Lesson 1
The Cycle of Research

Definitions of Terms

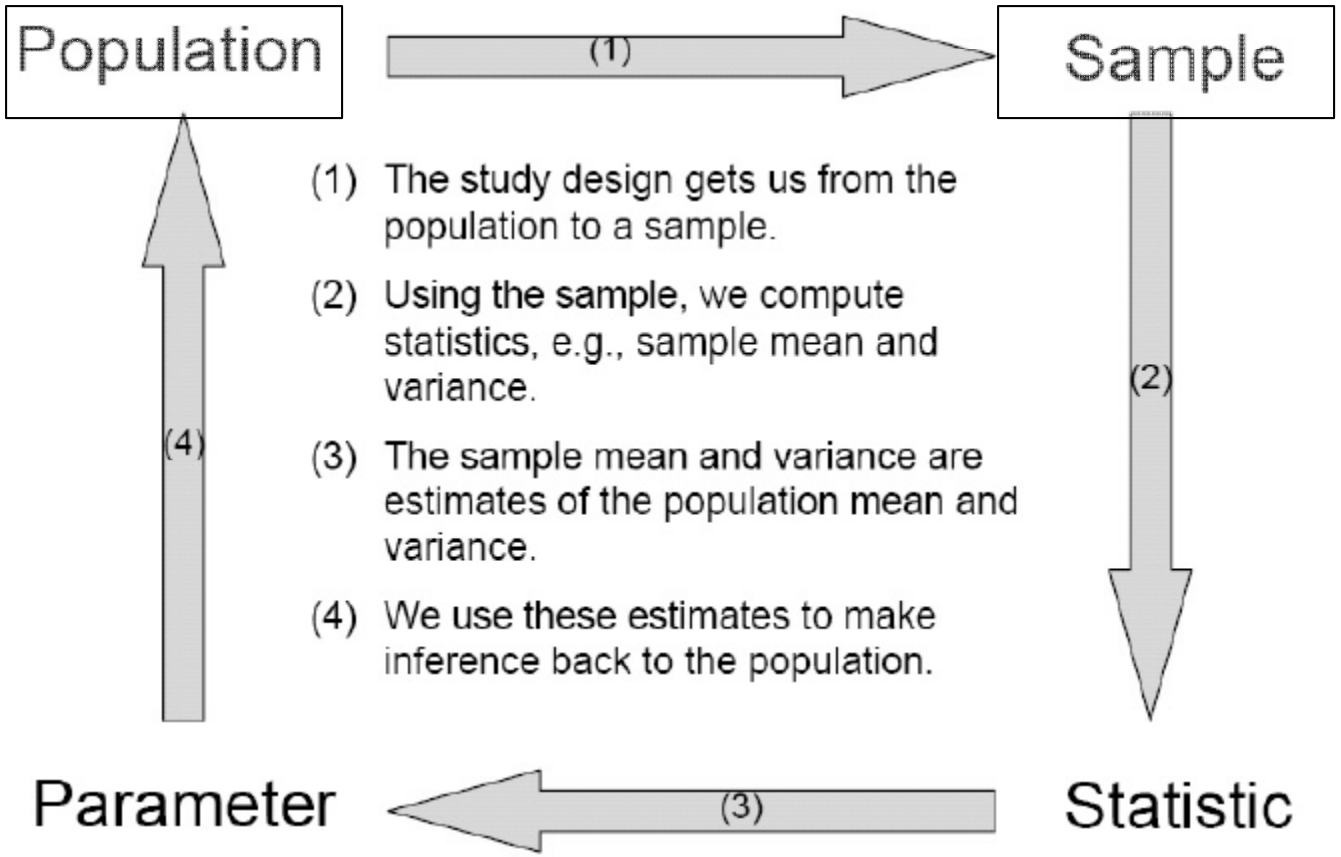
Study Design

Measurement Scales

Understanding the Research

- The description of any medical or health research study should provide answers to the following:
  - How were the data collected?
    - What was the sampling method?
    - What is the study design?
- In addition to understanding how the data were collected, the measurement scale of the data needs to be considered
- Both the type of data collected (measurement scale) and how the data were collected determine the appropriate descriptive statistics, graphs, and analysis methods to use.

Cycle of Research



Samples and Populations

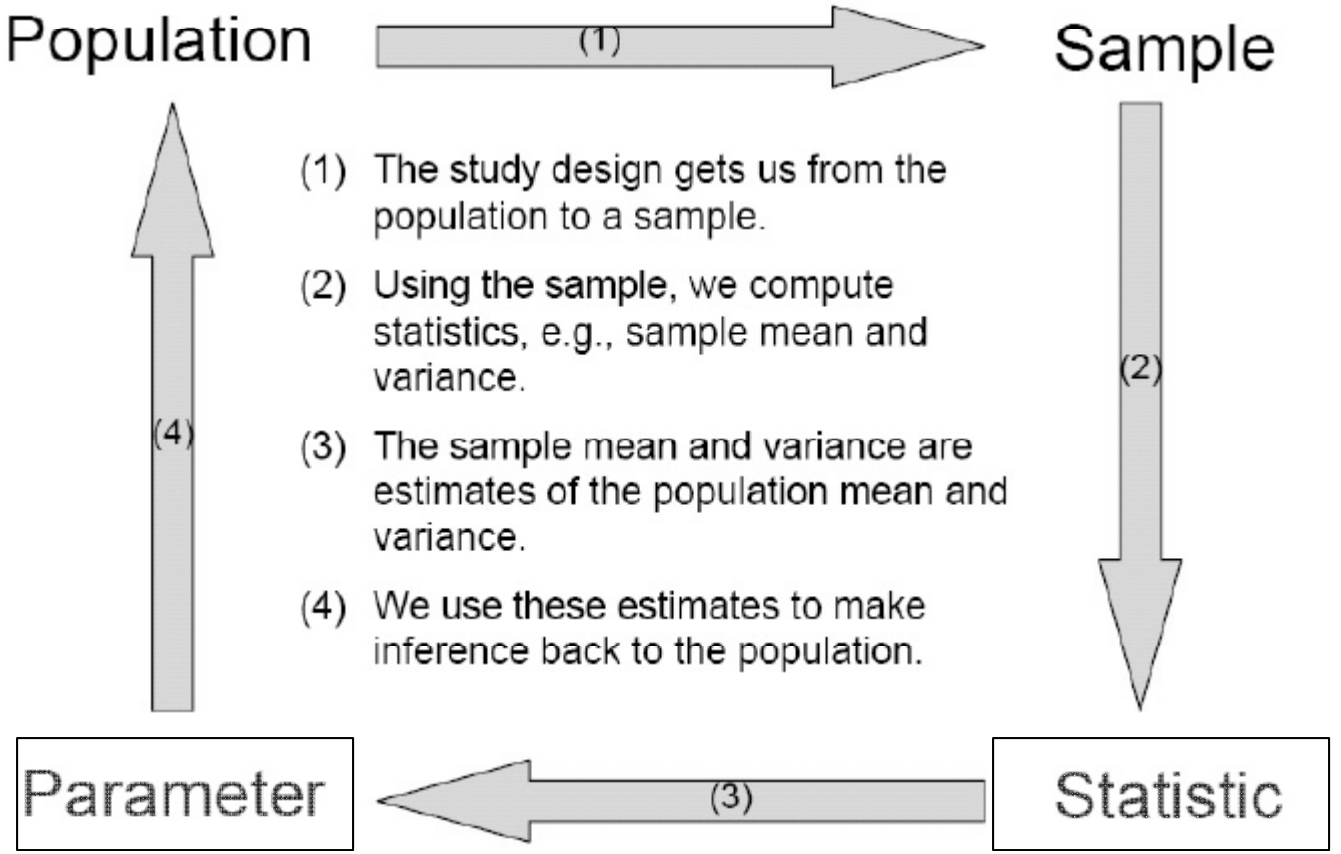
- Population: complete set of individuals that share some common characteristic(s).
  - the group we are interested in learning about.

- Sample: A subset of the defined population.
  - the group that is actually studied and from which data are collected.

Why study samples?

- If the goal is to make some conclusion(s) about the population, why not just collect data on the entire population?
  - It can be too expensive and/or time-consuming to collect data for the population
  - Study of the population may be impossible
  - Measurements from a sample may be more accurate than measurements from the population

Cycle of Research



Samples and Populations

- Statistic: A value obtained from a sample.
  - i.e. average weight of 1,000 boys sampled from the U.S.

- Parameter: A population variable, whose value is unknown.
  - i.e. average weight of ALL 10 year old boys in the U.S.

Random Sampling

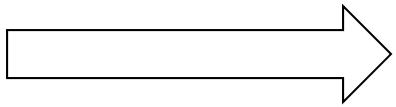
- Generalization of study results is only valid if the sample is representative of the larger population that it is drawn from.
- Random samples tend to be representative of the population from which the sample is drawn.
- Random samples are also called Probability samples because the probability of each subject (or unit) being included in the sample is known.

Simple Random Sample (SRS)

- Basic idea: Select a sample from the population so that:
  - Each subject in the population has equal probability of being selected
  - Selection of one subject does not influence the probability of selecting another subject
- Formal definition: A sample of size n selected by a procedure that gives every sample of size n the same probability of being selected
- SRS removes selection bias and results in a representative sample

Selection Bias

- Selection Bias: “A systematic tendency to favor the inclusion in a sample of selected subjects with particular characteristics while excluding those with other characteristics.”
  - Pocket Dictionary of Statistics

Garbage in  Garbage out

Bias (in general): a systematic deviation from the truth

How to obtain a Simple Random Sample (SRS):

Obtaining a simple random sample requires that all possible subjects are identified.

Any of the following methods will result in a random sample of size n

- Place an ID for each subject in a hat and randomly select a sample of size n
- Use a table of random numbers to select a sample of size n from the list
- Use a computer program to select a random sample of size n from the list

Sampling with and without replacement

- Sampling with replacement:
  - After a subject is selected, return the ID to the list
  - Possible for a single subject to be selected twice
- Sampling without replacement
  - Once selected the subject is not available for future selection
  - Most often, sampling without replacement is used
- For large populations, the difference between these is negligible

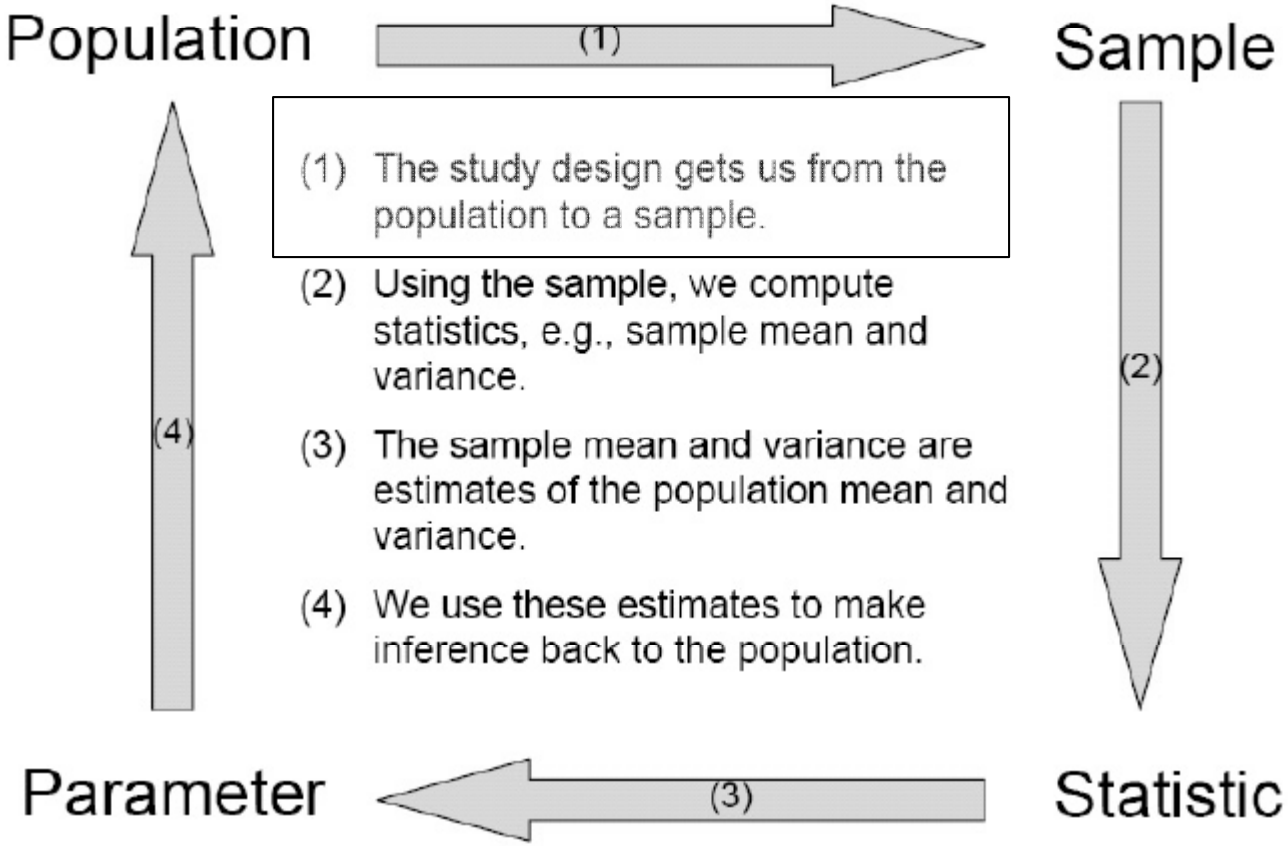
Stratified Random Sampling

- For stratified random sampling the population is divided into groups with similar characteristics. These groups are called ‘strata’. A simple random sample is selected from each strata
- Possible stratification variables are
  - Gender
  - Clinic
  - Age groups
- This is a useful method when the subgroups vary in the study measurements
- Requires that information about the stratification variable(s) is available for all subjects

Consider the sampling Method

- When reading the medical literature consider the following:
  - Is the study sample representative of the population of interest?
  - What was the sampling method? If a random sample was used, the statistical inference is valid
  - If complete population data were collected, no inference is necessary. Descriptive statistics completely represent the population when complete population data are available.

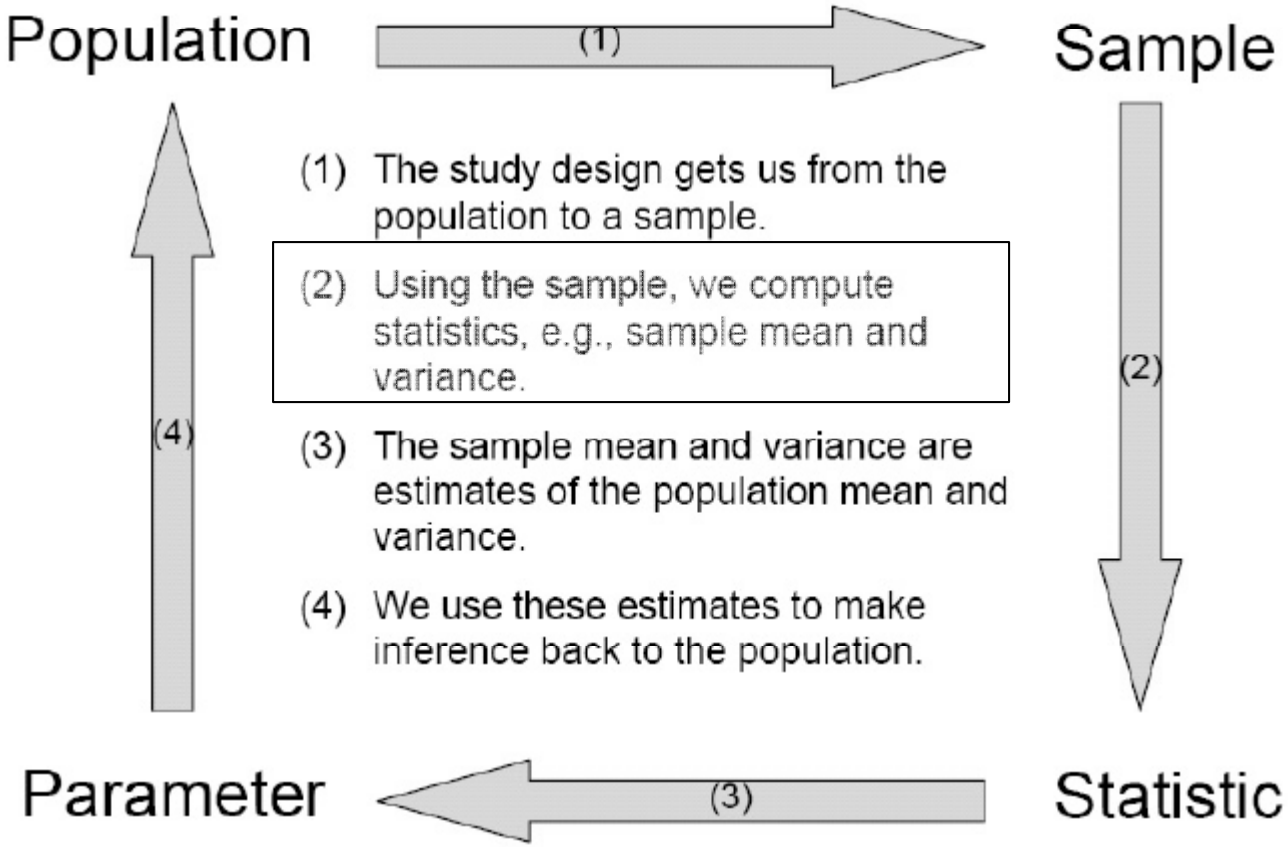
Cycle of Research

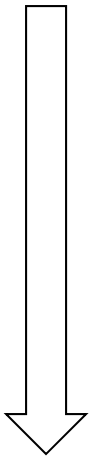


⟹ Study Design

- Twelve 18-24 year old males in a jogging club participated in a study of brain gycogen among physically fit individuals.
  - The population of interest is (potentially) 18-24 year-old males.
  - The sample is this group of twelve joggers.
  - The sample design is how these twelve where selected and studied.

Cycle of Research





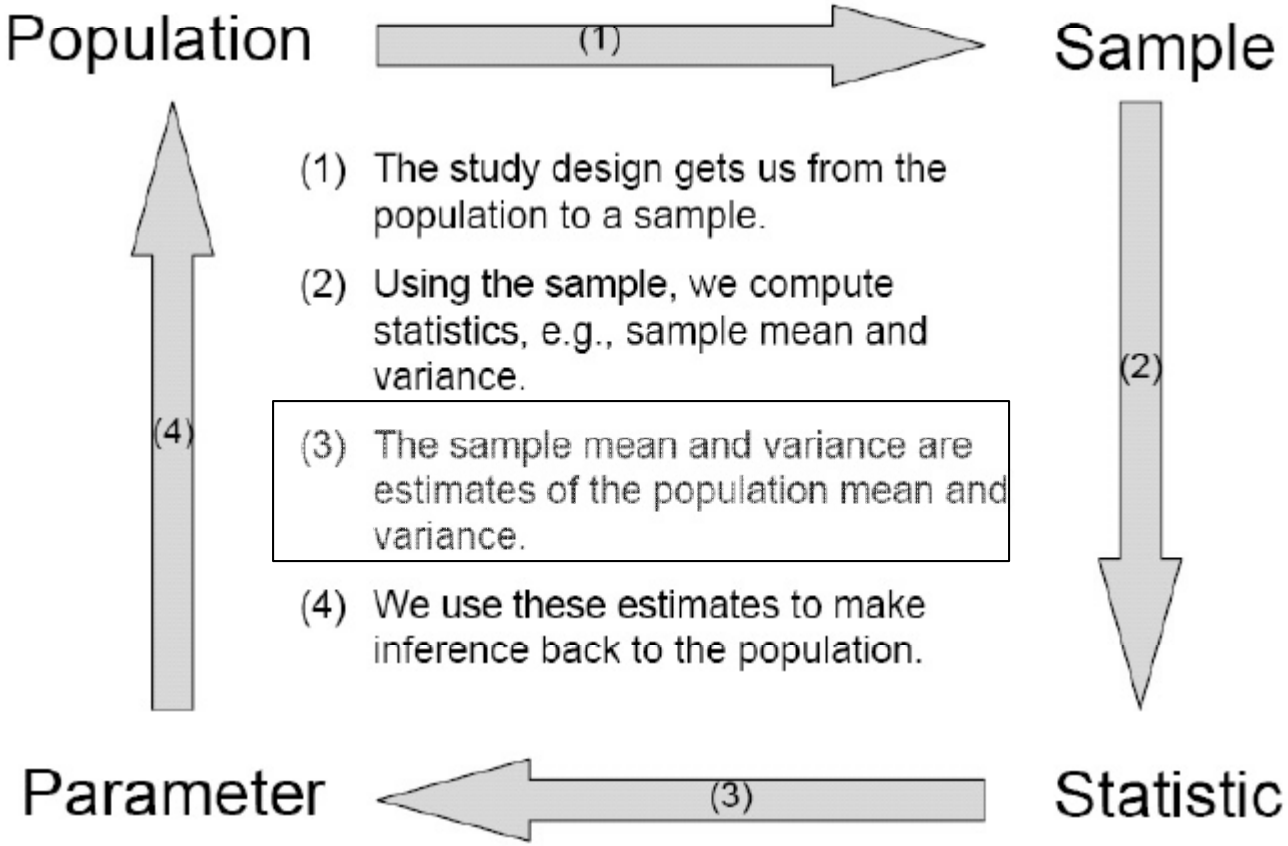
Calculation

- The twelve males were weighed, resulting in weights in pounds of :

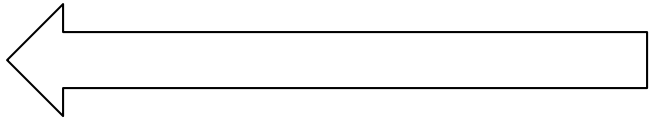
{129,134,136,140,141,142,144,155,158,162, 165,191}

  - The statistics how obtained were the average and standard deviation of the weights.

Cycle of Research

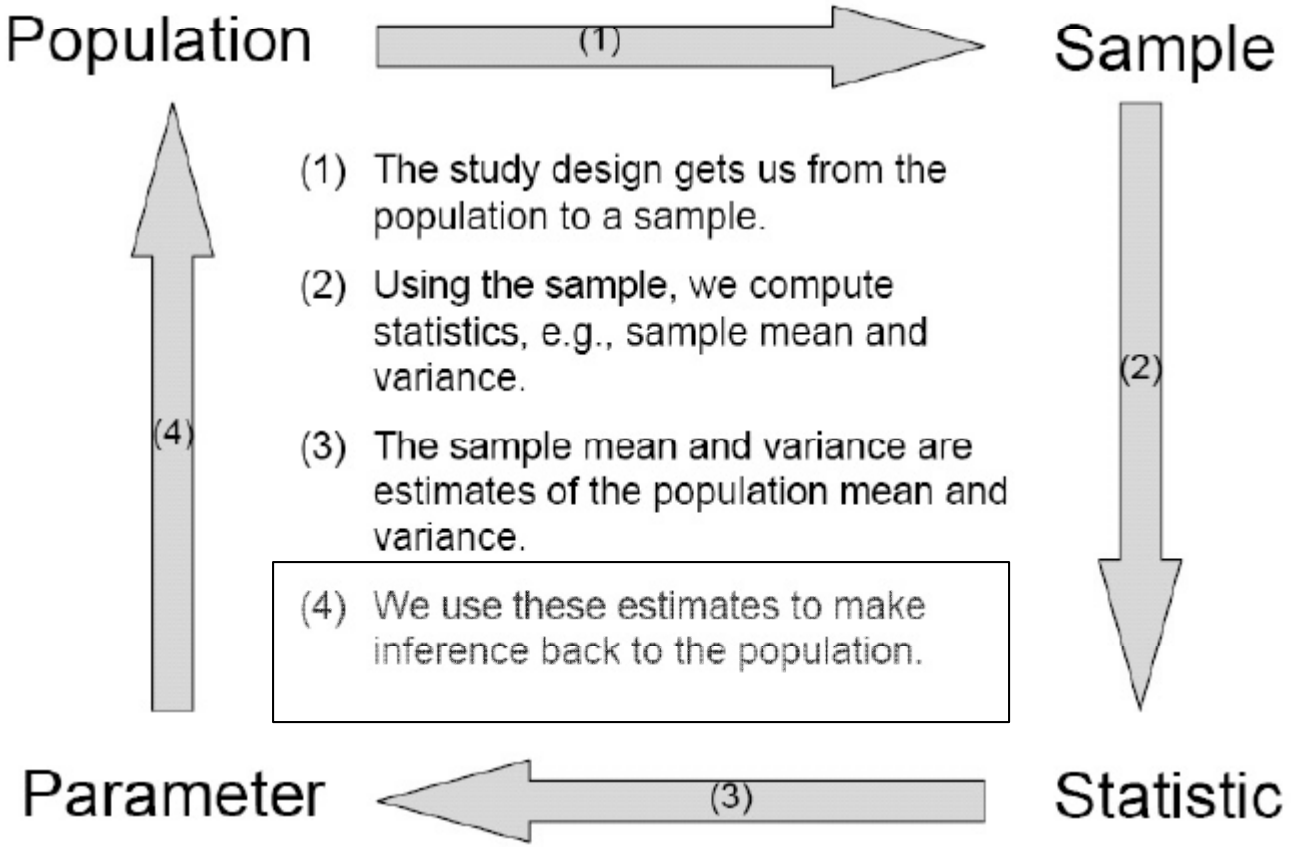


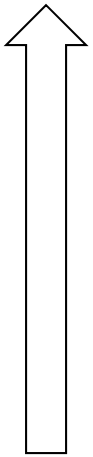
Estimation



- The sample statistics are our estimates of the population values:
- Average: 149.75 pounds
- Standard Deviation: 17.35 pounds

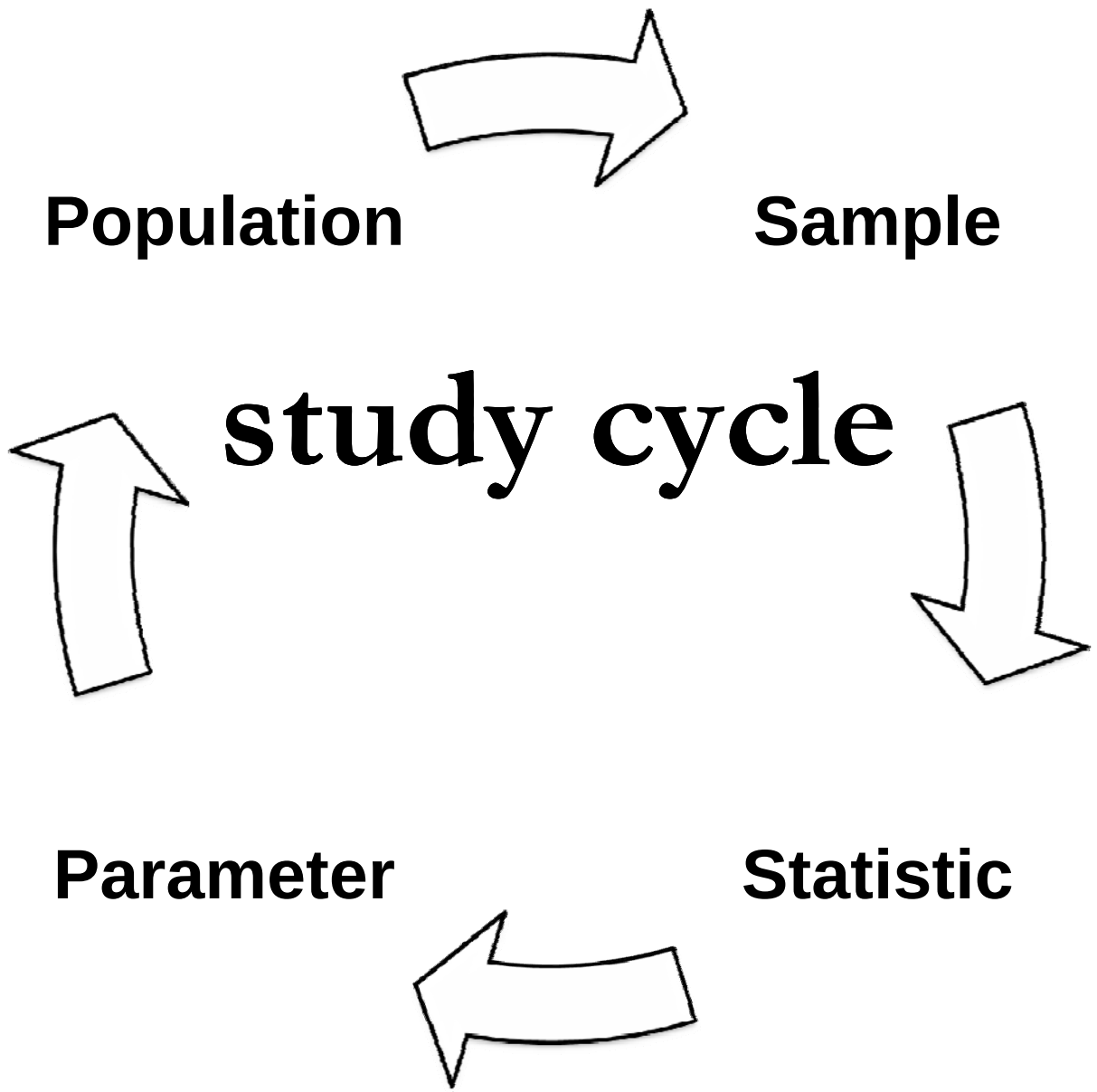
Cycle of Research





Inference

- By comparison with national health data from the CDC, we conclude that young athletic males have lower average weights than young males overall, and thus their brain glycogen results may not be generalizable to young males overall.


Population
Sample
study cycle
Parameter
Statistic

Consider This Study

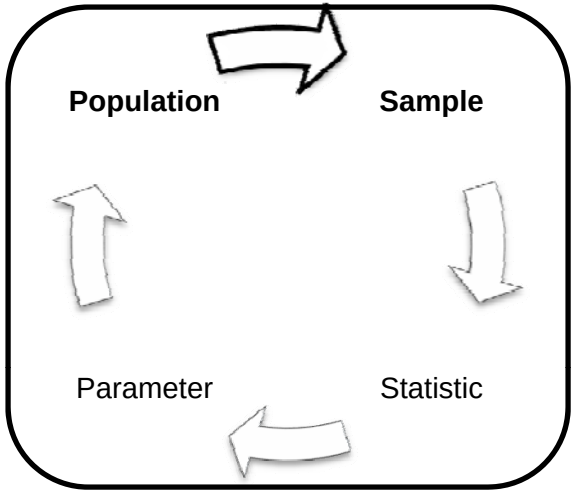
- Two thousand African-American women, ages 35 to 55 years, were surveyed to determine their adherence to recommended breast and cervical cancer screening schedules. At the time of the survey, each woman lived in the inner city of a southern metropolitan area.

Status	Household income <$5000	Education>high school	Not married or separated	Employed	Had Pap smear	had mammography
Yes	27%	68%	68%	80%	48%	35%
No	73%	32%	32%	20%	52%	65%

- The investigators concluded that African-American women in this age group who live in the inner city do not adhere to suggested screening schedules.

Question 1:

"___" corresponds to the <u>procedure</u> highlighted in this picture?

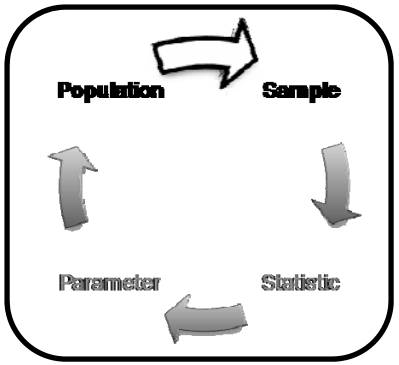


1. The investigators made the conclusion whether African-American women in this age group who live in the inner city adhere to suggested screening schedules.
2. The health and social-economic data of those women were recorded.
3. 2000 African-American women, ages 35 to 55 years and lives in the inner city , were surveyed in this study.

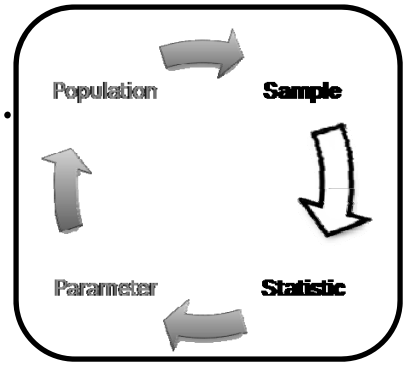
Question 2:
select the correct picture for the following procedure.

"Twenty-seven percent of the women had a household income under $5000."

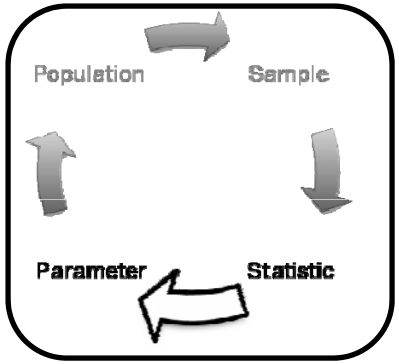
1.



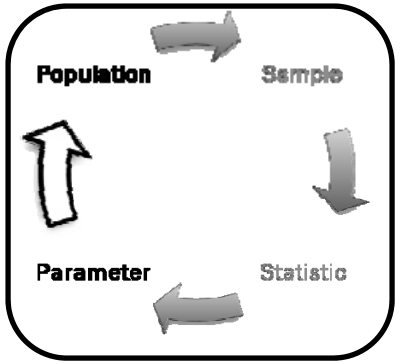
2.



3.



4.



Question 3:

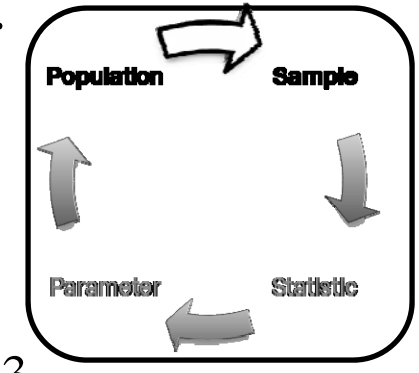
"Forty-eight percent had their most recent Pap smear" is ___?

1. Parameter
2. Statistic, used to estimate parameter
3. Statistic, also the parameter

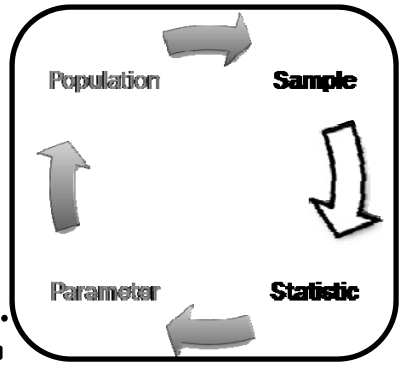
Question 4:
select the picture corresponding to the following statement.

"The investigators concluded that African-American women in this age group who live in the inner city do not adhere to suggested screening schedules."

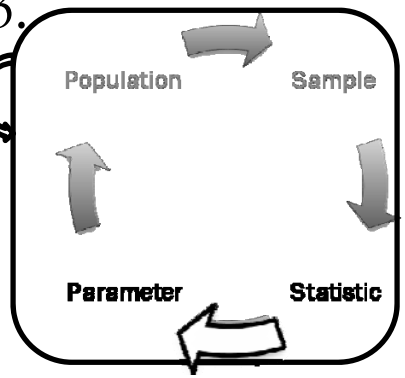
1.



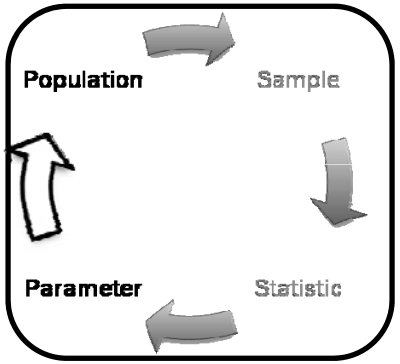
2.



3.



4.



More on "How was the Data Collected?"

Study Design

Classification of Study Designs

Study designs can be classified as either Observational or Experimental

Observational Studies: subjects are observed without any intervention by the investigator

- Case series
- Cross-Sectional
- Case-Control
- Cohort

Experimental Studies: a comparative study involving an intervention by the investigator

- Controlled studies
- Uncontrolled studies

Case Series

A case series is observational

- Subjects are observed and the outcomes or characteristics of the observation are described
  - Case series typically have a small number of subjects
- These are not planned studies so they are not always considered ‘studies’
- Observations from a case series may lead to research questions for more rigorously designed studies

Case-Series Example

- In 1981, Gottlieb et al. reported a rare form of pneumonia and unusual infections in 5 young men
  - Gottlieb, et al. *Pneumocystis carinii* pneumonia and *mucosal candidisasis* in previously healthy homosexual men: Evidence of a new acquired cellular immunodeficiency. *New England Journal of Medicine,* 1981; 305(24): 1425-1431
- Other similar case series were reported
- By 1982 this condition was named: Acquired Immune Deficiency Syndrome (AIDS)

Cross-Sectional Design

- Cross-sectional studies collect data from a sample of subjects at one point in time.
  - Provides a ‘snapshot’ of the characteristics of interest
  - Often large number of subjects

- Collected data can be used to explore relationships between variables at a point in time
  - Possible outcome variables: Disease, Death, Event
  - Explanatory variables: Exposure to some Risk Factor
  - Demographic characteristics: age gender, ethnicity, etc.

Cross-sectional Design: Cons

- Cross sectional studies are not useful for identifying the ‘cause-effect’ relationship between variables
- There is potential for bias in data collected by surveys
  - Volunteer bias
  - Non-response bias

Case-Control Design

- Case-Control studies are useful for exploring the relationship between disease (or some other outcome) and possible risk factors
- Subjects are selected based on the outcome
  - 'Cases' – already have the disease
  - 'Controls' – individuals from a similar population who do not have the disease
- Determine **past** exposure to suspected risk factor(s) for the disease or condition from
  - Interviews with subjects or family members
  - Medical records
- These are often called Retrospective studies because the outcome already occurred prior to inclusion in the study.

Case-Control Design Example

- Inskip, PD et al. Cellular telephone use and brain tumors. *New England Journal of Medicine* (2001), 344(2), 79-86
- Cases and Controls were identified
  - 782 cases with various types of brain tumors
  - 799 controls with nonmalignant conditions
- Exposure to cell phone use was evaluated by subject interview
- Study conclusion: “There was no indication of higher brain tumor risk among persons who had used hand-held cellular phones compared to those who had not used them.”

http://www.cancer.gov/cancertopics/factsheet/brain-tumors-cell-phones

Case-Control Design

- Drawbacks of case-control studies:
  - Misclassification Bias: It isn't always easy to determine disease status
  - Recall Bias: Subjects don't always accurately remember exposure history
  - Cannot be used to estimate the prevalence of a condition.
  - Selection of appropriate controls can be difficult
- Advantages of case-control studies:
  - Can obtain an adequate number of cases for studies of rare diseases
  - Economical and quick compared to prospective studies

Cohort Design

- Cohort studies follow a group of subjects (cohort) ***over time*** to determine disease or mortality status.
  - Sometimes called follow-up or longitudinal studies
- Members of the cohort can be classified as 'exposed' or 'not exposed' to one or more risk factors at the beginning of the study. For example:
  - smokers and non-smokers
  - overweight and normal weight
- The study follows the cohort over time to investigate the association between exposures and outcomes.

Cohort Study Design

- Most cohort studies are prospective
  - Risk factor exposure is determined in the present and information about the outcome is collected at a future time
- Cohort studies can also be retrospective
  - Exposure and outcome have already occurred
  - Requires having complete and accurate records to “follow” subjects from exposure to outcome
  - These are also called ‘historical cohort studies’

Cohort Study Example

- **NIH: Division of Epidemiology, Statistics & Prevention Research (DESPR): National Childrens Study**

- The National Children's Study is the largest and longest study of environmental effects on children's health ever conducted in the United States. The Study will follow 100,000 children from before birth to age 21 to examine the effects of environmental influences on their health and development.

- For more information about the Study and its progress, visit the Study Web site at http://www.nationalchildrensstudy.gov.

Cohort Design

- Drawbacks of cohort studies:
  - Time-consuming and costly.
  - Not suitable for very rare outcomes
  - Historical cohort design requires complete and accurate medical records.
  - Long follow-up time required can result in loss to follow-up (or drop out) of some subjects in the study

Experimental Studies

- Experimental Studies involve an intervention by the investigator. Some examples of interventions are:
  - New drug
  - New surgical procedure
  - Education / counseling program
  - Exercise program

Experimental Studies

- Clinical Trials are experimental studies of a medical treatment or procedure that involve human subjects.

- Two types of clinical trials: those with control subjects and those without control subjects.

Controlled Clinical Trials

- Control subjects in a clinical trial provide a comparison group for the subjects who receive the intervention
- Control subjects receive one of the following
  - No intervention
  - Placebo
  - Standard treatment (if one exists)
  - Intervention at a later time

Types of Controlled Trials

- Concurrently assign study subjects to two groups: an intervention group and a control group and follow both over the same period of time to observe outcomes

- Self-controlled study design: Each study subject is observed before and after the intervention so each subject is their own control.

- Cross-over study design: subjects are assigned to one of two groups
  - Intervention first followed by the placebo
  - Placebo first followed by the intervention

Randomized Controlled Clinical Trials (RCCT)

- In a RCCT Subjects are *randomly* assigned to the intervention group or the control group
  - Each subject has equal probability of being in control or treatment group
  - Reduce selection bias and produce groups that have comparable baseline characteristics.
- These trials provide the strongest evidence for a cause-effect relationship between the intervention and the outcome.
- Drawbacks of randomized controlled clinical trials
  - Time-consuming
  - Expensive

Randomized Controlled Clinical Trial Example

- MRFIT: Multiple Risk Factor Intervention Trial
- 12,866 men with RF for CHD randomly assigned to
  - Intervention: counseling to reduce risk factors
  - Control: usual physician care
- Endpoint: death from CHD
- 6-8 years of follow-up
- Study results
  - Risk factor levels declined in both groups but slightly more in the special intervention group.
  - Mortality from CHD and from all causes was not significantly different between the two groups. (JAMA, 1982; 248:1465-77)

Blinded studies

- Single-blind
  - Subjects do not know whether they are in treatment or control group
- Double-blind
  - Subjects *and* investigators are not aware of treatment assignment
- Triple-blind
  - Subjects, investigators *and* statistician do not know treatment assignment
- Goals of blinding: to promote objective assessment of study results and reduce bias
- Placebos are used to keep the study blinded

Random Assignment

Observational studies use random ***sampling*** but controlled clinical trials use random ***assignment***

- In a randomized clinical trial, subjects are *selected* based on inclusion and exclusion criteria
  - The selection process is not necessarily random
- However, selected subjects are randomly assigned to either a treatment or control group
  - Goal is to have comparable groups in all characteristics *except* the treatment
  - With large enough sample size, random assignment usually results in comparable groups

Bias in Clinical Trials

The following describe possible sources of bias in a Clinical Trial

- Procedure Bias
  - Treatment group receives more attention than control group
- Recall bias
  - Subjects in one group are more likely to remember events than subjects in another group
- Compliance bias
  - Patients comply with one treatment more than another

Consider when reading medical literature:

- For an observational study
  - Is the study sample representative of the population of interest?
  - What was the sampling method? If random sampling was used, statistical inference is valid.
  - What are potential sources of bias?
- For a clinical trial
  - What are inclusion / exclusion criteria?
  - How were random assignments made?
  - What are potential sources of bias?

Problem 1

- Twenty-five male volunteers, ages 18 to 24 years, participated in a study to compare effects of small doses of melatonin and a placebo to induce sleep at bedtime.
- Subjects who took melatonin fell asleep in 5 or 6 minutes; volunteers who received the placebo took 15 minutes or more to fall asleep. The researchers concluded that melatonin in small doses may be an effective sleep aid.

1. What is the sample?

1. Twenty-five male volunteers
2. Male who ages 18 to 24 years
3. All males

2. The interested population is ____?

1. Twenty-five male volunteers
2. Male who ages 18 to 24 years
3. All males

3. What is the outcome of interest?

1. The age of volunteers
2. The time to fall asleep
3. The doses of melatonin
4. The doses of a placebo

4. What is the study design?

1. Cross-sectional
2. Case-control
3. Cohort
4. Clinical trial

Problem 2

- Oncologists at the National Cancer Center Institute studied the effects of interleukin-2 (IL-2), an immune system protein, in patients with spreading melanoma or unresponsive, spreading kidney cell cancer. Their goal was to determine if the use of IL-2 spurs the patient's immune system to fight the cancer and halt its spread.
- 300 patients are randomized to treatment or placebo, 20 in the treatment group experienced total remission and remained in remission for at least 8 years. 5 in the placebo group experienced total remission and remained in remission for at least 8 years.. The investigators concluded the use of IL-2 marks a significant advance in the treatment of these two types of cancer.

1. The 300 patients in this study are the___?

1. Population
2. Sample
3. Selected to treatment or placebo groups by someone

2. This study is _____?

1. Retrospective and observational
2. Retrospective and experimental
3. Prospective and experimental
4. Prospective and observational

3. What is the study design?

1. Cross-sectional
2. Case-control
3. Cohort
4. Clinical trial

Problem 3

- A study was conducted relating to the epidemiology of breast cancer and the possible involvement of dietary fats, vitamins and other nutrients.
- It included 2024 breast cancer patients admitted to Roswell Park Memorial Institute, Erie County, New York, from 1958 to 1965. Another group of 1463 people was chosen from patients having no neoplasms and no pathology of gastrointestinal or reproductive systems. The primary factors being investigated were vitamins A and E (measured in international units per month).

1. What is the exposure of this study?

1. Vitamins A and E
2. Whether patients get breast cancer
3. Whether patients have neoplasms or pathology of gastrointestinal or reproductive systems

2. This study is _____?

1. Retrospective and observational
2. Retrospective and experimental
3. Prospective and observational
4. Prospective and experimental

3. What is the study design?

1. Cross-sectional
2. Case-control
3. Cohort
4. Clinical trial

What type of data were collected?

Measurement Scales

Measurement Scales

The two broad categories of data are Categorical and Numerical

- Categorical data
  - Also called ‘Qualitative’
  - Examples: gender, blood type, race, disease status, etc.
  - Data consist of the counts in each category
- Numerical data
  - Also called ‘Quantitative’
  - Differences between the numbers have meaning on a numerical scale
  - Examples: weight, height, survival time, blood pressure

Types of Categorical Data

- Nominal – two or more categories with no natural order
  - Examples: Gender, Race, Blood group
- Ordinal – two or more categories that have a natural order
  - Ex. Apgar scores, tumor stage, Social class
  - Numerical assignments are relative (> or <) and do not represent any interval relationship between categories
- Binary – a special case of either Ordinal or Nominal data
  - Binary data is categorical data with only two categories

Contingency Tables

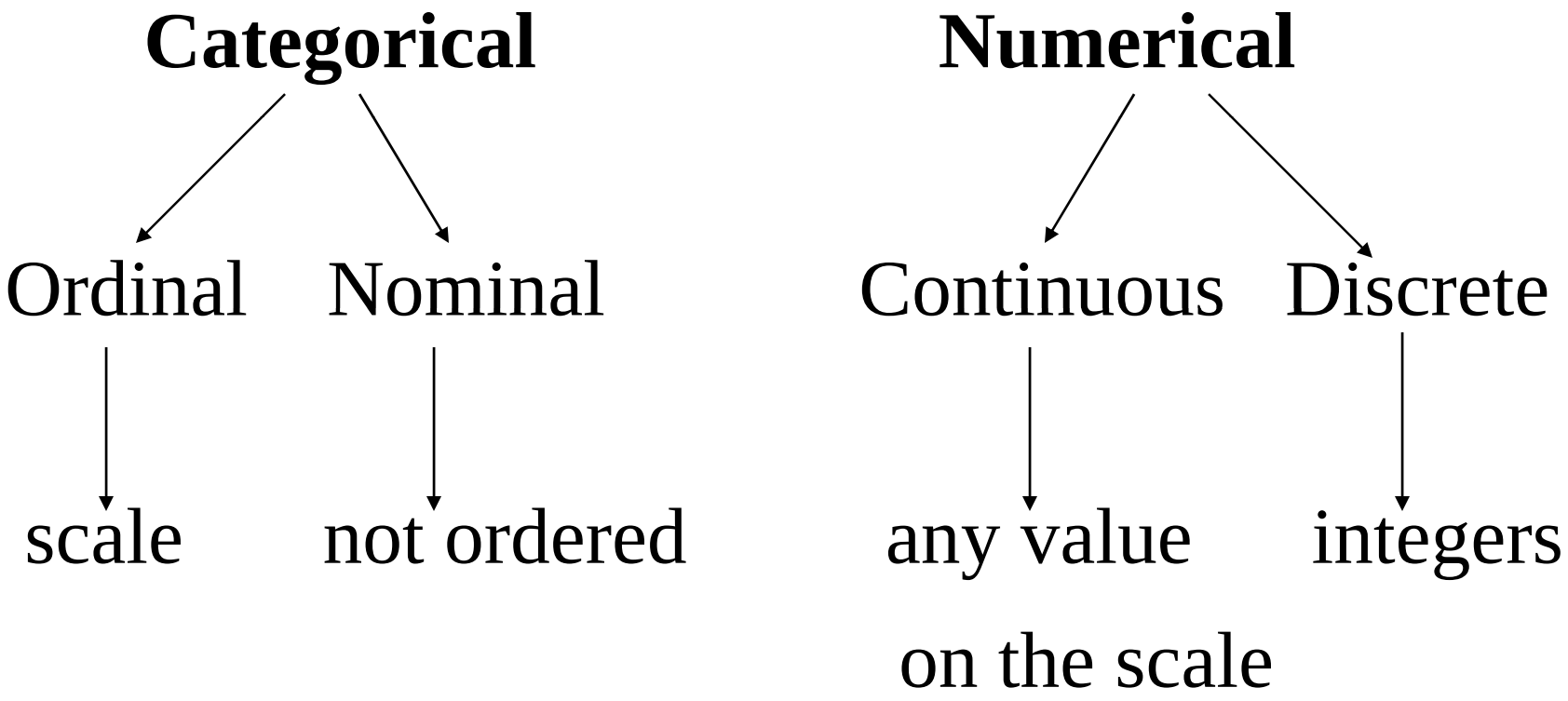
Data for two categorical characteristics are often displayed in contingency tables. The cell counts are the number of subjects with each combination of the characteristics.

	Lung Cancer	No Lung Cancer
Smoker	**85**	**15**
Ex-smoker	**15**	**10**
Non-smoker	**0**	**75**

Numerical Scales

- Continuous scale – data that can take on any real value on the measurement scale
  - Blood pressure, temperature, age, weight, height etc. can all be measured on a continuum
- Discrete scales – data that are only integers
  - Number of children in a family, number of births in a year, number of accidents in a month etc.
- Arithmetic operations are meaningful for numerical data

Diagram of Measurement Scales



Question 1:

Is this variable qualitative or quantitative?

Infant birth weight (grams)

1. Qualitative
2. Quantitative

Question 2:

Is this variable categorical or continuous?

Household income
(below normal, normal, above normal)

1. Categorical
2. Continuous

Question 3:

Is this variable qualitative or quantitative?

Number of children in the family

1. Qualitative
2. Quantitative (Discrete)
3. Quantitative (Continuous)
4. None of these.

Question 4:
Is this variable quantitative?

Social Security *number*

1. True
2. False

Question 5:

Rate your instructor's teaching ability.

1=Poor, 2=Fair, 3=Good, 4=Very Good, 5=Excellent

Is this variable categorical or continuous?

1. Categorical
2. Continuous
3. Either
4. Neither

Question 6

Rate your instructor's teaching ability.

1=Poor, 2=Fair, 3=Good, 4=Very Good, 5=Excellent

Is this variable Nominal, Ordinal or Binary?

1. Nominal
2. Ordinal
3. Binary
4. None of these.

Question 7

OUTCOME WHEN TOSSING A COIN.

Heads or tails

Is this variable Nominal, Ordinal or Binary?

1. Nominal
2. Ordinal
3. Binary
4. None of these.

Question 8

Number of heads when tossing a coin 30 times.

Is *this* variable Binary?

1. No
2. Yes

Question 8

Number of heads when tossing a coin 30 times.

What is *this* variable?

1. Quantitative (Discrete)
2. Quantitative (Continuous)
3. Nominal
4. Categorical Binary

Relationships between Numerical and Categorical Data

- Numerical data can be converted to categorical data
  - For example, age is a numerical continuous variable but it can be summarized as ‘age groups’: 0-5, 6-10, 11-19, 20-39, 40-59, 60-79, 80+
  - When represented as ‘age groups’, age is a categorical (ordinal) variable
- Ordinal data can be coded numerically
  - The numerical coding of ordinal data does not mean that the ordinal data are now numerical, however
  - The codes represent the order of the categories, not the interval relationship between categories

Next topics: Summary Statistics and graphs

- For Numerical Data
  - Statistics: Mean, standard deviation, median, range, percentile
  - Tables and graphs: frequency distribution, histograms, box-plots, scatter-plots
- For Categorical Data
  - Statistics: proportion, percent, rates, relative risk, odds ratio
  - Tables and graphs: bar charts, pie charts, contingency tables

Readings and Assignments

- Readings:
  - Chapter 2: Study Designs in Medical Research
  - Chapter 3: Scales of Measurement pgs 26-27
  - Chapter 4: Pages 68-72
  - Optional: Chapter 13 pgs 332 - 335
- Practice Exercises for Lesson 1
- Try Intro to Excel Module on your own:

http://www.biostat.umn.edu/~susant/FALL09PH6414LAB.html

- Start Homework 1: Problems 1 and 2
- New vocabulary: look up any unfamiliar terms in the *Pocket Dictionary of Statistics* or the text glossary