Lecture 03
Data visualization

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

A statistician's manifesto

(From T. Hastie, via J. McAuliffe, via Jordan Boyd-Graber)

- **Understand the ideas behind the statistical methods**, so you know how to use them, when to use them, when not to use them.
- Complicated methods build on simple methods. **Understand simple methods first**.
- The results of a method are of little use without **an assessment of how well or poorly it is doing**.

Three Rules of Data Analysis

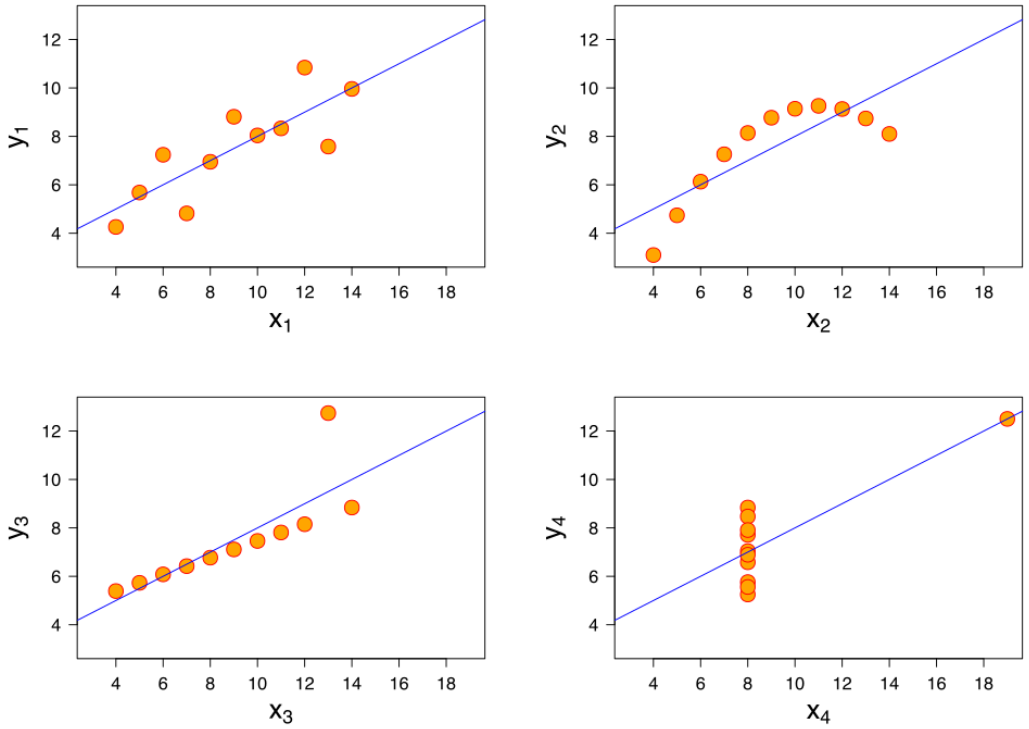
- Make a picture
- Make a picture
- Make a picture

- To think, show, and tell…
- And to make it cool!
- E.g., https://d3js.org

Why visualization matters

- We're lazy and don't like to read
- Some information isn't easy to describe verbally
- Statistical summaries can be misleading
- Aesthetically-pleasing visuals are engaging

- Anscombe's quartet:
- https://en.wikipedia.org/wiki/Anscombe%27s_quartet



Slide credit: Tal Yarkoni

Textbook's example data

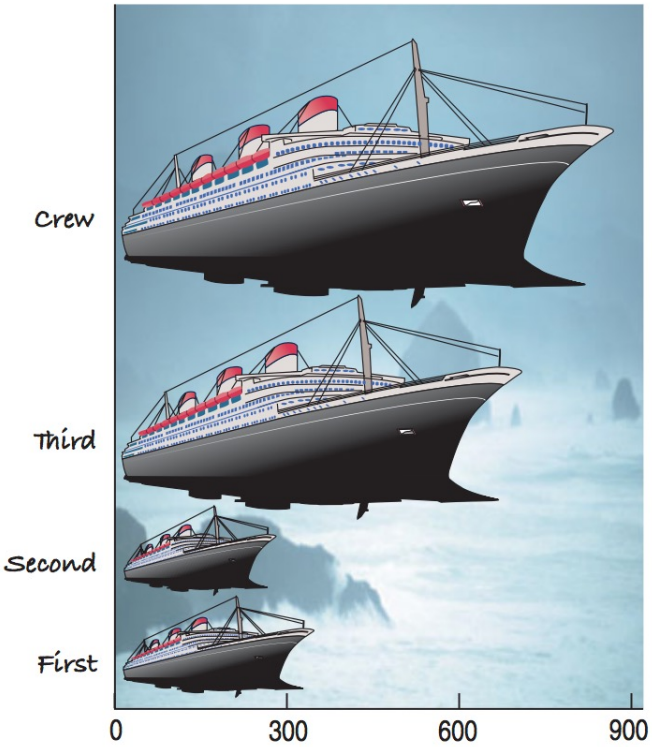
- Frequency/contingency table of "ticket class" and "Survival" for the *Titanic* passengers

Frequency table

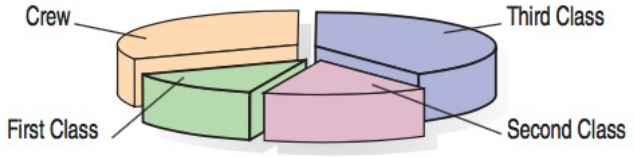
Class	Count
First	325
Second	285
Third	706
Crew	885

Contingency table

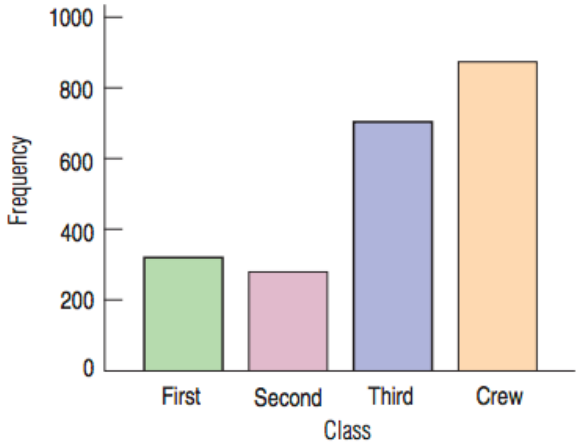
		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

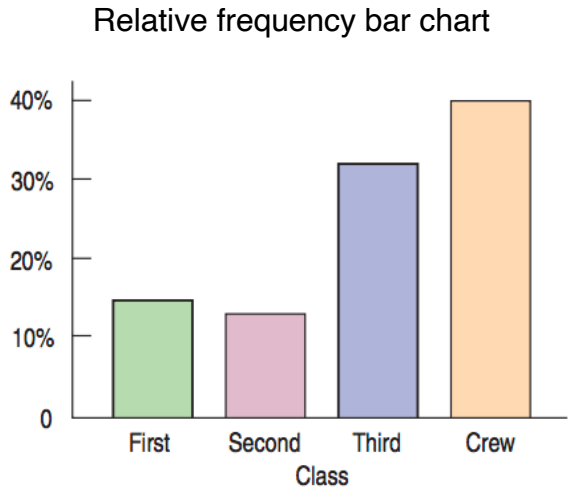


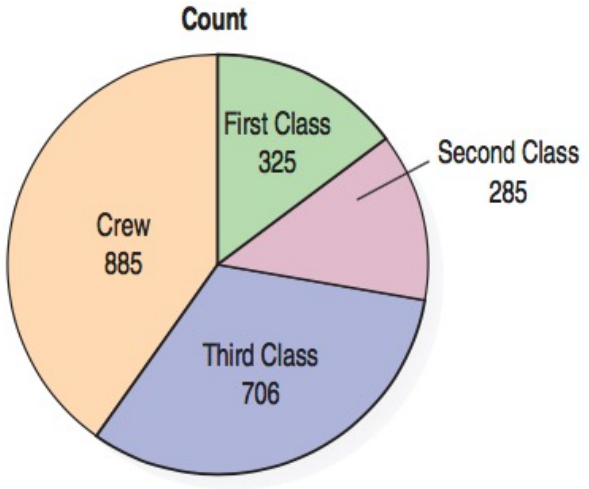
- What's wrong with this plot?
- "Area principle": the area occupied by a part of the graph should correspond to the magnitude of the value it represents.



Bar Charts, Pie Charts







Marginal distribution, conditional distribution

Contingency table

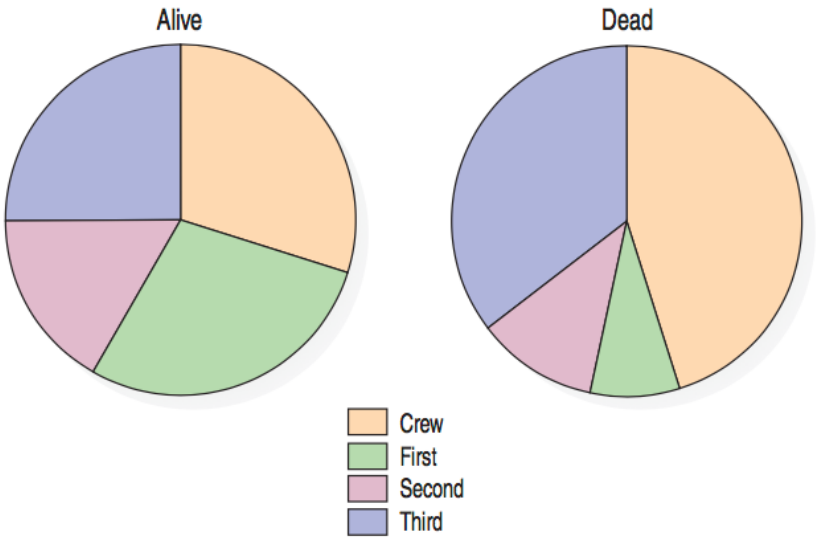
Marginal distribution of "Survival"

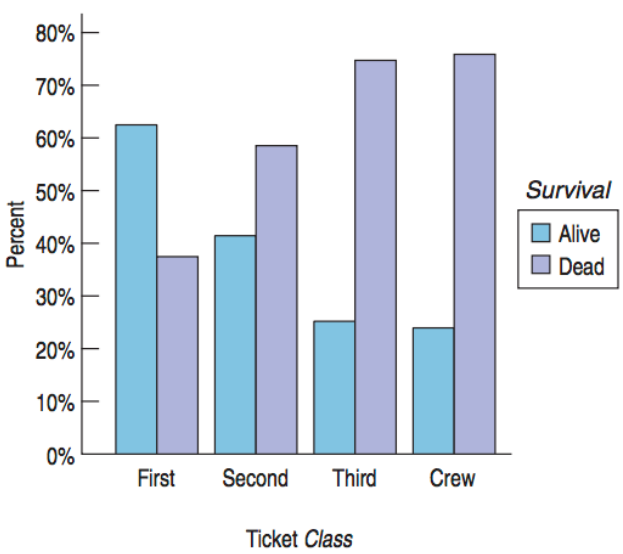
Survival	Class	First	Second	Third	Crew	Total
	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

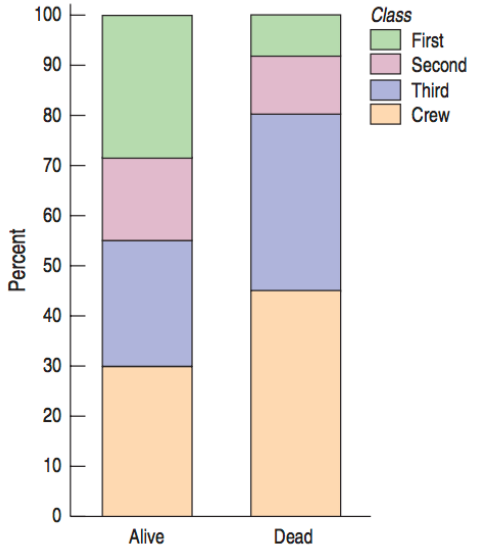
Conditional distribution
(satisfy a condition on another variable)

Marginal distribution of "Class"

Charts for conditional distribution







- **The variables are independent**: "when the distribution of one variable tells us nothing about the distribution of the other variable"

Simpson's Paradox

- Two pilots, Moe and Jill
- Who is the better pilot?

Proportion of on-time flights

		Time of Day		
		Day	Night	Overall
Pilot	Moe			100 out of 120 83%
	Jill			94 out of 120 78%

Simpson's Paradox

- Two pilots, Moe and Jill
- Who is the better pilot?
- Now who is the better pilot?

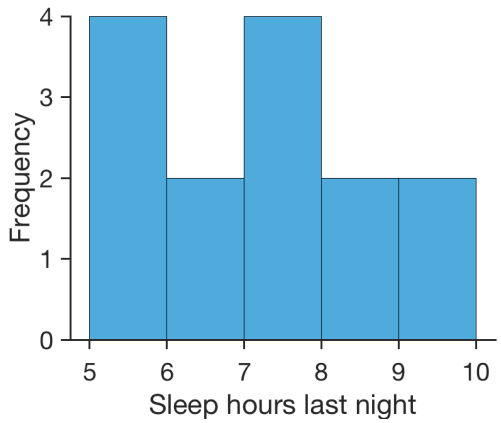
Proportion of on-time flights

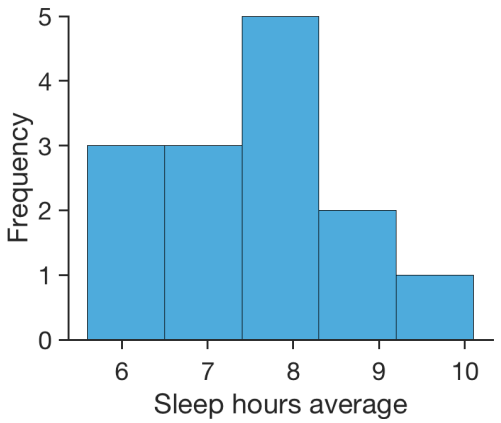
		Time of Day		
		Day	Night	Overall
Pilot	Moe	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

- Don't average over things *unfairly*
- *Different* numbers in *different* categories for Day and Night
- *Unfair* to average across categories

Histograms for quantitative data

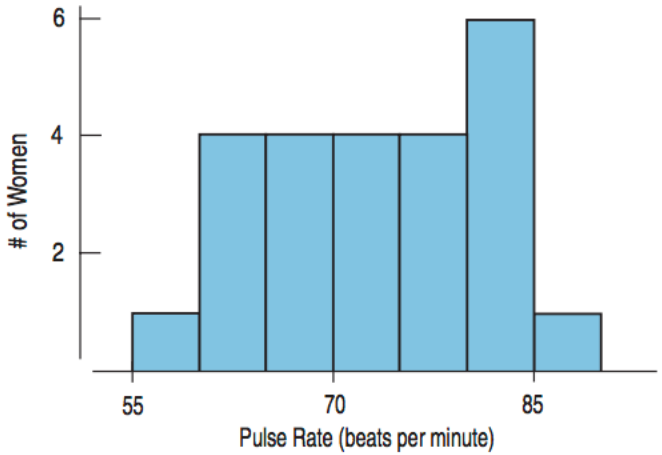
	sleep_hours_last_night	sleep_hours_average
Indiv 1	7	6
Indiv 2	5.5	6.5
Indiv 3	9	7
Indiv 4	5	7
Indiv 5	5	6
Indiv 6	7	8
Indiv 7	8	7
Indiv 8	7	8
Indiv 9	8	10
Indiv 10	10	9
Indiv 11	6	7.5
Indiv 12	7	9
Indiv 13	6	8
Indiv 14	5	8





Stem/leaf, dot plot

Histogram

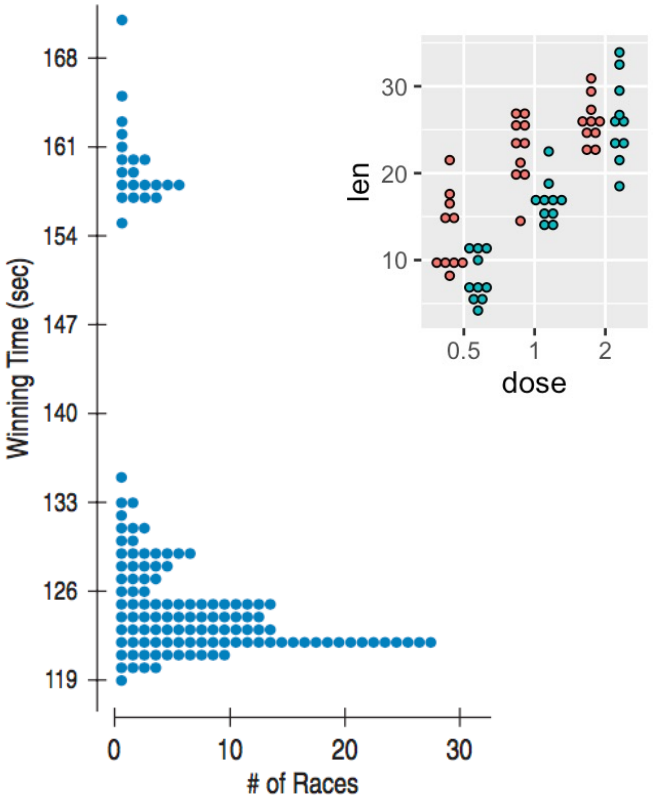


Stem-and-leaf display

8	8
8	000044
7	6666
7	2222
6	8888
6	0444
5	6

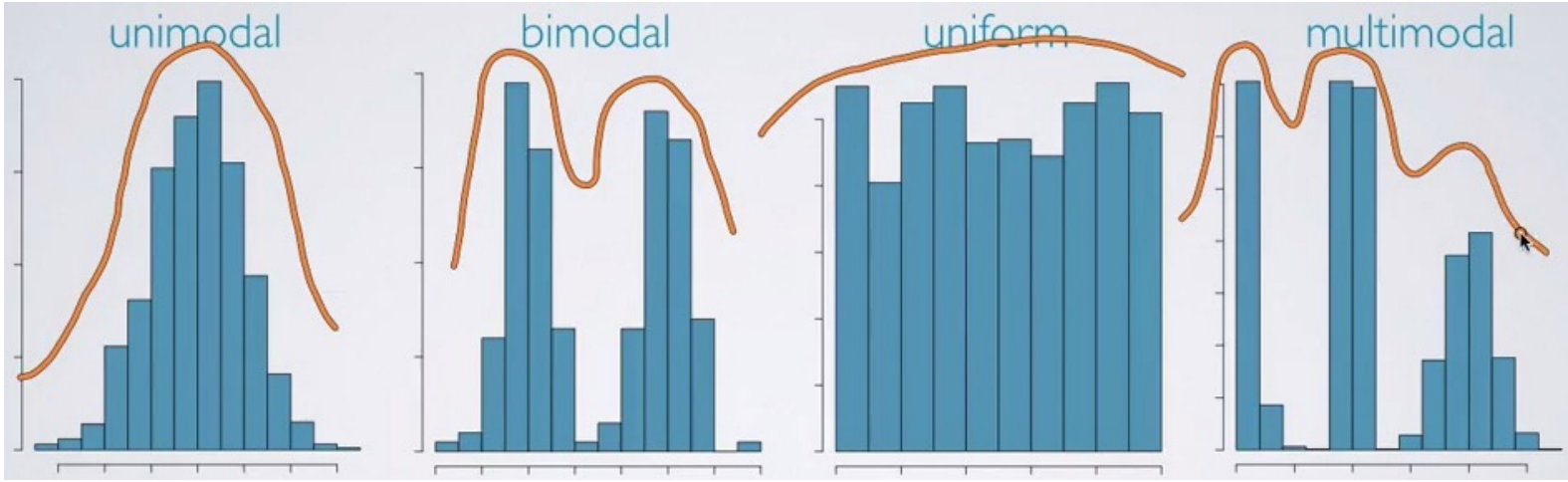
Pulse Rate
(8|8 means 88 beats/min)

Dotplots



Shape

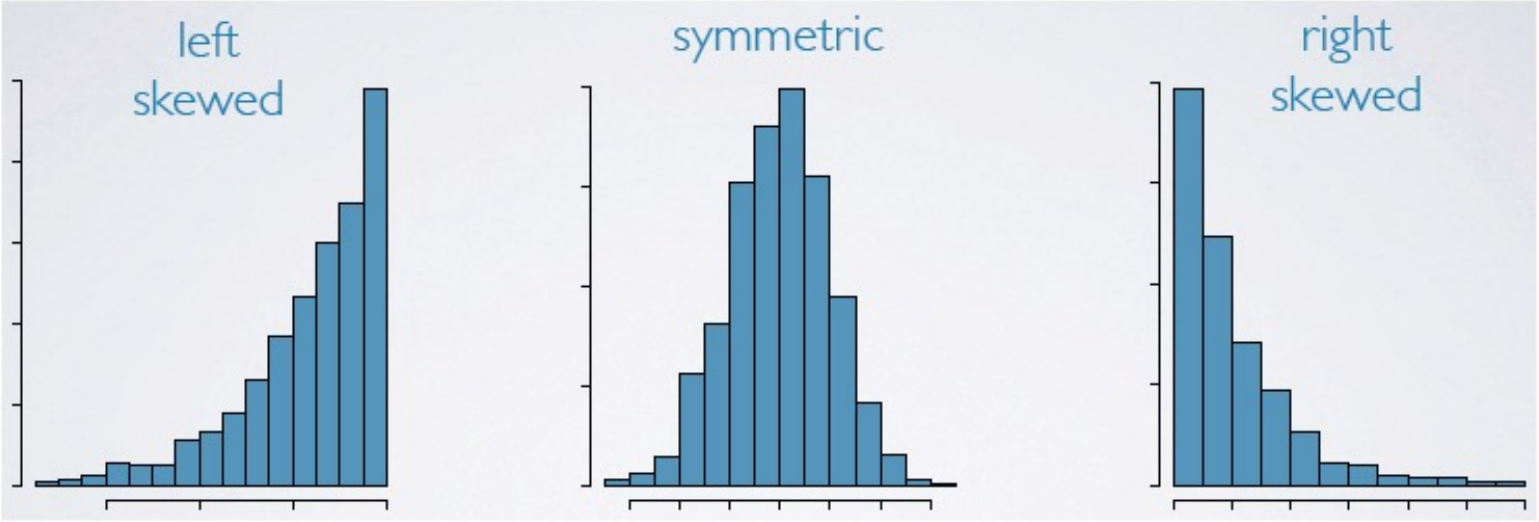
- Modes?



http://researchhubs.com/post/ai/data-analysis-and-statistical-inference/visualizing-numerical-data.html

Shape

- Modes?
- Symmetric?



http://researchhubs.com/post/ai/data-analysis-and-statistical-inference/visualizing-numerical-data.html

Center

- Median

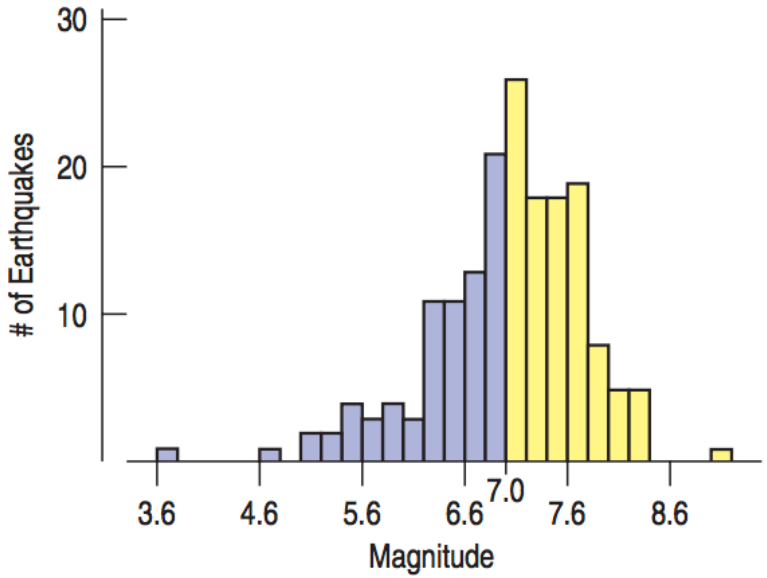


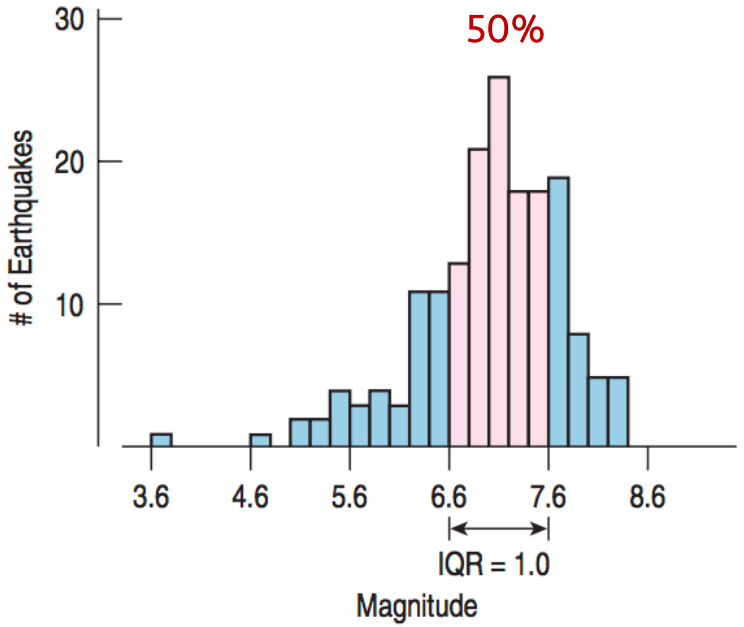
FIGURE 4.10
Tsunami-causing earthquakes (1981–2005).
The median splits the histogram into two halves of equal area.

- 176 earthquakes
- Median: (176+1)/2 = 88.5th value in the sorted data
- ".5" = average of the two values (88th and 89th)

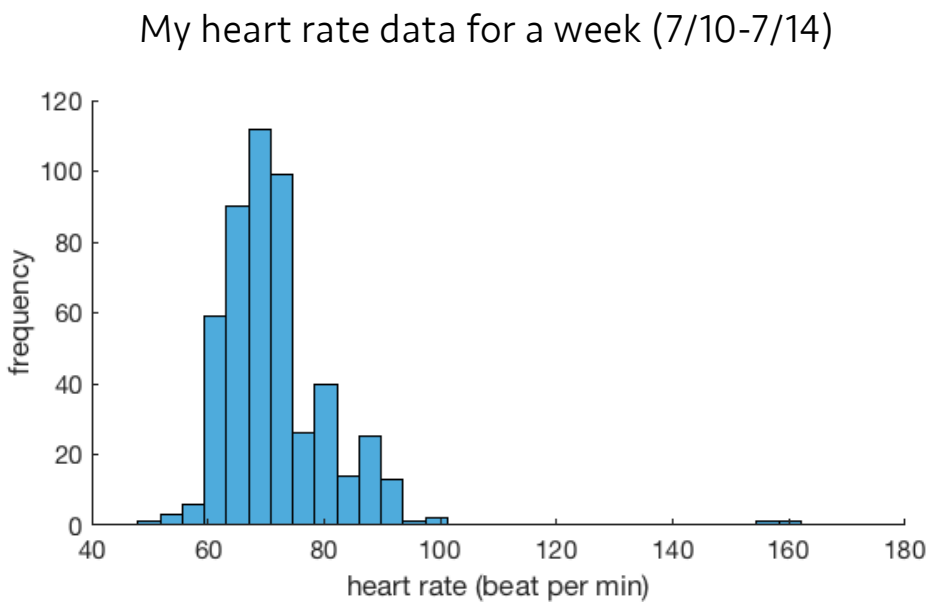
- If there was 221 earthquakes
- Median: (221+1)/2 = 111th value in the sorted data

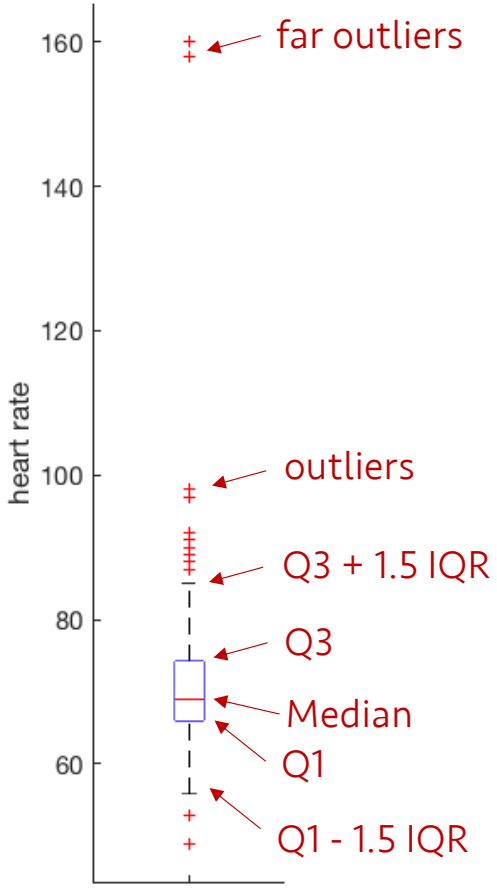
Spread

- Range
  - Range = max − min
- Interquartile range
  - Interquartile range (IQR) = upper quartile − lower quartile



Boxplots and 5-Number Summaries

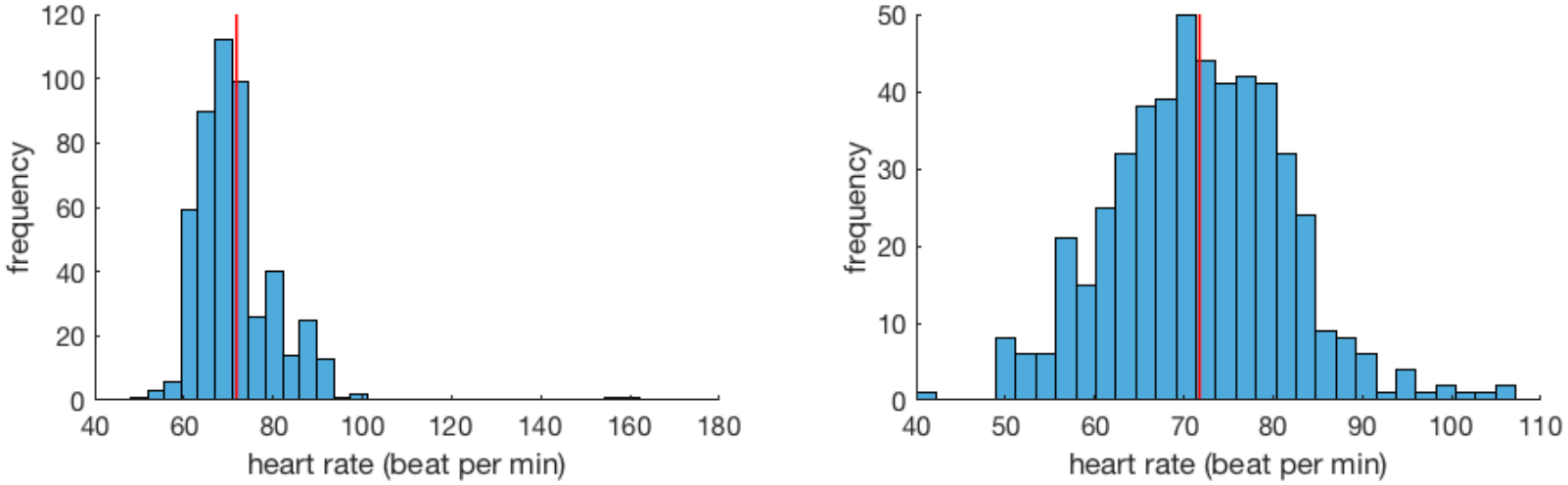




Max	160
Q3	74
Median	69
Q1	66
Min	49

Center of Symmetric Distribution: Mean

$$\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}$$



- If the histogram is symmetric and there are no outliers, the mean will be preferable.
- However, if the histogram is skewed or has outliers, the median might be better.

Quiz!

With the following data, which of the following is WRONG summary statistics? *

Data: 2, 26, 29, 21, 23, 23, 12, 20, 6, 22

- range = 27
- Q3=12
- median = 21.5
- IQR=11
- mean = 18.4

c.f.

Percentile = $100 \times \frac{i-0.5}{N}$

i = rank after sorting the values in an ascending order
N = the number of values

Spread of Symmetric Distribution: Standard Deviation

Variance

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

Standard Deviation

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

Mean = 17

Original Values	Deviations	Squared Deviations
14	$14 - 17 = -3$	$(-3)^2 = 9$
13	$13 - 17 = -4$	$(-4)^2 = 16$
20	$20 - 17 = 3$	9
22	$22 - 17 = 5$	25
18	$18 - 17 = 1$	1
19	$19 - 17 = 2$	4
13	$13 - 17 = -4$	16

Add up the squared deviations: $9 + 16 + 9 + 25 + 1 + 4 + 16 = 80.$
Now divide by $n - 1$: $80/6 = 13.33.$
Finally, take the square root: $s = \sqrt{13.33} = 3.65$

Key Points

Chapter 3: Displaying categorical data

- Bar chart for categorical data
- Pie chart for proportions of whole
- Faithful reporting and the area principle
- Contingency tables
- Simpson's paradox

Chapter 4: Displaying quantitative data

- Histograms, Stem-leaf, dot plots
- Shape (mode, symmetrical)
- Center (median, mean)
- Spread (range, IQR, variance, standard deviation)
- Box plots