More from Chapter 7

Sanford Weisberg

Univ. of Minnesota

February 11, 2009

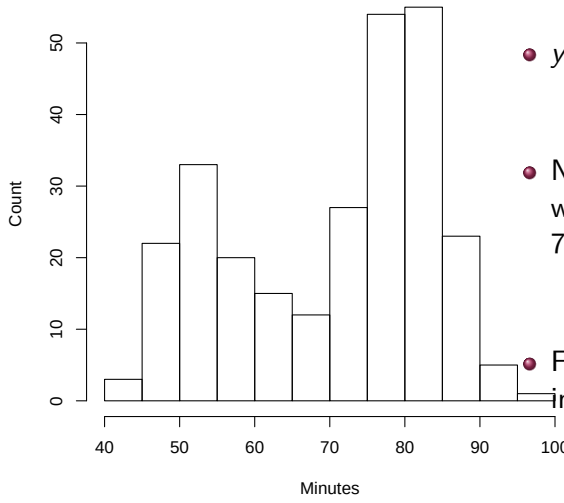
How much is a million, billion, and trillion?

1. How long is a **million** seconds?
2. How long is a **billion** seconds?
3. How long is a **trillion** seconds?

"Vaccines don't cause autism, court finds" *StarTribune*, Feb. 13 2009

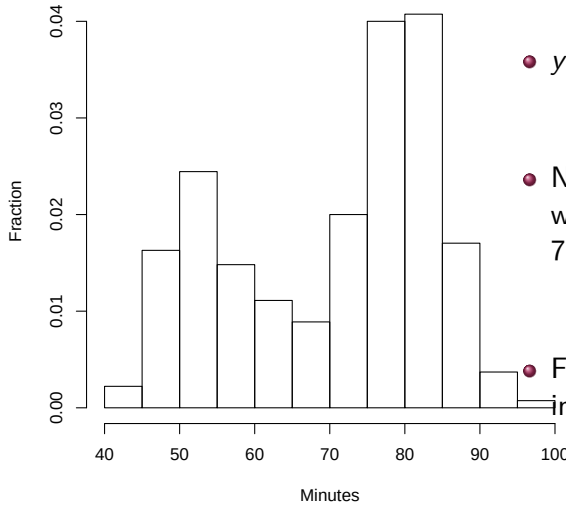
- "Thousands of families were hoping to win compensation and vindication through three test cases [with claims of a vaccination-autism link] presented to the court. They contended that a combination of the measles-mumps-rubella vaccine plus other shots triggered autism."
- "Science years ago concluded there's no connection, but Thursday's rulings in a trio of cases still have far-reaching implications."
- "The head of a consumer group that questions vaccine safety said she still felt there were the possibility of a link."

Frequency Histograms



- *x*-axis
- *y*-axis
- NUMBER of eruptions with wait between 60 and 70 min:
- FREQUENCY of eruptions in (60, 70) min:

Density Histograms: Divide Count by sample size



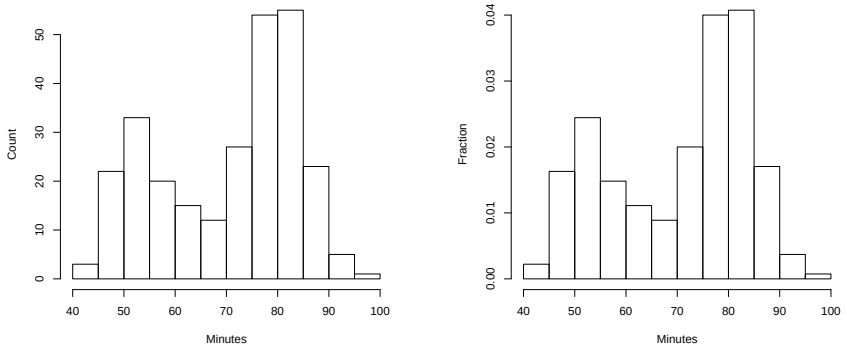
- x-axis
- y-axis
- NUMBER of eruptions with wait between 60 and 70 min:
- FREQUENCY of eruptions in (60, 70) min:

Frequency *vs* Density Histograms

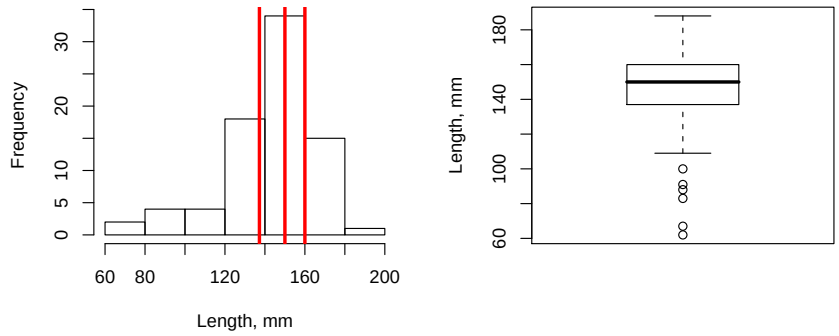
For histograms when the bins on the horizontal axis are equal widths,b

Frequency histograms have bar height equal to the number of observations in the interval

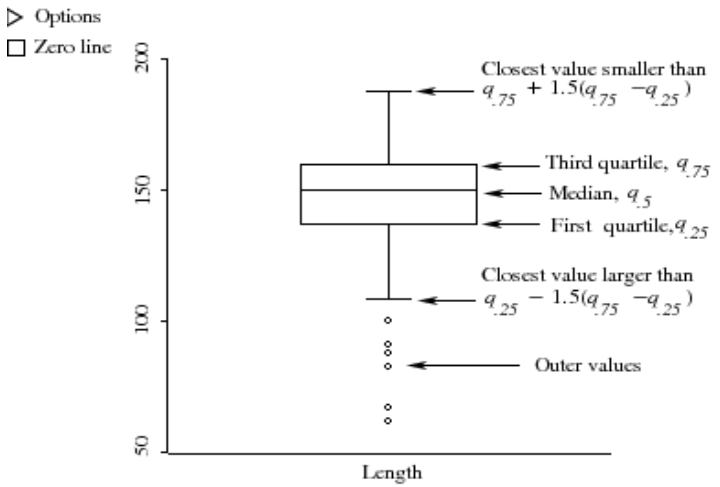
Density histograms have bar height equal to the fraction of observations in the interval.



Boxplots: Lake Mary Bluegills



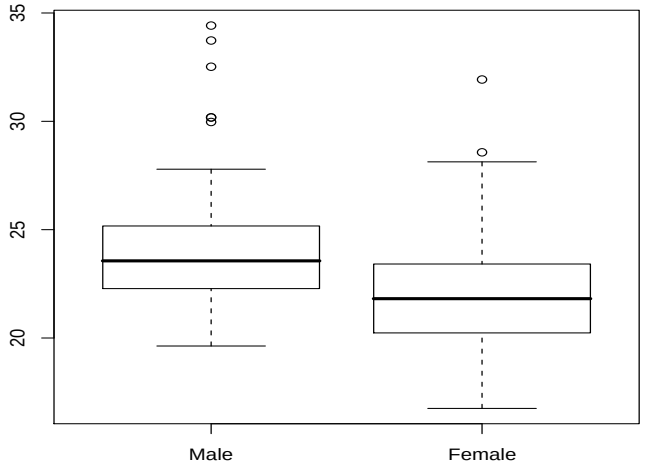
Boxplots (invented by J. W. Tukey)



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

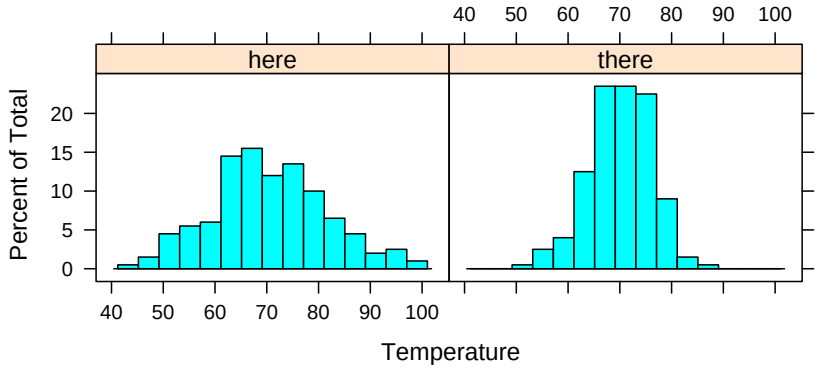
Body Mass Index of Elite Australian Athletes

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



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

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
$42/6 = 7$		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%

Empirical rules

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 with 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%