Describing Data

Let's say you are considering buying a house in a certain neighbourhood. When you are deciding to buy in the neighbourhood your realtor (sensing your snobbiness) mentions to you that the average income in this neighbourhood is $100,000. You decide to buy the house.

A year later, the same realtor knocks on your door. Now he is acting as a representative of the neighbourhood taxpayers' association. He wants you to sign a petition to decrease property taxes. After all, he says, the residents can't afford an increase in property taxes since the average family income in the neighbourhood is only $25,000.

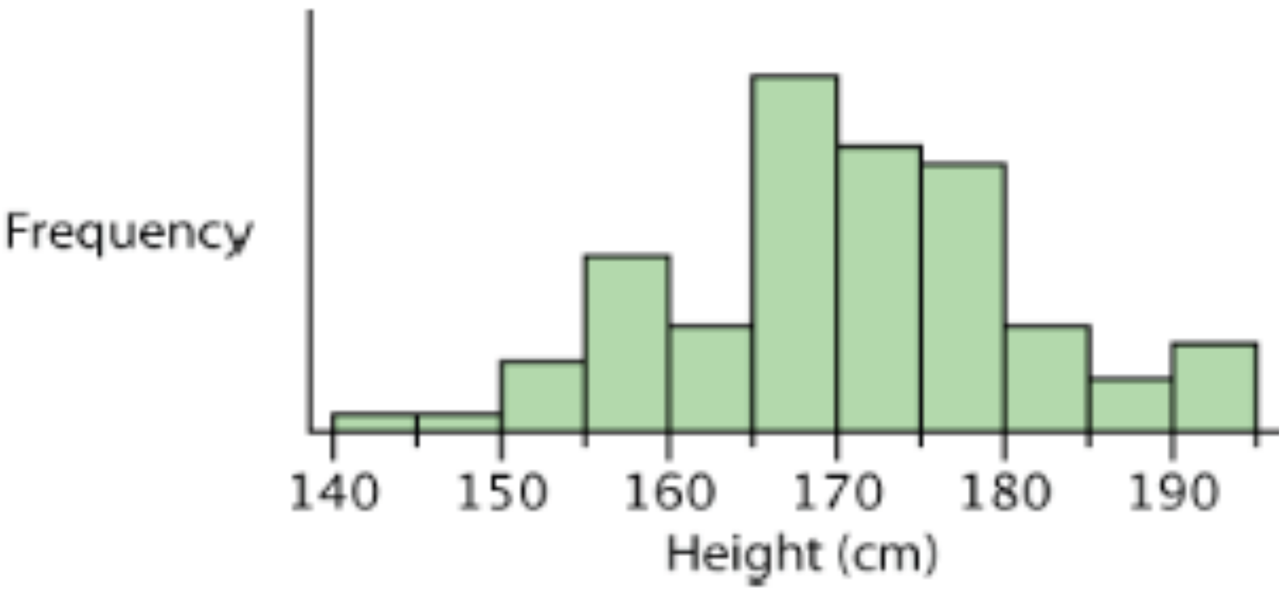
The two common descriptions over data:

- **Location** (central tendency or moment)
  - Where is the weight of the data?

- **Spread**
  - How far apart is the smallest and the largest data points?

(2 other less common descriptors: **Skew**, **Kurtosis**)

Describing Data



Mean salary of everyone present in a greasy spoon:

waiter	$35000
Cook	$30000
Dishwasher	$25000
Customer 1	$80000
Customer 2	$50000
Customer 3	$30000
Customer 4	$45000

Mean:

waiter	$35000
Cook	$30000
Dishwasher	$25000
Customer 1	$80000
Customer 2	$50000
Customer 3	$30000
Customer 4	$45000
Customer who is a Software Engineer	$1000000000

Median:

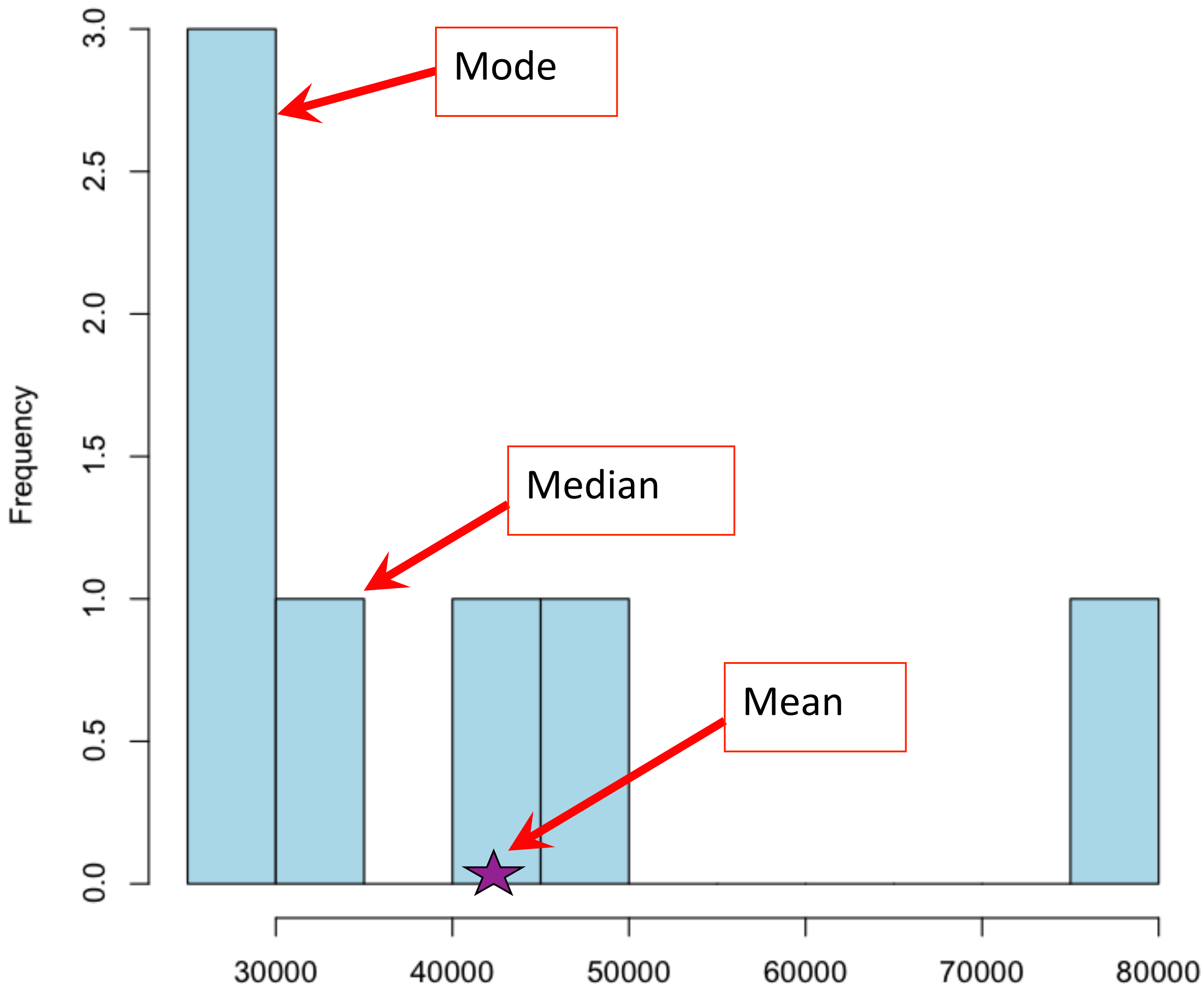
Midpoint of ordered data

$35000
$30000
$25000
$80000
$50000
$30000
$45000

Order data

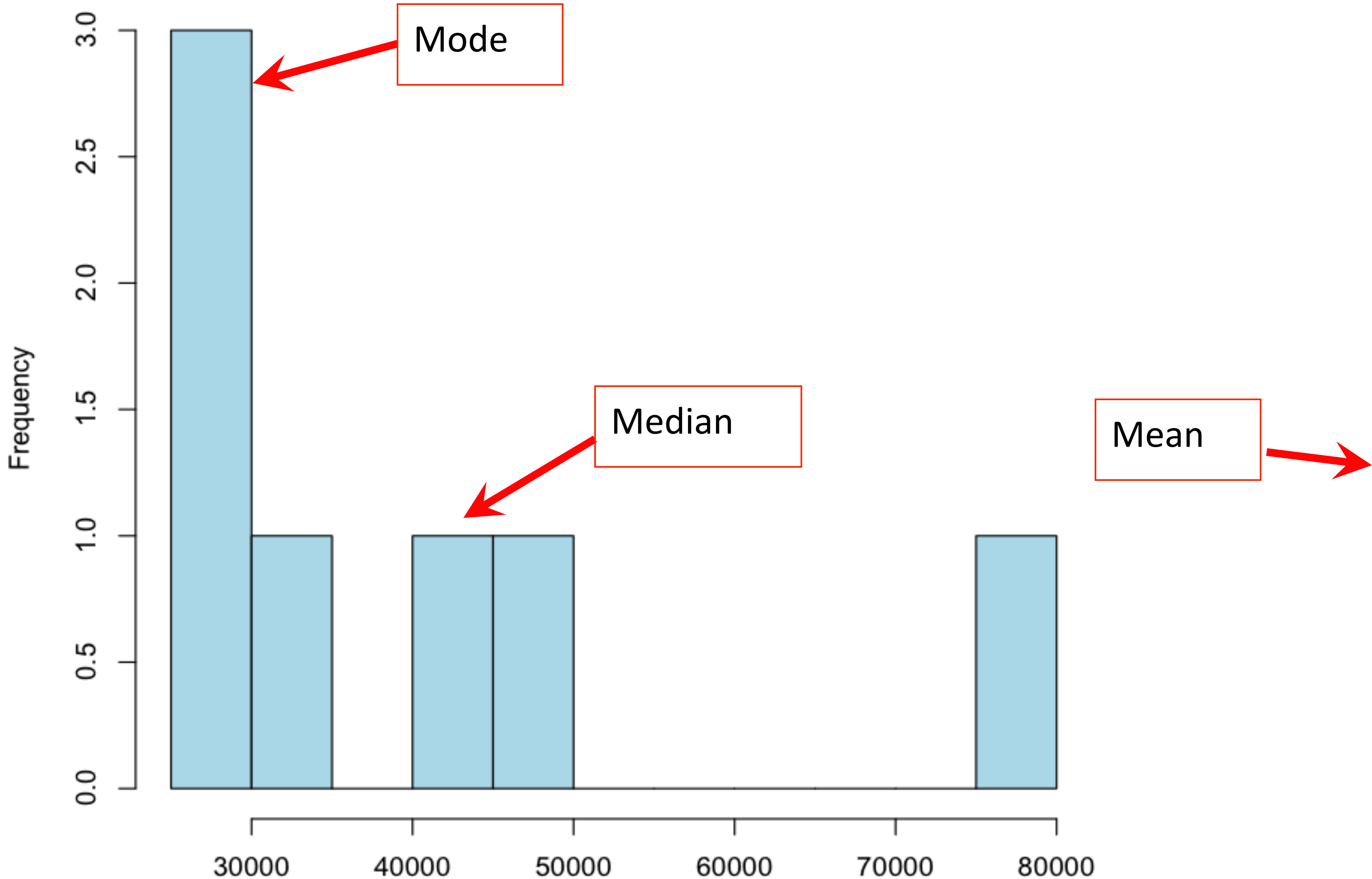
$25000
$30000
$30000
$35000
$45000
$50000
$80000

Describing Data



Describing Data

If the software engineers annual wage was included:



Summary of location descriptors :

- Mean, Mode and Median usually give you slightly different information and have different benefits

- If data are skewed (or have outlier), median is a more fair reflection of the data

- Median (and its measure of spread, interquantile range) give quick information about the data without having to calculate anything

- Mean can be an artificial abstract whereas median is ‘real’

- Why use the mean at all?
  - It allows you to answer questions about populations as a whole and do hypothesis testing
- Proportions are used on categorical data and ‘behaves’ similarly to a mean

Spread:

- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

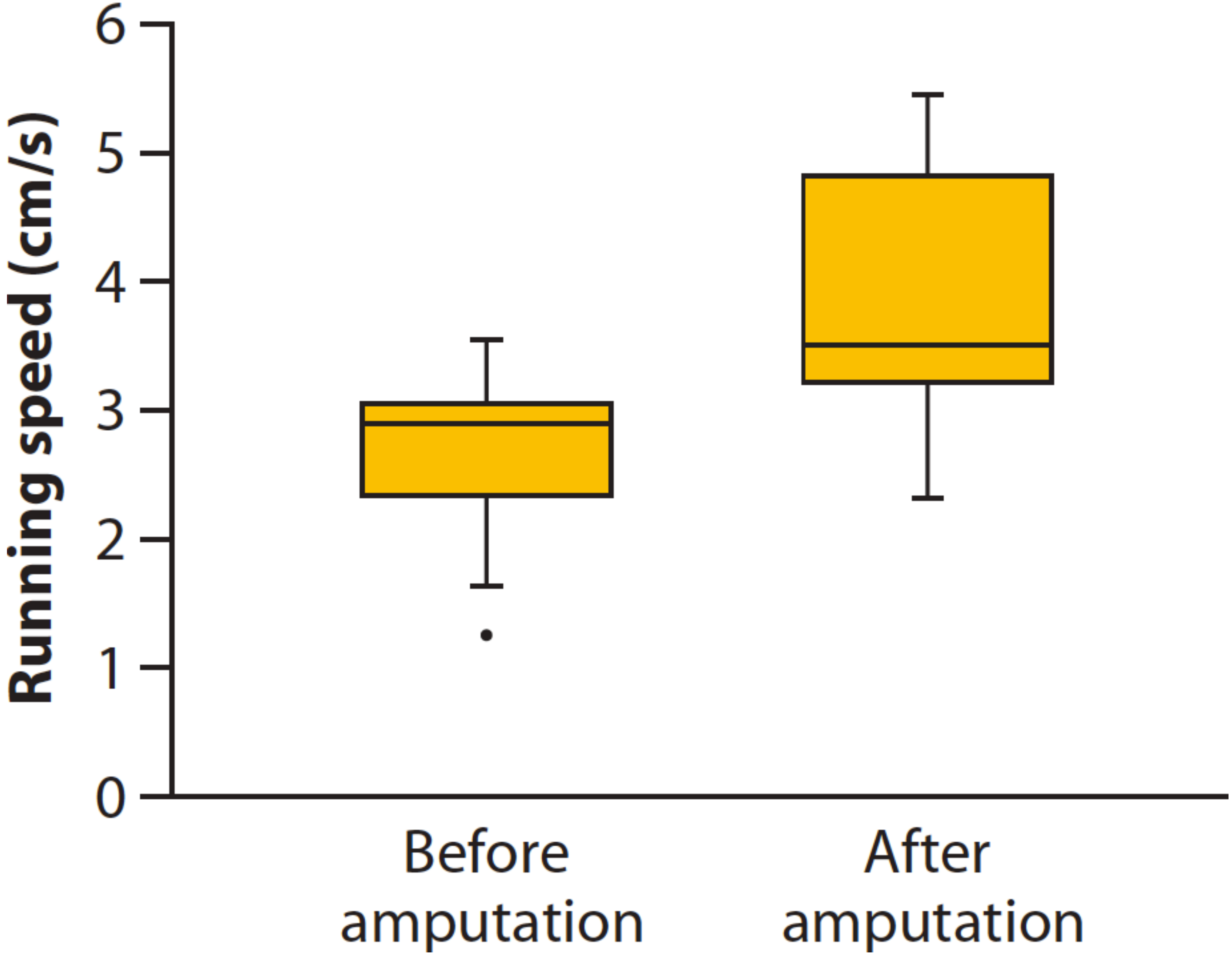
Interquartile Range:

- Divide data into four equal parts and see how far apart the extreme groups are

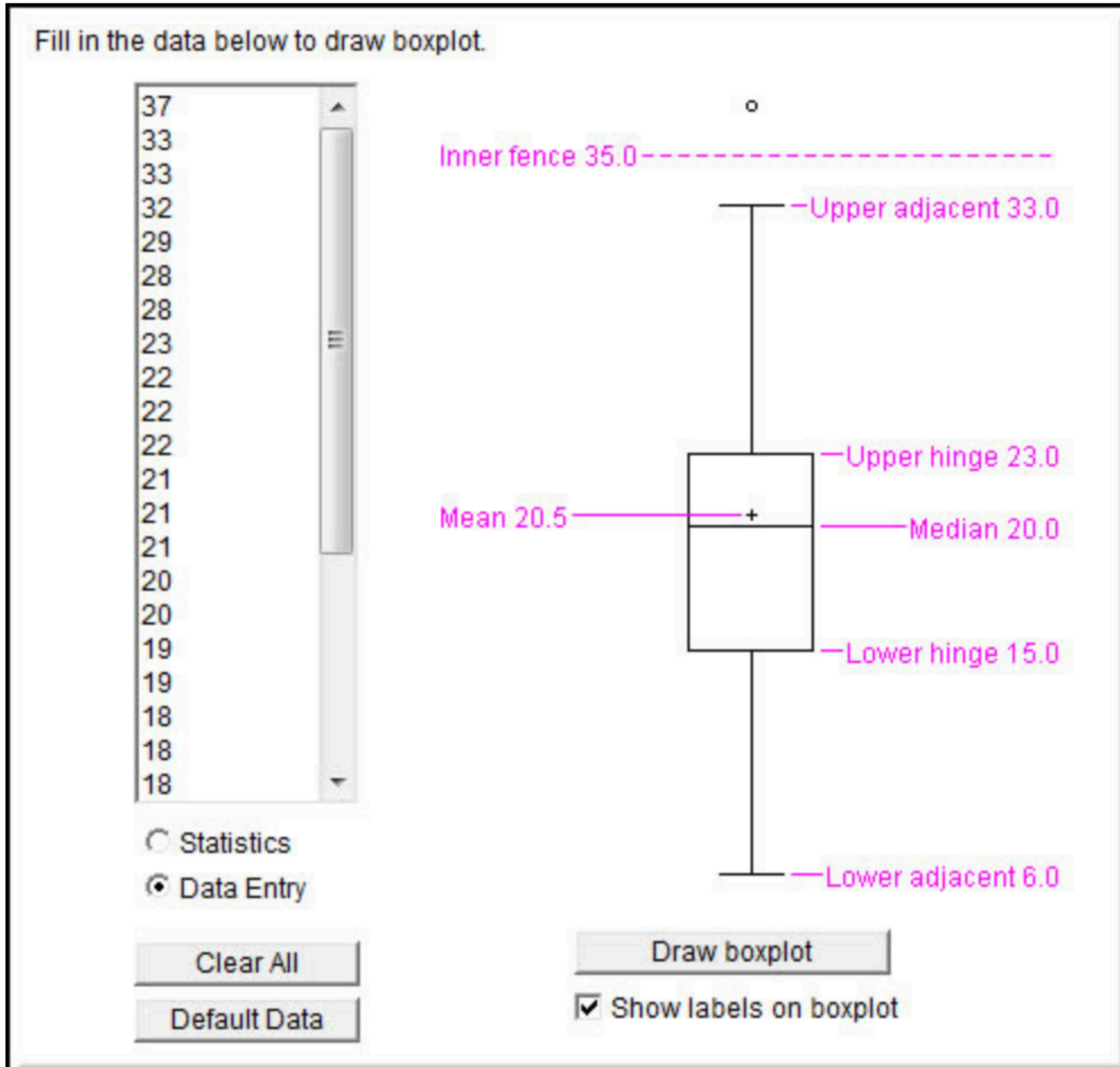
- Interquartile range = 3rd quartile - 1st quartile

- Ex. Box and whiskers plots
  - Displays median and interquartile range
  - Q_1 and Q_3

Box Plots:


Running speed (cm/s)
6
5
4
3
2
1
0
Before amputation
After amputation

Box Plots: http://onlinestatbook.com/2/graphing_distributions/boxplot_demo.html



Coefficient of Variation (CV):

$$CV = 100\% \frac{s}{\overline{Y}}$$

Coefficient of Variation is the standard deviation expressed as a percentage of the mean

Quantiles of a Frequency Distribution:

Percentile: *The percentile of a measurement specifies the percentage of observations less than or equal to it; the remaining observations exceed it.*

Ex: *50th percentile is the measurement that splits the frequency distribution into equal halves*

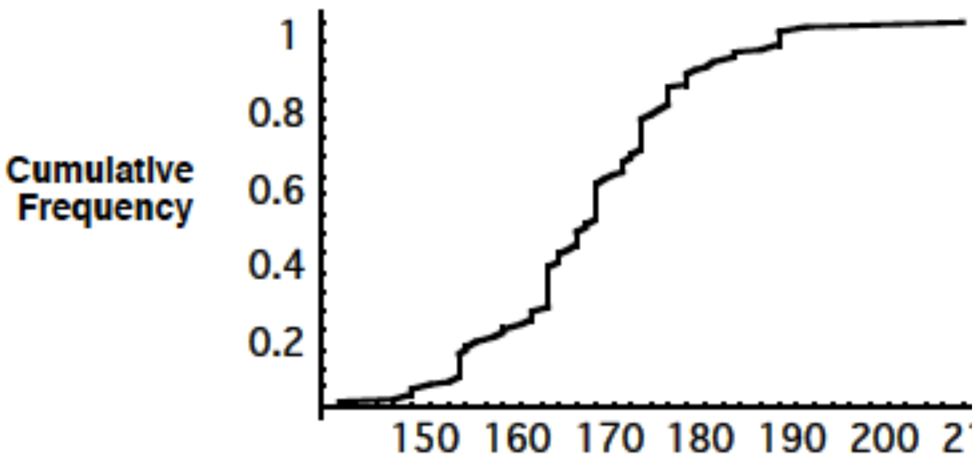
Ex: *10th percentile is the measurement where 10% of the data are less than or equal to it (the other 90% are greater than it)*

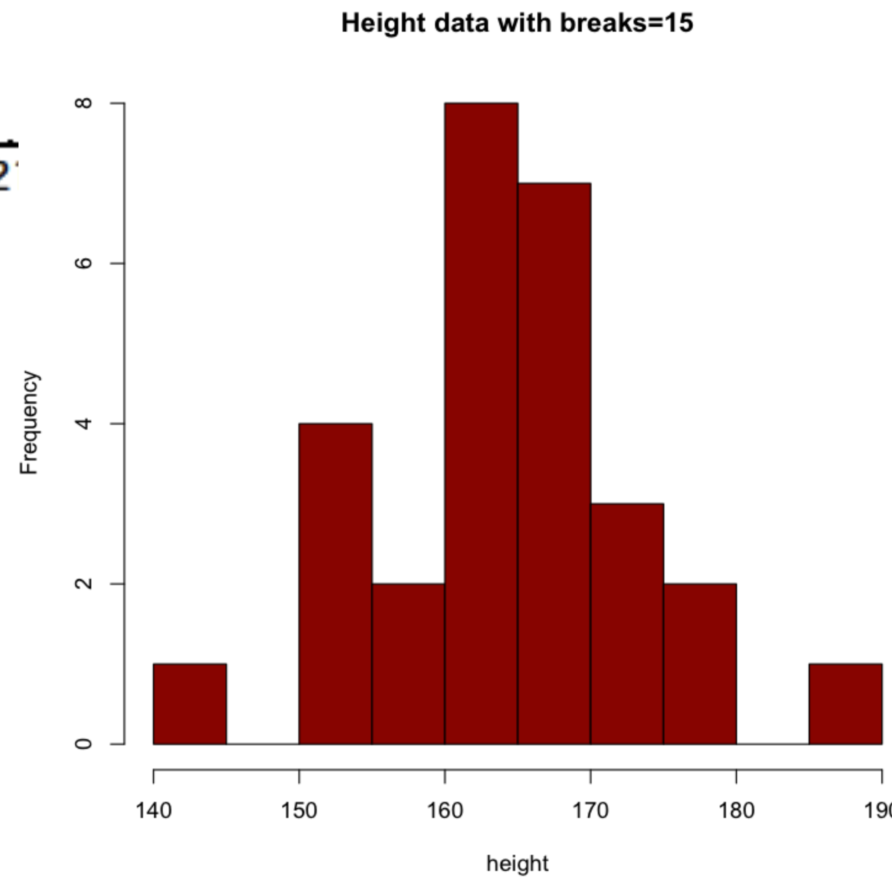
Quantile: X/100

Ex: 0.5 quantile = 50th percentile

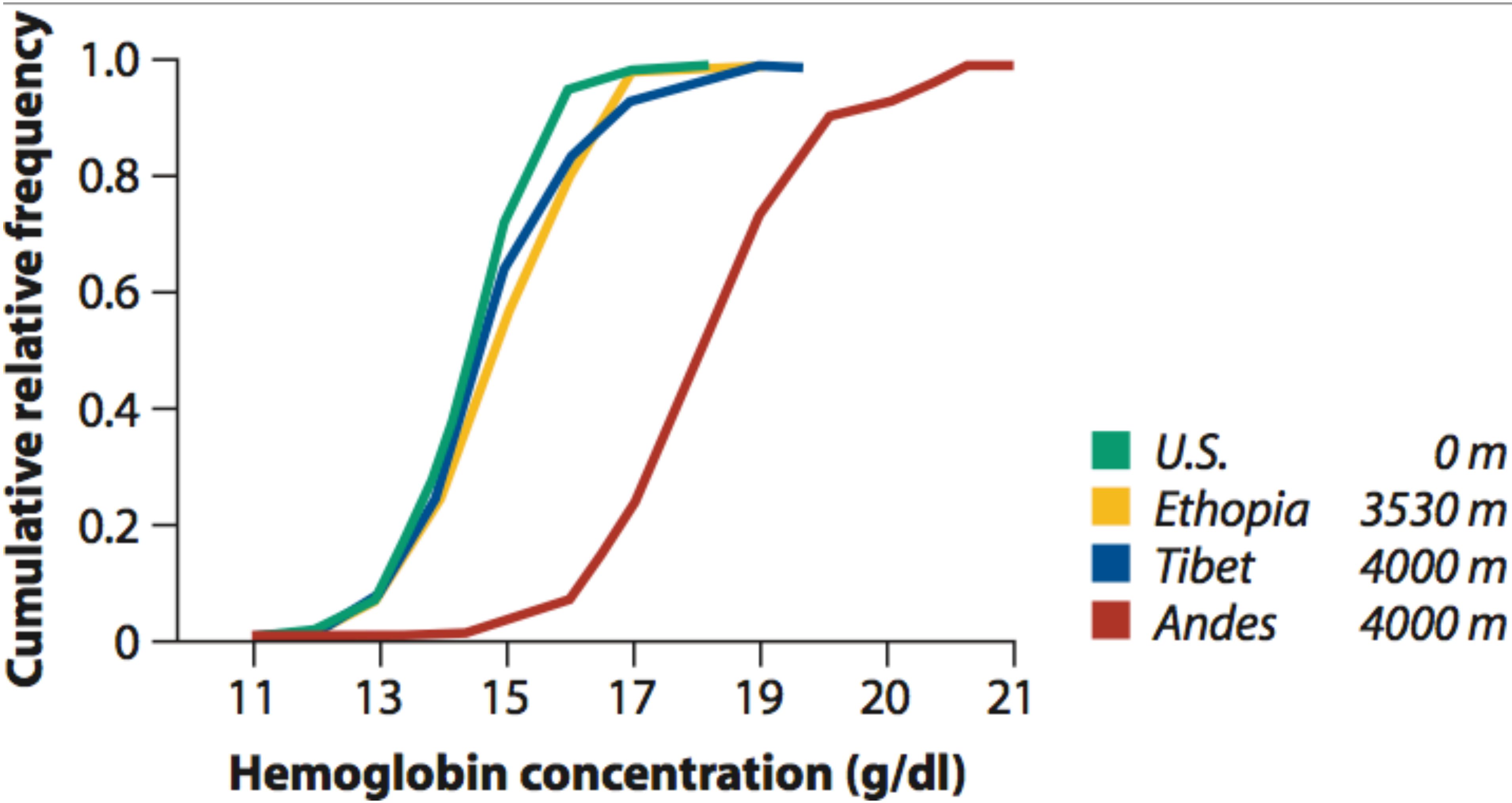
Ex: 0.10 quantile = 10th percentile

Graphing Data





Multiple Cumulative Frequency Distribution:



Proportions

- The most important descriptor for categorical variables
- Similar to arithmetic mean

$$\hat{p} = \frac{NumberCategory}{n}$$

Manipulating Means:

1. $E[X+Y] = E[X]+E[Y]$
2. $E[X+c] = E[X]+c$
3. $E[cX] = cE[X]$
4. $E[XY] = E[X]E[Y]$, iff X and Y are independent

Manipulating Variance:

1. $Var[X+Y] = Var[X]+Var[Y]$, iff X and Y are independent

2. $Var[X+c] = Var[X]$

3. $Var[cX] = c^2Var[X]$