Outline

§1.1 Populations and Samples

§1.2 Pictorial and Tabular Methods in Descriptive Statistics

- Bar chart

Statistics

Statistical concepts and methods helps us understand the world around us. We are constantly exposed with collections of facts, or data.

- **Statistics** is the science of understanding **data** and of making decisions in the presence of **variability** and **uncertainty**; it is a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

Example. How many legs does a cat have? – not (much) variability.

How much time do you study each day? – “variable”.

Will it snow tomorrow? – uncertain.

An investigation will typically focus on a well-defined collection of objects constituting a population of interest.

- The overall group of objects about which conclusions are to be drawn is called the (statistical) **population**.

When desired information is available for all objects in the population, we have a **census**; a census is usually impractical or infeasible due to constraints on money, time etc.. Instead, a subset of the population is selected.

- A subset of the population is called a (statistical) **sample**; the number of objects, denoted by n, is the **sample size**.

Example. Suppose we are interested in heights of UNB students. The population would be “*all* (?) UNB students”.

The term population is also used to refer to the characteristic of interest. In this example, we may say that *the collection of heights of all UNB students is the population*.

We hope to draw conclusion about the population based on a **representative** sample.

In **(simple) random sampling**, each subset of size n has the same chance of being chosen, resulting a likely representative sample.

We are usually interested only in certain characteristics of the objects in a population.

- A **variable** is any characteristic whose value may change from one object to another in the population.

Age, gender, heights etc. of a person are variables.

Data results from making observations either on a single variable or simultaneously on two or more variables.

A **univariate data set** consists of observations on a single variable.

We have **bivariate data** when observations are made on each of two variables. Our data set might consist of a (height, weight) pair for each basketball player on a team, with the first observation as (72, 168), the second as (75, 212), and so on.

Multivariate data arises when observations are made on more than one variable (so bivariate is a special case of multivariate).

An investigator who has collected data may wish simply to summarize and describe important features of the data. This entails using methods from **descriptive statistics**, such as contructing a histogram, or simply calculate an average.

More often, the researcher would like to use sample information to draw some type of conclusion (make an inference of some sort) about the population. Techniques for generalizing from a sample to a population are gathered within the branch of statistics called **inferential statistics**.

Statistics may be roughly classified into

- **descriptive statistics** methods to describe or picture a data set.
- **inferential statistics** methods by which conclusions are drawn about a large group of objects (population) based on observing only a portion (sample) of the objects in the larger group.

Random sampling model:

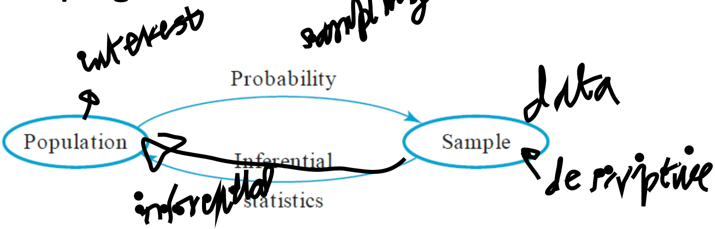


Figure 1.2 The relationship between probability and inferential statistics

Outline

§1.1 Populations and Samples

§1.2 Pictorial and Tabular Methods in Descriptive Statistics

Bar chart

Notation

- We denote variables by letters from the end of the alphabet. For example, variable X denotes the height of a randomly selected student. For an observed value, we use lowercase x. (The distribution of X is the population distribution.)

- For a sample of size n, the individual observations will be denoted by $x_1, x_2, \ldots, x_n$.

- Before the sample is taken, the random sample consists of $X_1, X_2, \ldots, X_n$. (These are independent and identically distributed (iid) as X.)

Types of variables

Statistics deals with data:

A **data table**:

Subject	Age	Gender	Weight	Diabetes?
1	27	M	182	No
2	45	M	205	Yes
3	34	F	136	No
4	52	M	177	Yes
5	42	F	143	No
6	29	F	125	No
7	67	M	176	Yes
8	43	F	147	No
⋮ n = 110	⋮	⋮	⋮	⋮

Types of variables

A **categorical** (or qualitative) variable records which of several categories a person or a thing is in. E.g., blood type, letter grade.

A **numerical** (or quantitative) variable records the amount of something. E.g., age in years, body height.

Types of variables

For some categorical variables, the categories can be arrayed in a meaningful order; such a variable is called a **ordinal variable**. E.g., letter grade.

A categorical variable that is not ordinal is called a **nominal variable**. E.g., blood type.

A, B, AB, O

Types of variables

A numeric variable that takes value on a continuous scale is a **continuous variable**. E.g., body height.

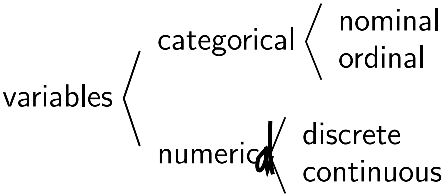
A numeric variable for which we can list the possible values is a **discrete variable**. E.g., age in years.

age intervals

regarded as continuous in analysis.

Types of variables

Variables: Summary



variables

categorical

nominal

ordinal

numerical

discrete

continuous

Types of variables

Example. We measure the birth weights of 20 babies.

The sample is 20 babies or 20 birth weights

The sample size is $n =$ 20

The variable is birth weight

The type of the variable is categorical/[✓ numerical].

continuous

Types of variables

Example. Five hundred potential voters are polled on which candidate out of five they will vote in an upcoming election.

The sample is 500 voters or their polling results.

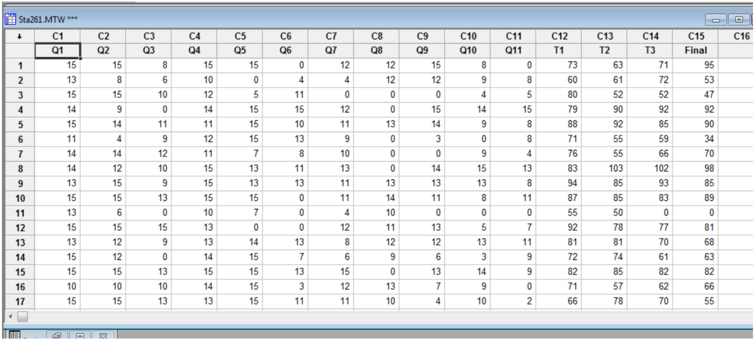
The sample size is $n =$ 500

The variable is one of the candidates

The type of the variable is [✓ categorical] / numerical.

nominal

How to summarize and/or visualize data?



Sta261.MTW ***

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	T1	T2	T3	Final	
1	15	15	8	15	15	0	12	12	15	8	0	73	63	71	95	
2	13	8	6	10	0	4	4	12	12	9	8	60	61	72	53	
3	15	15	10	12	5	11	0	0	0	4	5	80	52	52	47	
4	14	9	0	14	15	15	12	0	15	14	15	79	90	92	92	
5	15	14	11	11	15	10	11	13	14	9	8	88	92	85	90	
6	11	4	9	12	15	13	9	0	3	0	8	71	55	59	34	
7	14	14	12	11	7	8	10	0	0	9	4	76	55	66	70	
8	14	12	10	15	13	11	13	0	14	15	13	83	103	102	98	
9	13	15	9	15	13	13	11	13	13	13	8	94	85	93	85	
10	15	15	13	15	15	0	11	14	11	8	11	87	85	83	89	
11	13	6	0	10	7	0	4	10	0	0	0	55	50	0	0	
12	15	15	15	13	0	0	12	11	13	5	7	92	78	77	81	
13	13	12	9	13	14	13	8	12	12	13	11	81	81	70	68	
14	15	12	0	14	15	7	6	9	6	3	9	72	74	61	63	
15	15	15	13	15	15	13	15	0	13	14	9	82	85	82	82	
16	10	10	10	14	15	3	12	13	7	9	0	71	57	62	66	
17	15	15	13	13	15	11	11	10	4	10	2	66	78	70	55	

Frequency distributions

A **frequency distribution** is a display of the frequency, or number of occurrences, of each value in the data set.

The information can be presented in a tabular form, or, more vividly, with a graph.

Descriptive statistics can be divided into two general types:

graphical

- **Pictorial**: Summarize a data set using **visual** techniques such as a graph, a table.
- **Numerical**: Calculate a few numerical measures to summarize the key feature of the data set.

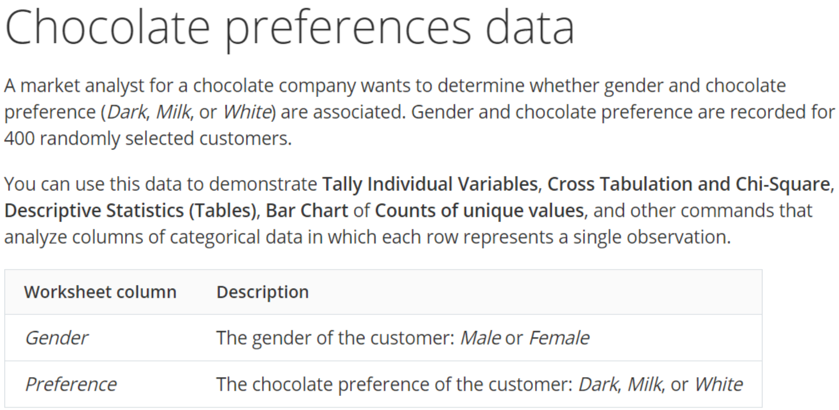
Outline

§1.1 Populations and Samples

§1.2 Pictorial and Tabular Methods in Descriptive Statistics
Bar chart

Bar chart

A **bar chart** is a graphic showing the categories that a categorical variable takes on and the number of observations in each category for the data in the sample.



Chocolate preferences data

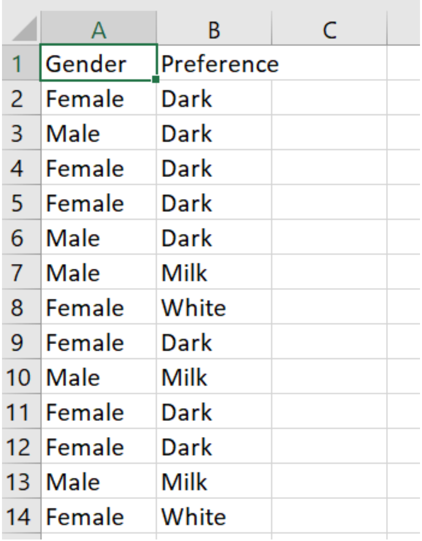
A market analyst for a chocolate company wants to determine whether gender and chocolate preference (*Dark*, *Milk*, or *White*) are associated. Gender and chocolate preference are recorded for 400 randomly selected customers.

You can use this data to demonstrate **Tally Individual Variables**, **Cross Tabulation and Chi-Square**, **Descriptive Statistics (Tables)**, **Bar Chart** of **Counts of unique values**, and other commands that analyze columns of categorical data in which each row represents a single observation.

Worksheet column	Description
Gender	The gender of the customer: *Male* or *Female*
Preference	The chocolate preference of the customer: *Dark*, *Milk*, or *White*

https://support.minitab.com/en-us/datasets/tables-data-sets/chocolate-preferences-data/

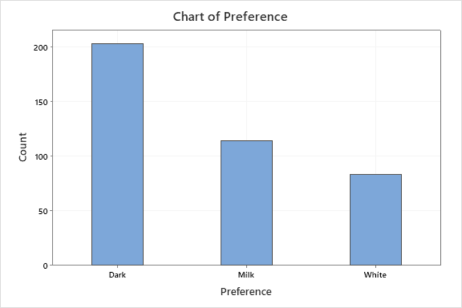
A portion of the data look like this:



Gender	Preference
Female	Dark
Male	Dark
Female	Dark
Female	Dark
Male	Dark
Male	Milk
Female	White
Female	Dark
Male	Milk
Female	Dark
Female	Dark
Male	Milk
Female	White

Bar chart

With a bar chart, we can visualize the distribution across different categories.



Bar chart: Do it by hand

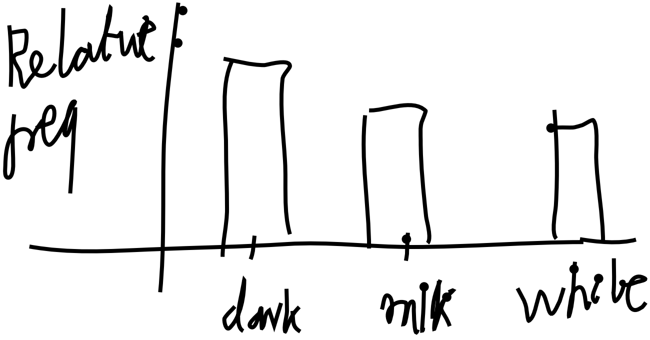
- Arrange categories in the horizontal axis.
- The vertical axis for the frequency of each category, or relative frequency.

times this cat occurs in the data.

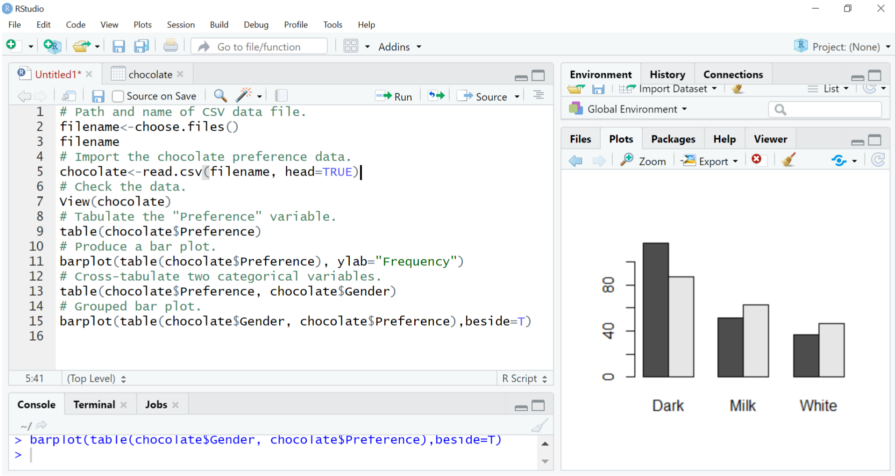
$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

(sample) Proportion

Bar chart: Do it by hand



Bar chart: Do it in R



```
# Path and name of CSV data file.
filename<-choose.files()
filename
# Import the chocolate preference data.
chocolate<-read.csv(filename, head=TRUE)
# Check the data.
View(chocolate)
# Tabulate the "Preference" variable.
table(chocolate$Preference)
# Produce a bar plot.
barplot(table(chocolate$Preference), ylab="Frequency")
# Cross-tabulate two categorical variables.
table(chocolate$Preference, chocolate$Gender)
# Grouped bar plot.
barplot(table(chocolate$Gender, chocolate$Preference),beside=T)
```

```
> barplot(table(chocolate$Gender, chocolate$Preference),beside=T)
>
```