

Chapter 7, Summarizing and Displaying Measurement Data

Sanford Weisberg

Univ. of Minnesota

September 28, 2011

Data are rarely helpful without summarization

scores on Quiz #1 (1.5 points per question)

```
[1] 27.00 28.50 29.50 26.50 0.00 27.75 31.00 26.75 26.25
[10] 23.50 29.00 24.75 27.75 28.00 24.50 28.75 23.00 29.75
[19] 24.00 24.50 29.25 25.25 23.75 25.50 20.75 21.50 29.75
[28] 30.25 19.50 26.75 28.75 27.50 25.75 28.25 28.75 27.25
[37] 0.00 23.75 0.00 27.75 26.75 26.50 22.75 24.25 21.75
[46] 26.75 26.25 24.75 22.50 25.75 29.75 28.75 24.75 25.00
[55] 27.00 24.25 24.25 26.25 17.75 25.25 26.25 30.50 26.25
[64] 26.50 24.00 27.75 26.50 28.50 26.25 25.75 27.25 27.25
[73] 30.00 28.50 29.50 22.00 31.25 29.25 26.75 31.50 23.25
[82] 24.25 31.00 26.75 25.25 25.50 30.00 20.75 28.00 24.75
[91] 25.25 30.25 26.25 0.00 23.25 27.75 26.50 27.00 26.00
[100] 21.00 28.50 27.00 20.75 0.00 27.75 27.00 27.75 21.50
[109] 27.50 24.75 28.00 0.00 23.50
```

... is not very informative

What would you like to know?

- 1 How well did I do compared to others?
- 2 What is the “average”; am I above or below?
- 3 Is there a lot of variation in scores or a little?
- 4 The instructor: is the exam too hard or too easy? Does it differentiate among students?

Sort...

```
[1] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 17.75 19.50 20.75
[10] 20.75 20.75 21.00 21.50 21.50 21.75 22.00 22.50 22.75
[19] 23.00 23.25 23.25 23.50 23.50 23.75 23.75 24.00 24.00
[28] 24.25 24.25 24.25 24.25 24.50 24.50 24.75 24.75 24.75
[37] 24.75 24.75 25.00 25.25 25.25 25.25 25.25 25.50 25.50
[46] 25.75 25.75 25.75 26.00 26.25 26.25 26.25 26.25 26.25
[55] 26.25 26.25 26.50 26.50 26.50 26.50 26.50 26.75 26.75
[64] 26.75 26.75 26.75 26.75 27.00 27.00 27.00 27.00 27.00
[73] 27.25 27.25 27.25 27.50 27.50 27.75 27.75 27.75 27.75
[82] 27.75 27.75 27.75 28.00 28.00 28.00 28.25 28.50 28.50
[91] 28.50 28.50 28.75 28.75 28.75 28.75 29.00 29.25 29.25
[100] 29.50 29.50 29.75 29.75 29.75 30.00 30.00 30.25 30.25
[109] 30.50 31.00 31.00 31.25 31.50
```

- 1 6 students didn't take the exam: **outliers**
- 2 0 under 50%, excluding outliers
- 3 **Typical** value, takes a little work
- 4 Variation: The **Range = Maximum - Minimum.**

Outliers

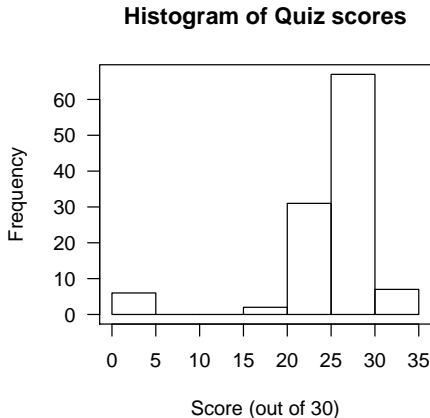
- 1 An **outlier** is a value “far removed” from the others.
- 2 Perhaps the suspected outlier is an error of some sort, and should be ignored.
- 3 Or... the outlier can be the most important observation and the others are unimportant.
- 4 See: Malcolm Gladwell (2008) *Outliers: The Story of Success*, <http://www.gladwell.com/outliers/index.html>.

Some of the “outliers” have probably dropped the course or were sick. Others, I don't know.

Shape: The frequency histogram

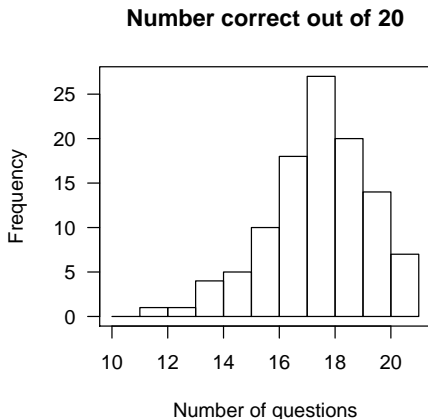
- Divide the **range of values** into equal width intervals or **bins**.
- The **height** of the bar above the interval is equal to the number of observations in the interval for a **frequency histogram**.

What does the histogram show?



Redraw with smaller bin width

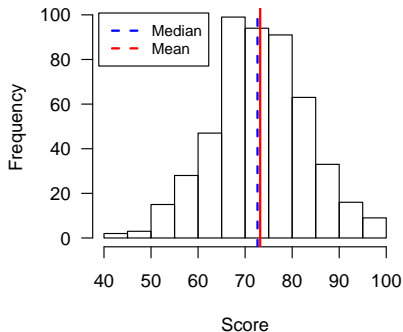
- Since all problems were worth 1.5 points, I'll divide scores by 1.5 to get number correct; this is easier to interpret.
- **Bin width** = 1 question
- The **height** of the bar above the interval is equal to the number of observations in the interval.



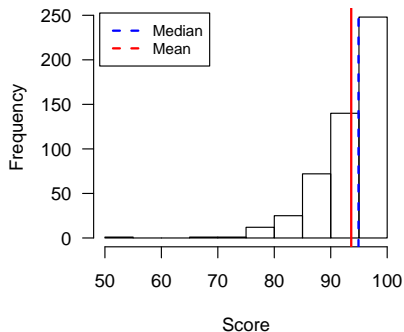
What does the histogram show?

Two more “exam” histograms

Symmetric



Skewed left



Measured volumes of casks in the Guinness Brewery

Casks under 3 or over 7 needed to be sent for repair.

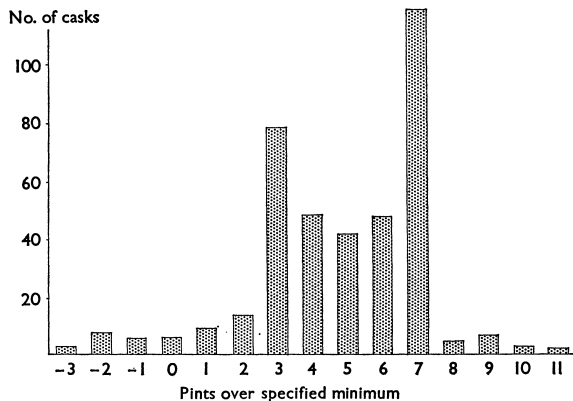
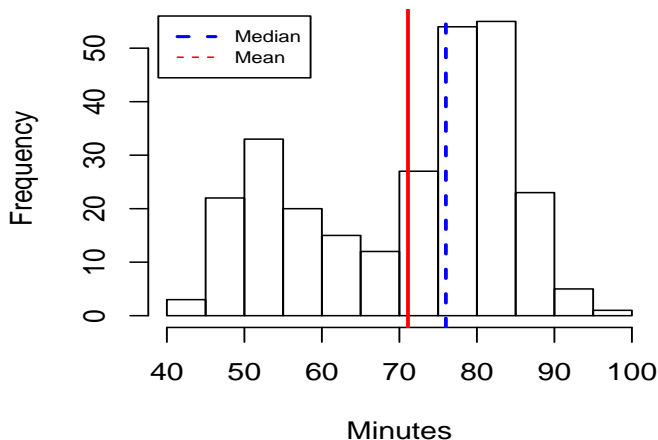


FIG. 4. Distribution of cask sizes.

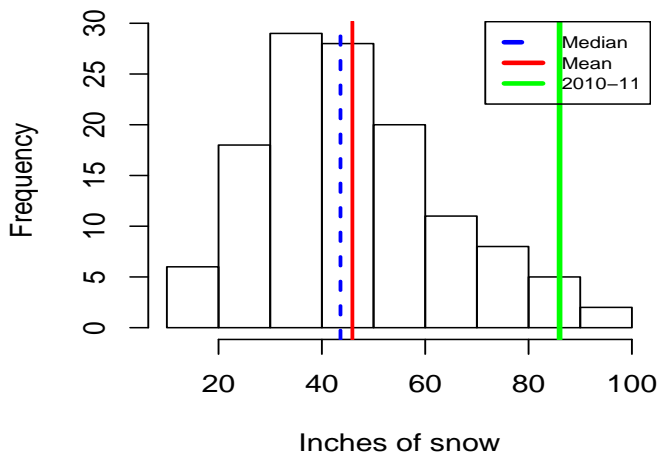
From Stella Cunliffe (1976), *Interaction J. Royal Stat. Soc. Ser A* 139,
<http://www.jstor.org/stable/2344381>

Time between Eruptions of Old Faithful Geyser



<http://www.stat.umn.edu/alr/data/oldfaith.txt>

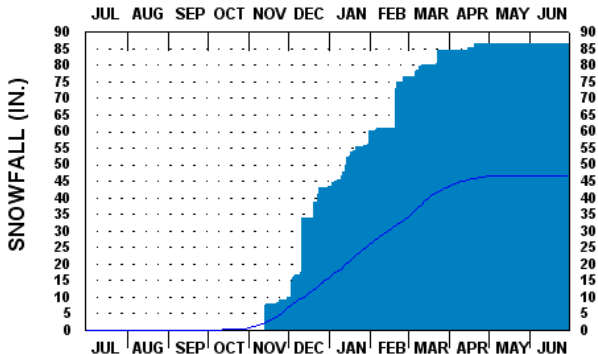
July-June Twin Cities Snowfall 1885-2010



<http://climate.umn.edu/text/historical/mgpsnow.pdf>

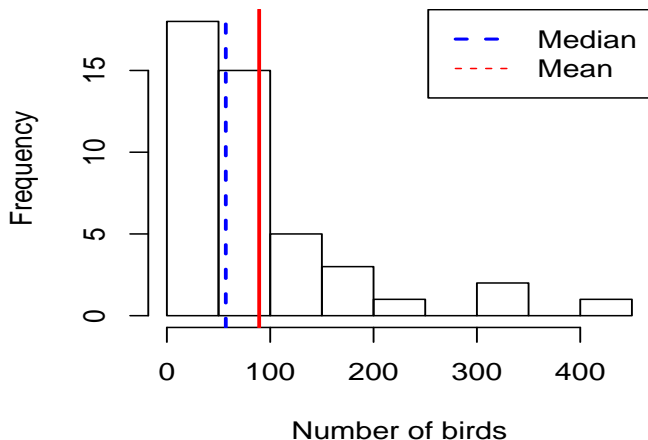
Remember Last Year?

SEASON-TO-DATE MPLS-ST. PAUL SNOWFALL FOR 2010-11 (BARS) VERSUS LONG-TERM AVERAGE (LINE TRACE)



<http://www.climatestations.com/images/stories/minneapolis/msp1011.gif>

Number of snow geese per flock in their summer range



<http://www.stat.umn.edu/alr/data/snowgeese.txt>

Stem Diagrams (invented by J. W. Tukey)

Stem diagrams

- ... provide a useful method for sorting values from smallest to largest.
- ... provide a picture that looks a lot like a histogram that can be interpreted as if it were a histogram with equal bin widths.
- ... are a common feature of most elementary stats books written in the last 30 years.

But...

- ... you will never see one in print except in a text book.
- ... you should know how to read them (as you would a histogram) but don't worry too much about actually drawing them.

The M words: Mean, Median and Mode

The mean

The mean is the same as the average. It is the most important “typical value”.

$$\text{Mean} = \frac{\text{Sum of values}}{\text{Number of values}}$$

Including the zeroes:

$$\begin{aligned}\text{Mean} &= \frac{27 + 28.5 + 29.5 + \cdots + 23.5}{113} \\ &= 24.9\end{aligned}$$

Without the zeroes:

$$\begin{aligned}\text{Mean} &= \frac{29.5 + \cdots + 23.5}{107} \\ &= 26.3\end{aligned}$$

The median

- The median is the “value in the middle”. Half the values are more, half are less.
- For an odd sample size (for example $n = 7$), it is the forth largest (or forth smallest) value.
- For an even sample size (say $n = 6$), it is the average of the two values in the middle.
- Except in trivial cases, it is insanely tedious to compute without first sorting the data.

Median with outliers = 26.5

Median without outliers = 26.5

When to use the mean and when the median

The Mean

Always use the mean unless there is a good reason to use the median.

The Median

Use the median only if the variable of interest is *skewed*.

Examples of skewed variables often include:

- Income: lots of relatively small values, a few big ones
- Population sizes (e.g., of bacteria)
- Waiting times (e.g., time until served at a call center)

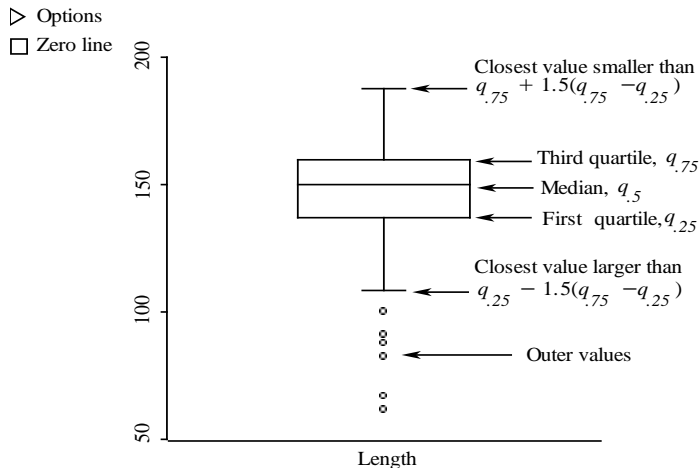
The Mode

The **mode** is the most frequent value.

It is almost never used in practice with measured variables.

We will not discuss it further.

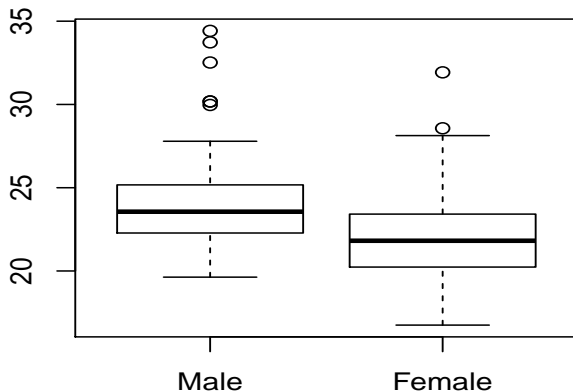
Boxplots (invented by J. W. Tukey)



From: R. D. Cook and S. Weisberg (1999) *Applied Regression Including Computing and Graphics*

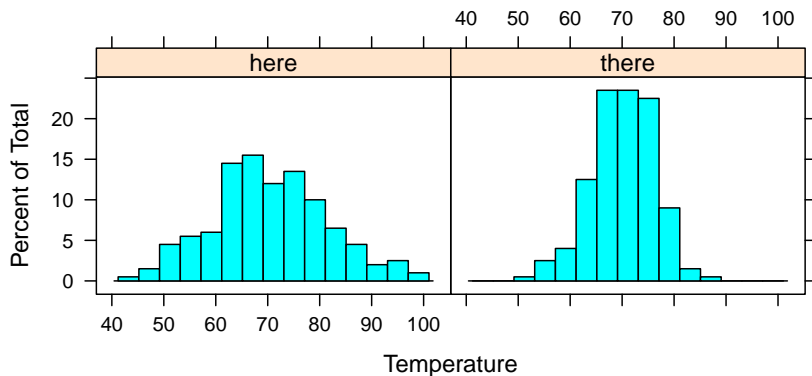
Body Mass Index of Elite Australian Athletes

$$BMI = (\text{Weight, kg})/(\text{Height, m})^2$$



<http://www.stat.umn.edu/alr/data/ais.txt>

Where would you prefer to live?



What's the difference?

Variation!

Definition

Statistics is the study of variation

Standard deviation

The most important measure of variation is called the *standard deviation*

Notation

The **sample standard deviation** is often written s or SD

The **population standard deviation** is often written σ , the Greek letter
sigma

The standard deviation

- 1 Compute the sample size n
- 2 Compute the mean
- 3 Subtract the mean from the numbers to get deviations
- 4 Square the deviations and add them up
- 5 Divide by $n - 1$
- 6 Take a square root of the answer

x	Dev	Dev^2
2	$2 - 7 = -5$	25
6	$6 - 7 = -1$	1
6	$6 - 7 = -1$	1
8	$8 - 7 = 1$	1
8	$8 - 7 = 1$	1
12	$12 - 7 = 5$	25
<hr/>		
$42/6 = 7$		54

$$s = \sqrt{54/5} = \sqrt{10.8} = 3.3$$

Example

Suppose you have \$1000 to invest in a stock fund.

Fund	Average return	SD of return
A	7%	10%
B	7%	4%
C	5%	1%
D	7%	0%

Return means your profit for the year: a 7% return equals

$$1000 + .07 \times 1000 = \$1070$$

at the end of the year.

“D” is known as the Bernie Madoff.

For populations and samples with “bell shaped” histograms:

- 2/3 of the time a new observation is within 1 SD of the mean
- 95% of the time a new observation is within 2 SD of the mean
- Almost always, a new observation is within 3 SD of the mean

... except for outliers and other odd stuff.

Example, continued

Fund	Ave.	SD	2/3 of time	95%	Almost always
A	7%	10%	-3% to 17%	-13% to 27%	-23% to 37%
B	7%	4%	3% to 11%	-1% to 15%	-5% to 19%
C	5%	1%	4% to 6%	3% to 7%	2% to 8%
D	7%	0%	7%	7%	7%