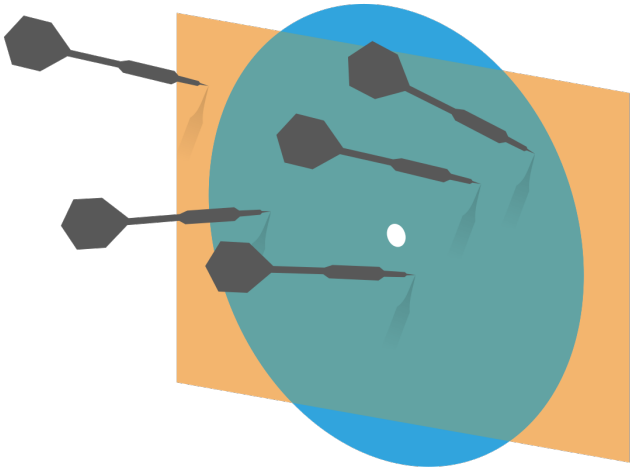
Probability and Statistics for Computer Science



Credit: wikipedia

"The statement that "The average US family has 2.6 children" invites mockery" – Prof. Forsyth reminds us about critical thinking

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 1.23.2020

Last lecture

- Course material and survey
- Meet staff team
- Overview of CS361
- Lecture 1 - Data Visualization & Summary (I)

Lecture videos and ClassTranscribe

- Lecture will be videotaped and accessible at https://mediaspace.illinois.edu/
- ClassTranscribe provides transcripts for lecture videos

  https://classtranscribe.illinois.edu/home

To be connected

We learned

- Visualizing with

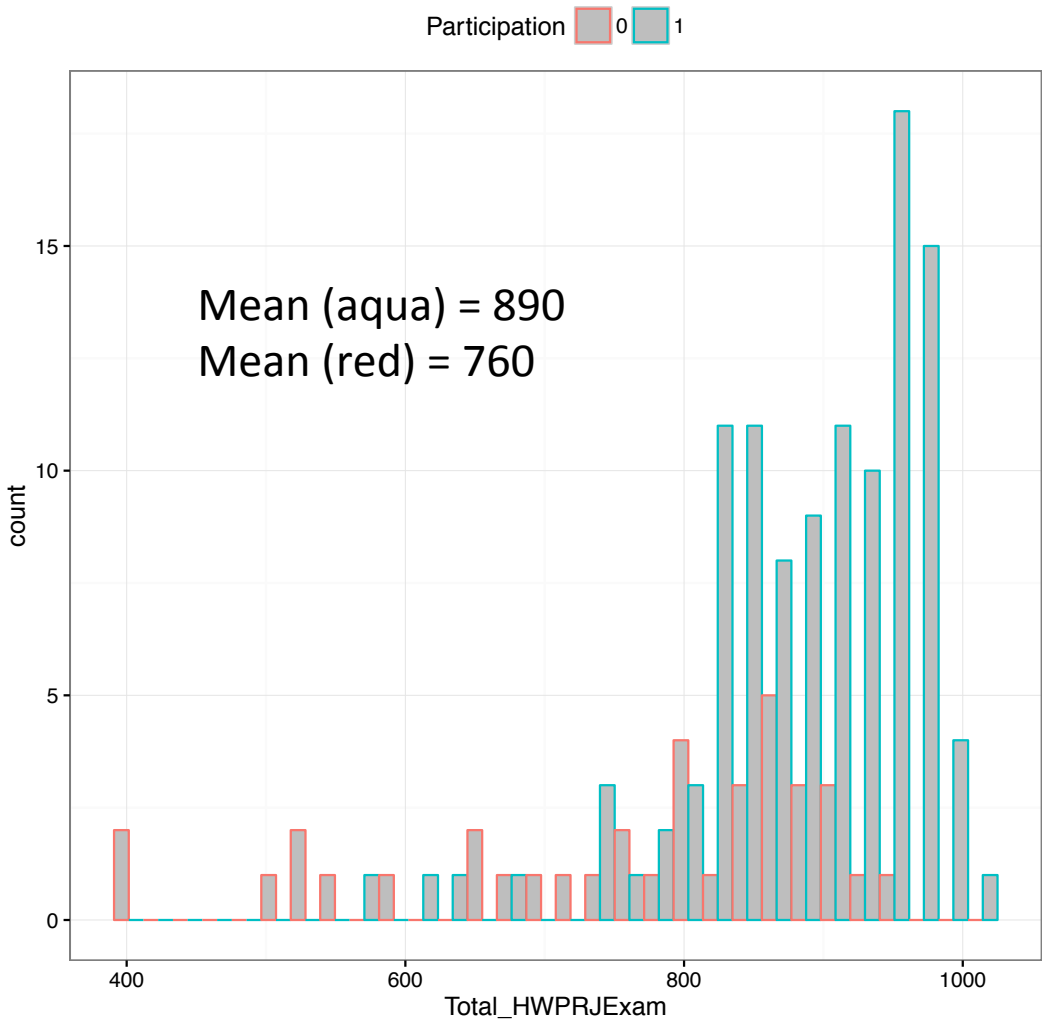
  Tables, bar charts, histograms

- Summarizing with location parameter

  Mean

Visualizing Data with Histogram (III)

- Conditional histogram



Data: Combined Score (HWs, Prj and Exams) grouped by students with full participation or **not full** in CS361 fall 2019

Today: More Summary (descriptive statistics & Data Visualization)

- Mean
- Standard deviation
- Variance
- Standardizing data
- Median, interquartile range,
- box plots and outliers

- Visualizing & Summarizing ***relationships***
  - Heatmap
  - 3D bar
  - Time series plots
  - Scatter plots
  - Correlation coefficient

Summarizing 1D continuous data

For a data set {x} or annotated as $\{x_i\}$, we summarize with:

- Location Parameters
  - Mean
  - Median
  - Mode
- Scale parameters
  - Standard deviation and variance
  - Interquartile range

Summarizing 1D continuous data

- Mean

$$mean(\{x_i\}) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

It's the centroid of the data geometrically,
by identifying the data set at that point, you find the center of balance.

Properties of the mean

- Scaling data scales the mean

$$mean(\{k \cdot x_i\}) = k \cdot mean(\{x_i\})$$

- Translating the data translates the mean

$$mean(\{x_i + c\}) = mean(\{x_i\}) + c$$

Less obvious properties of the mean

- The signed distances from the mean sum to 0

$$\sum_{i=1}^{N}(x_i - mean(\{x_i\})) = 0$$

- The mean minimizes the sum of the squared distance from any real value

$$\underset{\mu}{argmin} \sum_{i=1}^{N}(x_i - \mu)^2 = mean(\{x_i\})$$

Prove $\sum_{i=1}^{N} (x_i - \text{mean}(\{x_i\})) = 0$

$$\text{LHS} \quad \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \text{mean}(\{x_i\})$$

$$\sum_{i=1}^{N} x_i - N \cdot \text{mean}(\{x_i\})$$

$$\text{mean}(\{x_i\}) = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{LHS} \quad \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0$$

Prove: $\operatorname{Argmin}_{\mu} \sum_{i=1}^{N} (x_i - \mu)^2 = \text{mean}(\{x_i\})$

$$\frac{d}{d\mu} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$= \sum_{i=1}^{N} \frac{d}{d\mu} (x_i - \mu)^2$$

$$= \sum_{i=1}^{N} \frac{d}{dg}(g^2) \frac{dg}{d\mu} \qquad g = x_i - \mu$$

$$= \sum_{i=1}^{N} 2g \cdot (-1)$$

$$= \sum_{i=1}^{N} 2(x_i - \mu)(-1) = 0$$

$$\hat{\mu} = \text{mean}(\{x_i\})$$

$$\frac{d(f(g(x)))}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}$$

Qs:

- What is the answer for

  mean(*mean*($\{x_i\}$)) ?

  A. *mean*($\{x_i\}$) B. unsure C. 0

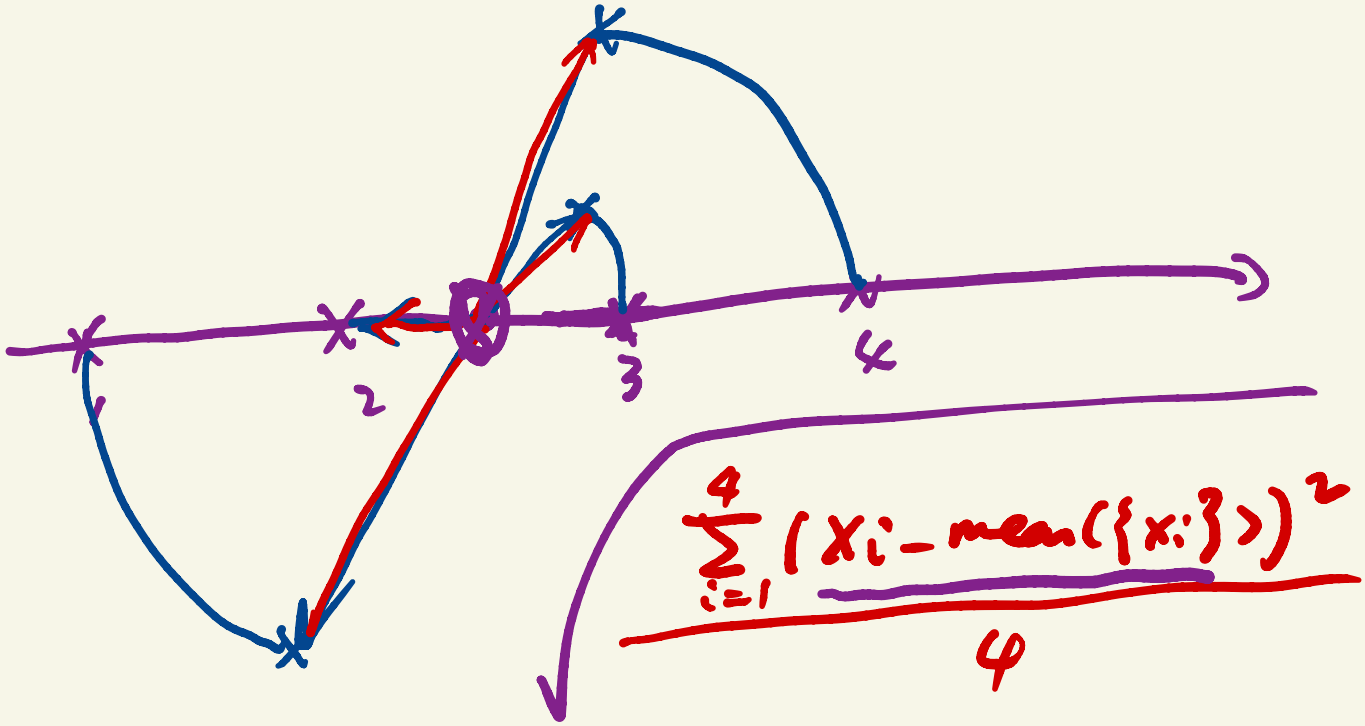
- Recall in which application in Lecture 1 were the means of experiments compared?

Standard Deviation (σ)

- The standard deviation

$$std(\{x_i\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - mean(\{x_i\}))^2}$$

$$= \sqrt{mean(\{(x_i - mean(\{x_i\}))^2\})}$$


1
2
3
4
$\frac{\sum_{i=1}^{4} (x_i - \text{mean}(\{x_i\}))^2}{4}$

Can a standard deviation of a dataset be -1?

A. YES

B. NO

Properties of the standard deviation

- Scaling data scales the standard deviation

$$std(\{k \cdot x_i\}) = |k| \cdot std(\{x_i\})$$

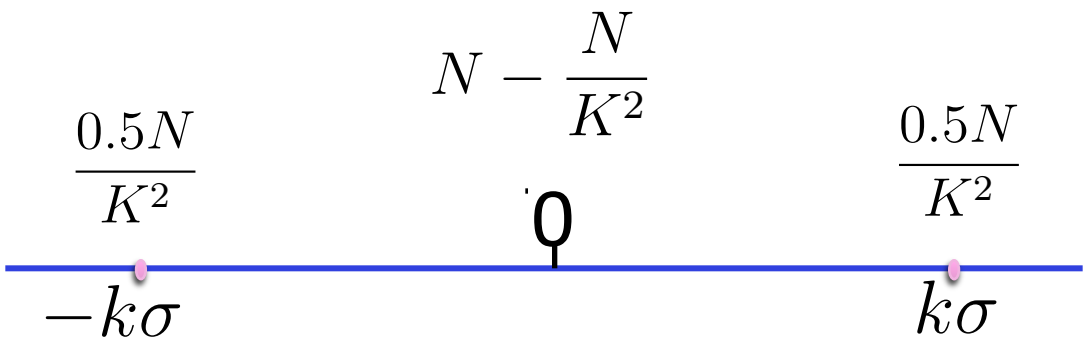
- Translating the data does **NOT** change the standard deviation

$$std(\{x_i + c\}) = std(\{x_i\})$$

Standard deviation: Chebyshev's inequality (1st look)

- At most $\frac{N}{k^2}$ items are k standard deviations (σ) away from the mean

- Rough justification: Assume mean =0



$$std = \sqrt{\frac{1}{N}[(N - \frac{N}{k})0^2 + \frac{N}{k^2}(k\sigma)^2]} = \sigma$$

Variance (σ^2)

- Variance = (standard deviation)^2

$$var(\{x_i\}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - mean(\{x_i\}))^2$$

- Scaling and translating similar to standard deviation

$$var(\{k \cdot x_i\}) = k^2 \cdot var(\{x_i\})$$

$$var(\{x_i + c\}) = var(\{x_i\})$$

Q: Standard deviation

- What is the value of

  std(*mean*($\{x_i\}$)) ?

  A. 0 B. 1 C. unsure

Standard Coordinates/normalized data

- The *mean* tells where the data set is and the *standard deviation* tells how spread out it is. If we are interested only in comparing the shape, we could

define:

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i)\}}$$

- We say $\{\widehat{x_i}\}$ is in standard coordinates

Q: Mean of standard coordinates

- μ of $\{\widehat{x_i}\}$ is:

  A. 1 B. 0 C. unsure

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i)\}}$$

Q: Standard deviation (σ) of standard coordinates

- σ of $\{\widehat{x_i}\}$ is:

  A. 1 B. 0 C. unsure

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i)\}}$$

Q: Variance of standard coordinates

- Variance of $\{\widehat{x_i}\}$ is:

  A. 1 B. 0 C. unsure

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i)\}}$$

Q: Estimate the range of data in standard coordinates

- Estimate as close as possible, 90% data is within:

A. [-10, 10]

B. [-100, 100]

C. [-1, 1]

D. [-4, 4]

E. others

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i)\}}$$

Standard Coordinates/normalized data to $\mu=0$, $\sigma=1$, $\sigma^2=1$

- Data in standard coordinates always has

  mean = 0; standard deviation =1;

  variance = 1.

- Such data is unit-less, plots based on this sometimes are more comparable

- We see such normalization very often in statistics

Median

- To organize the data we first sort it
- Then *if* the number of items N is odd

  median = middle item's value

  if the number of items N is even

  median = mean of middle 2 items' values

Properties of Median

- Scaling data scales the median

$$median(\{k \cdot x_i\}) = k \cdot median(\{x_i\})$$

- Translating data translates the median

$$median(\{x_i + c\}) = median(\{x_i\}) + c$$

Percentile

- k^{th} percentile is the value relative to which k% of the data items have smaller or equal numbers
- Median is the 50th percentile

Q: Scaling effect on percentiles

- Scaling data scales the percentile

  A. True B. False

Q: Translating effect on percentiles

- Translating data does **NOT** change the percentile

A. True B. False

Interquartile range

- iqr = (75th percentile) - (25th percentile)
- Scaling data scales the interquartile range

$$iqr(\{k \cdot x_i\}) = |k| \cdot iqr(\{x_i\})$$

- Translating data does **NOT** change the interquartile range

$$iqr(\{x_i + c\}) = iqr(\{x_i\})$$

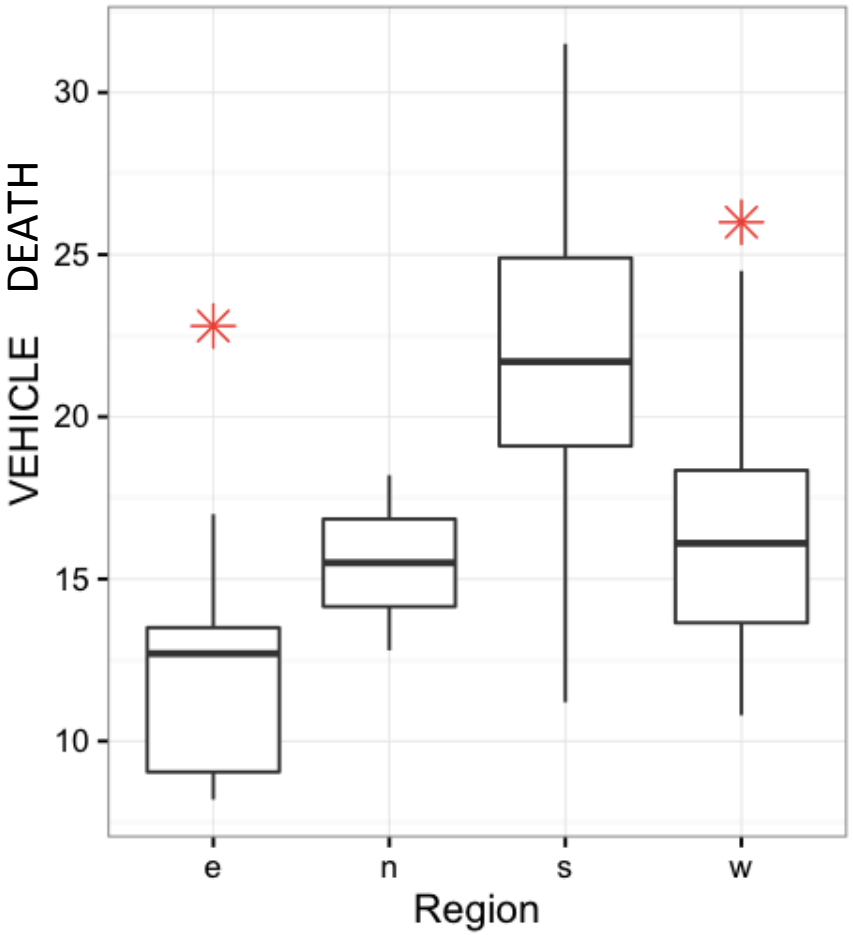
Summarizing 1D continuous data

- Location Parameters
  - Mean
  - Median
  - Mode
- Scale parameters
  - Standard deviation and variance
  - Interquartile range

Box plots

- Boxplots
  - Simpler than histogram
  - Good for outliers
  - Easier to use for comparison

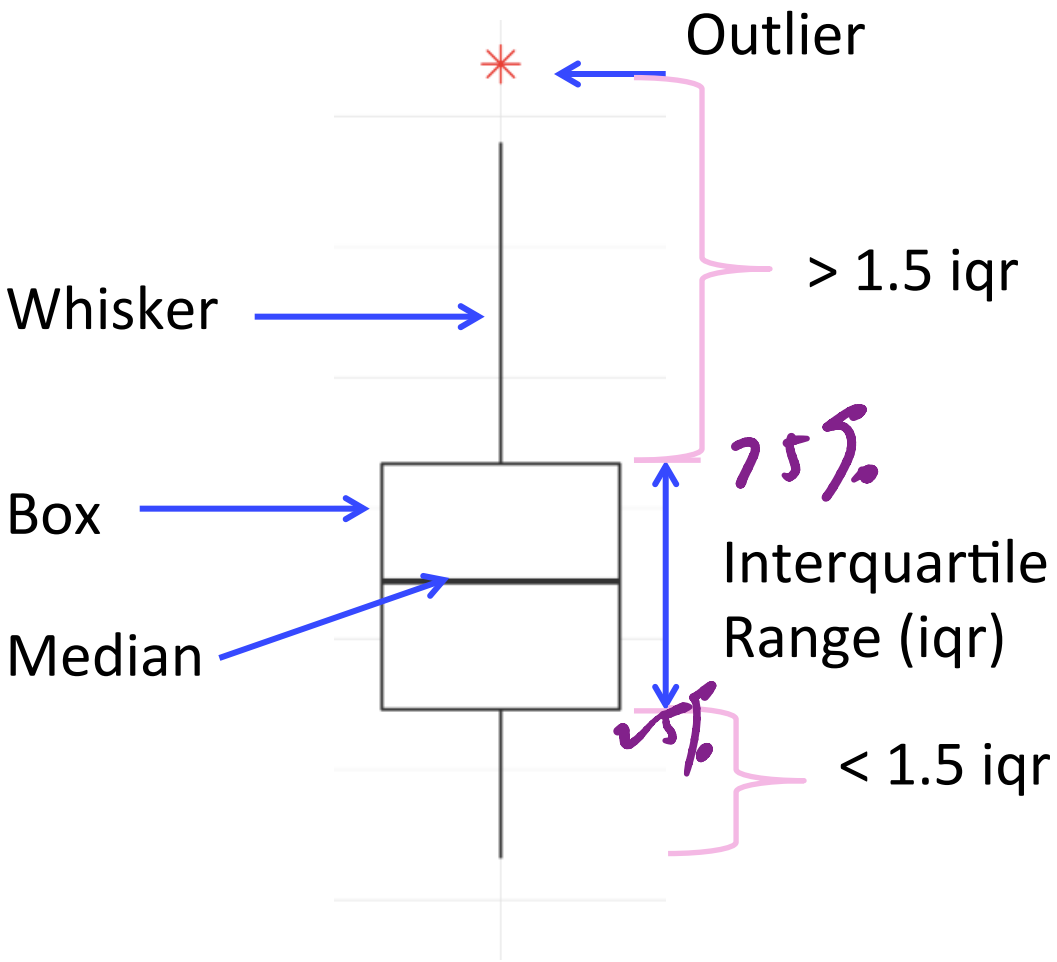
Vehicle death by region



Data from https://www2.stetson.edu/~jrasp/data.htm

Boxplots details, outliers

- How to define outliers? (the default)



Sensitivity of summary statistics to outliers

- mean and standard deviation are very sensitive to outliers
- median and interquartile range are not sensitive to outliers

Group Discussion

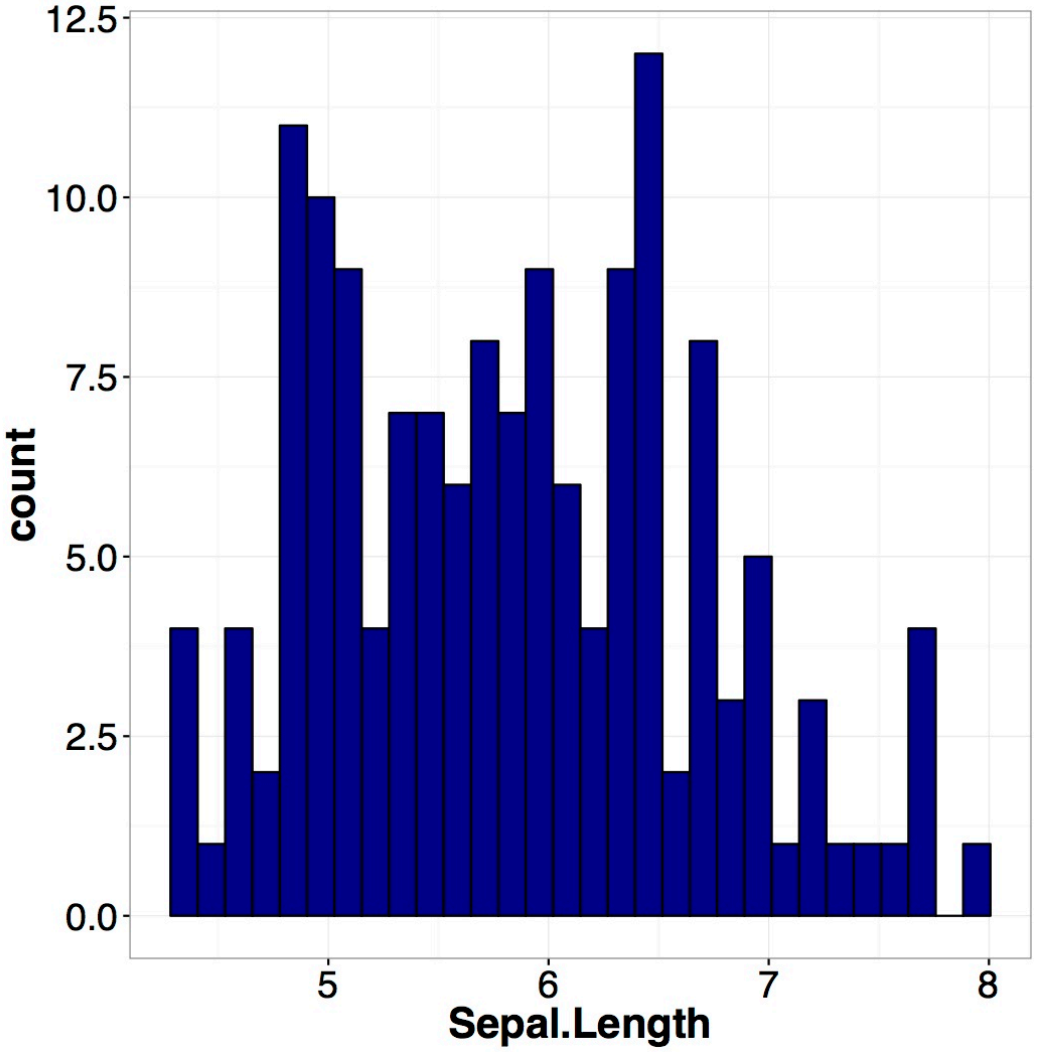
Modes

- Modes are peaks in a histogram
- If there are more than 1 mode, we should be curious as to why

Multiple modes

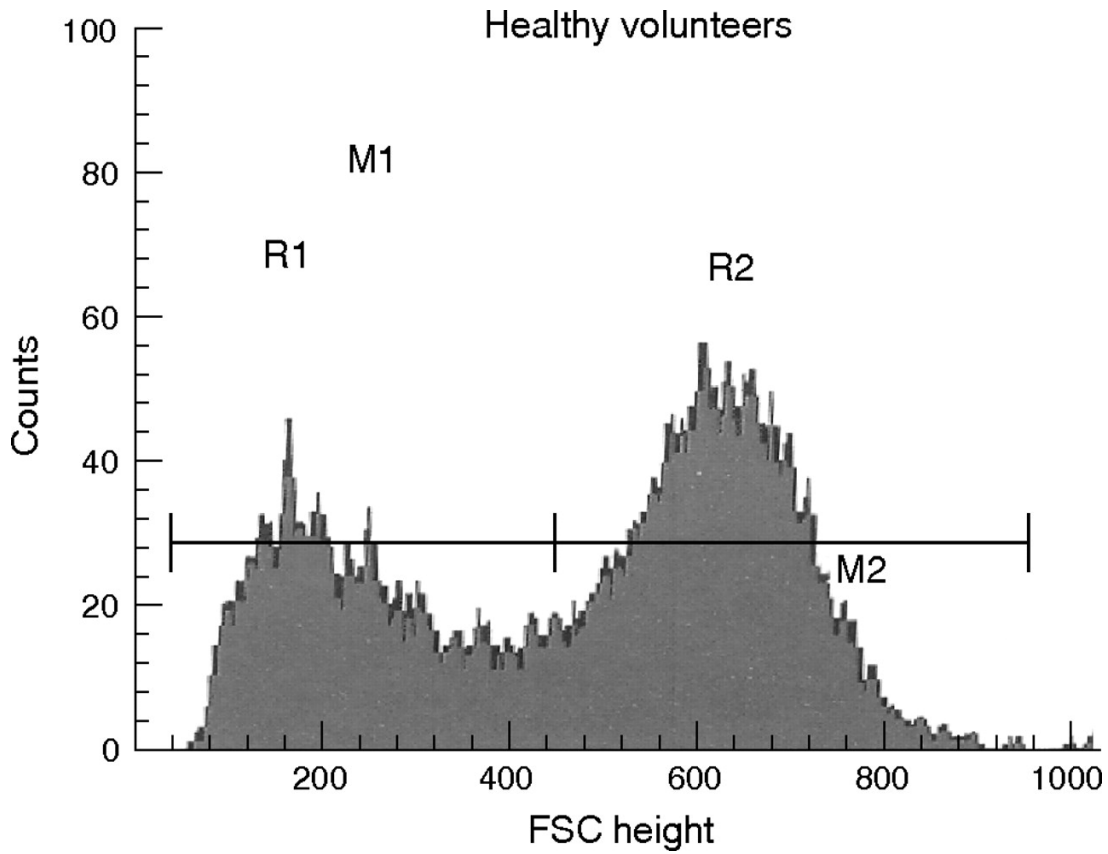
- We have seen the “iris” data which looks to have several peaks

Data: “iris”



Example Bi-modes distribution

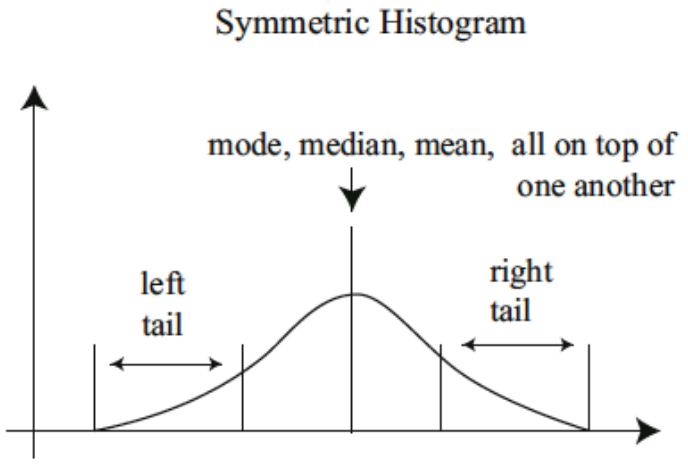
- Modes may indicate multiple populations

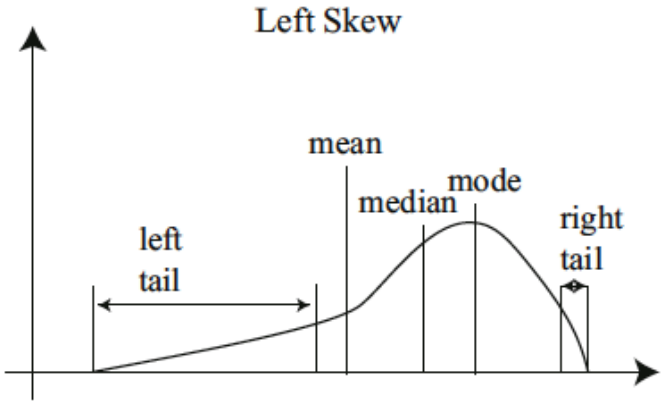


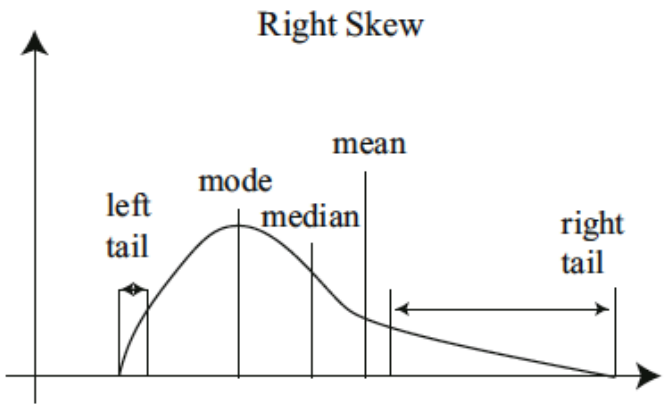
Data: Erythrocyte cells in healthy humans

Piagnerelli, JCP 2007

Tails and Skews







Credit: Prof.Forsyth

Assignments

- HW1, due on 1/30 Thurs.
- Reading Chapter 2 of the textbook
- Next time: Looking for relationship in data; correlation coefficient

Additional References

- Peter Dalgaard "Introductory Statistics" with R
- Charles M. Grinstead and J. Laurie Snell "Introduction to Probability”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

See you next time

See You!