STA 1013 : Statistics through Examples

Lecture 6: Review for Quiz 1

Hwiyoung Lee

September 11, 2019

Department of Statistics, Florida State University

Overview

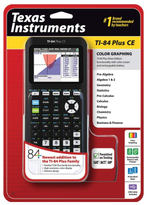
1. Calculator

2. Review

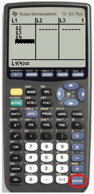
Calculator

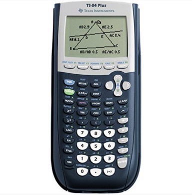
Calculator

TI 84 plus CE



TI 83, 84 plus : **OKAY !!**





Calculator last class

- **CelSheet** app
- **Pie & Bar** chart
- **SortA**
- **randIntNoRep**

Are **not necessary** in the Quizzes, and Final.

- Analyzing data and Intepreting statistical plots and result
- It is not always convenient to use the calculator

Review

Some definitions on Statistisc and Data

- **Statistics** : Science of collecting, summarizing, analyzing, and interpreting of data
- **Data** : Information
  1. Observed measurement : height, temperature, GPA, $\cdots$
  2. Descriptions : marital status, gender, ethnicities, $\cdots$
- **Observation, Data Point** : A single collected data value
- **Variable** : characteristic or property of an individual or unit whose value may change from one observation to another.

Population and Sample

Population : Entire overall group we are interested in
Parameter : A numerical measurement that describes a characteristic of a population

- Usually unknown
- Example : $\mu, \sigma^2, \pi, \rho, \cdots$

Sample : Subset of the population that we collect data on
Statistic : A numerical measurement that describes a characteristic of a sample

- known (can be calculated)
- Example : $\bar{X}, s^2, p, r, \cdots$
- statistics are used as estimates for the corresponding population parameters

Primary Data & Secondary Data

Primary Data : Data collected by the investigator himself/ herself for a specific purpose.

- The investigator collects data specific to the problem under study
- If required, it may be possible to obtain additional data during the study period
- Cost of obtaining the data is often the major expense in studies

Secondary Data : Data collected by someone else for some other purpose, or published elsewhere

- The data is already there → It can be gathered quickly and inexpensively
- Data may be outdated
- The investigator cannot decide what is collected
- Most often obtaining additional data is not possible

Types of variable

Categorical variable :

- Result in categorical responses.
- Also called **Nominal**, or **Qualitative variable**
- Never used directly in calculations
- Examples :

  Gender : Male, Female
  Animal species : Dog, Cat, Fish, Bird, $\cdots$

Quantitative variable :

- Result in numerical responses, can be used directly in calculations

Types of variable

Note :

- If numbers are used only as labels for categories, then they are considered **categorical variable**.
- If Quantitative variable is used for the purpose of categorization, then they are considered Categorical (Qualitative) variable in that context.

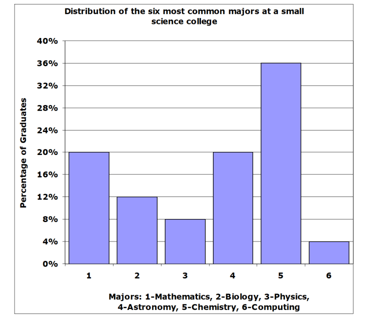
Distributions

Distribution of a variable tells us what values the variable can possibly take and how often it takes these values. The distribution of a variable refers to the way its values are spread over all possible values.

Summarize a distribution with a table or a graph

- **Categorical variable :**
  1. Frequency Table, Relative Frequency Table
  2. Bar chart
  3. Pie chart
  4. Dot plot
- **Quantative variable :**
  1. Stem and Leaf plot
  2. Histogram (Topic 3)
  3. Box plot (Topic 3)

Practice : Bar-chart



1. If 6 students were Biology majors, what was the total number of students?
2. How many students were Mathematics majors?

Shapes of Stem-Leaf plot

Example : STA1013 Test Result

Symmetric Data

2	8
3	7 9
4	3 6
5	0 1 7 8
6	1 1 2 5 7
7	2 3 5 7
8	0 1
9	3 4
10	0

Left Skewed Data

2	8
3	
4	5
5	0
6	1 2
7	2 3 5
8	0 1 2 8 9
9	3 4 5 6 8
10	0 0

Right Skewed Data

2	9 9
3	1 2 8 9
4	2 3 5 6 5
5	0 1 3 4
6	1 2
7	2
8	0
9	3
10	0

Practice

The split-stemplot below shows the customer service times (in minutes) at a supermarket chain. Remember that the first occurrence of a split stem carries leaves 0-4, the second occurrence carries leaves 5-9.

0	2 2 3 3 3 4 4 4
0	5 5 6 6 6 6 7 7 7 8 8 8 8 9 9
1	0 0 1 1 1 1 1 1 1 2 2 2 3 3 3 4 4
1	6 6 7 7 8 8 8 8 9 9
2	1 2 3
2	5 8
3	1 1
3	6
4	
4	5
5	2

Practice

A. The shape of the distribution is

1. bell-shaped 2. left-skewed 3. right-skewed 4. uniform

B. Choose the best answer to fill in the blank, that is, the answer that would make the statement closest to the truth.
Roughly half the customers had to wait more than (   ) before being served.

1. 4 mins 2. 2 mins 3. 1 min

C. Circle the words that would make true statements.

1. Most customers had to wait a (long / short) time and few had to wait a (long / short) time.
2. There (are / are no) service times that are exceptionally short compared to the other service times.
3. There (are / are no) service times that are exceptionally long compared to the other service times.

Types of Statistical Studies

Observational Studies : researchers observe or measure characteristics of the subjects, but do not attempt to influence or modify these characteristics.

Experiments : researchers apply some treatment and observe its effects on the subjects of the experiment.

Goal of Experiments : study **the effects of some treatment**.

Types of Observational Studies

- **Cross-sectional study** :
  - The most familiar observational studies are those in which data are collected all at once.
  - Data are observed, measured, and collected at one point in time, not over a period of time.
- **Retrospective study**
  - Uses data from the past, such as official records or past interviews.
  - Especially valuable in cases where it may be **impractical** or **unethical** to perform an experiment.
- **Prospective study**
  - Set up to collect data in the future from groups that share common factors.

Some Terminologies on Experiments

Two types of groups

- **Treatment group** : the group of subjects who receive the treatment being tested.
- **Control group** : the group of subjects who do not receive the treatment being tested.

Two types of variables of interest : When cause and effect may be involved

- **Explanatory variable, Independent variable**
  : variable that may explain or cause the effect
- **Response variable, Dependet variable**
  : variable that responds to changes in the explanatory variable

Confounding variable

Confounding variable

- A study suffers from **confounding** if the effects of different variables are mixed
- So we cannot determine the specific effects of the variables of interest.
- The variables that lead to the confusion are called **confounding variables.**

Sampling methods

- **Simple Random Sampling**
- **Systematic Sampling**
- **Cluster Sampling**
- **Stratified Sampling**

Sampling Methods

1. Simple Random Sampling

- In most cases, the best way to obtain a representative sample is by choosing **randomly from the population.**
- A random sample is one in which every member of the population has an equal chance of being selected to be part of the sample.
- Use the **random number generator**.

2. Systematic Sampling

- Select some starting point, then select every kth element in the population.
- Have to avoide hidden periodicity in population

Sampling Methods

3. Cluster Sampling

- Divide the population into homogeneous subgroup (cluster)
- Then randomly select some of those clusters (Sampling units are group)
- Choose all members from those selected clusters
- This can reduce travel and other administrative costs.

4. Stratified Sampling

- We use this method when we are concerned about differences among subgroups, or strata, within a population.
- First identify the strata and then draw a **random sample within each stratum**.
- The total sample consists of all the samples from the individual strata

Weight data

ind	1	2	3	4	5	6	7
Gender	M	M	F	F	M	F	F
Weight	195.6	200.4	165.6	165.3	191.7	169.3	153.2
ind	8	9	10	11	12	13	14
Gender	M	M	F	M	F	M	F
Weight	189.5	170.4	149.3	185.3	150.3	179.6	160.3
ind	15	16	17	18	19	20	21
Gender	M	F	F	M	M	F	M
Weight	198.4	163.2	166.3	197.3	201.3	168.2	198.4

We will draw samples from the population

Examples of Sampling methods

Simple Random Sampling :

- Random number generator : (3, 5, 9, 12, 15, 20)

Systematic sample :

- Start from 3 select every 3th obs : (3, 6, 9, 12, 15, 18, 21)

Cluster sample :

- Select cluster 3 and use every member in this cluster

Stratified sample :

- Select 3 from Man (Stratum 1),
  Select 3 from Female (Stratum 2)

Systematic Sampling will fail

Use the table given below to answer the question. We want to measure average weight of dogs. Which systematic sampling methods are appropriate ? (G : Golden Rietriever, P : Poodle, M : Maltese)

ind	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Breed	G	P	M	G	P	M	G	P	M	G	P	M	G	P	M	G	P	M	G	P	M	G	P	M

1. Select from 3 every 3rd observation
2. Select from 2 every 3rd observation
3. Select from 1 every 6th observation
4. Select from 2 every 4th observation
5. Select from every 12th observation