Lecture 3

Understanding asymmetry and variability in the data, Skewness, Correlation.

July 31

Content

- Understanding asymmetry and variability in the data
- Skewness
- Correlation

Estimates of Location - **Mean**

- Also known as the '**average**'.
- The sum of all values divided by the number of values.
- Denoted by greek letter 'mu' for the population
  - And x-bar for a sample out of the population
  - Can be calculated as
  - or

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

$$\frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

Estimates of Location - **Mean** - 2

- Very popular as easy to understand.
- But easily affected by ‘outliers’.
- **Outlier:** A data value that is very different from most of the data.
    - Synonyms: extreme value

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50

Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

Estimates of Location - **Median**

- Also known as the '**50th percentile** '.
- The middle number in the dataset.
- The value such that one-half of the data lies above and below.
- How to calculate this:
  - Arrange data in descending order.
  - Median = the number at position (n+1)/2 in the ordered list.

$$x_{(\frac{N+1}{2})}$$

Estimates of Location - **Median** - 2

- Not affected by extreme prices.
- **Robust:** Not sensitive to extreme values.
  - Synonyms: resistant

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50
Median	$ 6.00	$ 5.50

Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

Estimates of Location - **Mode**

- The value that occurs the most.

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50
Median	$ 6.00	$ 5.50
Mode	$ 3.00	-

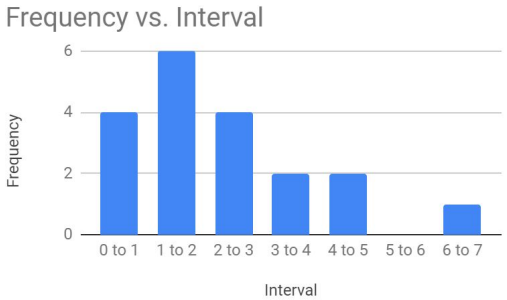
Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

Estimates of Asymmetry

- Skewness: Tells us if the data is concentrated on one side or not.

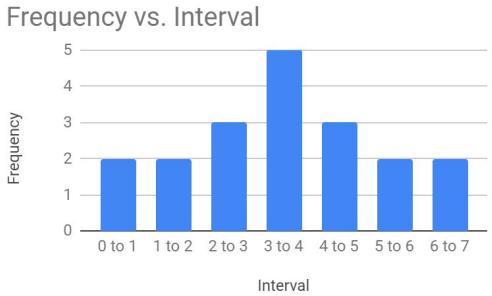
Dataset 1
1
1
1
1
2
2
2
2
2
2
3
3
3
3
4
4
5
5
7

Interval	Frequency	
0 to 1	4	
1 to 2	6	
2 to 3	4	
3 to 4	2	
4 to 5	2	
5 to 6	0	
6 to 7	1	
Mean	**Median**	**Mode**
2.79	2.00	2.00



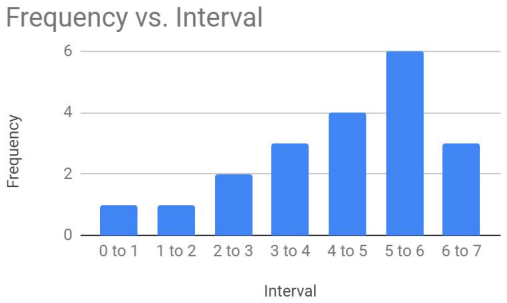
Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency	
0 to 1	2	
1 to 2	2	
2 to 3	3	
3 to 4	5	
4 to 5	3	
5 to 6	2	
6 to 7	2	
Mean	**Median**	**Mode**
4.00	4.00	4.00



Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
6
7
7
7

Interval	Frequency	
0 to 1	1	
1 to 2	1	
2 to 3	2	
3 to 4	3	
4 to 5	4	
5 to 6	6	
6 to 7	3	
Mean	**Median**	**Mode**
4.90	5.00	6.00



Estimates of Variability

- Variance
- Standard Deviation
- Coefficient of Variation

Estimates of Variability - **Variance**

- Sample statistic is an approximation of the population parameter.
- 10 different samples will get you 10 different measures.

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

Sample Variance $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n-1}$$

Estimates of Variability - **Variance** -2

Population			
1		**Mean**	3.00
2		**Population variance**	2.00
3		**Sample variance**	2.50
4			
5			

observation mean

$$\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}$$

Population variance formula

Sample Variance

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{4}$$

Estimates of Variability - **Variance** -3

Population			
1		**Mean**	3.00
2		**Population variance**	2.00
3		**Sample variance**	2.50
4			
5			

Imaginary population			
1		**Mean**	3.20
1		**Population variance**	2.96
1			
2			
3			
4			
5			
5			
5			
5			

Estimates of Variability - **SD**

- Variance can have a very high value.
- SD more meaningful.

Population SD

$$\sigma_s = \sqrt{\sigma^2}$$

Sample SD

$$s_s = \sqrt{s^2}$$

Estimates of Variability - **Coeff of variation**

- Also known as relative standard deviation

$$\frac{SD}{Mean}$$

Population CV

$$c_v = \frac{\sigma}{\mu}$$

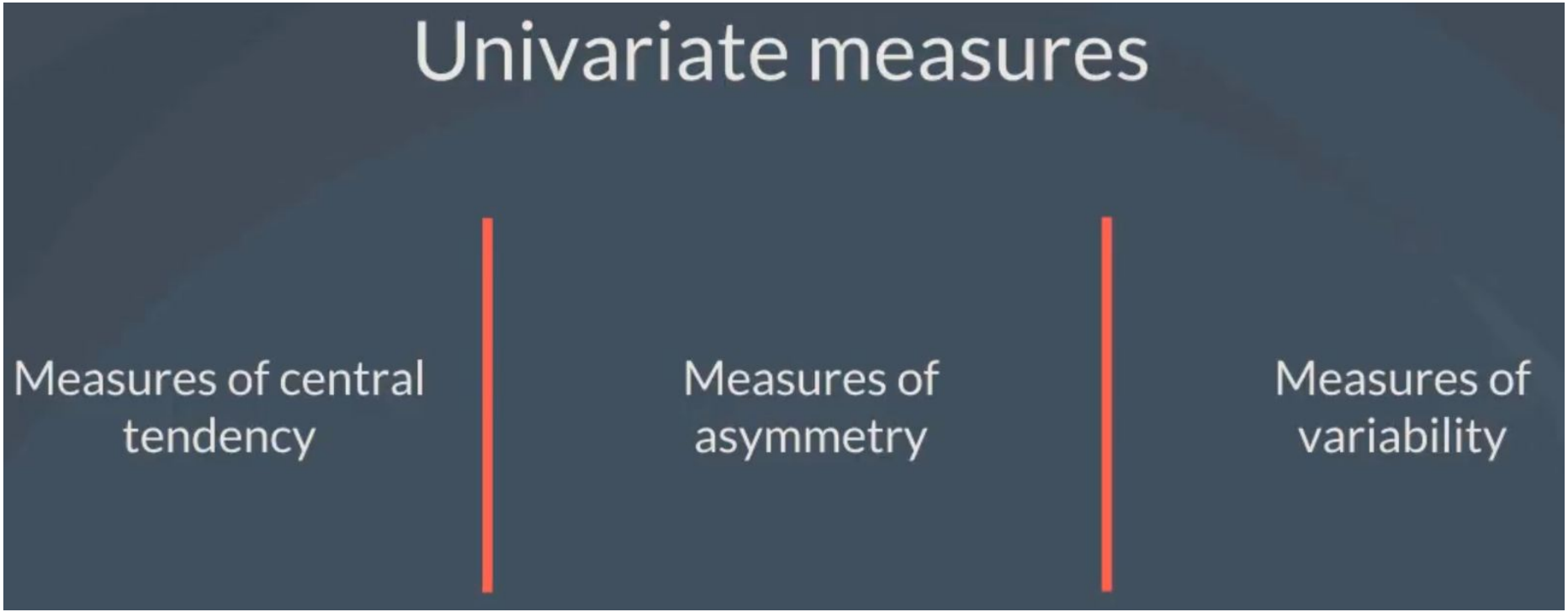
Sample CV

$$\hat{c}_v = \frac{s}{\overline{x}}$$

Estimates of Variability - **Coeff of variation** -2

NY Dollars	Pesos
$ 1.00	MXN 18.81
$ 2.00	MXN 37.62
$ 3.00	MXN 56.43
$ 3.00	MXN 56.43
$ 5.00	MXN 94.05
$ 6.00	MXN 112.86
$ 7.00	MXN 131.67
$ 8.00	MXN 150.48
$ 9.00	MXN 169.29
$ 11.00	MXN 206.91

	Dollars	Pesos
Mean	$ 5.50	MXN 103.46
Sample variance	$2 10.72	MXN2 3793.69
Sample standard deviation	$ 3.27	MXN 61.59
Sample coefficient of variation	0.60	0.60


Univariate measures
Measures of central tendency
Measures of asymmetry
Measures of variability

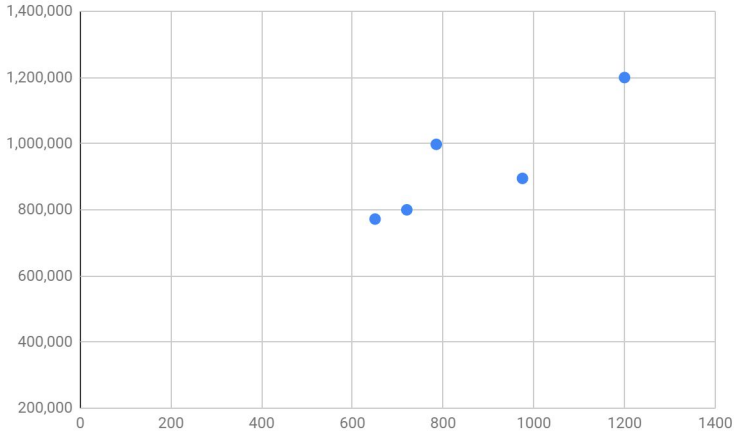
Measure of Relationship

- Covariance
- Linear correlation coefficient

Measure of Relationship - Covariance

- Let's look at houses data

Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000



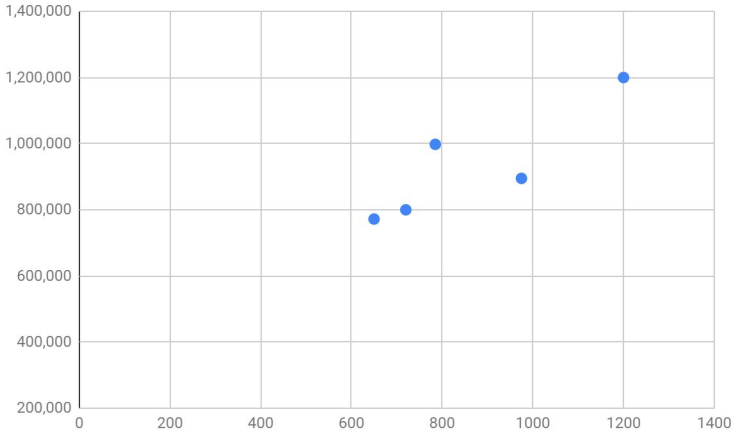
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})}{n - 1}$$

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$

Measure of Relationship - Covariance -2

	$(x-\bar{x})*(y-\bar{y})$
	34,776,000
	- 5,265,000
	89,178,000
	19,418,000
	- 4,142,000
Sum	133,965,000
Sample size	5
Cov. Sample	33,491,250

	Size (ft.)	Price ($)
	650	772,000
	785	998,000
	1200	1,200,000
	720	800,000
	975	895,000
Mean	866	933,000



$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})}{n-1}$$

Measure of Relationship -

- Similar issue as before.
  - Covariance can be massively high positively or negatively.
  - How to get a measure of value?
  - Solution: Correlation coefficient

- Correlation adjust covariance, so that the relationship between the variables is more easy to interpret.

Measure of Relationship -

- Similar issue as before.
  - Covariance can be massively high positively or negatively.
  - How to get a measure of value?
  - Solution: Correlation coefficient

- Correlation adjust covariance, so that the relationship between the variables is more easy to interpret.

$$\frac{Cov(x, y)}{StDev(x) * StDev(y)}$$

Measure of Relationship - Correlation coefficient

- Similar issue as before.
  - Covariance can be massively high positively or negatively.
  - How to get a measure of value?
  - Solution: Correlation coefficient

- Correlation adjust covariance, so that the relationship between the variables is more easy to interpret.

$$\frac{Cov(x,y)}{StDev(x) * StDev(y)}$$

Measure of Relationship - Correlation coefficient

	$(x-\bar{x})*(y-\bar{y})$
	34,776,000
	- 5,265,000
	89,178,000
	19,418,000
	- 4,142,000
Sum	133,965,000
Sample size	5
Cov. Sample	33,491,250
Correlation coeff.	0.87

	Size (ft.)	Price ($)
	650	772,000
	785	998,000
	1200	1,200,000
	720	800,000
	975	895,000
Mean	866	933,000