Ch 1 Sampling and data

Ch 1.1 Definitions of statistics, probability and key terms.

Statistics **is** the science of collecting, analyzing, interpreting and presenting data.

Two main branches of statistics:

- descriptive statistics: organizing and summarizing data.
- Inferential statistics: Draw conclusion from data.
-

Why should we study statistics?

- To be able to read and understand various statistical studies performed in their fields—requires a knowledge of the vocabulary, symbols, concepts, and statistical procedures
- To conduct research in their fields—requires ability to design experiments which involves collection, analysis, and summary of data.
- To become better consumers and citizens.

Probability: A tool to study randomness, it deals with the chance of an event occurring. Rare event rule is used to draw conclusion from data. An event is "significant" if it has 5% or lower chance of occurring.

Terms: Population vs sample
Population: the target group of person, things, or objects under study. It takes times and money to study the entire population.
Sample: A subset of the larger population to gain information about the population. (A census is a collection of data from every member of the population.)

Terms: Parameter vs Statistic
Parameter: A numerical characteristic of the population.
Statistic: A numerical characteristic of a sample. The value varies from sample to sample.

Sample is collected so we can use the statistic to estimate the corresponding parameter of the population. Sample must be representative to give an accurate estimation of the parameter.
Ex. Use sample mean (average) to predict population mean. Use sample proportion to predict population proportion.

Terms: Variable and Data
Variable: a characteristic or measurement for each member of the population.
Data: the actual values of the variables.

Ex1.
a) 37 students are randomly selected to find the proportion of students who prefer asynchronous mode of teaching.
What is the population?
What is the sample?

b) Of the 37 student surveyed, 14% prefer aynchronous teaching.
The number 37% is a parameter or statistic?

c) The average annual income for all residents in a city is $34,000.
The number $34,000 is a parameter or statistic?

d) A sample of 45 college graduates are surveyed and their average annual income is $34,000.
The number $34,000 is a parameter or statistic?

Ex2. Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year by surveying a randomly selected group of students who graduated last year.
a) Cumulative GPA of a student who graduated last year. _____
b) Cumulative GPA of a sample of students graduated last years are: 3.65, 2.80. 1.5, 3.9. _____
c) The group of students being surveyed. _____
d) College Registrar published that average cumulative GPA of all students who graduated from the college last year is 3.1. _____
e) All students graduated last year in the local college
f) Average cumulative GPA from the sample of students is 3.32 _____

Ex2. A study on car safeties are performed. A sample of 75 cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. For each car crash, the head injury condition (yes or no) of the dummy is recorded.

a) Population is
b) sample is
c) Parameter is
d) Statistics is
e) Variable is
f) Data is

Try Ex 1.4 in textbook

Chapter 1.2 Data, sampling and Variation

Types of Variable: Quantitative vs Categorical.
Categorical: name, label or a result of categorizing attributes. Also known as qualitative variable.
Quantitative: counts or numerical measurement with units.

Types of Quantitative data: Discrete vs Continuous
Discrete: counts or numbers that takes on finite values.
Continuous: measurement data that can have infinitely many possible values between two values not limited by measurement device.

Level of measurement: (from 1.3):

- Nominal scale level: categorical data with no natural ordering
- Ordinal scale level: categorical data with a natural ordering, difference and mean are not meaningful.
- Interval scale level: quantitative data with no natural zero. Zero value is a mark only. Difference is meaningful but ratio is not meaningful.
- Ratio scale level: quantitative data with natural zero. Difference and ratio is meaningful.

Note: Nominal and Ordinal data are usually summarized by proportion of the categories. There is no meaning to the mean.

Ex1 classify the data as quantitative or categorical, discrete or continuous if it is quantitative. Classify by level of measurement also.
a) The number of students in a zoom meeting.
b) The major of study of a student.
c) The student id of a student.
d) Jersey number of a basketball player.
e) The duration (in minutes) it takes to finish a homework.
f) Grade of an exam.
g) Daily high temperature of a city.

Try Ex 1.9 in textbook

Percent and Decimal:

Decimal → Percentage:

Move decimal point to the right by 2 places.

Example: 0.75 = 75% 0.0178 = 1.78%

Percentage → decimal.

move decimal point to the left by 2 places.

Example: 25% = 0.25, 4% = 0.04

0.1% = 0.001

Find percentage:

$\frac{part}{whole}$ → decimal → percent

Find Percent of an amount: To find percent of an amount, replace the % with decimal notation, interpret "of" to be multiplication.

Example:

6% of 1200 = 0.06 × 1200 = 72

12% of 2418 = 0.12× 2418 =290.16

Exact value of 12% of 2418 adults = 290.16 (no rounding)

Actual value of 12% of 2418 adults = 290 adults

Summarizing Categorical data.

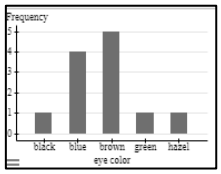
Proportions (relative frequencies) are used to summarize Categorical data. A proportion will be calculated for each category and presented as a table

Ex 1.

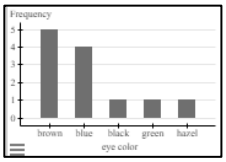
DVC-2020	frequency	rel. freq
Part-time	12586	
Full-time	7503	
Total	20089	

eye color	frequency	rel. freq
black	1	
blue	4	
brown	5	
green	1	
hazel	1	
total		

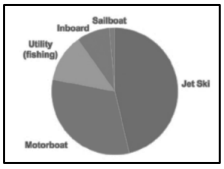
Graphs for categorical data.



1) **Bar Graph**: Use bars of equal width to show frequencies of categories. May or may not be separated by spaces.



2) **Pareto graph:** A Bar graph where bars are arranged in descending order of frequencies, so it is easier to compare difference categories



3) **Pie Chart** – show categorical data as slices of a circle. each slice is proportional to the frequency count for the category. Show composition of the whole.

How to draw a bar graph, pie chart:

use Excel, select the category and frequency or relative frequency column, insert chart.

When Categories proportion does not add up to 100%
A bar graph is appropriate, but a pie chart is not.
A total percentage more than 100% indicates subjects are in more than one category.
A total percentage less than 100% indicates a missing category such as the "other "category.
Example.

Categories	percent
full time students	40.90%
transfer to 4-year	48.60%
age under 25	61.00%
total	150.50%

race	percent
Asian	36.10%
Black	5.80%
Hispanic	17%
White	24.50%
total	83.40%

Types of sample:

1) Simple Random sample.
Any group of *n* individuals is equally likely to be chosen as any other group of *n* individuals if the simple random sampling technique is used.

2) Random sample:
Any subject has equal chance of being selected.

3) Non-random sample:
Not all subjects have an equal chance of being selected.

A good sample should have the same characteristic as the population it is representing. Random sample or Simple Random sample can achieve this goal.

Sampling method:

1) Simple Random sampling:
Sample are selected one by one using random procedures such as selecting names from a hat or generated by random number generator.

2) Stratified sampling:
divide the population into groups called strata and then take a proportionate number from each stratum.

3) Cluster sampling:
divide the population area into groups or clusters. The randomly select some of those clusters and choose all members from those selected clusters.

4) Systematic sampling: Select some starting point and then select very kth (such as 50^{th}) element in the population.

5) multistage sampling – use some combination of the preceding sampling methods.

6) Convenience sampling:. Use data that are very easy to get. This will produce a non-random sample
Voluntary response sample (self-selected sample) is a convenience sample.

Sampling without replacement: You do not replace the subject you select before selecting the next subject.

Sampling with replacement: Once a member is picked, that member goes back into the population and thus may be chosen more than once. This guarantee that all subjects has the same chance of being selected.

In practice, simple random sampling is done without replacement and survey are typically done without replacement. If the population is small, sampling without replacement becomes an issue.

Ex. Classify the following method of sampling:

a) Select 4 students by selecting first 4 who arrive first.

b) Select 100 customers by selecting every 50^{th} in a customer database.

c) Select 10 students randomly from each grade in a high school.

d) Select a sample of restaurants by randomly 10 streets and select all restaurants in the 10 streets.

e) Select 100 voters' response by posting a survey online.

Sampling error, non-sampling error, sampling bias.

Due to randomness of sampling, sample variation will occur, and the difference are known as sampling error. When sample size increase, sampling error will decrease. Sampling error can be analyzed.

Non sampling error occurs when the process of sampling is not random. Non sampling error cannot be analyzed.

Sampling bias occurs when some subjects in the population are not likely to be selected as others. There can be incorrect conclusion drawn from these sample.

Guideline for evaluating a statistical study:

1) Problems with samples: Bias sample is not representative of the population.

2) Self-selected sample (voluntary response sample): response only by subject who choose to participate. This usually only include subjects with strong opinion of the matter. Internet survey and call-in survey are examples of voluntary response sample.

3) Sample size issues: Small samples are unreliable but are unavoidable such as car test and medical test.

4) undue influence: Questions in survey are worded to influence response.

5) Non-response: high non-response rate make it a voluntary response sample.

6) Self-funded or self-interest study: A study performed by a person or organization in order to support their claim.

7) Causality: Correlation does not imply causation. It may be due to a confounding variable.

8) Misleading use of data: exaggerate difference by using non-zero axis.

Try 1.11 and 1.12

Ch 1.3 Grouped Frequency Distribution Table (GFDT)

Quantitative data can be summarized into a frequency table by classifying data into classes. Class can have a range of non-overlapping value with equal class width (difference between class lower class limits)

Terms related to GFDT:

lower limits: lower bound of each class .

upper limits: upper bound of each class.

class midpoints: (lower + upper)/2

class width: difference between 2 consecutive lower limits.

class boundaries: values between 2 classes.

Ex.Given GFDT below: find lower limits, classwidth, class midpoints.

no. siblings	frequency
0	8
1	12
2	5
3	2
4	1
5	1
more than 5	1

grade	freq.
60-69	2
70 - 79	3
80 - 89	12
90 - 99	25

Relative frequency and cumulative frequency can be evaluated for the classes. Because of rounding the relative frequency may not be sum to 1 but should be close to one.

Rounding review:

If the number **place** you are **rounding** is followed by 5, 6, 7, 8, or 9, **round** the **number** up.

If the **number place** you are **rounding** is followed by 0, 1, 2, 3, or 4, **round** the **number** down.

Ex1. Round to three decimal places:

a) 0.1278, b) 0.1283, c) 0.1239, d) 0.1298 e) 5/6

Ex2. Round to 1 decimal place of a percent.

a) 0.1184 b) 45.677% c) 52/89

Ex3. Round to the nearest whole number.

a) 12% of 781 b) 15.2% of 2344

Relative frequency for a class = $\frac{frequency\ for\ the\ class}{sum\ of\ all\ frequencies}$

cumulative frequency =sum of the frequencies for that class and all previous classes

Ex1. Find relative and cumulative frequency for service time for a fast food restaurant given in the following GFDT.

Time (sec)	freq.	Rel freq	class	cum freq
75 - 124	11		less than 125	
125 - 174	24			
175 - 224	10			
225 - 274	3			
275 - 324	2			

Classes with overlapping class limits:

When frequency table has classes with overlapping limits at the end points, the common convention is lower limit ≤ data < upper limit. or the classes are assigned so all data values fall between the limits.

Ex2. Find the percent of town with rainfall less than 9.01 in.

Rainfall (in)	Freq
2.95 - 4.97	6
4.97 - 6.99	7
6.99 -9.01	15
9.01 - 11.03	8
11.03 - 13.05	9
13.05 - 15.07	5

Frequency table where the class is time such as years.

Ex3. Find percent of crashes occurs after 2015.

year	no. of crashes
2013	30203
2014	32744
2015	35485
2016	37806
2017	37473
2018	36560

Graph a GFDT from data:

https://www.socscistatistics.com/descriptive/frequencydistribution/default.aspx

-Find the minimum data value.

- Enter data in a column in the input frame.
- Click Generate.
- select number of classes and the lowest class limits that should include the minimum data value and a nice value.
- Click Edit frequency table for the new table.

Ex1. Construct a GFDT from the data below:

22	42	29	61
33	54	38	75
41	73	50	32
53	27	57	40
71	37	74	52
26	46	30	69
35	56	39	83
	73	51	

Use 7 classes and start with a "nice" good lowest limit.

Use socialscience calculator, Input data to input frame. Click generate, then change class size to 7 and lowest class value to 20.

Then click Edit frequency table.

Frequency Distribution Table

Class	*Count*	*Percentage*
20 - 29	4	13.3
30 - 39	7	23.3
40 - 49	4	13.3
50 - 59	7	23.3
60 - 69	2	6.7
70 - 79	5	16.7
80 - 89	1	3.3

Ch 1.4 Experimental Design and Ethics.

Types of statistical study:

1) Observational study: we observe and measure specific characteristic of the subjects. A survey study is an observational study. We don't attempt to modify the individuals being studied.

Types of observational study:

a) Cross-sectional study – data are collected at one point in time, not over a period of time.

b) retrospective study – data are collected from a past time by going back in examination of records.

c) prospective study – data are collected in the future from cohort groups.

2) Experimental study: we apply treatments and proceed to observe its effect on individuals.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the response variable.

Observational study can demonstrate an association but no causal relationship. Controlled experiment can demonstrate a causal relationship.

Feature of good of experimental design:

Control: Use a control group with placebo, blinding or double blinding to reduce placebo effect (also known as power of suggestion).

Randomization: subjects are randomly assigned to control and treatment groups. (**Completely Randomized Design**). Control and Treatment groups should be as similar as possible. **Matched-pair design** can be used.

Replication: use large sample size in both control and treatment group.

Blinding: subjects are unaware if they are in a control or treatment group.

Double-blinding: one in which both the subjects and the researchers involved with the subjects are blinded.

Ex1. Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks. Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

population:

sample:

experimental units:

explanatory variable

response variable

treatments:

Ex 2. A survey shows that students who eat breakfast have higher average GPA. The researcher conclude that breakfast can cause an increase in academic performance. Discuss if the conclusion is valid or not and explain.

Ex 3. Classify if the study is an experiment or observational study:
https://www.sciencedaily.com/releases/2008/07/080707081834.htm

Make comment about the headline of the article: *"PTSD Causes Early Death From Heart Disease, Study Suggests."*

Ethics of statistic researcher

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.