

1

Chapter 1: Introduction

Material from Devore's book (Ed 8), and Cengagebrain.com

Populations and Samples

An investigation of some characteristic of a **population** of interest.

Example: Say you want to study if a specific generic drug has the same amount of active ingredient as the brand (eg: Aleve vs generic ibuprophen)

Population:

- 1) All generic ibuprophen pills
- 2) All Aleve pills

Characteristic of interest:

amount of active ingredient (ibuprophen)

Populations and Samples, cont.

What statisticians need to do:

- 1) Learn about **the distribution of the characteristic** (amount of active ingredient) in each population
- 2) Evaluate the claim given to us by the manufacturers (“do generic drugs contain the same amount of active ingredient as the brand ones?”)
- 3) How? Constraints on time, money, and other resources usually make a complete census infeasible.
- 4) Answer: a subset of the population—a **sample**—is selected in some manner
- 5) **Sample statistics** and exploratory data analyses (EDA) are performed to “learn” about the characteristics of interest

Populations and Samples, cont.

The following samples of amounts (mg) in pills were collected (8 per group):

Brand: 5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5

Gener: 5.3 4.1 7.2 6.5 4.8 4.9 5.8 5.0

What can we find out?

EDA: Histograms, frequencies, central values (means, medians, modes), spread (observed range, standard deviation, variance)

Frequencies

The **relative frequency** (sometimes called the “density”) of a group of values is the fraction or proportion of times the values in that group occur, relative to all the values:

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

The **absolute frequency** of a **group** of values is the number of times the values in that group occur in the sample (ie, the numerator above).

Frequencies

- Measurements that can take on infinitely many values: group them (make histograms, discuss frequencies).
- Measurements that take on a finite number of values: talk about frequency of a single value (can group for simplicity).

Example: A data set consists of 200 observations on x = the number of courses a college student is taking this term. If 70 of these x values are 3, then

frequency of the x value 3: 70

relative frequency of the x value $\frac{70}{200} = .35$

Example 2

Charity is a big business in the United States. The Web site charitynavigator.com gives information on roughly 5500 charitable organizations.

Some charities operate very efficiently, with fundraising and administrative expenses that are only a small percentage of total expenses, whereas others spend a high percentage of what they take in on such activities.

Example 2 – sample data

cont' d

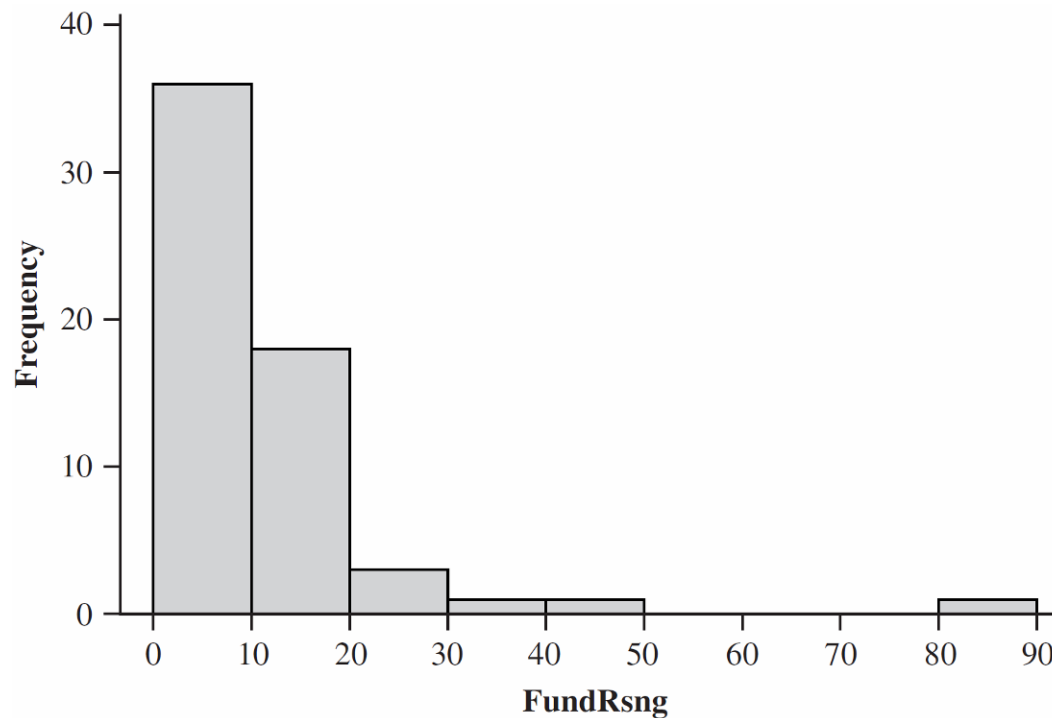
Here are the data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

| | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 6.1 | 12.6 | 34.7 | 1.6 | 18.8 | 2.2 | 3.0 | 2.2 | 5.6 | 3.8 |
| 2.2 | 3.1 | 1.3 | 1.1 | 14.1 | 4.0 | 21.0 | 6.1 | 1.3 | 20.4 |
| 7.5 | 3.9 | 10.1 | 8.1 | 19.5 | 5.2 | 12.0 | 15.8 | 10.4 | 5.2 |
| 6.4 | 10.8 | 83.1 | 3.6 | 6.2 | 6.3 | 16.3 | 12.7 | 1.3 | 0.8 |
| 8.8 | 5.1 | 3.7 | 26.3 | 6.0 | 48.0 | 8.2 | 11.7 | 7.2 | 3.9 |
| 15.3 | 16.6 | 8.8 | 12.0 | 4.7 | 14.7 | 6.4 | 17.0 | 2.5 | 16.2 |

Example 2 - histogram

cont' d

We can see that a substantial majority of the charities in the sample spend less than 20% on fundraising:



Histogram Shapes

Histograms come in a variety of shapes.

- **Unimodal** histogram: single peak
- **Bimodal** histogram: two different peaks
- **Multimodal** histogram: many different peaks

Bimodality: Can occur when the data set consists of observations on two quite different kinds of individuals or objects.

Multimodality:

Symmetric histograms

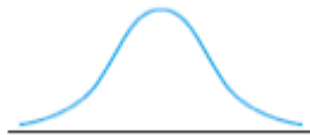
Positively skewed histograms

Negatively skewed histograms

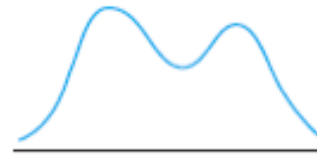
Examples

cont' d

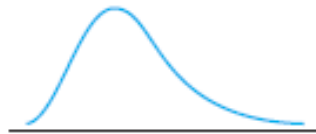
“Smoothed” histograms, obtained by superimposing a smooth curve on the rectangles, that illustrate the various possibilities.



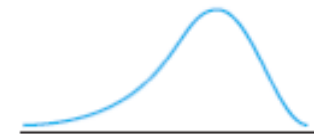
(a) symmetric unimodal



(b) bimodal



(c) Positively skewed

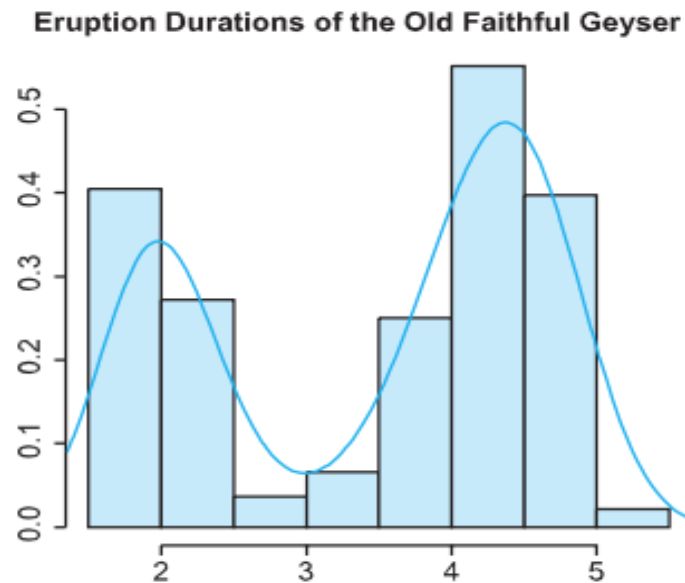


(d) negatively skewed

Smoothed histograms

Example 3

Histogram (with area adjusted to one) of the Old Faithful geyser's eruption durations (272 eruption durations in minutes) with a smooth histogram imposed, note the bimodal shape.



R CODE FOR HISTOGRAMS

Sample Statistics

- Histograms & other visual summaries of samples are excellent tools for informal learning about population characteristics.
- However, for a deeper analysis of the data, the calculation and interpretation of certain summarizing numbers (numbers that might serve to characterize the data) are required.
- These *sample numerical summaries* are called “**Sample Statistics**”

Sidenote about Inferential Statistics:

Sample statistics are great for describing the data (i.e. the sample) but they do not tell us anything rigorous yet (i.e. we cannot make any mathematically sound conclusions about the population).

Statistical inference is about making statistically rigorous statements and conclusions based on sample statistics.

More specifically: Having obtained a sample from a population, an investigator uses sample information to draw some type of conclusion (i.e an inference) about the underlying population.

Techniques for rigorously generalizing from a sample to a population are called **inferential statistics**.

We'll do this later in the course, back to sample statistics...

Examples of numerical summaries of the sample:
“Center” of the sample

Sample **center**: An important characteristic of a set of numbers.

3 popular types of “center” :

1. Mean
2. Median
3. Mode

The Mean

For a given set of numbers x_1, x_2, \dots, x_n , the most familiar measure of the center is the *mean* (arithmetic average).

Sample mean \bar{x} of observations x_1, x_2, \dots, x_n :

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Sample and Population Mean

- \bar{x} = **the sample mean** (represents the average value of the observations in a sample)
- The greek letter μ denotes the **population mean**, this is the arithmetic average of the of all the elements of the population which can (in theory) be calculated.

Difference between \bar{x} and μ ?

More general definitions of μ later.

The Sample Mean

- Main problem with sample mean: outliers
- Despite this, sample mean is most common measure (outliers unlikely)
- Outliers not so unlikely: measures that are less sensitive

The Median

- **Median** = “middle”: Middle value when observations are ordered smallest to largest.
- \tilde{x} = sample median (observations are denoted by x_1, \dots, x_n)
- $\tilde{\mu}$ = population median

The Median

Calculate **sample median**:

- Order the n observations smallest to largest (repeated values included)
- Find the middle one:

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

Example 1, cont

Brand: 5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5

Generic: 5.3 4.1 7.2 6.5 4.8 4.9 5.8 5.0

Brand sample median: 5.1 5.5 5.6 5.8 5.8 6.0 6.2 6.5
median (brand)

5.8

Generic sample median: 4.1 4.8 4.9 5.0 5.3 5.8 6.5 7.2
median (generic)

5.15

The Median

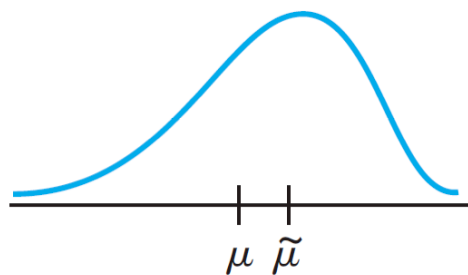
The population mean μ and median $\tilde{\mu}$. will not generally be identical. If the population distribution is positively or negatively skewed, as pictured below, then $\mu \neq \tilde{\mu}$.

In this case: which population characteristic most important?

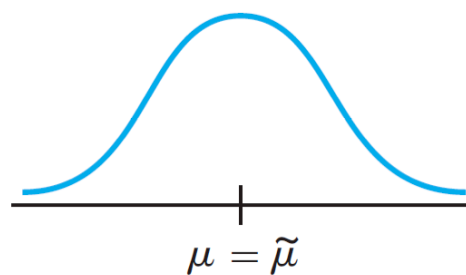
The Median

The population mean μ and median $\tilde{\mu}$ will not generally be identical. If the population distribution is positively or negatively skewed, as pictured below, then $\mu \neq \tilde{\mu}$.

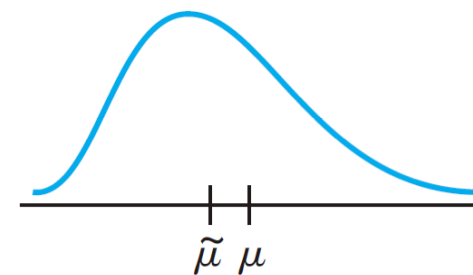
In this case: which population characteristic most important?



(a) Negative skew



(b) Symmetric



(c) Positive skew

Three different shapes for a population distribution

Other Sample Measures of Location: Quartiles, Percentiles, and Trimmed Means

Median: divides data set into **two** parts of equal size.

Finer measures of location: divide data into more than two such parts.

- **Quartiles:** divide the data set into **four** equal parts (how is this calculated?)
- **Percentiles:** A data set can be even more finely divided using percentiles (examples? What does a “percentile” mean?)

Other Sample Measures of Location: Quartiles, Percentiles, and Trimmed Means

- The mean: sensitive to a single outlier
- The median: impervious to many outlying values.
- Extreme behavior (sensitive, impervious) can be undesirable, so use alternative measures.

A **trimmed mean** is a compromise between the mean and the median. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.

Summary: types of Data

So far we have talked about finite and infinite populations.

To be more specific, we can have the following types of data:

- 1) Finite number of values
- 2) Infinite but countable number of values
- 3) Uncountably infinitely many values
- 4) Finite number of categories (“labels”) which are ordered
- 5) Finite number of categories (“labels”) which aren’t ordered

The 4th group of data are called “ordinal”.

The 5th group of data are simply called “categorical”.

Categorical Data and Sample Proportions

Categorical data: represent with frequency distribution or relative frequency distribution (provides tabular summary).

Example: Survey individuals who own digital cameras to study brand preference (e.g. Canon, Sony, Kodak)

Ordinal or categorical?

Categorical Data and Sample Proportions

Consider sampling a dichotomous (2 categories) population —such as “did vote” or “did not vote” in the last election, “does own a digital camera” or “does not own a digital camera”, etc.

If we let x denote the number in the sample falling in category 1, then the number in category 2 is $n - x$. The relative frequency or

sample proportion in category 1 is

$$x / n$$

sample proportion in category 2 is

$$1 - x / n$$

Categorical Data and Sample Proportions

Let's denote a response that falls in category 1 by a "1" and a response that falls in category 2 by a "0".

A sample of $n = 10$ random people yielded the following responses: 1, 1, 0, 1, 1, 1, 0, 0, 1, 1 (i.e., 7 voted, and 3 did not vote).

The sample mean for this sample is (since number of 1s = $x = 7$)

$$\frac{x_1 + \cdots + x_n}{n} = \frac{1 + 1 + 0 + \cdots + 1 + 1}{10} = \frac{7}{10} = \frac{x}{n} = \text{sample proportion}$$

What can be said about the sample proportion and the sample mean?

Categorical Data and Sample Proportions

The sample proportion of observations in category 1 is the actually also the sample mean.

Thus a sample mean can be used to summarize the results of a dichotomous sample.

More than 2 categories?

Categorical Data and Sample Proportions

We often use p to represent the **proportion** of those in the entire population falling in the category.

As with x / n , p is a quantity between 0 and 1, and while x / n is a sample characteristic, p is a characteristic of the population.

Some people prefer to use p_k for sample proportion of category k , and the Greek π_k for the population proportion of category k . (Yay! Greek letters help understand notation!)

Variability

So far, we've learned

- About the center of our sample (3 common centers - how are they each useful?)
- To visualize the sample distribution (Why is this useful?)

Next: Quantifying the variability of the data in the sample. What does “variability” mean?

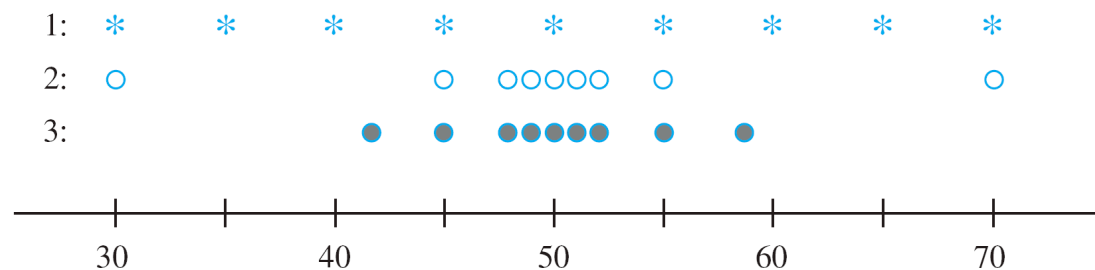
Measures of Variability

Measure of the center = only partial information about a data set or distribution. Different samples or populations may have the same measures of center, but differ from one another in other important ways. Example?

Measures of Variability

Measure of the center = only partial information about a data set or distribution. Different samples or populations may have the same measures of center, but differ from one another in other important ways. Example?

Figure below shows dotplots of three samples with the same mean and median, yet the extent of spread about the center is different for all three samples.

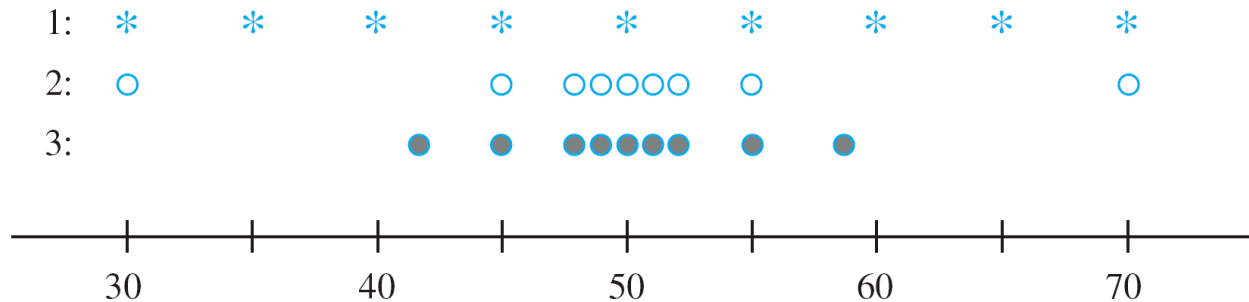


Samples with identical measures of center but different amounts of variability

Measures of Variability for Sample Data

Simplest measure of variability: The **range**

The value of the range for sample 1 is much larger than it is for sample 3, reflecting more variability in the first sample than in the third.



Samples with identical measures of center but different amounts of variability

Measures of Variability for Sample Data

Defect of the range: depends on *only* the two most extreme observations (how many are disregarded?). What is happening in samples 1 and 2?

A more robust measure of variation takes into account **deviations from the mean**

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}.$$

Measures of Variability for Sample Data

Can we combine the deviations into a single quantity by finding the average deviation?

Measures of Variability for Sample Data

Can we combine the deviations into a single quantity by finding the average deviation? No:

$$\text{sum of deviations} = \sum_{i=1}^n (x_i - \bar{x}) = 0$$

-- the average deviation will always be zero:

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n} \sum x_i\right) = 0$$

How can we prevent negative and positive deviations from counteracting one another when they are combined?

Measures of Variability for Sample Data

Working with the absolute values:

$$\sum |x_i - \bar{x}|/n.$$

Why is this a problem? What is a different solution?

Measures of Variability for Sample Data

Working with the absolute values:

$$\sum |x_i - \bar{x}|/n.$$

Squared deviations

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2.$$

Rather than use the average squared deviation, $\sum(x_i - \bar{x})^2/n$, in samples we divide the sum of squared deviations by $n - 1$ rather than n . WHY?

Measures of Variability for Sample Data

The **sample variance**, denoted by s^2 , is given by

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation**, denoted by s , is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

Note that s^2 and s are both nonnegative. The unit for s is the same as the unit for each of the x_i .

Example 5

www.fueleconomy.gov contains a wealth of information about fuel efficiency (mpg). Consider the following sample of $n = 11$ efficiencies for the 2009 Ford Focus equipped with an automatic transmission:

| Car | x_i | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|-----|--------------------|------------------------------|------------------------------------|
| 1 | 27.3 | -5.96 | 35.522 |
| 2 | 27.9 | -5.36 | 28.730 |
| 3 | 32.9 | -0.36 | 0.130 |
| 4 | 35.2 | 1.94 | 3.764 |
| 5 | 44.9 | 11.64 | 135.490 |
| 6 | 39.9 | 6.64 | 44.090 |
| 7 | 30.0 | -3.26 | 10.628 |
| 8 | 29.7 | -3.56 | 12.674 |
| 9 | 28.5 | -4.76 | 22.658 |
| 10 | 32.0 | -1.26 | 1.588 |
| 11 | 37.6 | 4.34 | 18.836 |
| | $\sum x_i = 365.9$ | $\sum (x_i - \bar{x}) = .04$ | $\sum (x_i - \bar{x})^2 = 314.106$ |

$\bar{x} = 33.26$

Example 5

The numerator of s^2 is $S_{xx} = 314.106$, from which

$$s^2 = \frac{S_{xx}}{n - 1} = \frac{314.106}{11 - 1}$$

$$= 31.41,$$

$$s = 5.60$$

The size of a representative deviation from the sample mean 33.26 is roughly 5.6 mpg.

Population equivalents

Note that whereas s^2 measures sample variability, there is a measure of variability in the population called the *population variance*.

We will use σ^2 to denote the population variance and σ to denote the population standard deviation.

SUMMARIZING DATA IN R