



Lecture 25

Produced by Dr. Worldwide

Welcome to the 305

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

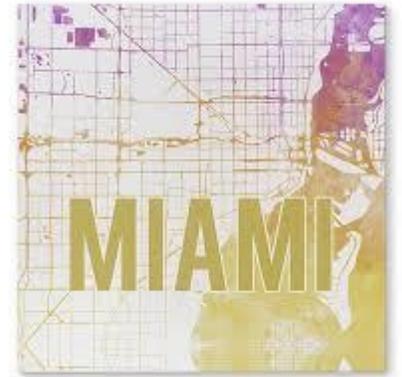
Ex: Sum of Dice



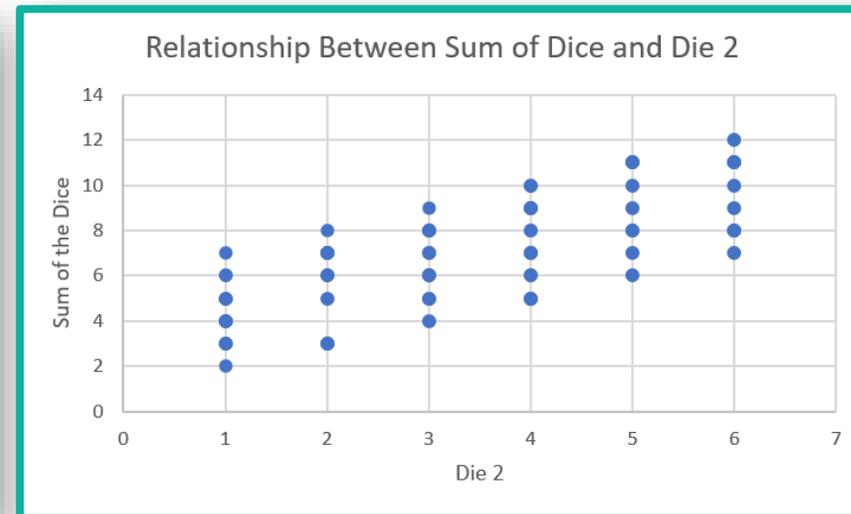
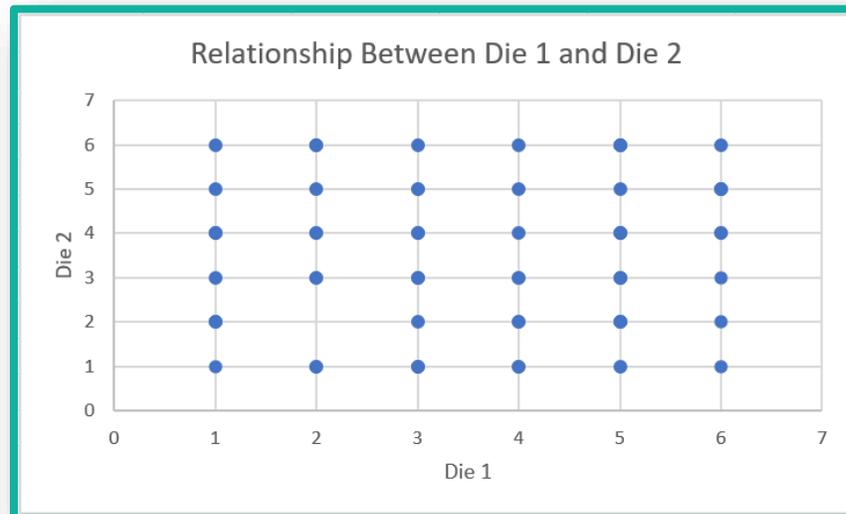
- Q: What kind of relationship exists between the outcomes of the two dice?
- Q: What kind of relationship exists between the outcome of the second die and the sum of the two dice?
- Download [SumDice-2.xlsx](#) from link [Sheet 1](#) on course website
- In Excel, create a scatterplot by using the [Insert](#) menu
- Optionally, use [Recommended Charts](#) to help you select [Scatter](#)
- Examine plots in tab named "50" and "100" for examples
- Investigate the plots to determine if your hypotheses were true



Ex: Sum of Dice

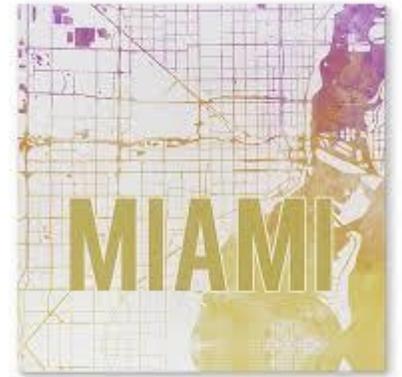


- Plots based on 100 observations from the population



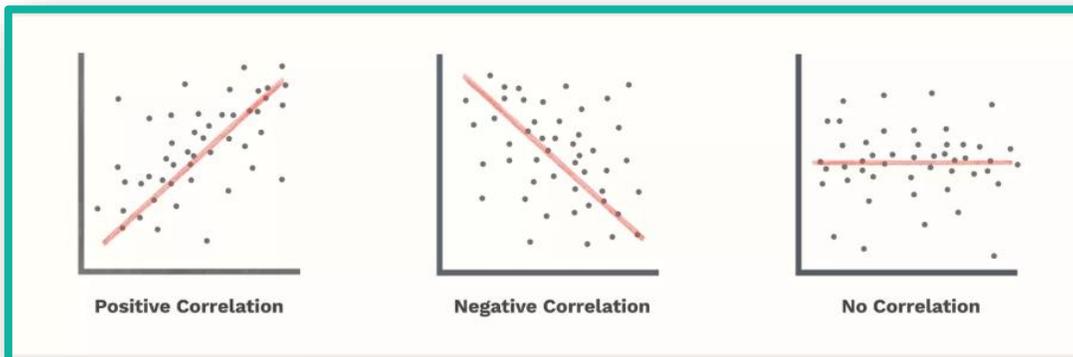
- Q: How would we quantify the difference between these relationships?

Descriptive Statistics

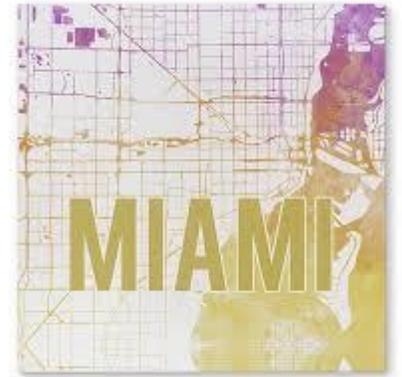


- The **sample correlation coefficient** measures the strength of **linear** relationship between two variables on a scale between -1 and +1
 - Close to 1 implies strong positive correlation
 - Close to -1 implies strong negative correlation
 - Close to 0 indicates no correlation
 - Formula

$$r = \frac{1}{n} \sum \left(\frac{x_i - \overline{x(n)}}{s_x} \right) \left(\frac{y_i - \overline{y(n)}}{s_y} \right)$$



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 $n=200$

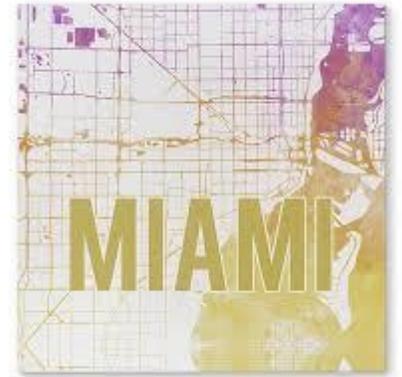
Sample correlation (W,X)	0.67376759
Sample correlation (Y,W)	-0.0530525

Descriptive Statistics

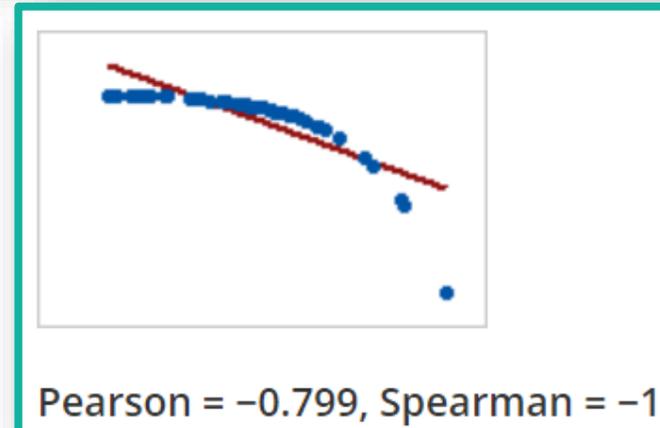
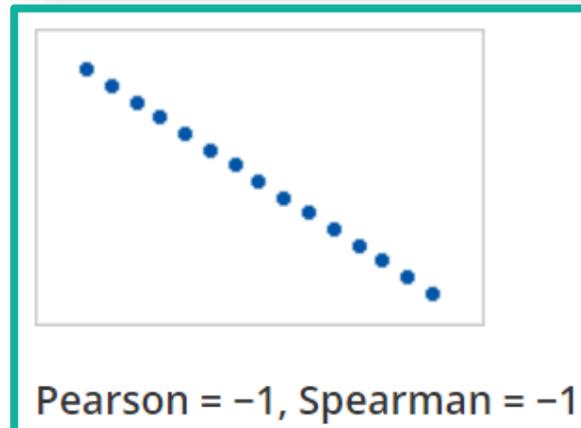
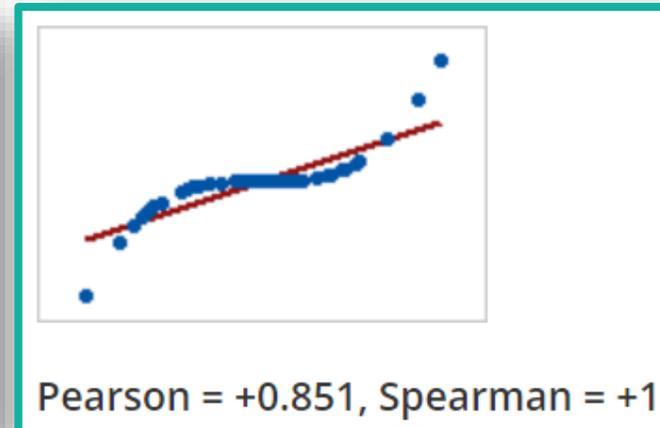
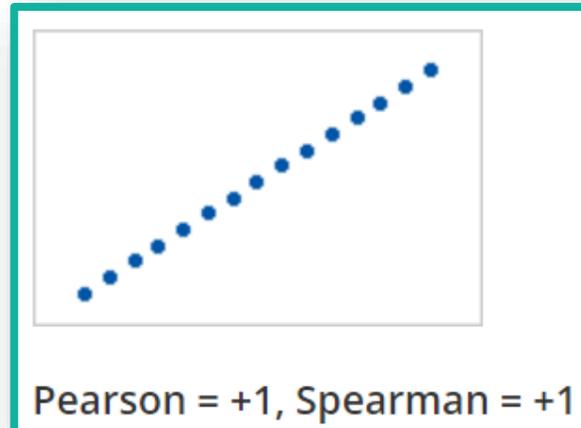


- Previous statistic is often called Pearson's correlation coefficient
- Assumptions for Pearson's correlation coefficient
 - Both variables are normally distributed (approximately bell-shaped)
 - Relationship can be expressed by a line
 - Data is equally distributed around the best-fitted line
- **Spearman's correlation coefficient** is the nonparametric version of the latter and evaluates the **monotonic** relationship between the ranked values
- **Monotonic** implies that variables change together but not at a constant rate
- Both correlation coefficients are between -1 and 1

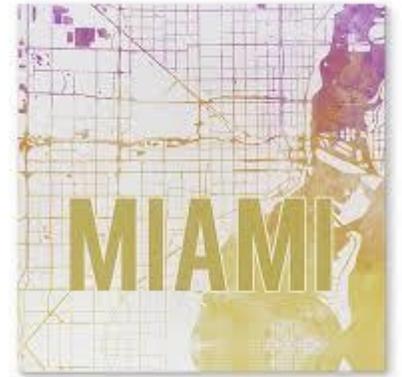
Descriptive Statistics



- Visual difference between Pearson and Spearman

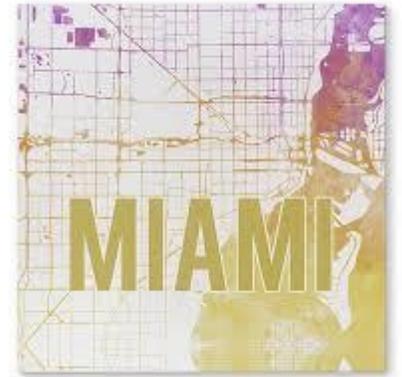


Descriptive Statistics

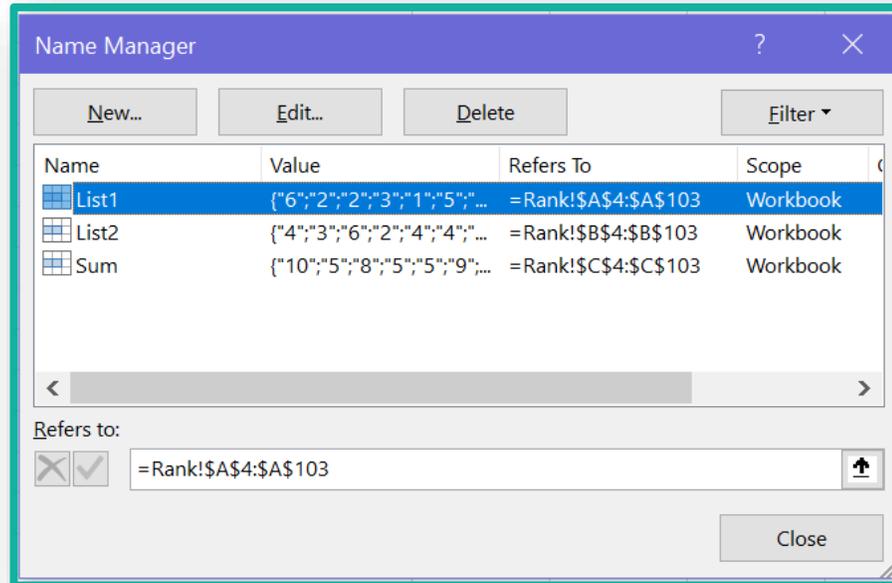


- Advantage of Spearman's is that it can be applied to discrete numeric and ordinal categorical data
- Formulation is based on ranking the observations for each of the variables and computing the Pearson correlation coefficient for the ranks
- When ranking, we handle ties by computing the average
=RANK.AVG(observation, variable, o=descending)
- Evaluation of Spearman's correlation coefficient
=CORREL(RANK.AVG(variable 1),RANK.AVG(variable 2))

Ex: Sum of Dice



- Calculation of Spearman's correlation in tab named "Rank"
 - Create variables for first die roll, second die roll, and sum of dice



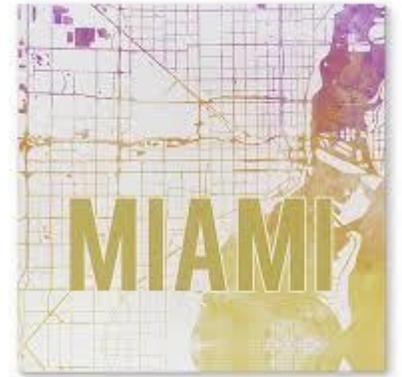
- Create columns of ranks using RANK.AVG
- Use CORREL function on ranked columns

Spearman's Correlation (W,X)	0.730623
Spearman's Correlation (Y,W)	0.070636

PivotTables



- The **PivotTable** is a powerful tool to calculate, summarize, and analyze data
- The purpose is to organize and summarize the data in a way that can be used to answer questions or visualize patterns
- Two tutorials provided on the course website
 - Tutorial from **Excel Easy** is found in **Link 1** on course website
 - Tutorial from **Microsoft Support** is found in **Link 2** on course website
- Many YouTube videos in addition to these two tutorials
- Companies use Excel's PivotTables as their main tool for summarizing data making competency in this area extremely marketable



Ex: Cancer Research

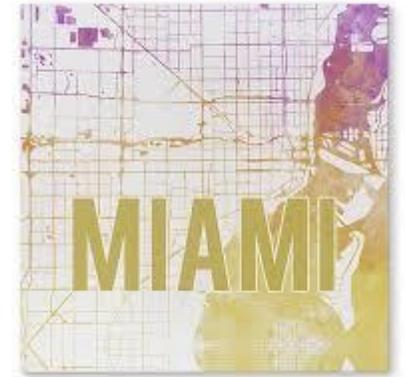
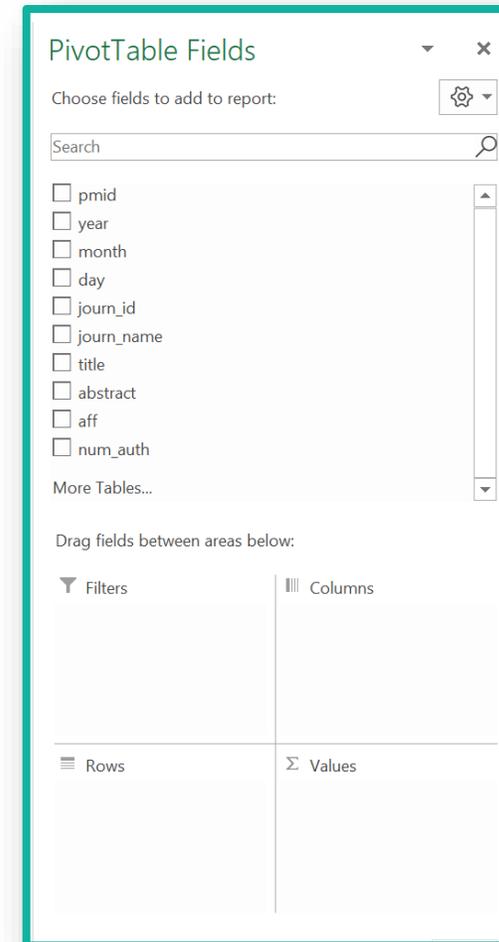


- Download [CancerResearch.xlsx](#) from link [Sheet 2](#) on course website
- Results form a PubMed search on the topic “Non-small lung cancer”
- Dataset contains 10 fields
 - Article ID number (*pmid*)
 - Year of publication (*year*)
 - Month of publication (*month*)
 - Day of publication (*day*)
 - Journal ID number (*journ_id*)
 - Journal title (*journ_name*)
 - Article title (*title*)
 - Article abstract (*abstract*)
 - Author’s affiliations (*aff*)
 - Number of authors (*num_auth*)

Ex: Cancer Research



- Select all data: Select cell A1 (top-left) and use the shortcut Ctrl+Shift+Down+Right to automatically select all the data
- When selecting the data include column names in selection
- Go to **Insert** menu to find **PivotTable** in the far left
- By default, this operation will generate a new tab
- Menu bar is used to customize the PivotTable



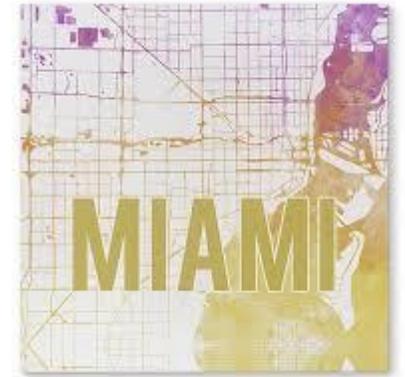
Ex: Cancer Research



- Aspects of the menu bar
 - **PivotTable Fields:** Box containing all the variables from selection
 - **Filters:** Box where you can select fields to filter the rows
 - **Columns:** Field(s) used to define the columns of the table
 - **Rows:** Field(s) used to define the rows of the table
 - **Values:** Type of summary statistic that the table should display
- Possible summary statistics
 - Sum
 - Count
 - Average
 - Max
 - Min
 - Product
 - StdDev
 - Var

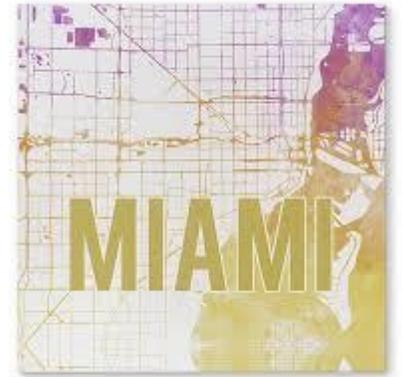


Ex: Cancer Research



- Q: When was the research on non-small cell lung cancer most active?
- Q: What journals published more papers on that topic?
- Q: What institutions have conducted the most research in this area?
- Q: What was the average number of authors for each journal?
- Q: What other questions could we explore in this dataset?
- Go to tab named "Pivot Table" and play around with the example or create your own tab and start from scratch

PivotCharts



- PivotCharts are visual representations of the PivotTable
- Create **PivotCharts** through the **Insert** menu after selecting data
- Different options in menu bar
 - **PivotChart Fields**: Box containing all the variables from selection
 - **Filters**: A box where you can select fields to filter the axis labels
 - **Legend(Series)**: Field(s) that will be used to create legends
 - **Axis (Categories)**: Field(s) used to define axis labels
 - **Values**: Type of summary statistic that the chart should summarize
- The default PivotChart is a barplot but many other options exist
- When chart is selected, go through the **Design** menu to find **Change Chart Type**





The End



Dale

