Lecture 24

Produced by Dr. Worldwide

Welcome to the 305

Descriptive Statistics

- Collection of data points is called a sample, and we interpret it as a subset of observations from some underlying random phenomenon

- We denote the *i*th point in the sample as X_i

- We denote the whole set of observations as $\{X_1, X_2, \cdots, X_n\}$

- To measure the center of the data, we compute three quantities
  - Sample mean: the average value of our observations

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - Sample median: the value that divides the bottom 50% by the top 50%
  - Mode: the most frequently occurring value (discrete or categorical only)

Descriptive Statistics

- Q: Why calculate the sample mean and sample median?

- To measure the spread of the data, we compute three quantities
  - Sample variance: the average squared distance between an observation and the sample mean

$$S_X^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}(n))^2$$

  - Sample standard deviation: more convenient than the sample variance

$$S_X(n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}(n))^2}$$

  - Range: the difference between the largest value and smallest value

Descriptive Statistics

- Percentiles are also helpful
  - The *k*th percentile is a value that divides the bottom k% from the top (1-k)%
  - The median is the 50th percentile
  - The 25th and 75th percentiles (Q1 and Q3) are useful for understanding the variability in the middle of the distribution
  - The interquartile range (IQR) is the difference between Q3 and Q1

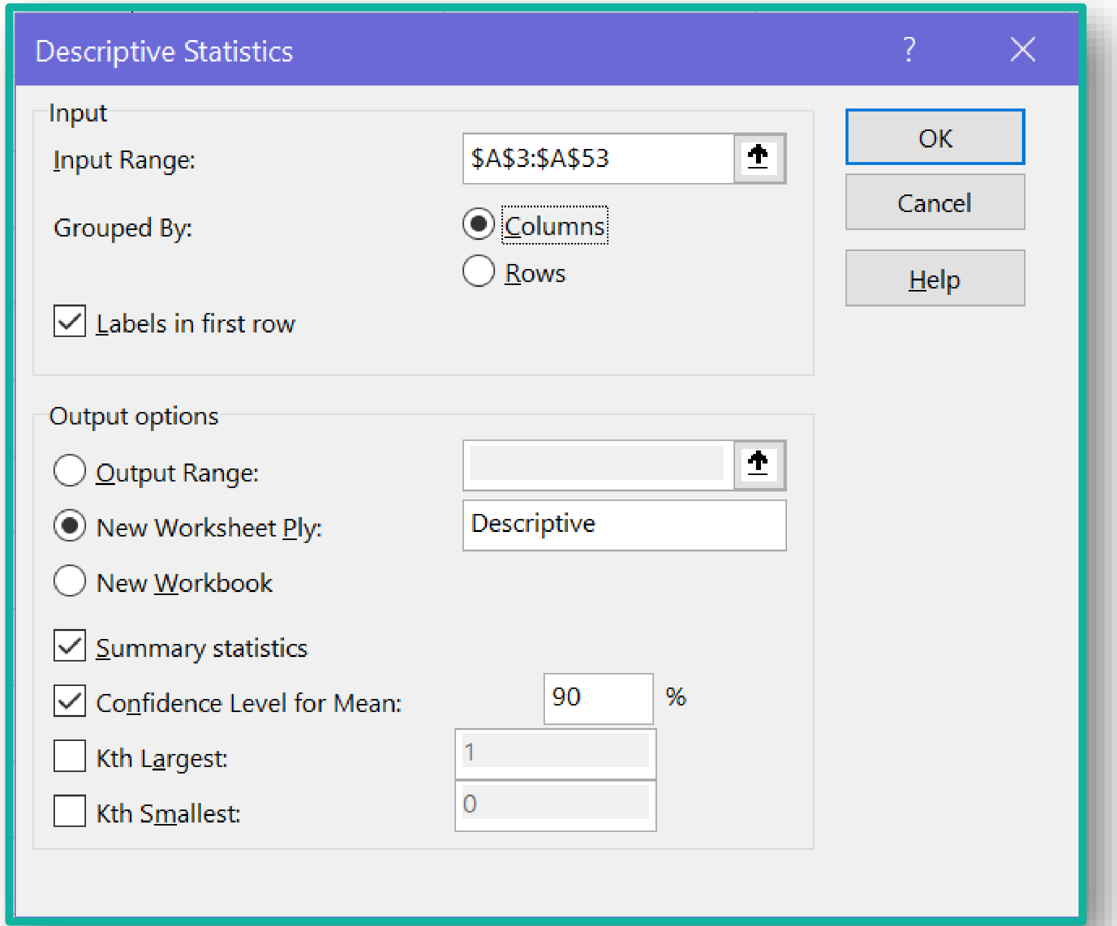
Ex: Starting Salaries

- Download Salaries-2.xlsx from link Sheet 1 on course website
- Analyze the formulas for these statistics in the Frequency sheet

Sample Mean	89.772812
Sample Median	89.73195
Sample Variance	38.9904019
Sample SD	6.24422949
Min	78.4019
Max	105.5129
Range	27.111
Q1	85.2254
Q3	93.963575
IQR	8.738175

Ex: Starting Salaries

- More information can be gathered using Data Analysis in the Data menu



Ex: Starting Salaries

- Results from the Analysis ToolPak

	A	B
1	*Salaries*	
2		
3	Mean	89.772812
4	Standard Error	0.883067403
5	Median	89.73195
6	Mode	#N/A
7	Standard Deviat	6.244229491
8	Sample Varianc	38.99040194
9	Kurtosis	-0.155889833
10	Skewness	0.417303487
11	Range	27.111
12	Minimum	78.4019
13	Maximum	105.5129
14	Sum	4488.6406
15	Count	50
16	Confidence Leve	1.774590386

Ex: Sum of Dice

- Consider the random experiment of tossing 2 identical 6-sided fair dice and collecting the outcome of their sum
- We call the values of the first die toss are $\{Y_1, Y_2, \cdots, Y_n\}$
- We call the values of the second die toss are $\{W_1, W_2, \cdots, W_n\}$
- Create random variable $X_i = Y_i + W_i$ where $i \in \{1,2, \cdots, n\}$
- Q: What are the possible values of X?
- Q: What is the most likely value of X?
- Q: What is the probability $P(X = 2)$?

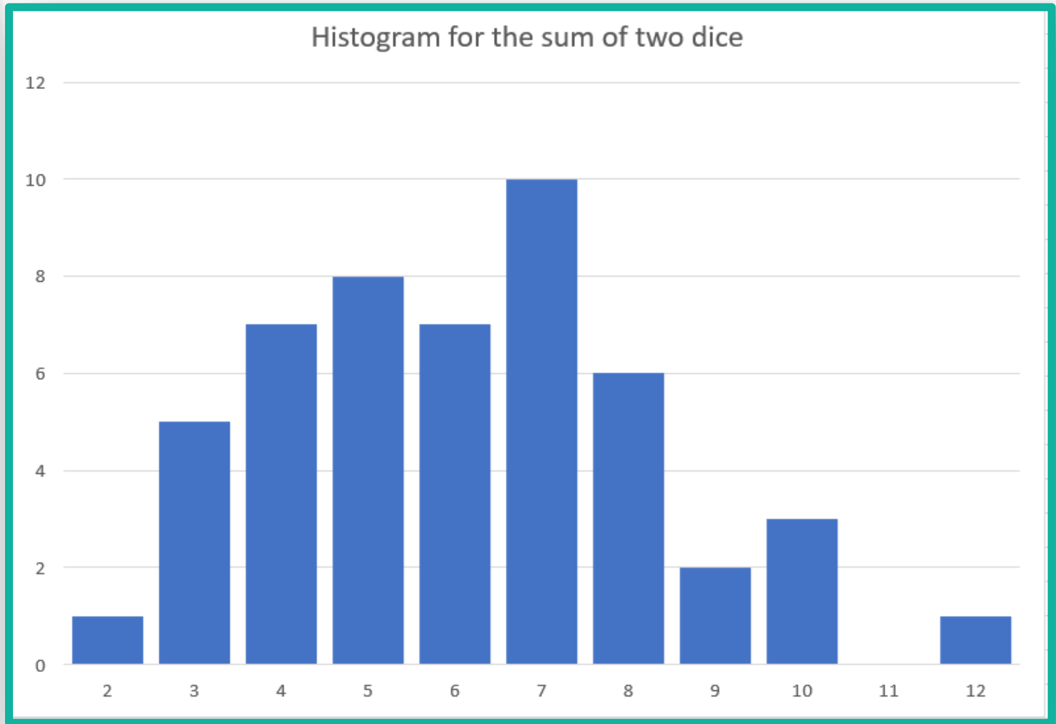
Ex: Sum of Dice

- Download SumDice.xlsx from link Sheet 2 on course website
- The tab named "50" contains 50 repetitions of this experiment
- Observations from both dice are contained in A4:A53 and B4:B53
- The values of X are contained in C4:C53
- The table in F4:H14 contains
  - Possible values for the sum of 2 dice
  - Frequency for each of the possible values
  - Relative frequency for each of the possible values
- Q: How is the relative frequency more useful than the frequency?

Ex: Sum of Dice

- Q: How well do you think this sample represents the population?

Bin	*Frequency*	*Relative Frequency*
2	1	0.02
3	5	0.1
4	7	0.14
5	8	0.16
6	7	0.14
7	10	0.2
8	6	0.12
9	2	0.04
10	3	0.06
11	0	0
12	1	0.02
Sample mean		6.1
Sample variance		4.744897959



Ex: Sum of Dice

- Update tabs “100” and “200” with frequency tables and histograms
- Q: How does the number of observations from our experiment effect the results?
- As we sample more from a population, the characteristics in the sample start matching the characteristics of the population
  - Statistic → Parameter
  - $E[X]$: $\bar{x} \to \mu$
  - $Var[X]$: $s^2 \to \sigma^2$
- There is always error between a sample and a population, but that error is removed as we increase our sample size

Descriptive Statistics

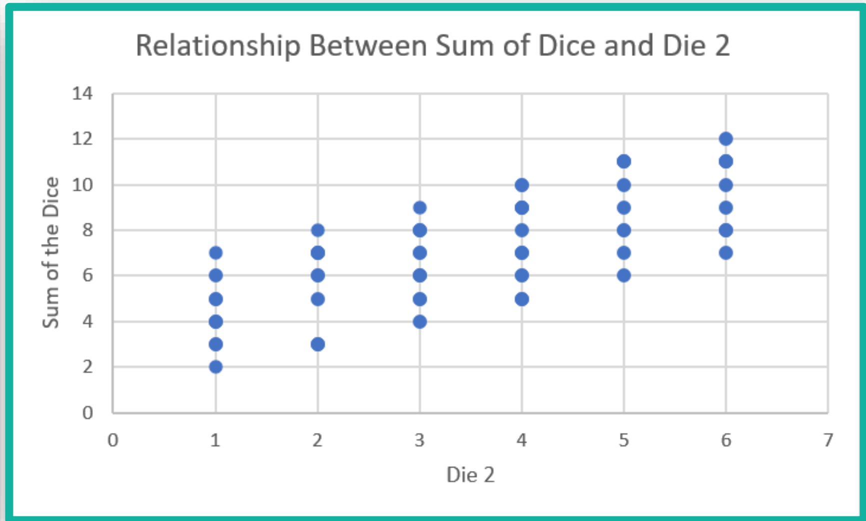
- Current methods are appropriate for univariate data
- Bivariate data contains observations from a pair of variables
- For bivariate data, the focus shifts to understanding the relationship between the two variables
- Descriptive statistics for bivariate data
  - Scatterplot
  - Covariance
  - Correlation
- Since the most widely used method for modeling relationships is linear regression, the scatterplot is often use to inspect if a linear relationship exists

Ex: Sum of Dice

- Q: What kind of relationship exists between the outcomes of the two dice?
- Q: What kind of relationship exists between the outcome of the second die and the sum of the two dice?
- In Excel, create a scatterplot by using the Insert menu
- Optionally, use Recommended Charts to help you select Scatter
- Take a moment to create scatterplots to capture both relationships on the tab named "200"
- Examine plots in tab named "50" and "100" for examples
- Investigate the plots to determine if your hypotheses were true

Ex: Sum of Dice

- Plots based on 100 observations from the population

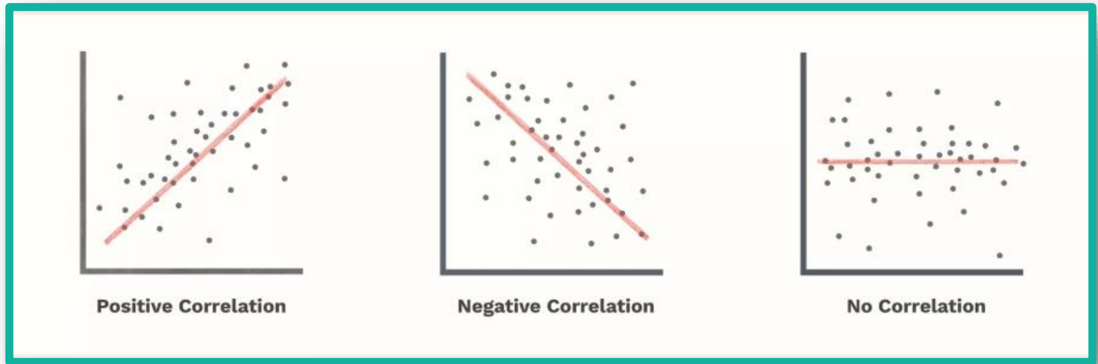


- Q: How would we quantify the difference between these relationships?

Descriptive Statistics

- The sample correlation coefficient measures the strength of linear relationship between two variables on a scale between -1 and +1
  - Close to 1 implies strong positive correlation
  - Close to -1 implies strong negative correlation
  - Close to 0 indicates no correlation
  - Formula

$$r_{X,Y} = \frac{1}{S_X(n)S_Y(n)} \sum_{i=1}^{n} (X_i - \overline{X}(n))(Y_i - \overline{Y}(n))$$



Ex: Sum of Dice

- Calculation of correlation using CORREL(variable 1, variable 2)
  - When n=50

Sample correlation (W,X)	0.675188473
Sample correlation (Y,W)	-0.110785404

  - When n=100

Sample correlation (W,X)	0.73263405
Sample correlation (Y,W)	0.06826615

  - When n=200

Sample correlation (W,X)	0.67376759
Sample correlation (Y,W)	-0.0530525

The End

Dale