Resistance

- Outlier – An extreme data point so different from the majority that is may warrant correction or removal
- How do outliers affect our descriptive statistics?
- Resistant Statistic – Its computed value is unaffected by outliers
- Non-resistant Statistic – Its computed value is affected by outliers

Statistic	Resistant or not
Mean	Non-resistant
Trimmed Mean	Non-resistant
Median	Resistant
Mode	Resistant
Range	Non-resistant
Variance	Non-resistant
Standard Deviation	Non-resistant

- The trimmed mean is more resistant than the mean
- Consider the datasets:
  - Daily pounds of produce sold = 45, 46, 49, 50, 52, 54, 55, **56**, 58
  - Daily pounds of produce sold = 45, 46, 49, 50, 52, 54, 55, **400**, 58

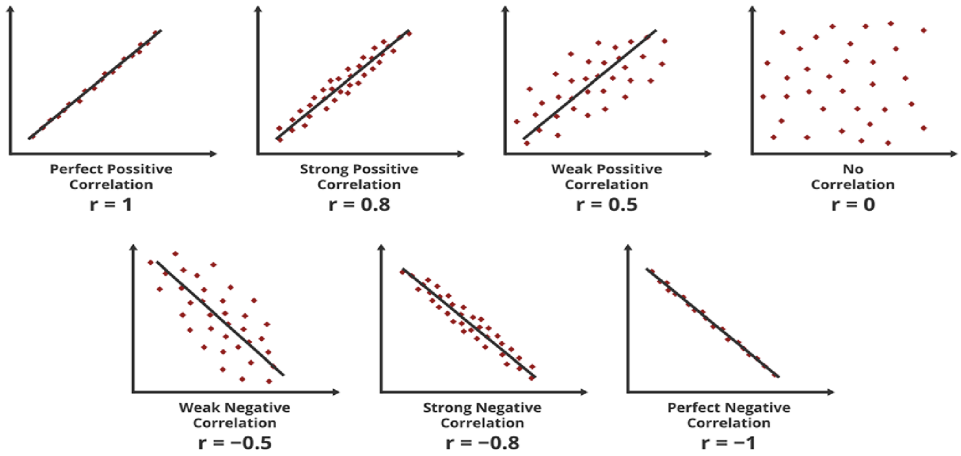
Statistic	With 56	With 400
Mean	51.4	85.8
80% Trimmed Mean	41.1	41.3
Median	51	51
Mode	49	49
Range	13	355
Variance	18.7	12204.0
Standard Deviation	4.3	110.5

Data Associations

- Association refers to the idea that there are relationships between attributes
- We will look at three primary types of relationships
  - Quantitative-by-quantitative
  - Quantitative-by-categorical
  - Categorical-by-categorical

Quantitative-by-quantitative Relationships

- Correlation – Looks at the statistical relationship of two attributes and if they change together
- Correlation coefficient (r) – Measures the strength and direction of the linear relationship between two quantitative attributes
  - Ranges from -1 to +1
  - Direction is given by (-/+)
  - Strength is given by the magnitude
    - Can be described as none, weak, moderate, strong, perfect



Pearson Correlation Coefficient

- This measures the linear correlation
- Consider the covariance:

$$\text{Covariance} = \frac{1}{N}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- This measures how two attributes change together
- Shows the sign of a linear relationship between attributes
  - Positive indicates that as one attribute increases, so does the other
  - Negative indicates that as one attribute increases, the other decreases

$$\text{Standard Deviation of } x\ (\sigma_x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{N}}$$

$$\text{Standard Deviation of } y\ (\sigma_y) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})}{N}}$$

$$\begin{aligned}\text{Pearson Correlation Coefficient } (r) &= \frac{\text{Covariance}(x,y)}{\sigma_x \sigma_y} \\ &= \frac{\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i}^{n}(y_i - \bar{y})^2}}\end{aligned}$$

- Consider the attributes:
  - A: 1, 2, 3, 4, 5
  - B: 7, 8, 9, 10, 11
  - $\bar{A} = 3$, $\bar{B} = 9$

$A_i - \bar{A}$	$B_i - \bar{B}$
-2	-2
-1	-1
0	0
1	1
2	2

$$\sum_i^n \left(A_i - \bar{A}\right)\left(B_i - \bar{B}\right) = 4 + 1 + 0 + 1 + 4 = 10$$

$$\sqrt{\sum_i^n \left(A_i - \bar{A}\right)^2} = \sqrt{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2} = \sqrt{10}$$

$$\sqrt{\sum_i^n \left(B_i - \bar{B}\right)^2} = \sqrt{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2} = \sqrt{10}$$

$$r = \frac{10}{\sqrt{10}\sqrt{10}} = 1$$

- Consider the attributes:
  - A: 1, 2, 3, 4, 5
  - B: 11, 9, 8, 10, 7
  - $\bar{A} = 3$, $\bar{B} = 9$

$A_i - \bar{A}$	$B_i - \bar{B}$
-2	2
-1	0
0	-1
1	1
2	-2

$$\sum_i^n \left(A_i - \bar{A}\right)\left(B_i - \bar{B}\right) = -4 + 0 + 0 + 1 - 4 = -7$$

$$\sqrt{\sum_i^n \left(A_i - \bar{A}\right)^2} = \sqrt{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2} = \sqrt{10}$$

$$\sqrt{\sum_i^n \left(B_i - \bar{B}\right)^2} = \sqrt{2^2 + 0^2 + (-1)^2 + 1^2 + (-2)^2} = \sqrt{10}$$

$$r = \frac{-7}{\sqrt{10}\sqrt{10}} = -0.7$$

- Consider the attributes:
  - A: 1, 2, 3, 4, 5
  - B: 33, 27, 24, 30, 21
  - $\bar{A} = 3$, $\bar{B} = 27$

$A_i - \bar{A}$	$B_i - \bar{B}$
-2	6
-1	0
0	-3
1	3
2	-6

$$\sum_i^n \left(A_i - \bar{A}\right)\left(B_i - \bar{B}\right) = -12 + 0 + 0 + 3 - 12 = -21$$

$$\sqrt{\sum_i^n \left(A_i - \bar{A}\right)^2} = \sqrt{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2} = \sqrt{10}$$

$$\sqrt{\sum_i^n \left(B_i - \bar{B}\right)^2} = \sqrt{6^2 + 0^2 + (-3)^2 + 3^2 + (-6)^2} = \sqrt{90}$$

$$r = \frac{-21}{\sqrt{10}\sqrt{90}} = -0.7$$

Introduction to the Dot Product

- When calculating $x_i - \bar{x}$ and $y_i - \bar{y}$, we created two vectors of values

$x_i - \bar{x}$	$y_i - \bar{y}$
-2	6
-1	0
0	-3
1	3
2	-6

- Let's call each vector v_x and v_y
  - $v_x = [-2, -1, 0, 1, 2]$
  - $v_y = [6, 0, -3, 3, -6]$
- Notice that the numerator $\sum_i^n (x_i - \bar{x})(y_i - \bar{y})$, is summing the product for each element in the two vectors. This summing of element-wise product is actually known as the dot product $(\cdot)$
- This numerator is calculating $v_x \cdot v_y$

Length of a vector

- So, the numerator is the dot product of vectors v_x and v_y, which represents $x_i - \bar{x}$ and $y_i - \bar{y}$, respectively
- In the denominator, $\sqrt{\sum_i^n (x_i - \bar{x})^2}$ can be written as $\sqrt{v_x \cdot v_x}$ which gives us the length of the vector x. This can also be written as $||v_x||$
- Similarly, $\sqrt{\sum_i^n (y_i - \bar{y})^2} = \sqrt{v_y \cdot v_y} = ||v_y||$

Cosine Similarity

- Putting all the pieces back together, we get

$$\frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}} = \frac{v_x \cdot v_y}{||v_x||\ ||v_y||}$$

- This is known as the cosine similarity
- The value ranges from -1 (exactly opposite) to 1 (the exact same)
- 0 indicates orthogonal vectors

Quantitative-by-categorical Relationships

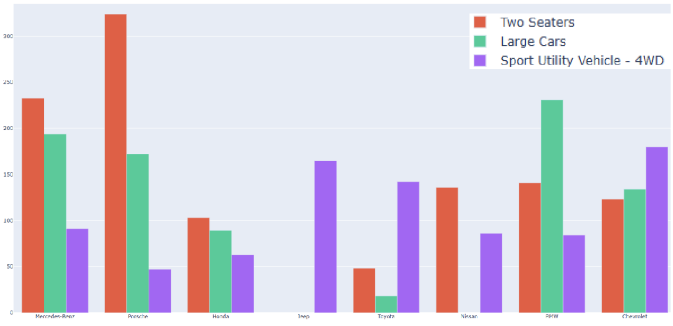
- We don't include new unique measures to compute
- We look at how a quantitative attribute varies among different values of a categorical attribute
- We can think of the calculations being conditioned on the value of the categorical attribute

Province	Mean Price	Median Price	Standard Deviation of Price
Nova Scotia	20,123	19,234	12,223
New Brunswick	19,231	18,230	12,330
Prince Edward Island	19,331	19,233	11,233
Newfoundland and Labrador	18,302	19,223	13,223

Categorical-by-categorical Relationships

- We don't include new unique measures to compute
- Looking to compare the distribution of values for one categorical attribute against values of another categorical attribute

	Two seaters	Large cars	SUV - 4WD
Mercedes-Benz	233	194	91
Porsche	324	172	47
Honda	103	89	63
Jeep	0	0	165
⋮	⋮	⋮	⋮



- This could answer questions like "How do the proportions of two seaters created vary by manufacturer?"