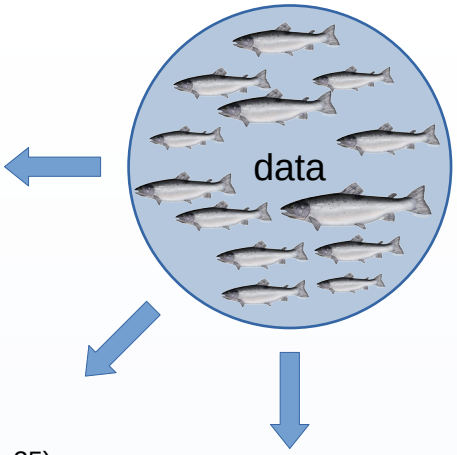
# EGM101 – Skills Toolbox

## Week 5, Part 5: Descriptive Statistics

# Summarizing Data

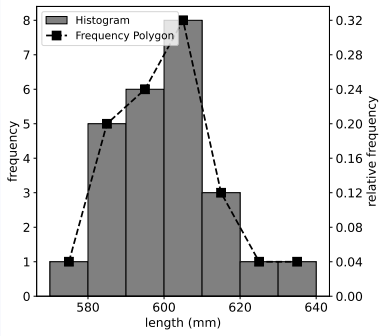
- So far, we have seen:
  - Stem-and-leaf plot
  - Tally chart
  - Frequency table
  - Histogram/frequency polygon
- Despite advantages, comparisons between datasets can be difficult
- Example questions:
  - Are there differences between datasets?
  - Is the population changing over time?

| Stem | Leaf |
|---|---|
| 57 | 4 |
| 58 | 4 5 7 9 9 |
| 59 | 1 2 3 7 7 8 |
| 60 | 1 4 4 5 5 6 7 8 |
| 61 | 5 6 9 |
| 62 | 3 |
| 63 | 5 |




Frequency table of sampled salmon lengths (n=25)

| Length (mm) | Frequency | Relative Frequency | Cumulative Relative Frequency |
|---|---|---|---|
| 570 – 579.99 | 1 | 1/25 = 0.04 | 0.04 |
| 580 – 589.99 | 5 | 5/25 = 0.20 | 0.24 |
| 590 – 599.99 | 6 | 6/25 = 0.24 | 0.48 |
| 600 – 609.99 | 8 | 8/25 = 0.32 | 0.80 |
| 610 – 619.99 | 3 | 3/25 = 0.12 | 0.92 |
| 620 – 629.99 | 1 | 1/25 = 0.04 | 0.96 |
| 630 – 630.99 | 1 | 1/25 = 0.04 | 1.00 |
| **Total** | **25** | **25/25 = 1.00** | **1.00** |



# Central Tendency

- Goal: describe the dataset using a single representative value
- Central Tendency: “where” the data are clustered
  - *Measures of location*
  - *Average*
- NB: all of the following are “averages”
  - Be specific!

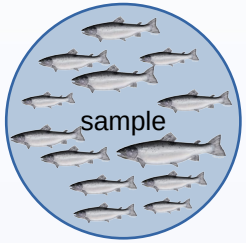
# The Arithmetic Mean

- Perhaps the most commonly-used measure
- Pros:
  - Fully representative of data (uses all values)
  - Easy to calculate
  - Can be algebraically manipulated
- Cons:
  - Sensitive to large outliers
  - Value might not be possible for data

| $x$: | $x_1$ | $x_2$ | $x_3$ | … | $x_{n-1}$ | $x_n$ |
|---|---|---|---|---|---|---|

sample: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

population: $$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$




$$\bar{x} = \frac{601+623+585+\ldots+589+597}{25}$$

$$= \frac{15024}{25} = 600.96$$

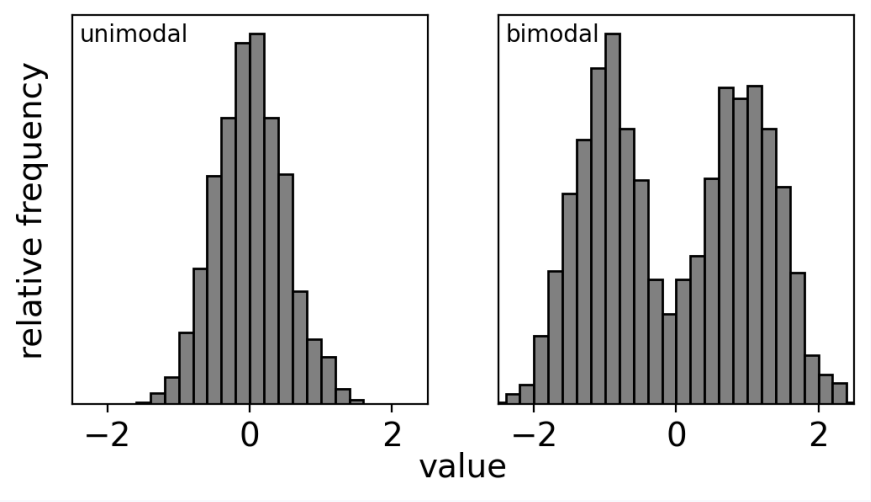
# Mode

- Most frequent value
- Straightforward to find for discrete observations
  - Less so for grouped, continuous
- Pros:
  - Modal value is possible
  - Unaffected by extremes
- Cons:
  - Not necessarily unique (bimodal)
  - Doesn't necessarily exist

$x$:

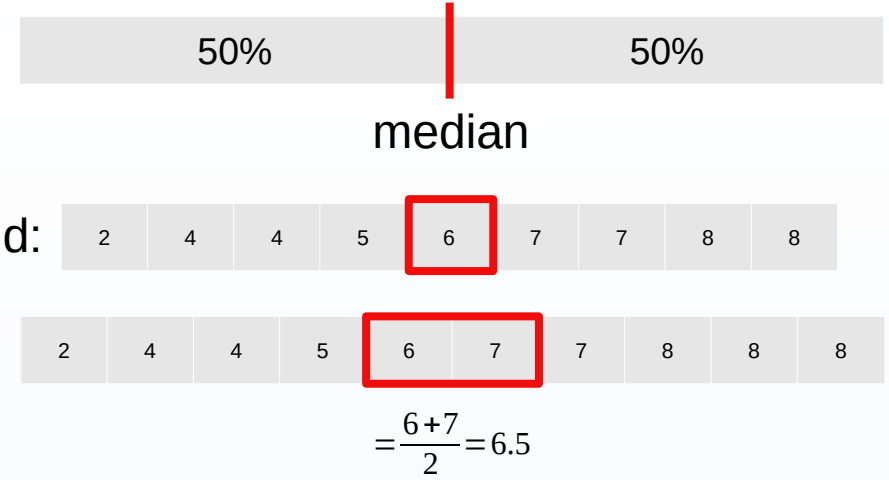
| 2 | 4 | 4 | 5 | 6 | 7 | 7 | 8 | 8 | 8 |
|---|---|---|---|---|---|---|---|---|---|

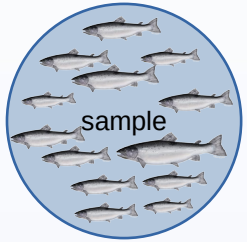
Mean: 5.9
Mode: 8



# The Median

- The exact middle of the distribution
  - 50% above, 50% below
- Straightforward to calculate, even for open-ended classes
- Pros:
  - Less affected by extreme values
  - Can calculate without measuring all values
- Cons:
  - Might not be possible value
  - Not based on all observations

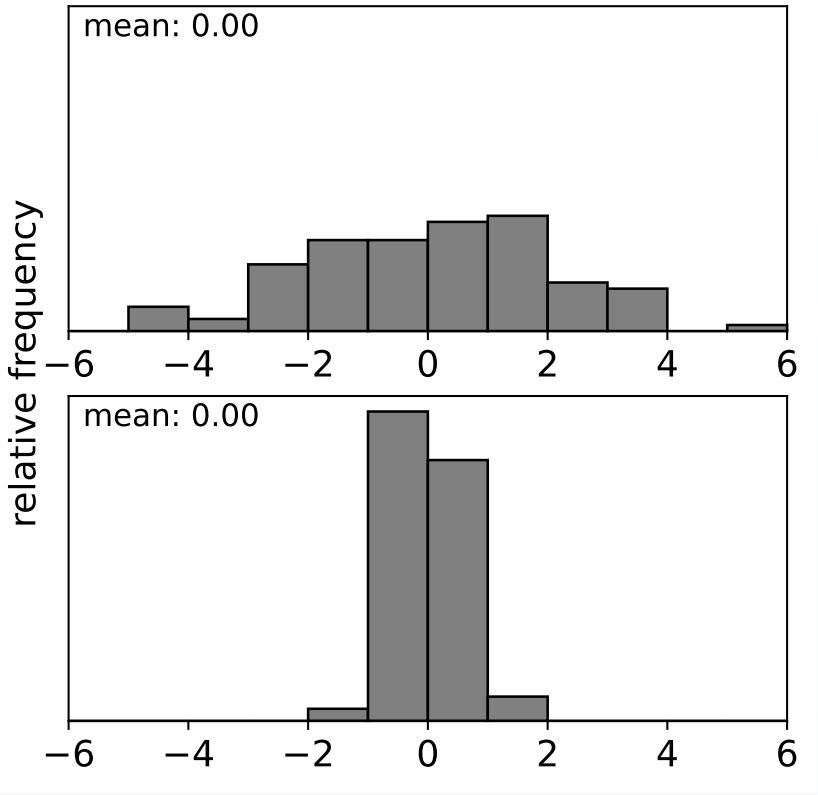



574 ... 598 601 604 ... 635

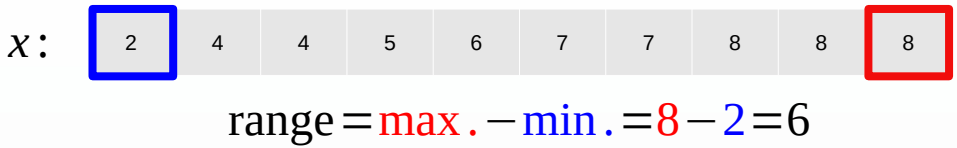
# Dispersion

- Problem: central tendency doesn't say much about sample
- Two datasets can have same mean (median/mode/etc.) and be very different
- Spread of data (dispersion) is also important



# Range

- Difference between highest, lowest value
- Pros:
  - Easy to calculate, understand
- Cons:
  - Extreme values can be misleading

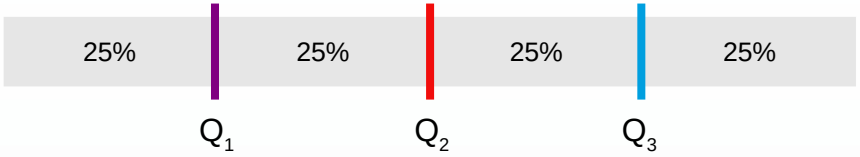



$$\text{range} = \max. - \min. = 8 - 2 = 6$$


$$\text{range} = 100 - 2 = 98$$

# Interquartile Range

- Quartile: 25% of the data
  - Median: $Q_2$
- Interquartile Range (IQR): difference between $Q_3$, $Q_1$
  - Middle 50% of data
- Semi-interquartile range (quartile deviation): IQR / 2
- Pros:
  - Can be used even if we don't have exact values (e.g., grouped frequency distribution)
  - Not affected by extreme values
- Cons:
  - Less useful for mathematical manipulation




$x$: 2 4 4 5 6 7 7 8 8 8

$$Q_2 = \frac{6+7}{2} = 6.5$$

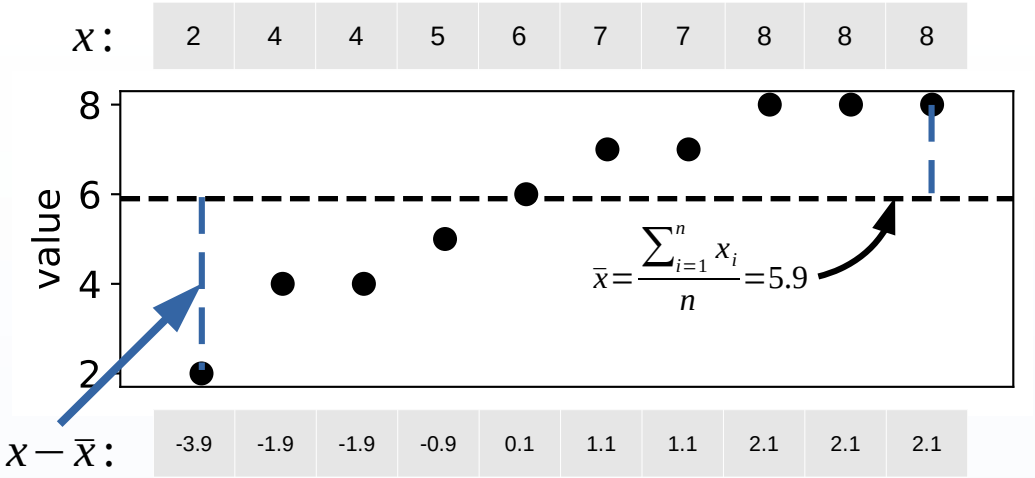
$Q_1$: 4 $Q_2$: 6.5 $Q_3$: 8

$$IQR = Q_3 - Q_1$$

$$= 8 - 4 = 4$$

# Mean deviation

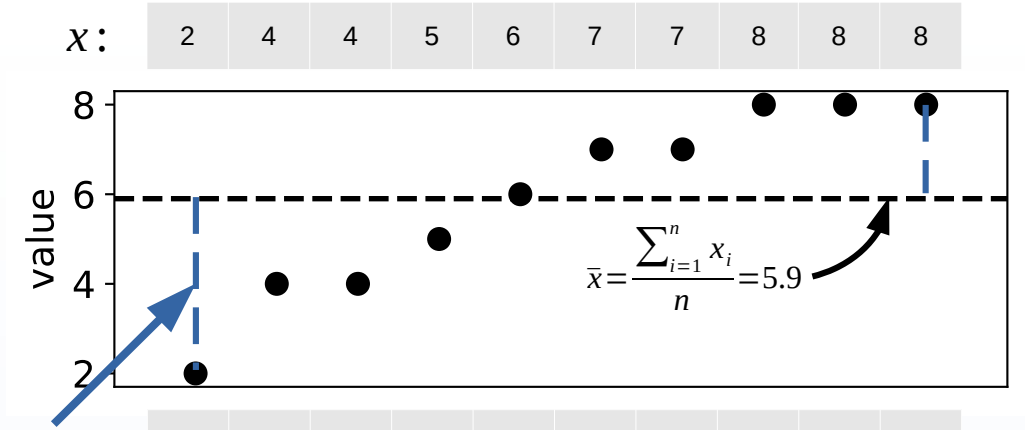
- Idea: subtract mean from dataset
- Problem: total deviation from mean is always 0.
- Solution: ignore sign (use absolute value)
- Pros:
  - Useful to compare variation of different datasets
- Cons:
  - Sensitive to extreme values
  - Grouped calculation can be more difficult




$$\text{MD} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = 1.72$$

# Variance

- Instead of absolute value, square the differences
  - More “weight” on larger differences
  - Makes some kinds of analysis easier
- For sample, use $n - 1$
  - Attempt to correct for sampling bias
- Pros:
  - Direction (±) doesn’t matter
  - Can be easily mathematically manipulated
- Cons:
  - More “weight” given to larger deviations (bad for skewed data)
  - Different units (squared)




| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$: | 2 | 4 | 4 | 5 | 6 | 7 | 7 | 8 | 8 | 8 |
| $x - \bar{x}$: | -3.9 | -1.9 | -1.9 | -0.9 | 0.1 | 1.1 | 1.1 | 2.1 | 2.1 | 2.1 |
| $(x - \bar{x})^2$: | 15.2 | 3.6 | 3.6 | 0.8 | 0.0 | 1.2 | 1.2 | 4.4 | 4.4 | 4.4 |

$MD = 1.72$

$\sigma^2 = 3.89$

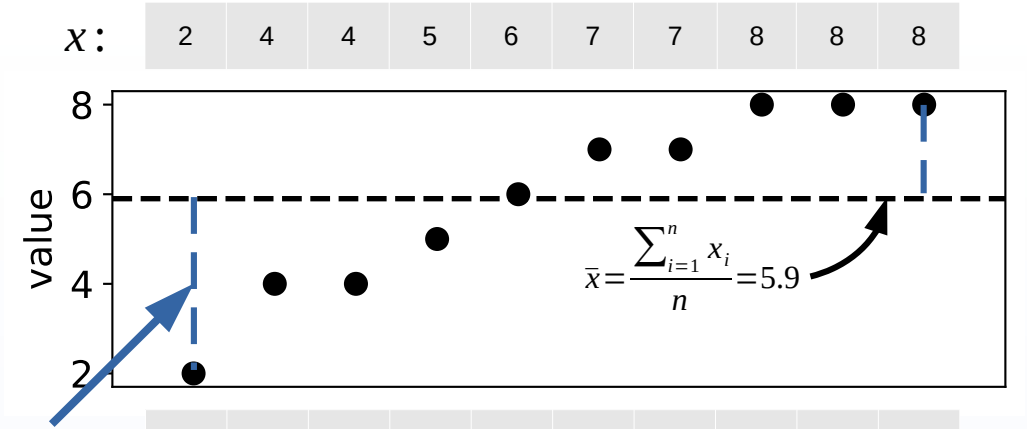
$s^2 = 4.32$

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

sample: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

# Standard Deviation

- Problem: variance has different units (squared)
  - Solution: take a square root
- Pros:
  - Same as variance, but easier to compare
- Cons:
  - More “weight” given to larger deviations
  - Dividing by $n-1$ for the sample isn’t as good of a correction




| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$: | 2 | 4 | 4 | 5 | 6 | 7 | 7 | 8 | 8 | 8 |
| $x-\bar{x}$: | -3.9 | -1.9 | -1.9 | -0.9 | 0.1 | 1.1 | 1.1 | 2.1 | 2.1 | 2.1 |
| $(x-\bar{x})^2$: | 15.2 | 3.6 | 3.6 | 0.8 | 0.0 | 1.2 | 1.2 | 4.4 | 4.4 | 4.4 |

$MD = 1.72$

$\sigma = 1.97$

$s = 2.08$

population: $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

sample: $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

# Coefficient of Variation

- Sometimes, we want to compare distributions with different units
  - e.g., $, £, €
- Want a dimensionless measure (no units):

population: $c_V = \frac{\sigma}{\mu}$ sample: $\hat{c_V} = \frac{s}{\bar{x}}$

- Applications:
  - Chemistry/Engineering
  - Economics
- Pros:
  - It's dimensionless (can compare quantities with different units)
- Cons:
  - Tends to infinity for small $\mu$
  - Not a reliable measure of certainty
  - Sensitive to extreme values

# Summary

- Goal: describe datasets using single values to facilitate comparison
  - Central tendency: describes “where” data are
  - Dispersion: describes how “spread out” data are
- Many, many ways to achieve goal, each with advantages/disadvantages

# Additional Resources

- Illowsky and Dean, Chapters 2.3, 2.5, 2.7
- Caswell, Chapters 7, 8
- Weiss, Chapters 3.1 – 3.2, 3.5
- Gorard, 2004 [Brit. J. Educ. Stud.]
- Means and Medians [UW iSchool]
- Average or Central Tendency [Khan Academy]
- Measures of Dispersion [Khan Academy]