

CHAPTER 1: INTRODUCTION

LESSON 1: INTRODUCTION

PART A: STRUCTURE OF THE TEXT (4TH EDITION OF TRIOLA'S ESSENTIALS)

Statistics

- 1) Designing an Experiment
(Chapter 1)
- 2) Collecting Data
(Chapter 1)
- 3) Describing Data
(Chapters 2, 3)
- 4) Interpreting Data
(Chapters 3, 7-11)

Probability (Chapters 4-6)

using

←-----

Population [of interest]

2) ↘ ↗ 4)

Sample

3)

Size: N elements (or members)
 All adult Americans?
 All registered voters in California?

Size: n elements (or members)
 For a poll? A scientific study?
 The Niensens?

The population must be carefully defined. For example ...

- Do “adult Americans” include illegal immigrants?
- Different polls of “likely U.S. voters” use different models for the purposes of screening poll respondents. Voter enthusiasm and voting history may be issues.

PART B: COLLECTING DATA; SAMPLING METHODS

If we manage to collect data from each element of the population, we have ourselves a census. Often, a census is impossible or impractical, so we collect data on only some elements of the population. These elements make up a sample from the population.

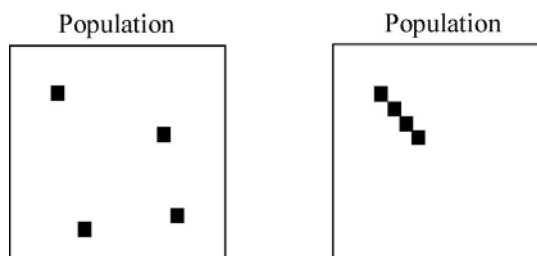
How do we select a sample so that it is representative of the overall population? [Section 1-4](#) details a number of methods. We will discuss related issues later, when we discuss polls in [Chapter 7](#).

A common problem in practice is the use of overly homogeneous samples in which the elements within the sample are much more similar than is the case within the overall population. For example, you wouldn't want to restrict your sample to a single state if you wanted to study the national popularity of the President.

We will typically assume that a sample from a population is a simple random sample (SRS). When constructing an SRS, each group of n elements in the population is equally likely to be our selected sample of n elements.

Example

If $n = 4$, then the two samples below (each represented by four black squares) are equally likely to be selected:



Note: When constructing a random sample in general, each individual element is equally likely to be among those selected for the sample, but some groups of n elements may be more likely to be selected than others. Selections could be linked.

In systematic sampling, the elements of the population are ordered, and, after some point, every k^{th} element is selected (where k is some integer greater than 1).

Example

We can select every third person that enters a particular bookstore on a particular day. Here, $k = 3$.

Skip Skip **Select** Skip Skip **Select** Skip Skip **Select**

CHAPTERS 2: DESCRIPTIVE STATISTICS I

How can we efficiently and effectively summarize data or compare data from two or more populations?

LESSON 2: FREQUENCY TABLES (SECTION 2-2)

Example

Let's summarize the ages (in years) at which the 43 U.S. Presidents became President (as of 2007). The complete list of 43 ages may not be so appealing to people!

The ages will be grouped into classes. [Triola](#) suggests using between 5 and 20 classes.

The frequency of each age class is the number of Presidential ages that lie within that class.

The relative frequency of each age class is obtained by dividing the corresponding frequency by N , which is the population size. Here, $N = 43$. The decision to round off relative frequencies to three decimal places was an arbitrary one.

Age classes	Frequency	Relative Frequency	Relative Frequency (as a percent)
35-39	0	0	0%
40-44	2	0.047	4.7%
45-49	6	0.140	14.0%
50-54	13	0.302	30.2%
55-59	12	0.279	27.9%
60-64	7	0.163	16.3%
65-69	3	0.070	7.0%
70+	0	0	0%
	Sum = N = 43	Sum = 1	Sum = 100%

Note on Rounding: [Triola](#) obsesses over this issue, but the rounding issue really depends on the application. Here, we have rounded down ages.

Note on Roundoff Error: The rounded off values in the “Relative Frequency” column actually add up to 1.001, but we shouldn’t worry about this. If the sum were something like 2, then we should worry!

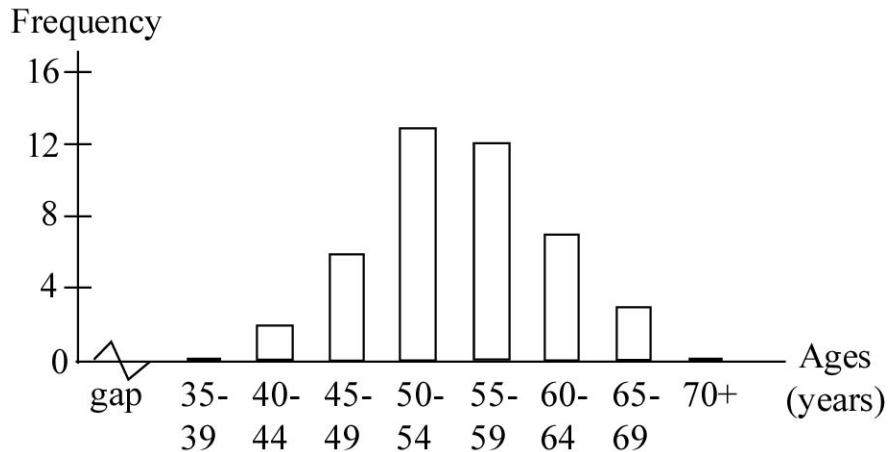
Note on Trailing Zeros: Why, for instance, do we write “0.140” as opposed to simply “0.14”? The trailing zero at the end of “0.140” indicates that 0.140 is accurate to three decimal places. Given that we are rounding off relative frequencies to three decimal places, the “0.14” might be mistaken to be an exact value.


Note on Classes: Observe that the classes are the same size, with the exception of the “70+” class. We typically avoid using classes of unequal size: for example, 40-49 and 50-54. Also, the “35-39” class was included because 35 is the minimum required age for the U.S. Presidency mandated by the Constitution.

Historical Notes: We are counting Grover Cleveland twice, because he served two nonconsecutive terms. The youngest President was Teddy Roosevelt, who was 42 when he succeeded William McKinley on his assassination. John Kennedy was the youngest elected President at the age of 43. Ronald Reagan was the oldest elected President at the age of 69.

LESSON 3: HISTOGRAMS (SECTION 2-3)

Here is a frequency histogram for the example in [Lesson 2](#):

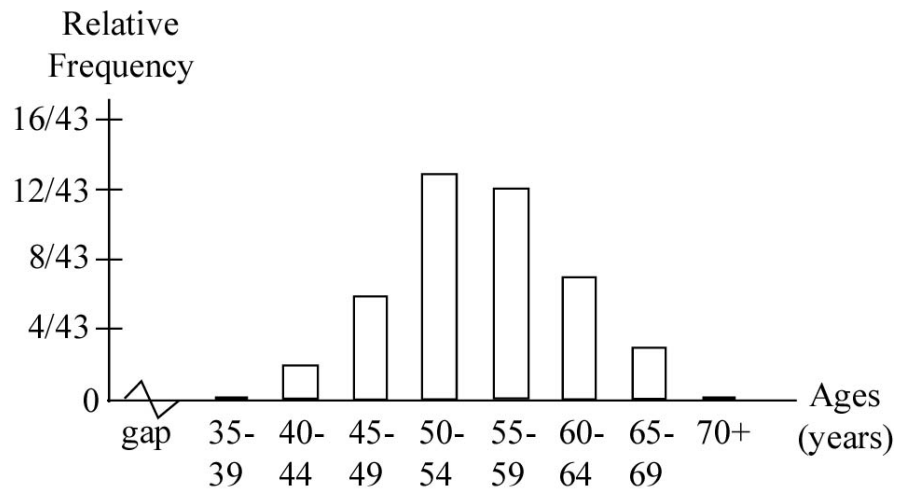


Observe that the histogram resembles a bell curve ; this is common for data involving natural measures such as ages and heights. We will be studying bell curves for much of this course.

Notes on Scaling: Observe that the tick marks on the Frequency axis are evenly spaced, and the highest labeled tick mark (16) is at least as high as the highest frequency (13) among the classes. The “gap” between 0 and 35 years is indicated on the Ages axis.

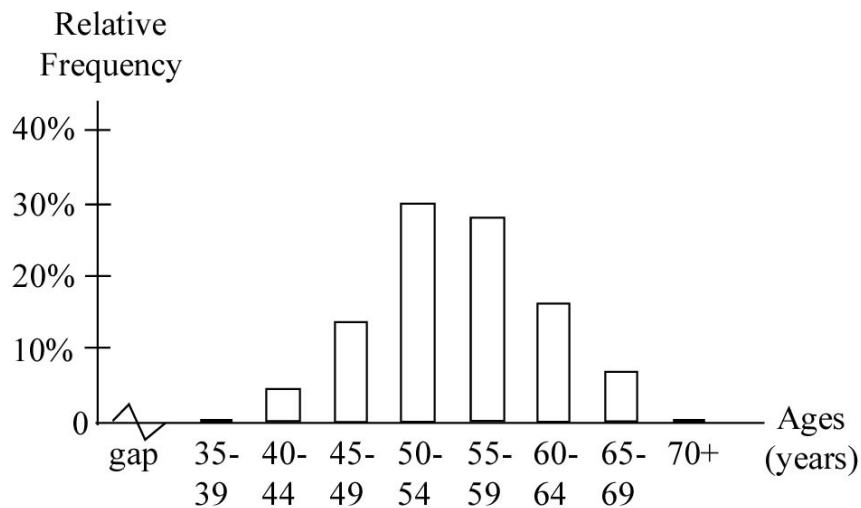
Note: Try not to stripe your bars. Stripes tend to lead to illusions of vibration, referred to as the Moiré effect.

To construct a corresponding relative frequency histogram, we rescale the vertical axis. We could simply divide the vertical tick marks by N (i.e., 43):



Note: For sample data, you would divide by n .

A better idea may be to use “nice” percents along the vertical axis:



The shape of a relative frequency histogram indicates the distribution of the data values. Here, we have a roughly bell-shaped distribution.

Think About It

If a die is rolled 100 times, would you expect a relative frequency histogram of the die results (numbered 1-6) to have the same shape?

LESSON 4: MORE STATISTICAL GRAPHICS (SECTION 2-4)

Look at [my MINITAB handout](#).

PART A: DOTPLOTS

These are like histograms with many “thin” classes, except that stacked dots replace the bars.

PART B: STEMLOTS, OR STEM-AND-LEAF PLOTS

These resemble sideways histograms, but you can use them to:

- recover the original data values
- sort the data values (i.e., place them in numerical order)

Example

Nine students take a test. Their scores are as follows:

77 93 73 51 74 85 82 73 100

Stage 1: Assign leaves to stems.

The leaf of a data value is the value’s last digit.

The stem consists of the other digits.

Note: If the data values are not integers, they must be written out to the same number of decimal places.

Write the stems in increasing order.

Comments

5		1	← Corresponds to 51
6			← No scores in the 60s
7		7343	← Correspond to 77, 73, 74, and 73
8		52	
9		3	
10		0	

Note: Do not skip the “6”; include it as a stem.
Otherwise, the shape of the distribution will be distorted.

Note: Include repetitions. There were two “73”s.

Stage 2: Sort the leaves for each stem (in increasing order).

5		1
6		
7		3347
8		25
9		3
10		0

Note: Stages 1 and 2 can be done simultaneously.

PART C: SCATTERPLOTS, OR SCATTER DIAGRAMS

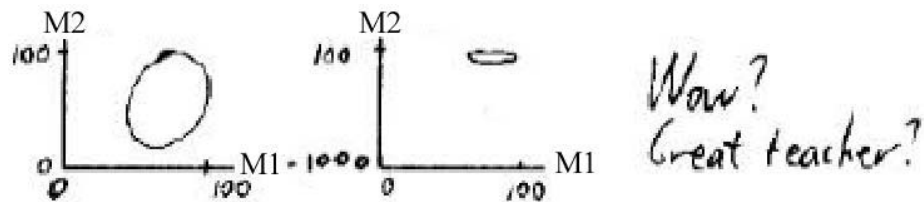
We will discuss these further in [Chapter 10](#).

These are for paired data of the form (x_i, y_i) .

Example

In the [MINITAB handout](#), student # i received a score of x_i on Midterm 1 and a score of y_i on Midterm 2. ($1 \leq i \leq N = 92$)

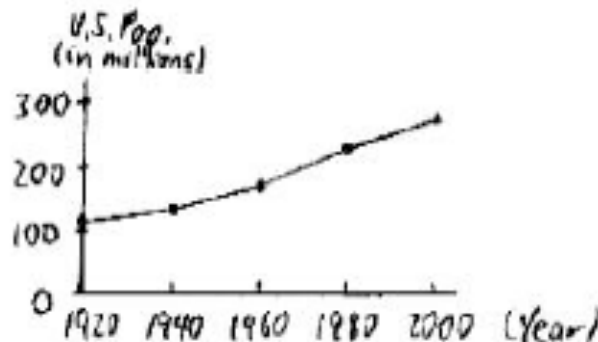
Beware of abuse! Rescaling axes can dramatically change the picture.



PART D: TIME-SERIES GRAPHS

Example

Year	U.S. Population (in millions)
1920	106
1940	132
1960	179
1980	227
2000	281



Thus far, we have been dealing with quantitative data, the most common form of numerical data.

We will now look at qualitative data.

PART E: PARETO CHARTS

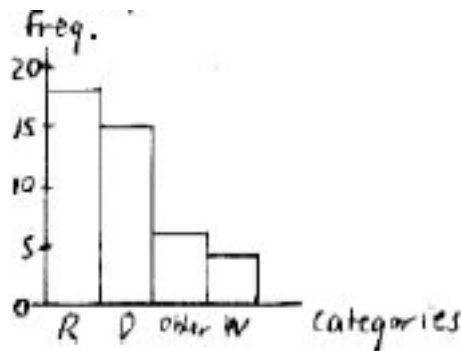
Example

Political parties of the 43 U.S. Presidents (as of 2007):

Party affiliation	Frequency
Democrats (D)	15
Republicans (R)	18
Whigs (W)	4
Other	6

Historical Note: The six “Others” included four Democratic-Republicans and two Federalists.

A Pareto chart resembles a histogram, except that the horizontal axis is organized by categories, not by quantitative measurements. The categories are listed in descending order of frequency.



PART F: PIE CHARTSExample (Presidential Parties again)

Again, we have:

Party affiliation	Frequency
Democrats (D)	15
Republicans (R)	18
Whigs (W)	4
Other	6

 $N = 43$.For each party, the relative frequency is given by: $\frac{\text{Frequency}}{43}$.The measure of the central angle of the corresponding pie slice is given by: $(\text{Relative Frequency}) \times (360^\circ)$. Remember that there are 360 degrees in a full counterclockwise revolution sweeping out the boundary circle.

Here, we will round off relative frequencies to three decimal places and angle measures to the nearest degree.

Party affiliation	Frequency	Relative Frequency	Measure of Central Angle
Democrats (D)	15	0.349 (34.9%)	126°
Republicans (R)	18	0.419 (41.9%)	151°
Whigs (W)	4	0.093 (9.3%)	33°
Other	6	0.140 (14.0%)	50°

