Statistics

Statistics is the science of gathering and summarizing data and using the results to make decisions. Statistics can be divided into two broad areas:

Descriptive statistics: Collecting, summarizing, and presenting sample data using numerical and graphical methods.

Inferential statistics: Making estimates, decisions, predictions, or other generalizations about a larger set of data based on sampling.

Descriptive statistics involves

Data collection

Data classification

Describing and analyzing the data (measures of central tendency, dispersion)

Evaluating and presenting information contained in the data

Inferential statistics involves

Making inferences concerning an unknown aspect of the population being analyzed

decision making and risk management

quality control and quality assurance

Example:

The average weight of a package, based on a sample of 25 packages is 3.2 pounds.

What is the weight of an entire shipment of 1000 packages?

Terms and Definitions

Population: the complete collection of measurements, objects or individuals under study

Parameter: numerical characteristic of a population

Census: a survey of the entire population

A **sample** is a portion or a subset taken from a population

A **statistic** is numerical characteristic of a sample.

Inference is the process of obtaining the value of a population parameter from a sample statistic

Sampling

Many methods of taking samples exist. Random samples reduce the chance of introducing a bias or sampling error.

Simple Random Sampling: All elements of a population have equal chance to be selected for the sample.

Systematic sampling: A random starting point is selected and then every k-th number of the population is selected.

(k-th element) Sampling Interval = Population size/ sample size
e.g. we need 4 samples from a card deck(52) i.e. 52/4 = 13 therefore every 13th card will be taken as sample

Stratified sampling: A population is divided into groups, called strata, and a sample is randomly selected from each stratum. Categories should be chosen in such way that variability within them is minimal, with maximum variability between them.

take the same number of samples from each group that share same characteristics

Cluster Sampling: A population is divided into clusters using naturally occurring geographic or other boundaries. Then clusters are randomly selected and a sample is collected from each selected cluster. Clusters should have maximum variability within.

Exercises:

1. The AGT Corporation has branches in three major cities with a total of 326 salespeople. The sales manager wants to obtain a sample of 40 of his staff to determine their average gross sales per month. For each of the following identify the type of sample (random/systematic/stratified/cluster) obtained:

a) One of the three branches is randomly selected, and 40 people are selected from this branch.
b) Sales employees are numbered 1-326 and random number table is used to produce a sample of 40.
c) Salespeople are listed alphabetically. One salesperson is selected randomly as a starting point, then every 8th person down on the list is selected to produce a sample of 40.
d) Salespeople are categorized by years of service. A proportional number is randomly selected from each category to obtain a sample of 40.
e) State why it would be more practical to obtain data values from a sample rather than from the whole population.

2. Match each of the following terms to its correct definition:

Terms	**Definitions**
1. Parameter	a. The complete collection of items under study
2. Inferential statistics	b. A number that describes a sample characteristic
3. Census	c. Procedures for collecting, classifying, summarizing, and presenting data
4. Statistics	d. A number that describes a population characteristic
5. Population	e. The science of gathering and summarizing data and using results to make decisions
6. Descriptive statistics	f. A subset of a population
7. Sample	g. The process of arriving at a conclusion about a population parameter on the basis of a sample statistics
8. Statistic	h. A survey of all elements in a population

Answers:

1.

a) Cluster
b) Simple Random Sampling
c) Systematic Sampling
d) Stratified Sampling
e) Time, Money, Resources, etc..

2. 1d, 2g, 3h, 4e, 5a, 6c, 7f, 8b

Data Collection

When collecting data we can record attributes, measurements, counts etc.. The type of data that we end up with depends on what type of things we record.

Data can be classified as either qualitative or quantitative.

Qualitative data involves categorical variables.

Examples: gender, political affiliation, citizenship

Quantitative data involves numerical variables that are either discrete or continuous.

A **discrete numerical variable** can be determined by counting.

Example: number of students in a class, number of laptops on a network

A **continuous numerical variable** can be determined by measuring

Examples: speed (km/h), weight (kg), temperature (°C)

Levels of Measurement

Nominal scales identify, classify objects into mutually exclusive and collectively exhaustive classes (there is no particular order).

Example: Statistics Canada lists the marital status of the Canadian population, 15 years and older.

Single	7,285,560
Married	14,614,564
Divorced	1,452,000
Widowed	1,527,075
Total:	24,879,199

Ordinal scales identify, classify, and put objects into the order.

Example: Students in a college are classified according to year ranking.

Class Rank	Number
1st Year	13
2nd Year	17
3rd Year	9
Post Graduate	5

Interval scales identify, classify, and put objects in order according to the equal distances between scale values. The "zero point" on an interval scale is arbitrary; and negative values can be used.

Example: Temperature on the Celsius scale. Suppose the highest temperature during last three days was: 19 degrees, 9 degrees, 6 degrees.

Ratio scales identify, classify, put objects in order according to the equal distances between the scale values, and contain absolute zero point.

Example: Weight, Length, Time, Mass and Money values…

Exercises:

1. Classify each of the following data types as *qualitative* or *quantitative*. If the classification is quantitative, identify if it is *continuous* or *discrete*.

a) Height measurements in cm.
b) Student Programs (CST, AST, EE, ME).
c) Web site hits per day.
d) Total sales of gasoline, per month, in Liters.
e) Total units sold, per month.
f) Vote cast in last election.
g) Average room temperatures.
h) Student letter grades (A^+, A, B^+, B, etc…).
i) Student grades in percentages.
j) Student GPAs.

2. Identify the level of measurement in each of the following:

a) What type of Laptop do you prefer?

1. Lenovo
2. HP
3. Apple
4. Acer
5. Other

b) How old are you? _____

c) What is your gender? (Circle one) Male Female

d) How would you describe your political ideology?

1. Very Liberal
2. Liberal
3. Somewhat Liberal
4. Middle of the Road
5. Somewhat Conservative
6. Conservative
7. Very Conservative

e) What is your occupation?

1. Lawyer
2. Homemaker
3. Blue-Collar worker
4. Teacher
5. Doctor
6. Manager
7. Other professional

f) How much data is stored on your laptop in GB? _____

g) What is the average temperature in Brampton for January in °C? _____

h) In a study of the quality of city life, an observer measures the number of motorized vehicles passing a particular spot on an expressway between the hours of 6 a.m. and 10 a.m., each day for 20 days. The numbers of vehicles are recorded in 5 categories:

1: 0-999
2: 1000-1999
3: 2000-2999
4: 3000-3999
5: 4000-4999

Answers:

1.

a) Quantitative, continuous
b) Qualitative
c) Quantitative, discrete
d) Quantitative, continuous
e) Quantitative, discrete
f) Qualitative
g) Quantitative, continuous
h) Qualitative
i) Quantitative, continuous
j) Quantitative, continuous

2.

a) Nominal
b) Ratio
c) Nominal
d) Ordinal
e) Nominal
f) Ratio
g) Interval
h) Ordinal

Frequency Distributions and Histograms

Frequency distributions organize data items into compressed form without obscuring essential facts and patterns. They also provide insight into patterns in data. A histogram is a graphical representation of the frequency distribution.

Example

The following raw dataset shows the height (in inches) of fifty people from the GTA:

{68,71,64,70,65,69,61,71,67,70,66,73,62,71,69,65,67,69,63,68,66,65,72,64,
66,71,67,70,74,67,68,67,63,70,66,65,72,63,73,66,70,67,68,69,64,70,65,70,69,64}

In this case the population would be all people in the GTA, and the sample would by the fifty people from the GTA whose heights were measured and are shown in the dataset. The type of data is quantitative, and continuous because each data point is a numerical value that was measured.

The following frequency distribution gives some insight into patterns in the data including the average height, and the variation in the heights:

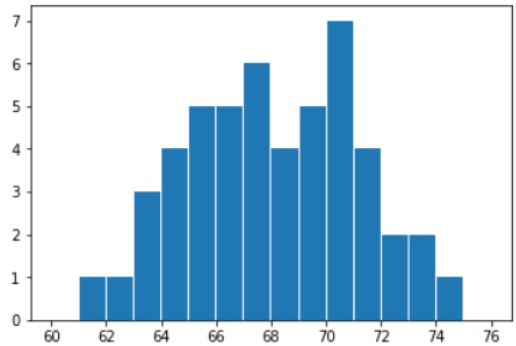
height (inches)	frequency
60	0
61	1
62	1
63	3
64	4
65	5
66	5
67	6
68	4
69	5
70	7
71	4
72	2
73	2
74	1
75	0

A histogram shows these insights visually:

```
data=[68,71,64,70,65,69,61,71,67,70,66,73,62,71,69,65,67,69,63,68,66,65,72,64,
      66,71,67,70,74,67,68,67,63,70,66,65,72,63,73,66,70,67,68,69,64,70,65,70,69,64]

import matplotlib.pyplot as plt

plt.hist(data,bins=16,range=[60,76],rwidth=0.95);
```



Exercises

1) Classify each of the following data types as qualitative or quantitative. If the classification is quantitative, identify if it is continuous or discrete.

a) Vehicle speeds.
b) Applied computing programs (CST, ISS, Mobile, CP, ...).
c) Web site hits per day.
d) Total volume sales per month in Litres.
e) Total customers per month.
f) Vote cast in last election (Liberal, PC, NDP, Green, ...)
g) Average daily room temperature.
h) Student letter grades (A+, A, B+, B, ...).
i) Student grades in percentages.
j) Student ID numbers.
k) Session times.

2) Manually compute the frequency distribution of the following quiz scores using five bins of width 2:

{1.0, 6.0, 5.0, 6.5, 3.0, 3.0, 5.0, 1.5, 9.5, 8.0, 6.0, 6.5, 5.0, 7.0, 7.5, 10.0, 8.0, 5.5, 7.5}

3) The following raw dataset shows the height (in inches) of fifty people from the GTA:

{68,71,64,70,65,69,61,71,67,70,66,73,62,71,69,65,67,69,63,68,66,65,72,64,
66,71,67,70,74,67,68,67,63,70,66,65,72,63,73,66,70,67,68,69,64,70,65,70,69,64}

Use a python jupyter notebook to generate three histograms of the heights using

a) 12 bins, and a range of 56 to 80 inches.
b) 7 bins, and a range of 61 to 75 inches.
c) 3 bins and a range of 60 to 75 inches.

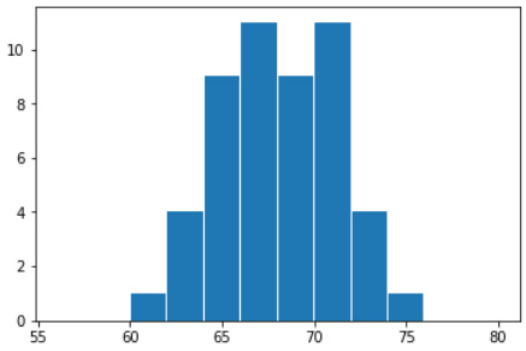
Answers

1) a) quantitative, continuous
   b) qualitative
   c) quantitative, discrete
   d) quantitative, continuous
   e) quantitative, discrete
   f) qualitative
   g) quantitative, continuous
   h) qualitative
   i) quantitative, continuous
   j) qualitative
   k) quantitative, continuous

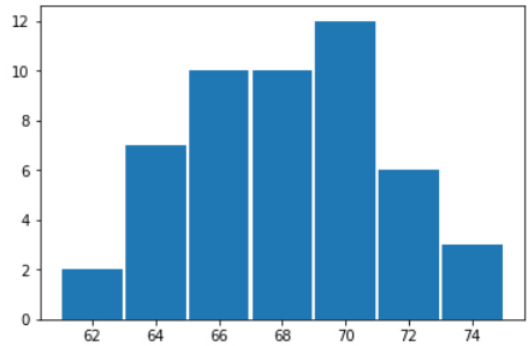
2)

height (inches)	**frequency**
0.0 to 1.9	2
2.0 to 3.9	2
4.0 to 5.9	4
6.0 to 7.9	7
8.0 or more	4

3) a)



b)



c)