MATH 1635, Statistics (1)

Chapter 1

I. Basic Definitions

1. Data- observations (such as measurements, genders, survey responses) that are collected.
2. Statistics- a.decision making involving incomplete information
   b.educated guess based on limited information
   c. science/study of how to collect, organize, and interpret numerical information
   d. a numerical measurement describing some characteristic of a sample.
   e. central theme--using information obtained from sample to reach decision or inference concerning an entire population from which the sample has been drawn
   f. A collection of methods for planning studies and obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusion based on data.
3. Population-set of all possible measurements, counts, or observations that are of interest in a particular study
4. Sample-a subset of the population
5. Random Sample-one in which every member of the population has equal chance of belonging
6. Descriptive Statistics-involving the organization, summary, and display of data
7. Inferential Statistics-involving using a sample to draw conclusions about a population. (Basic tool is probability).
8. Census-collection of data from every member of the population.

9. Parameter-a numerical measurement describing some characteristic of a population.
10. Quantitative vs. qualitative data-data representing counts/measurements vs. separation into categories based on nonnumeric characteristic.
11. Discrete vs. continuous data-data that are finite and countable vs. infinite and measureable (range of values without gaps or interruptions or jumps).

Examples:

1. To estimate the amount of time State College students who live off campus spend to commute to class each week, a random sample of 45 such students were surveyed
   (a) What is the population? (b) What is the sample?
   (c) Would this sample necessarily be representative of the time part-time off-campus students spend commuting to class each week? Why or why not?

2. An insurance company wants to know the proportion of medical doctors in New York involved in at least one malpractice suit in the last three years. They surveyed a random sample of 200 medical doctors in New York.
   (a) What is the population?
   (b) What is the sample?
   (c) Could the insurance company generalize and say this sample is representative of all medical doctors in rural areas of the country?

3. A company making radar detection devices maintains quality control by testing a random sample of 20 such devices produced each day.

(a) What is the population?

(b) What is the sample?

(c) If the manager decides to test a random sample of 100 devices on Monday and did no further testing for the rest of the week. Could she draw any conclusions about the quality of production for the entire week? Explain.

4. Which is/are true? Statistics:

I) is a science

II) include analyzing and interpreting data

III) does not include collecting data

(A) I + II

(B) II + III

(C) I + II + III

5. Which of the following is included in inference statistics?

I) poll

II) surveys

III) results of a class

(A) I

(B) I + II

(C) I + III

(D) All of the above

6. In statistics population refers only to:

I) people

II) bears

III) chairs

(A) I

(B) I + III

(C) I + II

(D) All of the above

7. Which of the following is a sample?

I) 10 people out of 25

II) 15 people out of 15

III) 100 people out of 125

(A) I

(B) II

(C) I + III

(D) I + II + III

8. A Roper poll asked 938 adults in the Unites States if they thought feeling financially secure was an important aspect of having money.

(a) The population is_____

(b) The sample is _____

II. Levels of Measurement

It is important to know the level of measurement that we collect so we can determine which types of computation are meaningful. Calculations appropriate for each level can be used at any higher level but should not be used for any lower level. Levels are listed below from lowest to highest.

1. Nominal Data (can be put into categories)
   a. consists of names, categories, qualities or labels.(ex. type of car you drive)
   b. can put data into categories but we are unable to determine if one piece of data is better or higher than another.
   c. when numbers are used as labels, such as on a football uniform, they are classified as nominal data. (not meaningful to average the numbers on the uniforms for the Philadelphia Eagles).
2. Ordinal Data (can be placed in rank order and be put into categories)
   a. designations or numerical rankings which can be arranged in ascending or descending order.(ex. TV ratings for #1 show, #2 show, etc.)
   b. can compare rankings as to which is higher but does not make sense to subtract one rank value from another. (ex. interview 3 candidates for a job and rank them in order of preference 1, 2, and 3. You can tell which candidate is ahead of the others but only in terms of order of preference, not in magnitude or degree of preference. Candidates who are ranked #1 and #2 could be close while #3 might be totally

unacceptable.)

3. Interval Data (can be subtracted to find the difference between two values, put in order and put in categories);do not have a natural zero starting point.
   a. data is numerical; 0 can be used to indicate a position in time or space, but the zero at this level does not correspond to "none" of the specific variable being measured. (ex. the year you graduated from high school...the "year 0" has no meaning.)
   b. differences between data values are meaningful (but does not make sense to compare one data value as being twice (or multiple of)another.
4. Ratio Data (values can be divided, subtracted, put in order and put into categories);there is a natural zero starting point;differences and ratios are both meaningful.
   a. the highest level of measurement (ex. number of gallon of gas you put in your car)
   b. there is a zero on this scale that is interpreted as "none" of the variable in question.(i.e. you can put 0 gallons of gas in your Car.). **a zero entry is an inherent zero**.
   c. **is meaningful to say one measure is two times or three times as much as another**

From Lowest to Highest Level of Measurement:

Nominal--Ordinal--Interval--Ratio

Examples:

1. At a used car lot the following information is obtained about one of the cars. What type of data/measurement are these?
   a. Make and Model: Ford Escort b. Model Year:1990
   c. Color: Blue d. Number of Cylinder:
   e. Gas Consumption: 35 mpg f. Sticker Price: $5817.00
   g. Condition: Clean and good
   h. Number of Miles on Odometer:36,719
   i Sales Person Claims: This is the best car for you

2. Categorize these measurements associated with student life according to level: nominal, ordinal, interval, or ratio.
   a. length of time to complete an exam
   b. time of first class
   c. class category: freshman, sophomore, junior, senior
   d. course evaluation scale: poor, acceptable, good
   e. score on last exam (based on 100 possible points)
   f. age of student

3. At a hospital nursing station the following information is available about a patient. What
   a. name: Jim Wood b. age: 19
   c. weight: 180 lbs d. height: 6'2"
   e. blood type: AB f. temperature: 96.8 degree F
   g. condition: fair
   h. date of admission: January 26, 1998 i. response to treatment: excellent

III. Methods of Data Collection

1. Sampling-measuring, counting, or noting a response from a subset of the population
   goal:to gain an accurate picture of the population and disturb the population as little as possible
2. Experimenting-imposing a treatment and observe its effect. (ex. study the effect of calcium supplement given to young girls on their bone mass)
   goal:to measure the effects of an intervention to understand how nature (or a population) responds to change
3. Simulation-a numerical facsimile of real-world phenomenon artificially producing outcomes when it is impractical to do real life experiments. (often performed on computers)
4. Census-measurement of the entire population of interest (ex. U.S. Bureau of Census conducts a census every 10 years.)
5. Voluntary Response Sample-means to gather data about people through questions;or self-selected sample in which respondents themselves decide whether to be included. (ex. telephone surveys) concerns: Are questions neutral? Are answers truthful? (Voluntary response sample often over-represent people with strong opinion. Hidden bias. Other variables.)
6. Potential Misuses of Statistics-via: voluntary response samples, small samples, graphs, pictographs, unclear percentages, loaded question, order of question, nonresponse, missing

data, correlation and causality, self-interest study, partial pictures, deliberate distortions.

*Be able to convert to: percentage of;fraction→ percentage; decimal→percentage;percentage→decimal;

Examples:

Categorize the style of gathering data (sampling, experiment, simulation,census)

(a) Look at all the apartments in a complex and determine the monthly rent charged for each unit.

(b) Give one group of students a flu vaccination and compare the number of times these students are sick during the semester with the number of students who get sick in a group who did not receive the vaccination.

(c) Select a sample of students and determine the percentage who are taking mathematics this semester.

(d) Use a computer program to show the effects on traffic flow when the timing of stop lights is changed.

IV. Obtain Data From Two Distinct Sources

A. Observational Study-observe and measure specific characteristic without modifying the subjects being studied. Ex. Gallup Poll

1. Cross-Sectional Study-data are observed, collected, and measured at one point in time.
2. Retrospective (or case-control) Study-data are collected from the past by going back in time(through records, interviews, etc.)
3. Prospective (or longitudinal) Study-data are collected in the future from groups sharing common factors (called cohorts).

B. Experiment-introduce treatment/change into the system and observe the effects on the subjects (these subjects are called experimental units).

*Confounding occurs in an experiment when one is not able to distinguish among the effects of different factors.

1. In designing experiment, there are three important considerations:

a. control the effects of variables by: blinding, randomized block design, completely randomized experimental design, or a rigorously controlled experimental design.

*A block is a group of subjects that are similar , but blocks are different in the ways that might affect the outcome of the experiment.

*Placebo Effect, **Hawthorne Effect** (when treated subject respond differently simply because they are part of the experiment), Experimenter/**Rosenthall Effect** (when experimenter unintentionally influences subjects through facial expression, tone of voice, or attitude, etc.)

b. use replication: repetition of an experiment on sufficiently large groups of subjects.

c. use randomization:

V. Types of Random Samples

***If incorrectly collected, the data may be completely useless.**

1. Simple Random Sample

a. every member of the population has an equal chance of belonging to a simple random sample.

b. insuring against a biased representation of the population.

c. random numbers can be generated with a calculator that has a

random number function, a computer software program, or from The Table of Random Numbers.

2. Probability Sample-involves selecting members from a population in such a way that each member has a known (but not necessarily the same) chance of being selected.
3. Stratified Random Sample
   a. members of the population are separated into subgroups (strata) having the same characteristics....members can be determined by gender, age, ethnicity, etc.
   b. a random sample is drawn from each subgroup.
   c. this method insures that each subgroup will be represented and no group will be omitted from the sample.
   d. often individual stratum are sampled in proportion to their membership in the population.
4. Systematic Sample
   a. members of the population are assigned in some order and assigned numbers.
   b. a first number is selected at random..from there, every kth number is selected.
   c. for example:to select a sample of 10 from a population of 529, divide 529/10 ~52.....so in this example, k=52....select a random number from 1 to 52...the measure or response for this number will be the first member of your sample.
   d. once the first member is selected, select every 52nd number for the sample....if the random value selected were 23, the sample will consist of measures or responses from population members number 23, 75, 127, 179, 231, 283, 335, 387, 439, and 491

e. once the first member is selected, others are automatically selected.....this method of sampling can be used when there is no danger of cyclical phenomena (ex. choose every 10th member of the population but do not choose every 4th season or every 7th day.)

5. Cluster Sample
   a. population is divided into sections or clusters from the same location. all clusters are likely to have similar characteristics so that when a sample is selected from one cluster, the sample is not biased....the unit for sampling is not an individual but a occurring subgroup
   b. one or more of the sections (clusters) is randomly selected
   c. sample is taken from the selected subgroups
   d. often used when clusters can be geographically determined (ex. the same zip code, same block, the same bank, same section of a course, etc.)
6. Convenience Sample
   a. sample consists only of the members of the population that are readily available and willing to participate
   b. this sort of sampling may produce misleading or biased results

*Sampling Error-difference between a sample result and the true population result;such an error results from chance sample fluctuations.

*Nonsampling Error-when sample data are incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective measure instrument, or copying the data incorrectly).

Examples: (Using the Random Numbers Table, TI-83 Plus)

1. There are 529 students enrolled in a statistics course for the semester at your school. You are asked to interview a sample of ten of these students for a research project. What steps would you take to select a simple random sample?

First assign a number to each student in the population. College I.D. numbers could be used, however it is probably easier to get the class lists and assign numbers 1 to 529.

Select a starting point (arbitrarily) on your Random Number Table

For example, if you arbitrarily picked the starting point (26907) on a random number generating table: 26907 52180 05538 56277 54190 10910 97564 11278 03772 83834 57300 21769 78972 05007 19561 91610 00432 08299 63480 04119

Draw a vertical line at intervals of every 3 digits to separate the digits into sets of 3 (since 529 is a set of 3 digit number). The first groups are 269, 075, and 218. Select the first ten three-digit numbers that are less than 530. The digits would group as :

269 **075** **218** **005** 538 562 775 **419** **010** 910 975 641
127 803 772 838 **345** 730 **021** 769 789 720 **500**

The random sample consists of the responses from the students who were assigned the numbers: 269, 75, 218, 5, 419, 10, 127, 345, 21, 500.

With a different starting point , the sample would be different.

2. Use a random-number table to get a list of six random numbers from 1 to 99. Explain.Use the numbers generated from Random Number Generating Table in the Appendix

   Arbitrarily pick 49587 as the starting point. What will be your result?

   49587 76612 39789 13537 48086 59483 60680 84675 53014

   What will be your result if your starting point is 06348?

3. Use a random-number table to get a list of eight random numbers from 1 to 598. Explain.

4. A die is a cube with dots on each face. The faces have 1, 2, 3, 4, 5, or 6 dots. Use a random-number table to simulate the outcomes of tossing a die 20 times. Explain.

5. Lotto is the name of the Colorado lottery. The Lotto boards consist of 42 numbers (from 1 to 42). To play you select six distinct numbers. Every week a drawing machine randomly selects six numbered Ping-Pong balls. If one of your boards contains all six winning numbers, in any order, you've hit the jackpot! You can pick your numbers any way you wish. However, suppose you want to use a random-number table to pick your six numbers. Describe how you would do so.