

## Chapter 7, Summarizing and Displaying Measurement Data

Sanford Weisberg

Univ. of Minnesota

February 11, 2009

## Data are rarely helpful without summarization

55 scores on Quiz #1

```
[1] 27 27 33 30 27 21 27 15 24 30 27 27 0 18 27 27 30 30
[19] 27 21 36 27 33 27 33 21 27 30 30 24 33 24 21 24 36 36
[37] 21 24 33 21 24 30 27 33 27 27 30 30 21 30 21 27 18 27
```

... 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 the data...

Sorting helps a little:

```
> sort(exam)
```

```
[1] 0 15 18 18 21 21 21 21 21 21 21 21 24 24 24 24 24 24
[19] 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 30
[37] 30 30 30 30 30 30 30 30 30 30 33 33 33 33 33 33 36 36 36
```

### What can you learn from sorting the data?

- 1 One student didn't take the exam: an **outlier**
- 2 No partial credit for problems
- 3 How many items?
- 4 Minimum and Maximum
- 5 **Typical** value
- 6 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*, Little-Brown.

Our “outlier” didn't turn in homework #2 and has probably dropped the course.

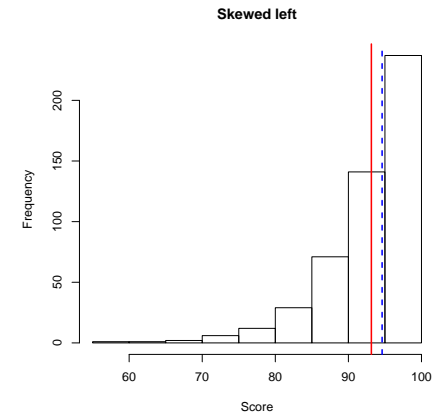
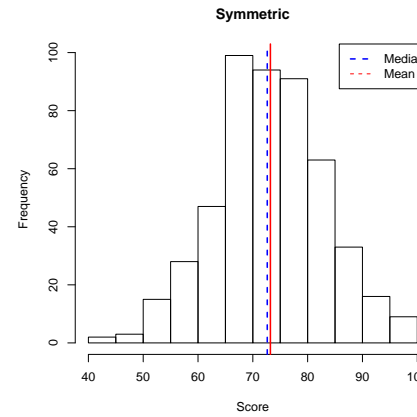
## Shape: The frequency histogram

- Divide the **range of values** into equal width intervals.
- 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



## Measured volumes of casks in the Guinness Brewery

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

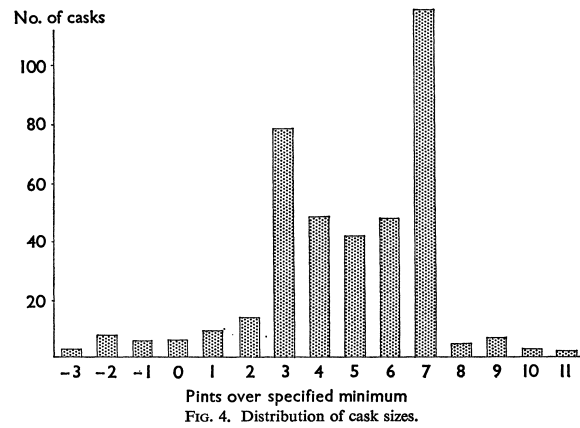
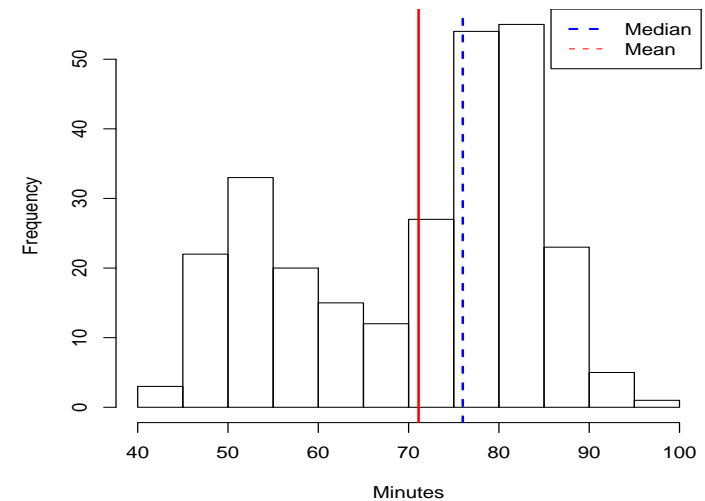


FIG. 4. Distribution of cask sizes.

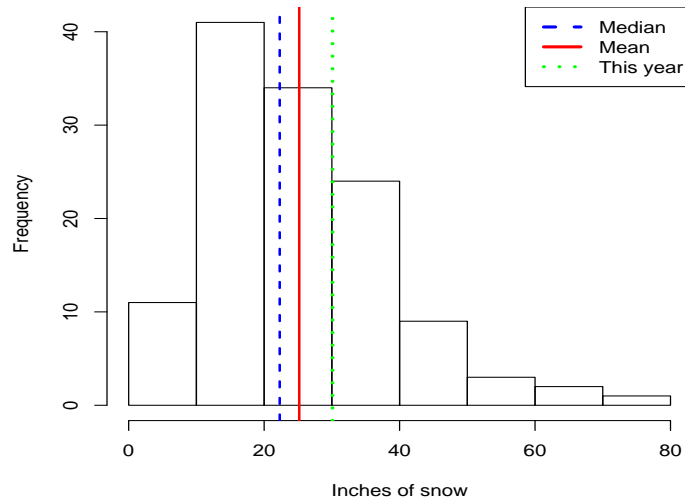
From Cunliffe, S. V. (1976), *Interaction J. Royal Stat. Soc. Ser A* 139, <http://www.jstor.org/stable/2344381>

## Time between Eruptions of Old Faithful Geyser



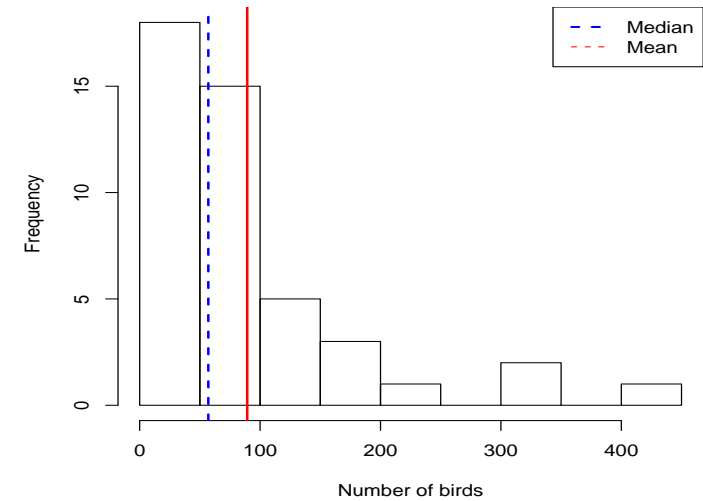
<http://www.stat.umn.edu/alr/data/oldfaith.txt>

## July-Jan. Twin Cities Snowfall 1885-2008



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

## Number of snow geese per flock in their summer range



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

## 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 outlier:

$$\begin{aligned} \text{Mean} &= \frac{0 + 15 + \dots + 36 + 36}{54} \\ &= 26.4 \end{aligned}$$

Without the Outlier:

$$\begin{aligned} \text{Mean} &= \frac{15 + \dots + 36 + 36}{53} \\ &= 26.9 \end{aligned}$$

## 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 fourth largest (or fourth 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.

$$\text{Median with outlier} = 27$$

$$\text{Median without outlier} = 27$$

## 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)

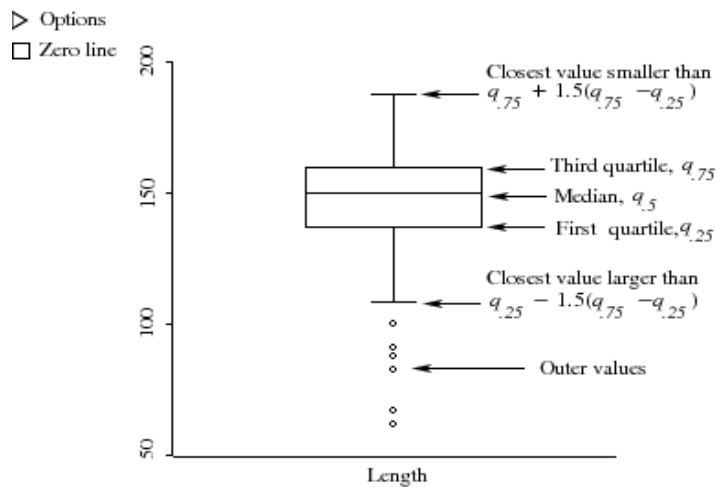
## The Mode

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

It is almost never used in practice with measured variables.

We will not discuss it further.

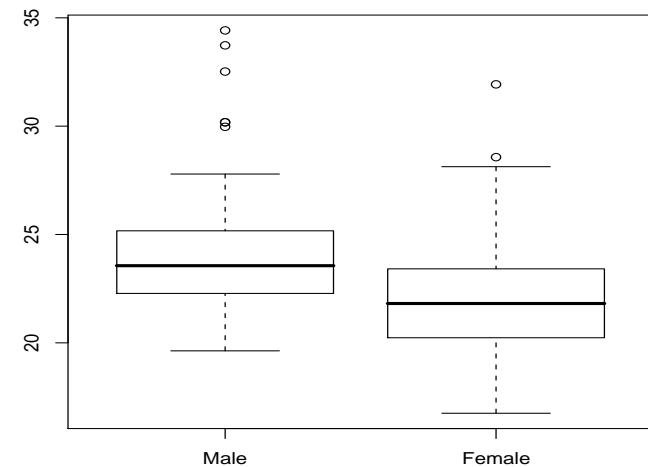
## Boxplots (invented by J. W. Tukey)



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

## Body Mass Index of Elite Australian Athletes

$$BMI = (Weight, kg) / (Height, m)^2$$



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