Statistical Methods
APPM 4570/5570, STAT 4000/5000

Week 1:
Intro to R and EDA

Populations and Samples

Objective: study of a characteristic (measurable quantity, random variable) for a **population** of interest.

Example: amount of active ingredient in a generic and a name brand drug

2 populations:

1) All generic pills
2) All specific brand pills

Characteristic of interest:
amount of active ingredient

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 (“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. Answer: a subset of the population–**a sample**–is selected in some manner

4) **Sample statistics** and exploratory data analyses (EDA) are performed to “learn” about (infer) the characteristics of interest

Populations and Samples, cont.

DATA from 2 random samples:
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 say based on these data?

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

Working with R: ingredient.csv

```
> ingredient = read.csv(file="ingredient.csv", header=TRUE,
                       sep=",")
> ingredient
```

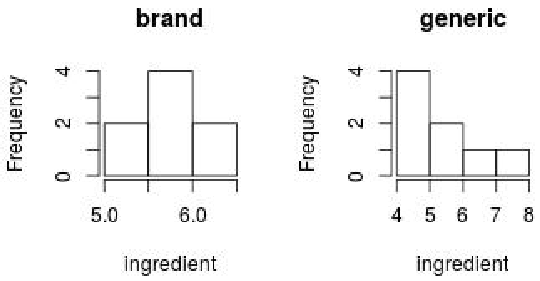
```
  brand generic
1  5.6    5.3
2  5.1    4.1
3  6.2    7.2
4  6.0    6.5
5  5.8    4.8
6  6.5    4.9
7  5.8    5.8
8  5.5    5.0
```

Graphical summaries: histograms

Objective: get a sense of the values in two groups visually

Histograms provide a quick way of visualizing the data values

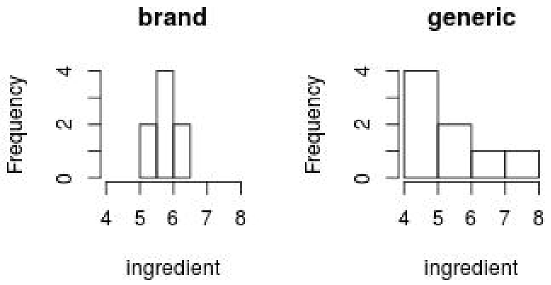
```
> jpeg(file='hist.jpg', width=350, height=200)
> par(mfrow=c(1,2))
> hist(ingredient$brand, main="brand",xlab="ingredient")
> hist(ingredient$generic, main="generic",xlab="ingredient")
> dev.off()
```



Graphical summaries: histograms

A lot easier to compare if the data are plotted on the same scale!

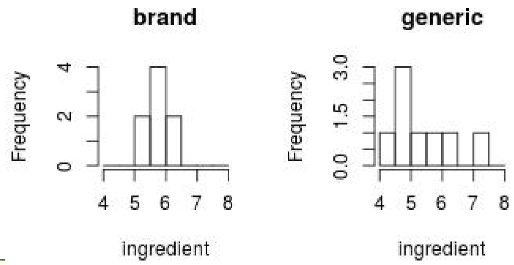
```
> hist(ingredient$brand, main="brand",xlab="ingredient", xlim=c(4,8))
> hist(ingredient$generic, main="generic",xlab="ingredient", xlim=c(4,8))
```



Graphical summaries: histograms

Even better if histograms have the same bin width!

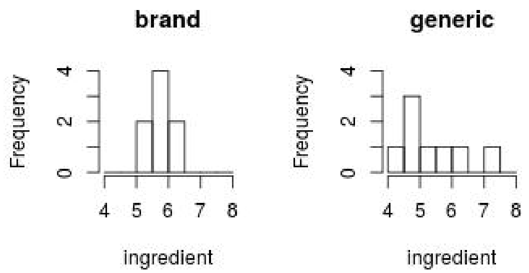
```
> hist(ingredient$brand, main="brand",xlab="ingredient", xlim=c(4,8), breaks =
seq(4,8,.25))

> hist(ingredient$generic, main="generic",xlab="ingredient", xlim=c(4,8), breaks =
seq(4,8,.25))
```



Graphical summaries: histograms

And even better if y axes have the same scale too!

```
> hist(ingredient$brand, main="brand",xlab="ingredient",breaks=seq(4,8,.5),ylim=c(0,4))
> hist(ingredient$generic,
                      main="generic",xlab="ingredient",breaks=seq(4,8,.5),ylim=c(0,4))
```

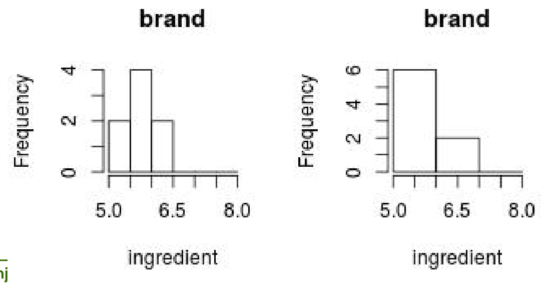


Graphical summaries: histograms

If you play with bin widths and starting values until you get the shape they want, disclose that – and show a few other ones too, for comparison.

Below are two shapes of the same data: symmetric and exponentially decaying.

```
> hist(ingredient$brand, main="brand",xlab="ingredient",breaks=seq(5,8,.5))
> hist(ingredient$brand, main="brand",xlab="ingredient",breaks=seq(5,8,1))
```



Frequencies – two types

Relative frequency ("density") of a **set** of values is the fraction or proportion of times the values in that set occur, relative to all the values:

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

Absolute frequency of a **set** of values is the number of times the values in that group occur in the sample – ie, the numerator above

Frequencies

In practice, we often group data values into "bins" in order to make histograms and discuss frequencies.

When there are only finitely many values possible, we can talk about a frequency of a single value – though we may still want to group them for simplicity.

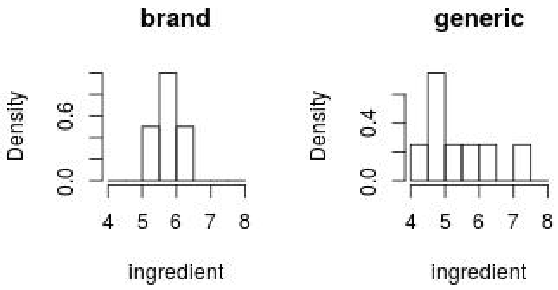
Suppose that our data set consists of 200 observations on *x* = the number of courses a college student is taking this term. If 70 of these *x* values are 3, then

frequency of the *x* value 3: 70

relative frequency of the *x* value 3: $\frac{70}{200} = .35$

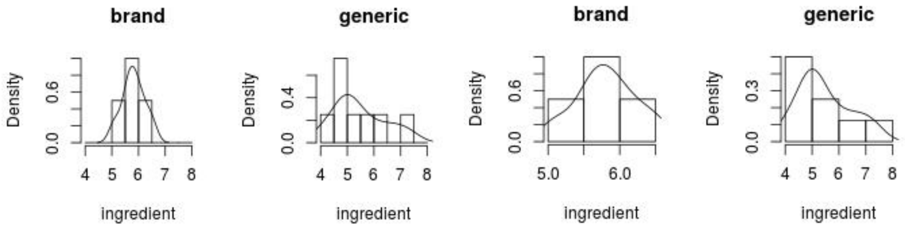
Histograms with relative frequency

```
hist(ingredient$brand, freq=FALSE,
           main="brand",xlab="ingredient",breaks=seq(4,8,.5),ylim=c(0,4))
> hist(ingredient$generic, freq=FALSE,
           main="generic",xlab="ingredient",breaks=seq(4,8,.5),ylim=c(0,4))
```



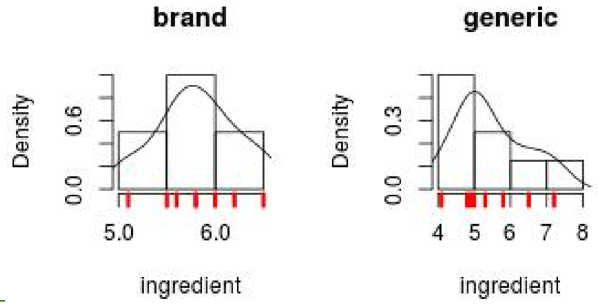
Histograms with a kernel smooth

```
 > hist(ingredient$brand, freq=FALSE, main="brand",xlab="ingredient",breaks=seq(4,8,.5))
> lines(density(ingredient$brand))
> hist(ingredient$generic, freq=FALSE,
                           main="generic",xlab="ingredient",breaks=seq(4,8,.5))
> lines(density(ingredient$generic))
```



Best to check all data if you can

```
 > hist(ingredient$brand, freq=FALSE, main="brand",xlab="ingredient",breaks=seq(4,8,.5))
> lines(density(ingredient$brand))
> rug(ingredient$brand,ticksize=-0.1,col='red',lwd=3)
> hist(ingredient$generic, freq=FALSE,
                           main="generic",xlab="ingredient",breaks=seq(4,8,.5))
> lines(density(ingredient$generic))
> rug(ingredient$generic,ticksize=-0.1,col='red',lwd=3)
```



Example 2 (Devore)

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

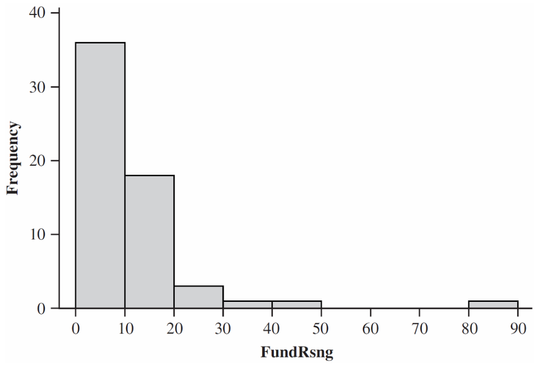
Here are the data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

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

A **unimodal** histogram only has a single peak.

A **bimodal** histogram has two different peaks.

A **multimodal** histogram has many different peaks.

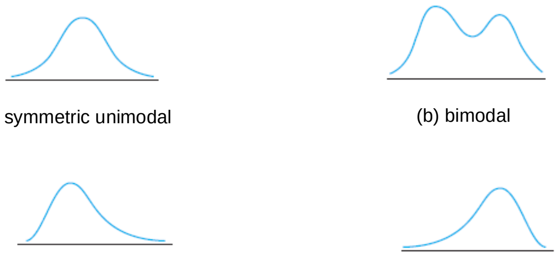
Bimodality can occur when the data set has observations on two well differentiated kinds of individuals or objects. Multimodality can occur when it has many well differentiated types of observations.

A histogram is **symmetric** if the left half is a mirror image of the right half.

A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail..... and **negatively skewed** if the stretching is to the left.

Skewness and multiple modes

cont'd



(a) symmetric unimodal

(b) bimodal

(c) Positively skewed

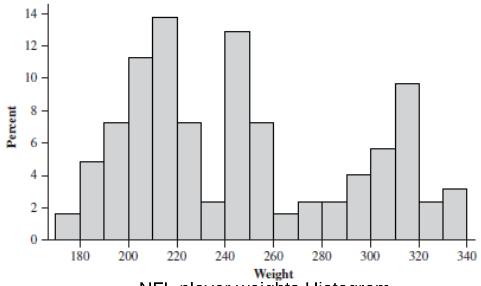
(d) negatively skewed

Smoothed histograms

Example 3 (Devore)

3 modes:

Histogram of the weights (lb) of the 124 players listed on the rosters of the San Francisco 49ers and the New England Patriots as of Nov. 20, 2009.



NFL player weights Histogram

Numerical summaries of samples

Histograms and other visual summaries of data are great tools for learning about population characteristics.

More formal data analysis often requires numerical summary measures (still considered exploratory data analysis), as well as inference.

These numbers summarize the data set and convey some of its salient features.

We call these sample summaries "Sample Statistics"

Inferential statistics

Sample statistics describe the data; but they do not tell anything rigorous yet

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

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

We'll do this later in the course

"Center" of the sample

Suppose, that our sample is of the form
x_1, x_2,. . ., x_n, where each x_i is a number.

How can we summarize this set of numbers? One important characteristic of a set of numbers is the sample **center**.

You've probably heard about 3 types of "center" notions:

1. Mean
2. Median
3. Mode

The Sample Mean

For a given set of numbers x_1, x_2,. . ., x_n, the most familiar and useful measure of the center is the ***mean***, or **arithmetic average** of the set.

It's the center of mass: the point at which the whole distribution (histogram) of the sample will balance

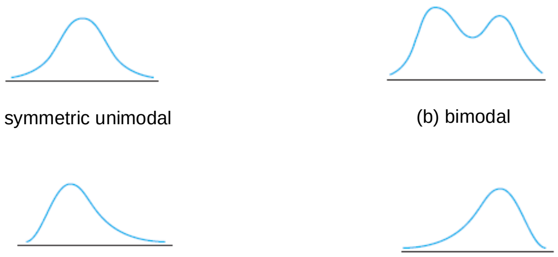
The **sample mean** x of observations x_1, x_2,. . ., x_n, is given by

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

Examples of means

cont'd

Where are the means of data represented by these histograms?

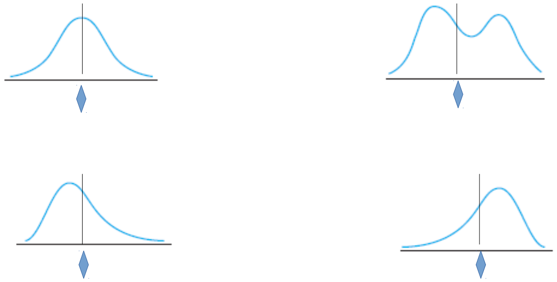


Smoothed histograms

Examples of means

cont'd

Where are the means of data represented by these histograms?



Means are affected by heavy tails and outliers - they are pulled towards them.

Population Mean

The average of all values in the population can (in theory) also be calculated.

This average is called the **population (true) mean** and is usually denoted by the Greek letter μ.

When there are N values in the population (a finite population), then μ = (sum of all N population values)/N.

We will give a more general definition for μ that applies to both finite and infinite populations later in the course. μ is an interesting and important (often the most important) characteristic of a population.

The Trimmed Mean

The sample mean can be greatly affected by even a single outlier (unusually large or small observation).

However it is still the most widely used measure, because there are many populations for which an extreme outlier in the sample would be highly unlikely

If not so unlikely, we might look for a measure that is less sensitive to outlying values: eg, discard top 10% of samples and bottom 10% of samples, and find the sample mean of what's left (that's called "trimmed mean").

The Median

Median means "middle"

The sample median is the middle value once the observations are ordered from smallest to largest.

When the observations are denoted by $x_1,\dots, x_n$, we will use the symbol $\tilde{x}$ to represent the sample median.

Analogous to the middle value in the sample is a middle value in the population, the **population (true) median,** denoted by $\tilde{\mu}$.

To find the sample median:

- order the n observations from smallest to largest (repeated values included – every sample observation appears in the ordered list): $x_1,\dots, x_n$

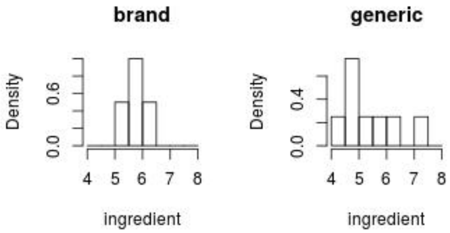
- then, find the middle one:

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

Example: ingredient

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

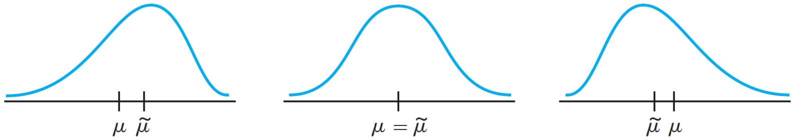
```
> mean(ingredient$brand)
[1] 5.8125
> median(ingredient$brand)
[1] 5.8

> mean(ingredient$generic)
[1] 5.45
> median(ingredient$generic)
[1] 5.15
```



Sample vs population median

The population mean μ and median $\tilde{\mu}$.will not usually be identical – only when the population distribution is symmetric

In that case, choose the summary of interest. For very heavy tailed situations, median is often a better – ie, a more representative – summary.



(a) Negative skew (b) Symmetric (c) Positive skew

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

The median divides the data set into two parts of equal size.

To obtain finer measures of location, we could divide the data into more than two such parts.

Quartiles divide the data set into four equal parts

Percentiles or in general quantiles divide the data into hundredths; the 99th percentile separates the highest 1% from the bottom 99%, and so on.

Quantiles: ingredient

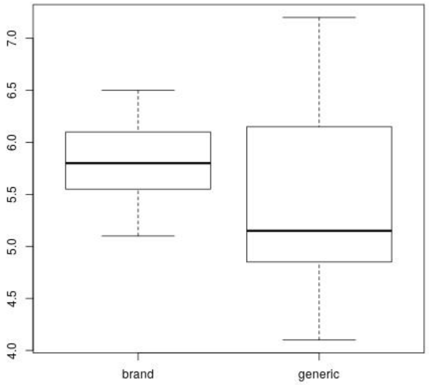
Brand:	5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5
Generic:	5.3 4.1 7.2 6.5 4.8 4.9 5.8 5.0

```
> quantile(ingredient$brand, c(.1,.25,.5,.75,.9))
  10%    25%    50%    75%    90%
5.38   5.58   5.80   6.05   6.29

> quantile(ingredient$generic, c(.1,.25,.5,.75,.9))
  10%    25%    50%    75%    90%
4.59   4.88   5.15   5.98   6.71
```

Box plots: ingredient

A nice way of visualizing quantiles is via the box plots (sometimes called box and whisker plots)

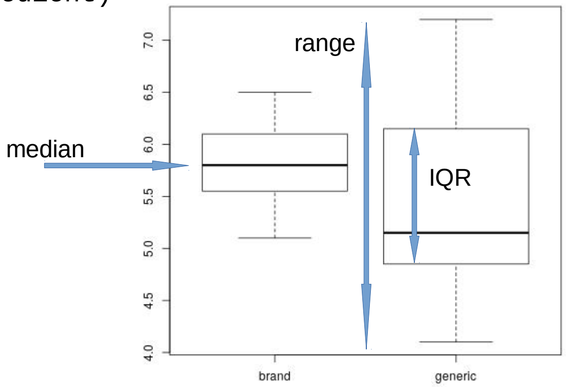
```
> boxplot(ingredient)
```



Box plots: ingredient

A nice way of visualizing quantiles is via the box plots (sometimes called box and whisker plots)

```
> boxplot(ingredient)
```



Mean vs median

The mean is quite sensitive to a single outlier, whereas the median is impervious to many outlying values.

that $\overline{x}$ and $\widetilde{x}$ are at opposite extremes of the centrality measures.

The mean is the average of all the data

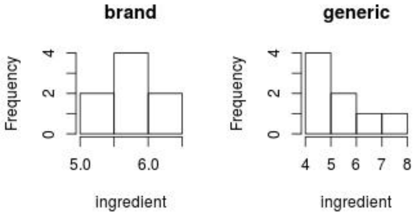
The median results from eliminating all but the middle one or two values and then averaging just those two.

Mode

The mode is the most frequent value (observation)

Unlike mean and median, mode works on numeric and categorical (label) data alike

When working with a numerical sample, usually this means finding the midpoint of the tallest bin of the histogram



Mode

```
> histbrand = hist(ingredient$brand)

> histbrand
$breaks
[1] 5.0 5.5 6.0 6.5
$counts
[1] 2 4 2
$density
[1] 0.5 1.0 0.5
$mids
[1] 5.25 5.75 6.25

> histbrand$mids[histbrand$density==max(histbrand$density)]
[1] 5.75
> histgen = hist(ingredient$generic)
> histgen$mids[histgen$density==max(histgen$density)]
[1] 4.5
```

Variability

So far, we've learnt to describe the center of our sample:

-mean

-median

-mode

We've also learnt to visualize the sample distribution

- histogram, box plot, rug plot

- there are many others (violin plots, beeswarm plots, dot plots, bar charts, … look these up on your own!)

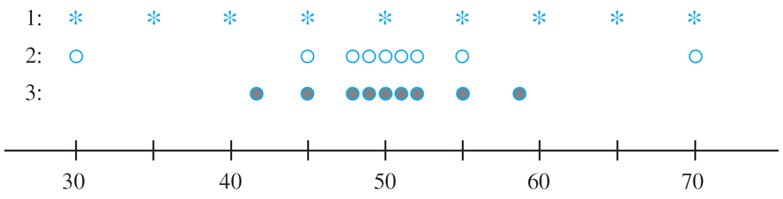
Next: what can we use to quantify the variability of the data in the sample? (And consequently, estimate the variability in the population)

Measures of Variability

Reporting a measure of center gives only partial information about a data set or distribution.

Different samples or populations may have identical measures of center, but differ from one another in other important ways.

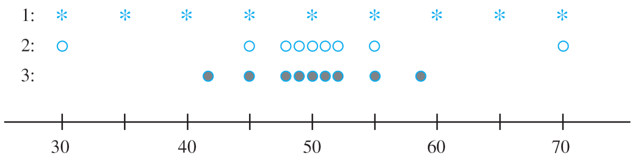
Figure below shows rugplots of three samples with the same mean and median, yet the spread is very different



Sample range

The simplest measure of variability in a sample is the **range,** which is the difference between the largest and smallest sample values.

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.

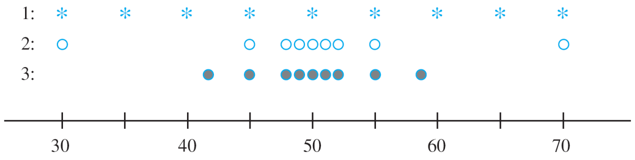


Samples with identical measures of center but different amounts of variability

Measures of Variability for Sample Data

A defect of the range, though, is that it depends on only the two most extreme observations and disregards the positions of the remaining *n* – 2 values.

Samples 1 and 2 in the last Figure have identical ranges, yet when we take into account the observations between the two extremes, there is much less variability or dispersion in the second sample than in the first.



Measures of Variability for Sample Data

Primary measure of variability involves the **deviations from the mean:**

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

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

One possibility is to work with the absolute values of the deviations and calculate the average absolute deviation

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

Because the absolute value operation leads to some theoretical difficulties, we'll consider instead the 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.*

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.

NB: we will use σ^2 (the square of the lowercase Greek letter sigma) to denote the population variance and σ to denote the population standard deviation.

Example 5 (Devore)

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 (Devore)

Effects of rounding account for the sum of deviations not being exactly zero. The numerator of s^2 is S_{xx} = 314.106, from which

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

$$= \frac{314.106}{11 - 1} = 31.41,$$

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

$$s = 5.60$$

Example 5

Effects of rounding account for the sum of deviations not being exactly zero. The numerator of s^2 is S_{xx} = 314.106, from which

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

$$= \frac{314.106}{11 - 1} = 31.41,$$

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

$$s = 5.60$$

R commands for variability measures

```
> range(ingredient$generic)
[1] 4.1 7.2
> range(ingredient$brand)
[1] 5.1 6.5

> sd(ingredient$brand)
[1] 0.4323937
> sd(ingredient$generic)
[1] 1.004277

> var(ingredient$brand)
[1] 0.1869643
> var(ingredient$generic)
[1] 1.008571

> sd(ingredient$brand)^2
[1] 0.1869643
```