Module 1A

Descriptive statistics: Location & Spread

Module 1 : Descriptive Statistics

Measurements of ***location*** and ***spread*** of data

Agenda:

- Mean, mode, median
- Variability, variation, range
- Simpson’s paradox
- Intuitions about uncertainty: Fermi Estimation
- Accuracy/Bias and Precision/Spread

You are considering buying a house in a certain neighbourhood. You find a potential house and, to appeal to perceived snobbiness as you are making your decision, your realtor mentions that the **average income in this neighbourhood is $100,000 per year.**

You buy the house.

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

How is this possible, if the realtor is telling the truth, and no one in the neighbourhood has moved or changed jobs in the last year?

The two common descriptions of data:

1. **Location:**
   - Central Tendency
   - Where is the weight of the data?

Average

2. **Spread:**
   - How far apart are the data points? Especially: how far apart are the largest and smallest data points?

Range

You will also see:

1. **Skew –** The third standardized moment; positive or negative skew. The shape of the distribution is not symmetric.
2. **Kurtosis –** The fourth standardized moment; sort of ‘peakness’ of the distribution (fatness of the tails)

A story about central location of the data

Waiter	$35,000
Cook	$30,000
Dishwasher	$25,000
Customer 1	$80,000
Customer 2	$50,000
Customer 3	$30,000
Customer 4	$45,000

“Average” is approx. **$42,143**

“Average” is $**125,000,037**

Waiter	$35,000
Cook	$30,000
Dishwasher	$25,000
Customer 1	$80,000
Customer 2	$50,000
Customer 3	$30,000
Customer 4	$45,000
Software or Social Engineer	$1,000,000,000

$35,000
$30,000
$25,000
$80,000
$50,000
$30,000
$45,000

Reorder → data

$25,000
$30,000
$30,000
$35,000
$45,000
$50,000
$80,000

(Arithmetic) **Mean** = $\frac{\sum_1^n x_i}{n}$

Median = middle value (odd), mean of middle value (even)

Mode = most frequent value

	Scenario 1	Scenario 2
mean	$42 143	$125,000,037
median	$35,000	$40,000
mode	$30,000	$30,000

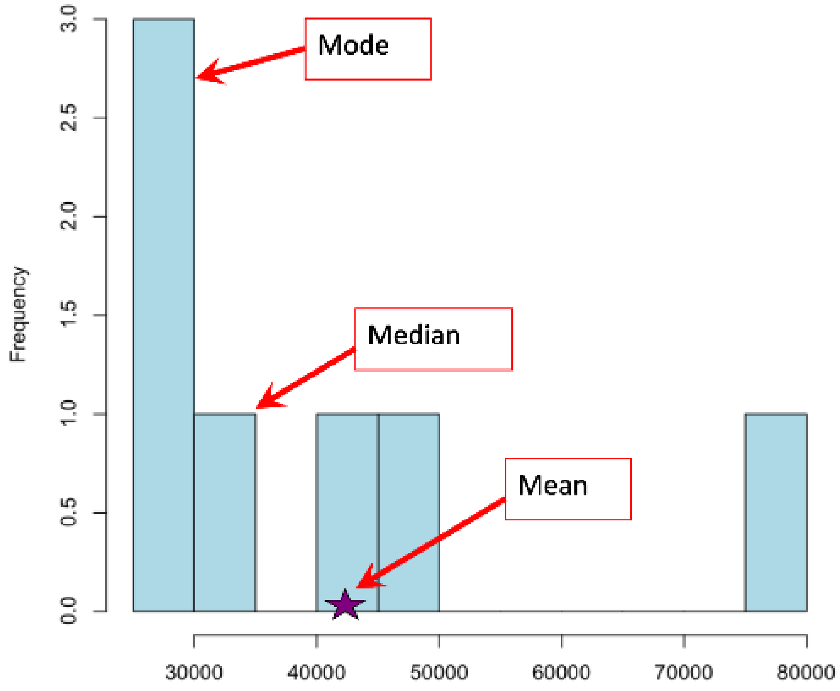
$35,000
$30,000
$25,000
$80,000
$50,000
$30,000
$45,000
$1,000,000,000

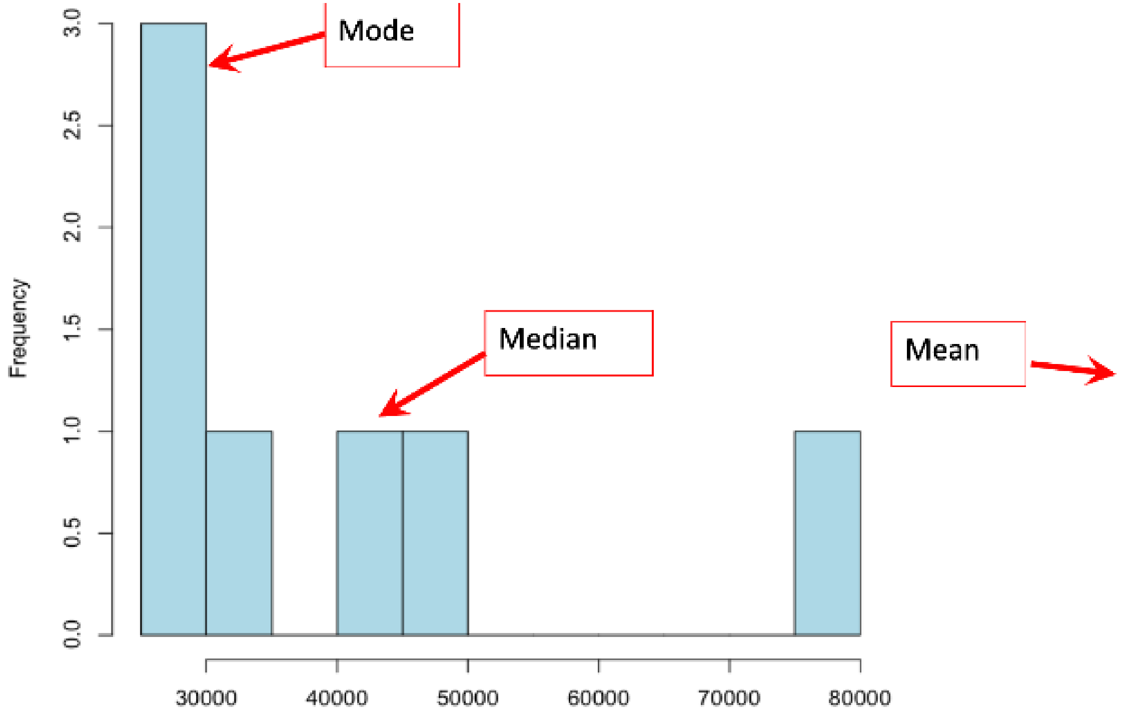
Reorder → data

$25,000
$30,000
$30,000
$35,000
$45,000
$50,000
$80,000
$1,000,000,000

Mean, Mode, and Median can give you **different information** and they have **different benefits**

- If the data are **skewed** or have an outlier**, median** is often a fairer reflection of the data
- **Median** can give **quick information** abut the data without having to calculate anything
- (arithmetic) mean can be a theoretical abstract (2.2 children per woman doesn't actually exist), but it **allows you to use normal distribution** to answer questions about the **whole population**





A story about spread (and shift of location) of the data

Spread of Data:

1. Variance

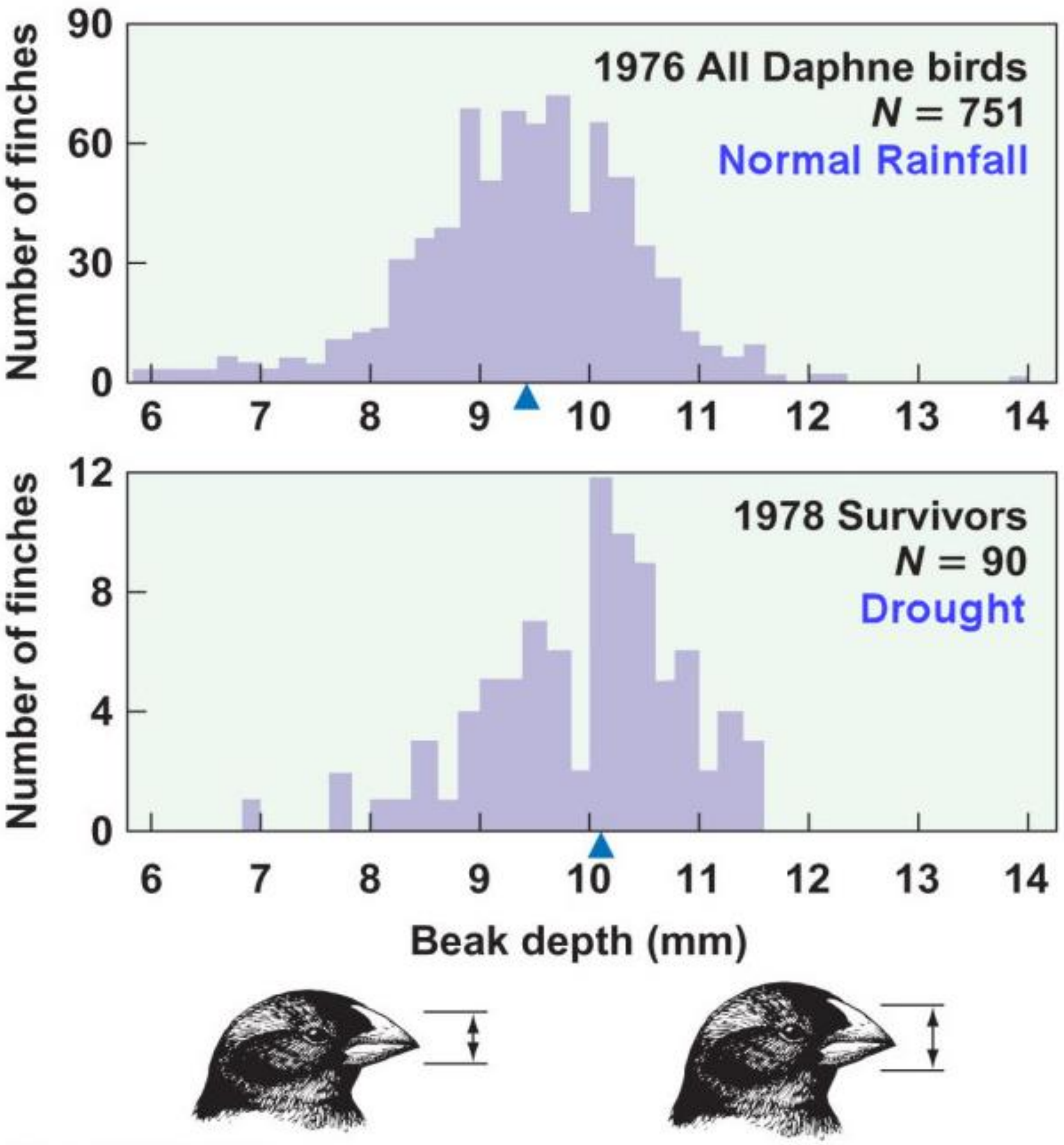
$$\sigma^2 = \frac{\sum(x_i - \bar{x}\,)^2}{n-1}$$

2. Standard Deviation
   - Same units as data
   - σ

3. Range
   - largest – smallest value

4. Interquartile Range
   - 25th to 75th percentile

Peter and Rosemary Grant and the Ongoing Evolution of Galapagos Finches


Number of finches
1976 All Daphne birds
N = 751
Normal Rainfall
Number of finches
1978 Survivors
N = 90
Drought
Beak depth (mm)