Variables
  - Confounding Variable
    - Vary systematically with Independent Variable
    - E.g., Income & Health, very often wealthy people are the healthy ones
    - Can (usually) solve by good design or by limiting scope of conclusion, or controlling statistically
- http://stattrek.com/experiments/experimental-design.aspx

Errors

Two types of Errors

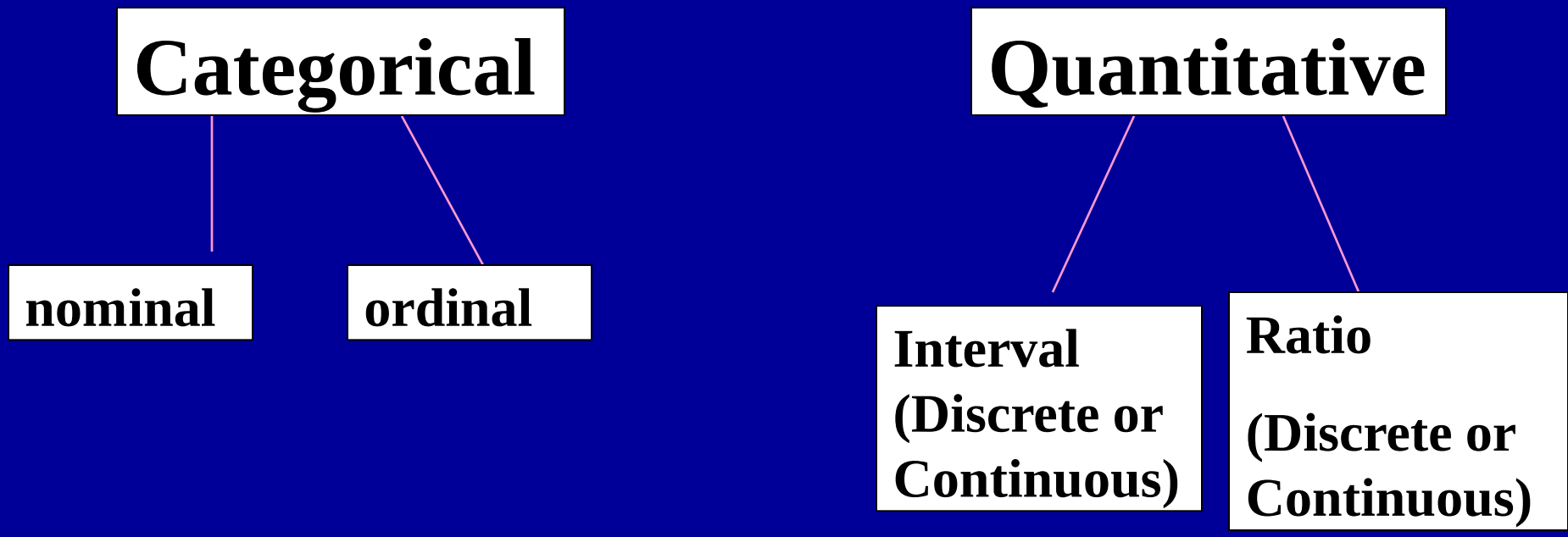
Random Errors

- You measure the mass of a ring three times using the same balance and get slightly different values: 17.46 g, 17.42 g, 17.44 g
- Take more data. Random errors can be evaluated through statistical analysis and can be reduced by averaging over a large number of observations.

Systematic Errors

- The cloth tape measure that you use to measure the length of an object had been stretched out from years of use. (As a result, all of your length measurements were too small.)
- The electronic scale you use reads 0.05 g too high for all your mass measurements (because it is improperly tared throughout your experiment).
- Systematic errors are difficult to detect and cannot be analyzed statistically, because all of the data is off in the same direction (either to high or too low). Spotting and correcting for systematic error takes a lot of care.

Types of Variables: Overview

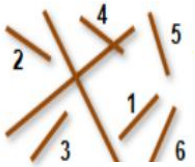


Data Types: What Are You Counting?

Nominal — **Mutually Exclusive Categories**

Desktop Laptop iPhone

Ford, GM, Honda



Ordinal — **Mutually Exclusive Categories, Ordered**

Freshman, Junior Senior;

Agree, Neutral, Disagree

Interval — **Mutually Exclusive, Ordered, Regular Intervals**

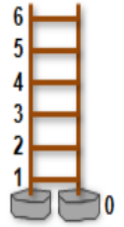
Degrees Celsius

Nielsen Ratings

Ratio — **Mutually Exclusive, Ordered, Regular Intervals, Proportional (Absolute Zero)**

Degrees Kelvin

$$$ Time



Clinical Data Example

- 1. Kline et al. (2002)
  - The researchers analyzed data from 934 emergency room patients with suspected pulmonary embolism (PE). Only about 1 in 5 actually had PE. The researchers wanted to know what clinical factors predicted PE.
  - I will use four variables from their dataset today:
    - Pulmonary embolism (yes/no)
    - Age (years)
    - Shock index = heart rate/systolic BP
    - Shock index categories = take shock index and divide it into 10 groups (lowest to highest shock index)

Categorical Variables

- Nominal variables – Named categories
  Order doesn't matter!

  - The blood type of a patient (O, A, B, AB)
  - Marital status
  - Occupation

Categorical Variables

- Ordinal variable – Ordered categories. Order matters!
  - Staging in breast cancer as I, II, III, or IV
  - Birth order—1st, 2nd, 3rd, etc.
  - Letter grades (A, B, C, D, F)
  - Ratings on a scale from 1-5
  - Ratings on: always; usually; many times; once in a while; almost never; never
  - Age in categories (10-20, 20-30, etc.)
  - Shock index categories (Kline et al.)

Quantitative Variables

- Discrete Numbers – a limited set of distinct values, i.e., whole numbers.

  - Number of new AIDS cases in CA in a year (counts) (ratio/interval)
  - Years of school completed (ratio)
  - The number of children in the family (cannot have a half a child!) (ratio)
  - The number of deaths in a defined time period (cannot have a partial death!) (ratio)
  - Roll of a die (Interval)

Quantitative Variables

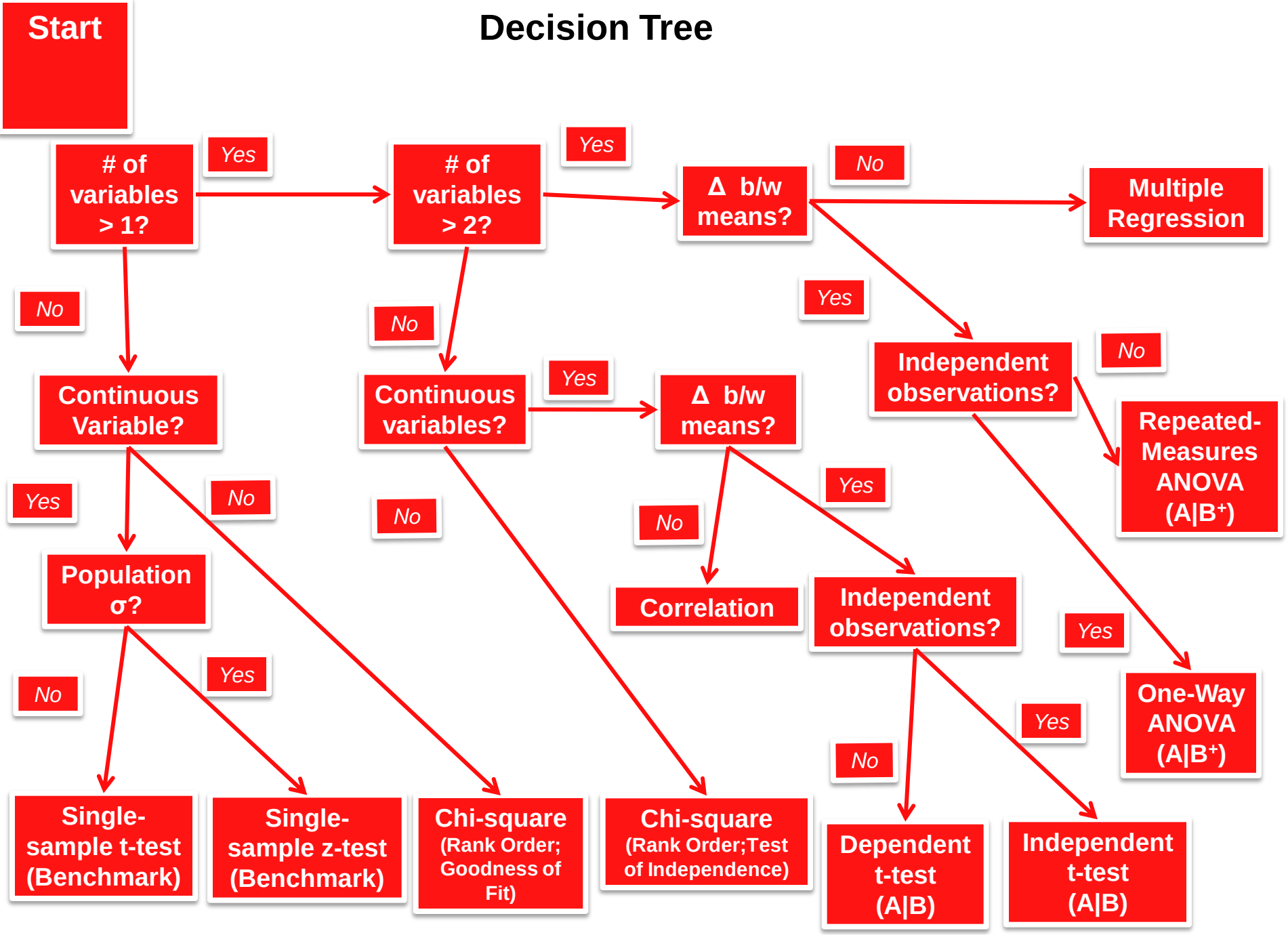
- Continuous Variables - Can take on any number within a defined range.
    - Time-to-event (survival time)
    - Age
    - Blood pressure
    - Serum insulin
    - Speed of a car
    - Income
    - Shock index (Kline et al.)

Summary

Experience Counts

Usability Test	Statistical Test	Data Type	# of Variables	Dependence / Independence	Sample Size
Rank Order	Chi-square	Nominal, Ordinal	1 or 2	Dependent	5 per cell
Benchmark	Single-sample t-test	Ratio, Interval	1	Independent	12
A|B (Ind)	Independent sample t-test	Ratio, Interval	2	Independent	12 x 2
A|B (Dep)	Dependent sample t-test	Ratio, Interval	2	Dependent	8-12
Survey	Correlation	Ratio, Interval	2	Dependent	8-12
A|B^{+}	ANOVA	Ratio, Interval	2 or more	Independent	8-12 / level
A|B^{+} (RM)	Repeated-Measures ANOVA	Ratio, Interval	2 or more	Dependent	8-12

Decision Tree



The End

- Next week: Bring a Bag of Chocolate Chip Cookies!

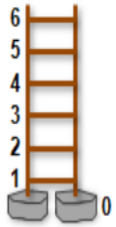
- Back Up Slides

What Are You Counting?

Ratio

Proportional
(Absolute Zero)

$$$
Time



Interval

Regular Intervals

Ratings
Scale

Ordinal

Ordered but Irregular

Lickert Scale 1-7

Agree Neutral
Disagree

Nominal

Labels or Names

Desktop Laptop
iPhone

Data Variables

Quantitative Variables: Discrete or Continuous

- **Discrete**
  - Nominal
    - Purely qualitative, no ordering is possible, category or name
    - E.g., Toyota, Ford, Honda; Breast Cancer, Throat Cancer, Lung Cancer
  - **Ordinal**
    - Where there is a sequential order, but the intervals are irregular
    - E.g., First, Second, Third; Freshman, Sophomore, Junior

- **Continuous (can be Discrete)**
  - **Scale**
  - Sequential order, and the intervals are regular, but values are not proportional, no Absolute Zero
    - E.g., degrees Fahrenheit or Celsius
  - Ratio
    - Regular intervals which are proportional, there is an Absolute Zero
    - E.g., degrees Kelvin, Height

Where Does Data Come From?

- Case Studies
- Experiments
- Naturalistic / quasi-experimental
- Longitudinal
- Cross-sectional
- Surveys

The Science of Observation

- Theory: Statement of relationship
  - Amongst events otherwise unrelated
  - For which there is already supportive data (a *theory* is more than an *idea*)
- Hypothesis: Statement of possible relationship between variables which follow logically (but sometimes unexpectedly!) from theory

Experiments

Looking for a Difference

Between-Groups

- Participants experience only one level of the independent variable.
  - E.g., one group performs their driving license exam after consuming alcohol, and the other group performs the test sober.

Within-Groups

- Participants experience all levels of the independent variable.
  - All members take the driving license exam twice, once after consuming alcohol, and again, sober.

The Science of Observation

- Experiment: Test of hypothesis with 2 critical elements
- (1) Manipulation
  - independent variable
  - dependent variable—measured
  - control group
  - experimental group
- (2) Randomization (controls for 3rd variable)
  - versus self-selection

Operational Definition

- DVs that aren't subject to biased responses
- Examples:

– Is a painting in a museum popular?

- There will be increased wear on the carpet near it.

– Did a dental flossing lecture work?

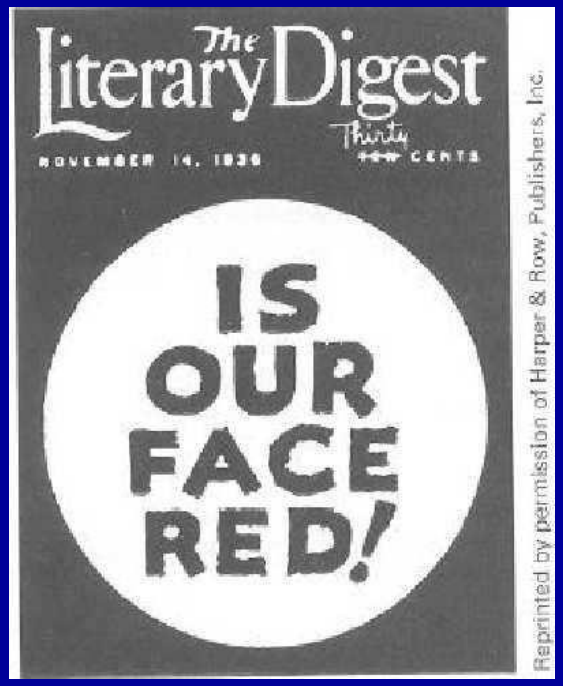
- Students will have cleaner teeth the next day.

– Did a safer sex intervention for commercial sex workers work?

- There will be more condoms discarded in the park they work in.

Random Sampling and Confounding Variables

- National polls in this country can be dated back to the early 1800's.
- The largest of these was *Literary Digest*, tried to
- determine the outcome of the 1936 presidential election (Franklin D. Roosevelt and Alfred Landon).
- They sent out 10 million questionnaires using lists from phone books, vehicle registration lists, and club memberships across the country, from which 2.4 million responded!
- Today, polls typically rely on the opinions of 500-1000 people.
- Potential problems???

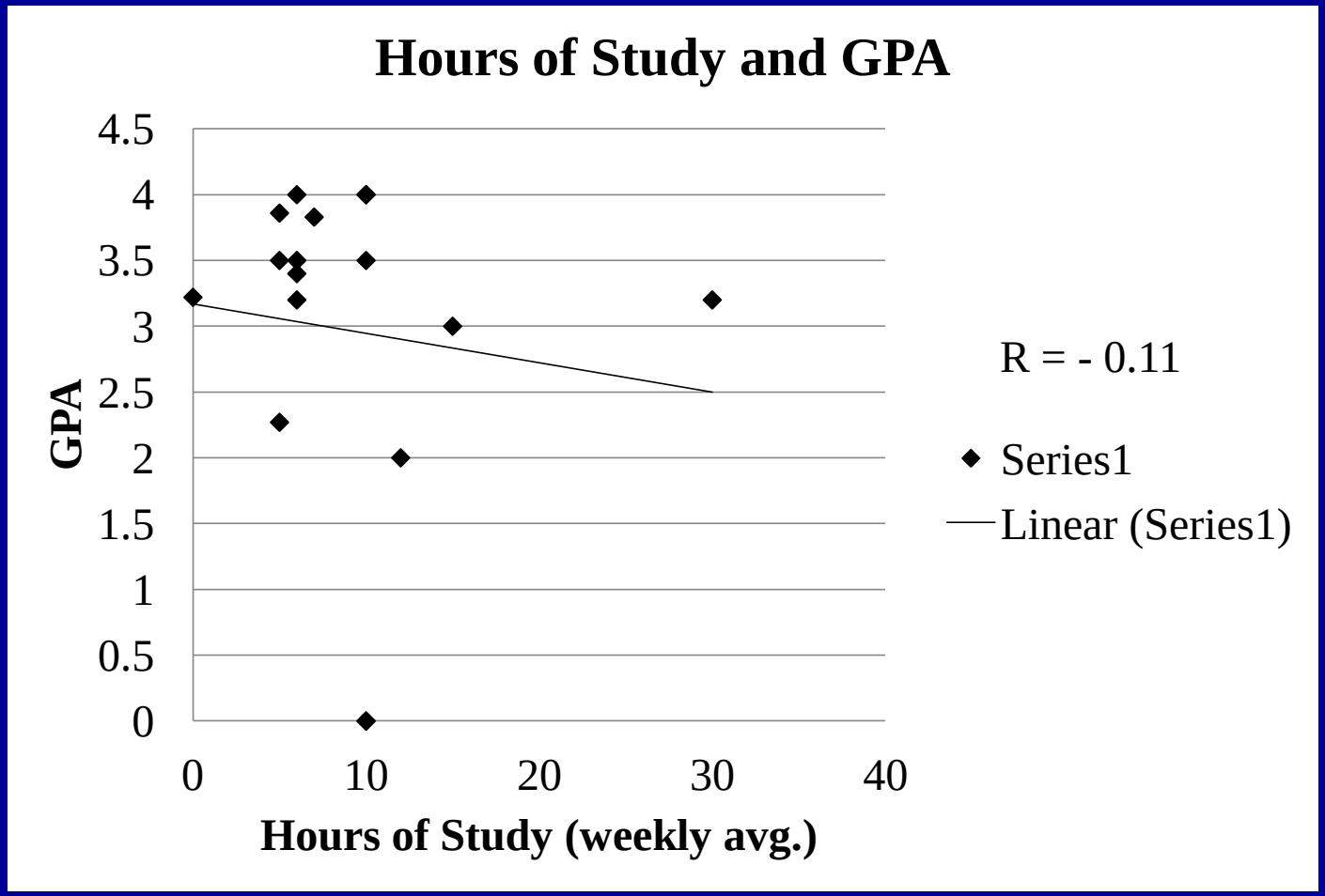


Correlation

Looking for a Relationship

- Correlation measures the strength of a relationship between two variables, and the direction of the relationship, *positive* or *negative*.

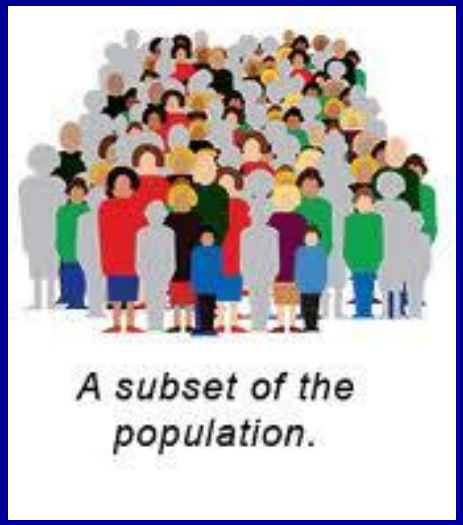
Data From Survey

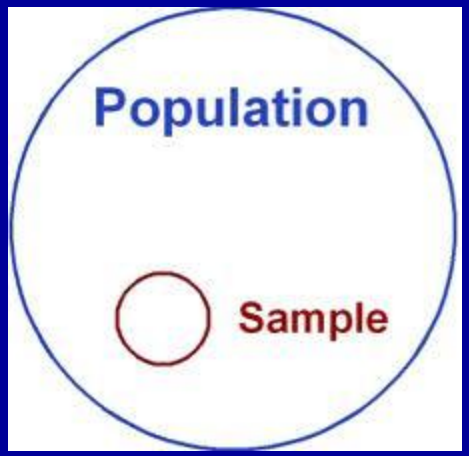


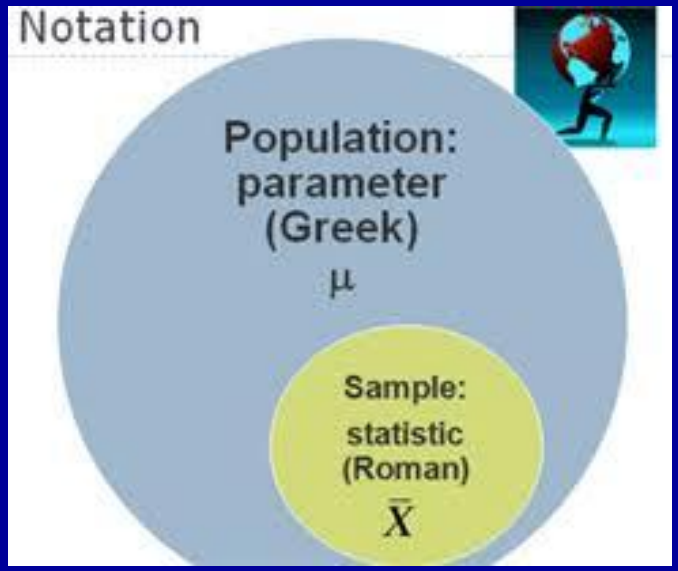
The Science of Observation

- Validity—able to draw accurate inferences
  - construct validity: e.g., describing what intelligence is and is not, "construct" refers to the "theory"
  - predictive validity: over time you find X predicts Y

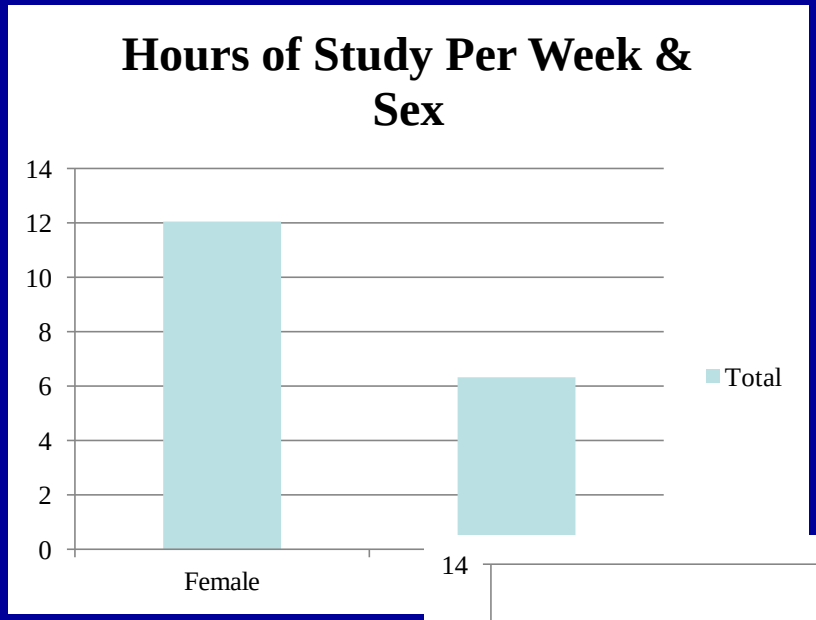
- Reliability—same result each time?

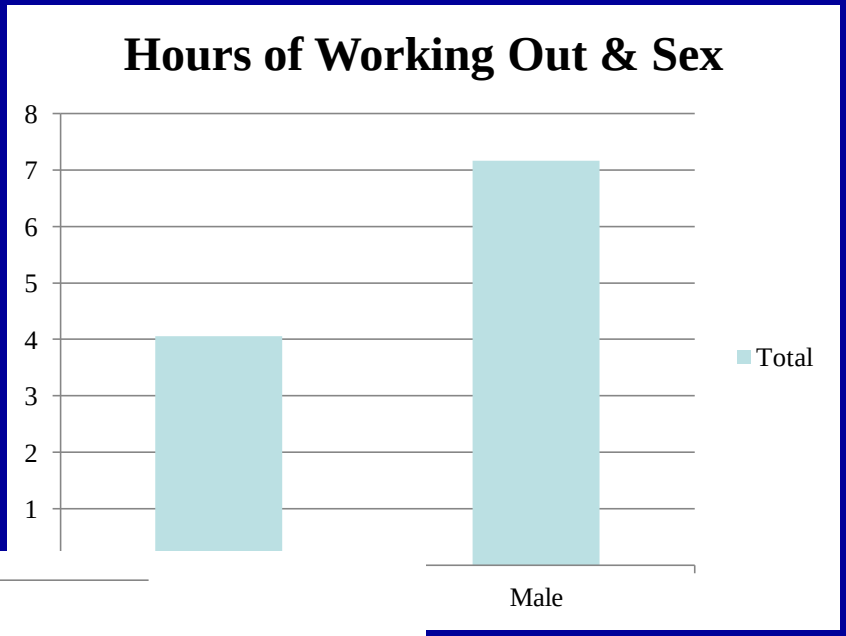

A subset of the population.

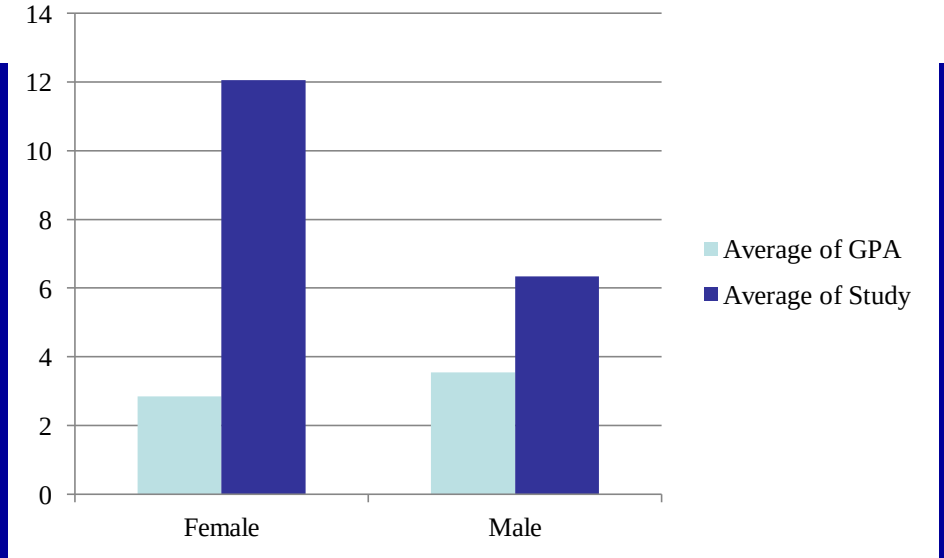

Population
Sample

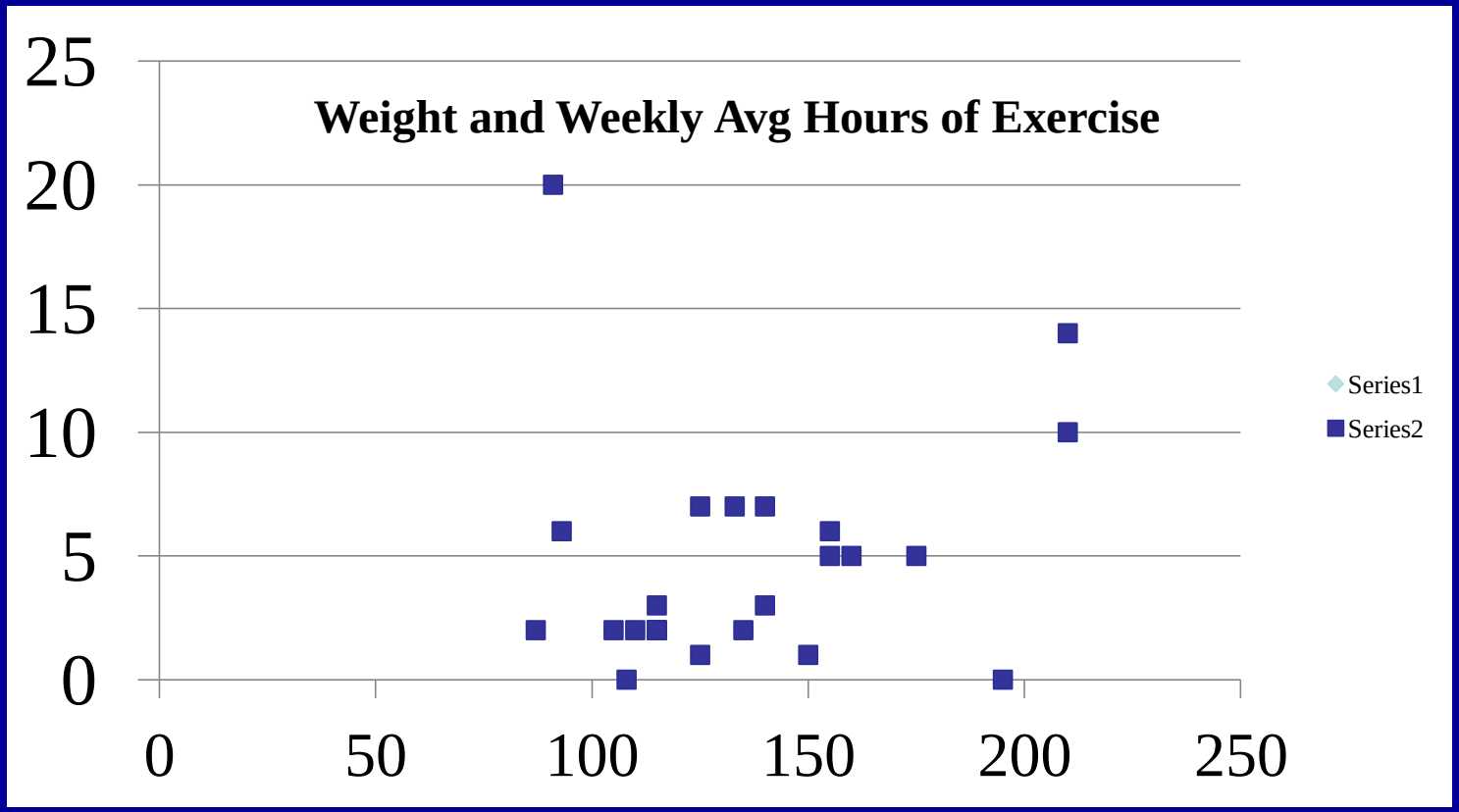

Notation
Population: parameter (Greek)
μ
Sample: statistic (Roman)
$\bar{X}$

Data From Survey








Weight and Weekly Avg Hours of Exercise
25
20
15
10
5
0
0
50
100
150
200
250
Series1
Series2