Introduction to Data Analytics

ITE 5201

Lecture5-Data Visualization-2

Instructor: Parisa Pouladzadeh
Email: parisa.pouladzadeh@humber.ca

Frequency Distributions

- After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.

- This is the goal of descriptive statistical techniques.

- One method for simplifying and organizing data is to construct a **frequency distribution.**

Frequency Distributions

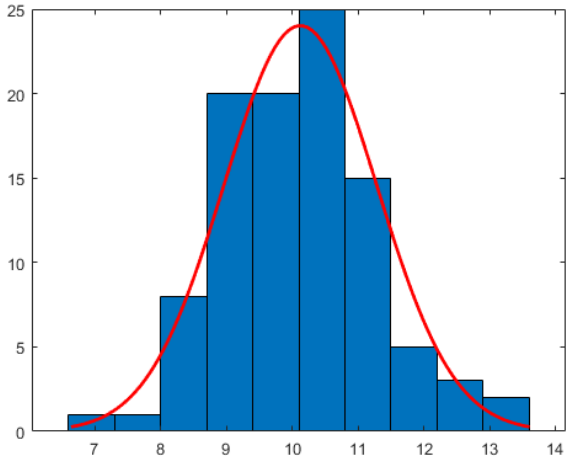
- A frequency distribution is an organized tabulation showing exactly how many individuals are located in each category on the scale of measurement.

- A frequency distribution presents an organized picture of the entire set of scores, and it shows where each individual is located relative to others in the distribution.

- In a frequency distribution graph, the score categories (X values) are listed on the X axis and the frequencies are listed on the Y axis.

Histograms

- A histogram shows a variable's distribution as a set of adjacent rectangles on a data chart.
- A graphical display of data using bars of different heights.
- It is similar to a Bar Chart, but a histogram groups numbers into ranges.
- Histograms represent counts of data within a numerical range of values.



Histograms graph

- Frequency distribution graphs are useful because they show the entire set of scores.
- At a glance, you can determine the highest score, the lowest score, and where the scores are centered.
- The graph also shows whether the scores are clustered together or scattered over a wide range.

Smooth curve

- If the scores in the population are measured on an interval or ratio scale, it is customary to present the distribution as a smooth curve.

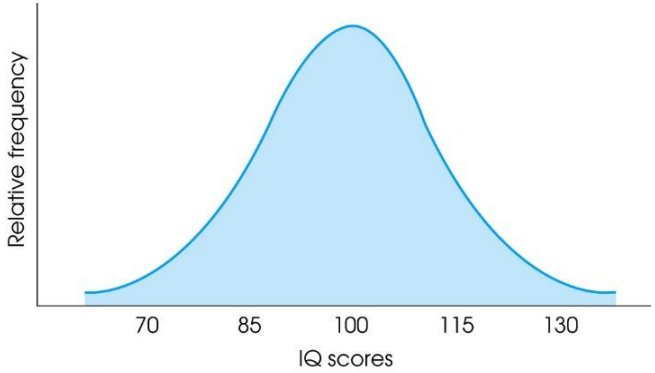
- The smooth curve emphasizes the fact that the distribution is not showing the exact frequency for each category.

Histograms and Bar graph

- A histogram represents the frequency distribution of continuous variables.
- A bar graph is a diagrammatic comparison of discrete variables.
- Histogram presents numerical data whereas bar graph shows categorical data.
- The histogram is drawn in such a way that there is no gap between the bars.

Shape

- A graph shows the shape of the distribution.
- A distribution is symmetrical if the left side of the graph is (roughly) a mirror image of the right side.



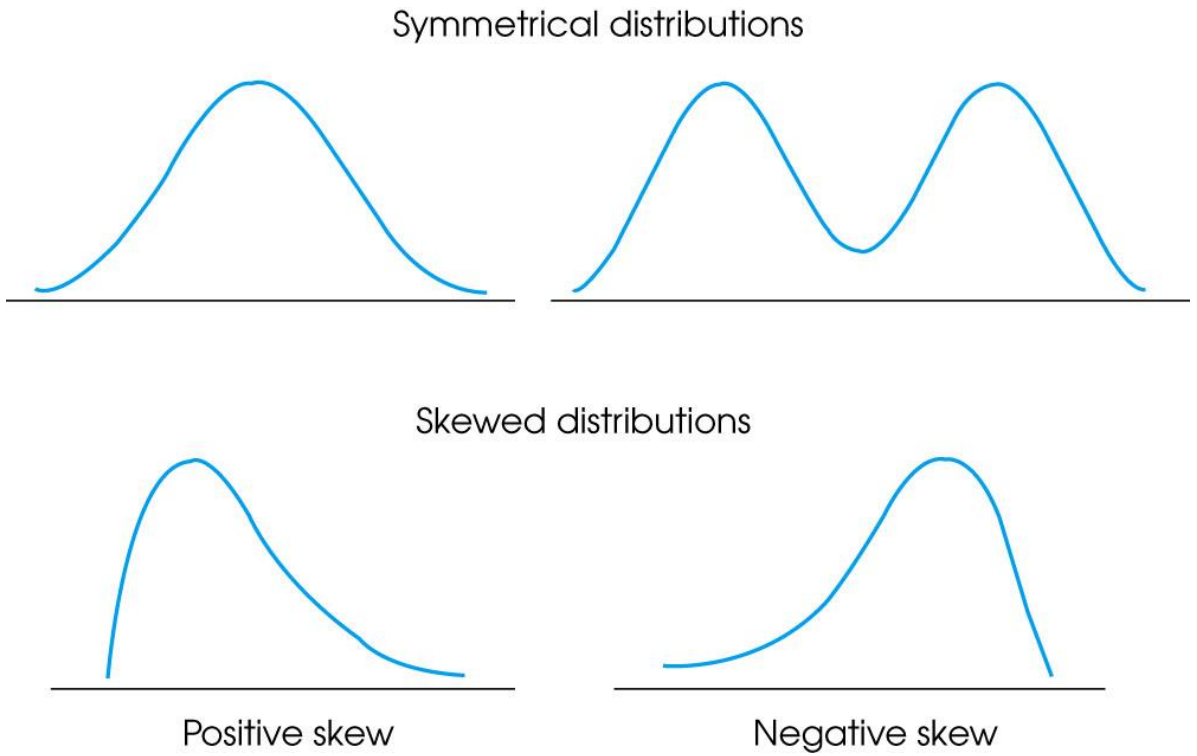
- On the other hand, distributions are skewed when scores pile up on one side of the distribution, leaving a "tail" of a few extreme values on the other side.

Positively, Negatively Skewed Distributions

- In a **positively skewed** distribution, the scores tend to pile up on the left side of the distribution with the tail points to the right.

- In a **negatively skewed** distribution, the scores tend to pile up on the right side and the tail points to the left.

Positively, Negatively Skewed Distributions



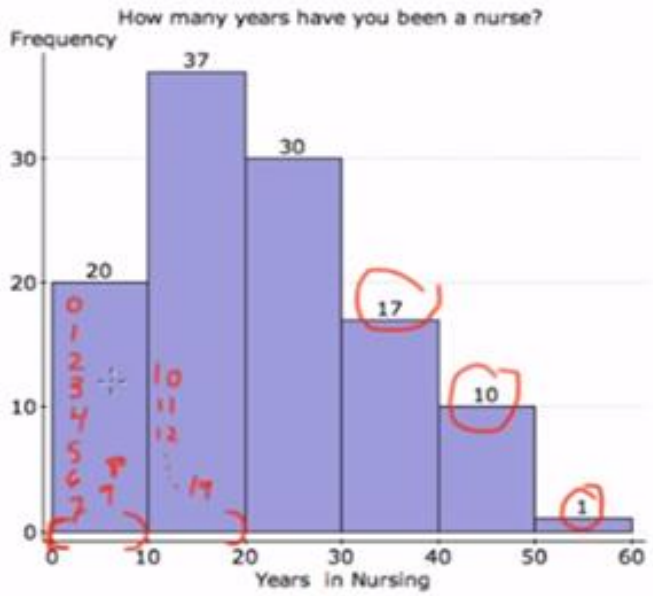
Example

How many classes are there?

What is the class width?

What is the upper and lower limit for the first class?

How many nurses were surveyed?



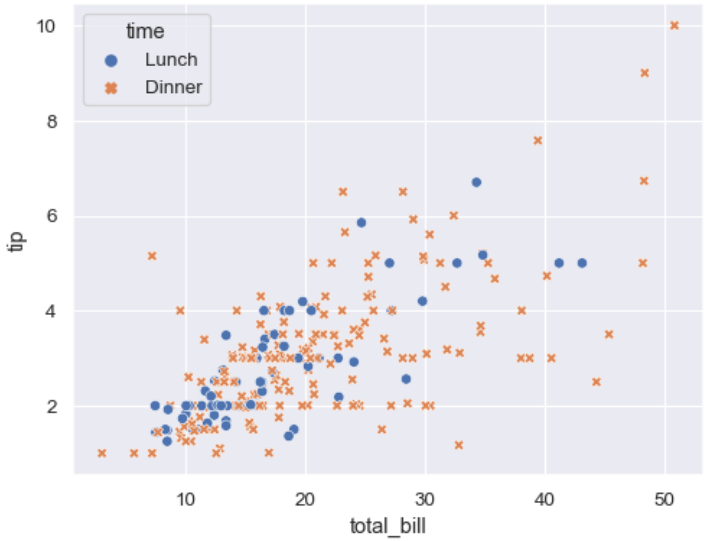
What percentage of the nurses have worked longer than 30 years?

Statistical plots

- Statistical plots allows viewers to :
  - Identify outliers
  - Visual distributions
  - Deduce variable types
  - Discover relationships and core relations between variables in a dataset.

Scatter plots

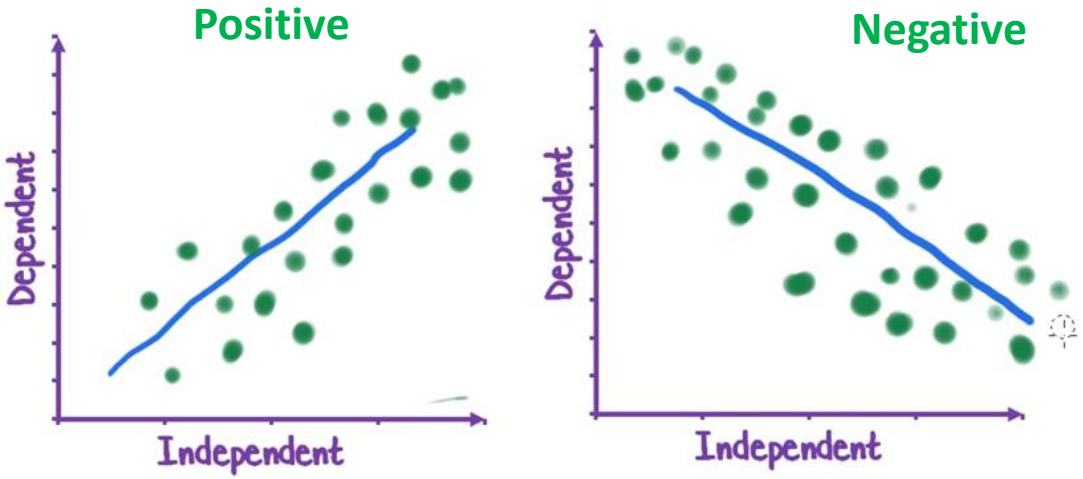
- Scatter plots are useful when you want to explore interrelations or dependencies between two different variables.
- These data graphics are ideal for visually spotting outliers and trends in data.
- Scatter plots are used when you want to show the relationship between two variables.
- Scatter plots are sometimes called correlation plots because they show how two variables are correlated.



Scatter plot

- Direction:
  - Positive: as one variable increases so does the other. Height and shoe size are an example; as one's height increases so does the shoe size.
  - Negative: as one variable increases, the other decreases. Time spent studying and time spent on video games are negatively correlated; as your time studying increases, time spent on video games decreases.
  - No correlation: there is no apparent relationship between the variables. Video game scores and shoe size appear to have no correlation; as one increases, the other one is not affected.
- Form:
  - Linear
  - Nonlinear
- Strength:
  - Weak
  - Moderate
  - Strong

Direction


Positive
Dependent
Independent
Negative
Dependent
Independent

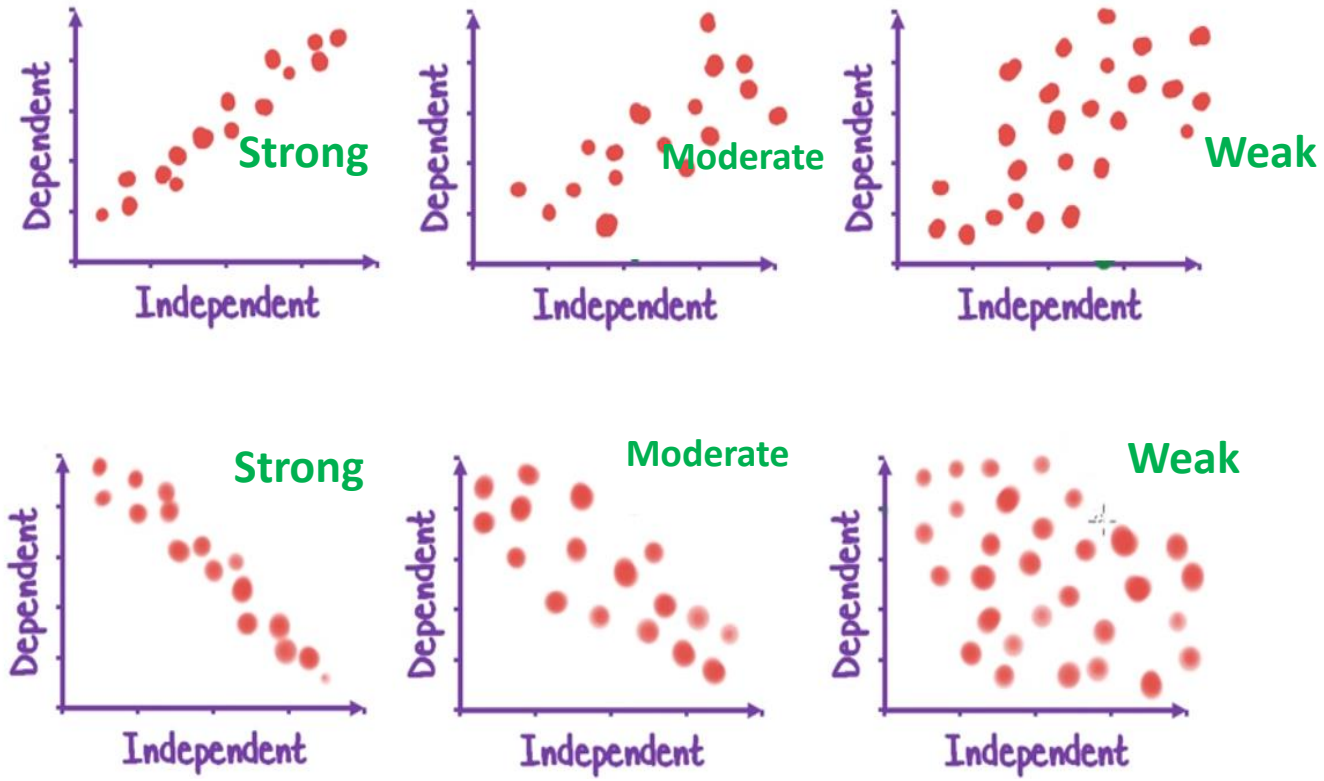
Form

Linear

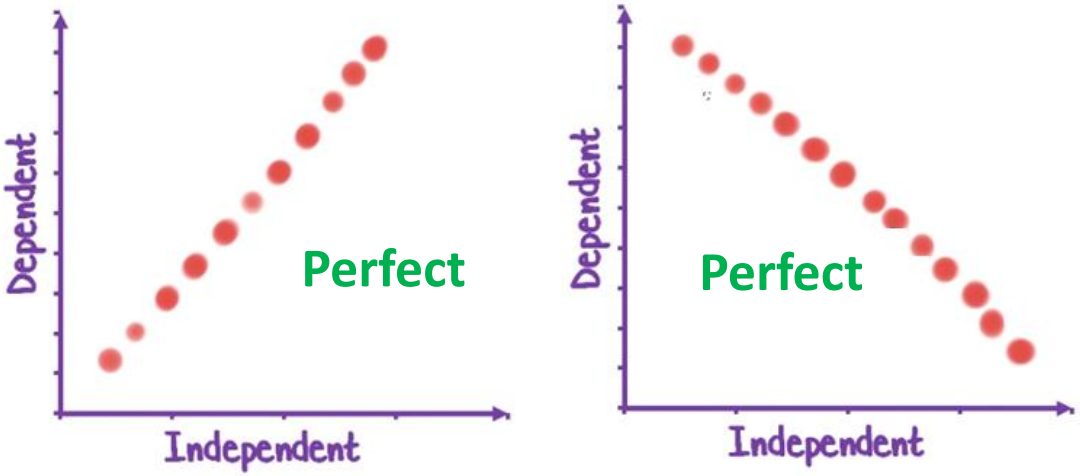
Linear

Non-Linear

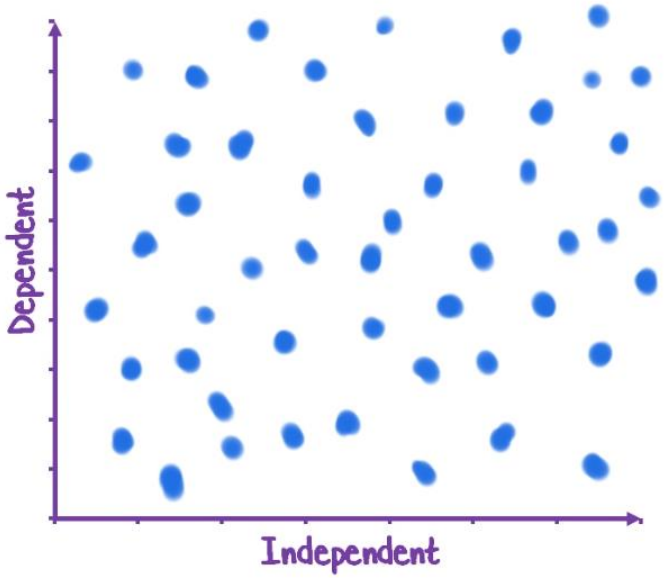
Strength


Dependent
Independent
Strong
Dependent
Independent
Moderate
Dependent
Independent
Weak
Dependent
Independent
Strong
Dependent
Independent
Moderate
Dependent
Independent
Weak

Strength-Perfect

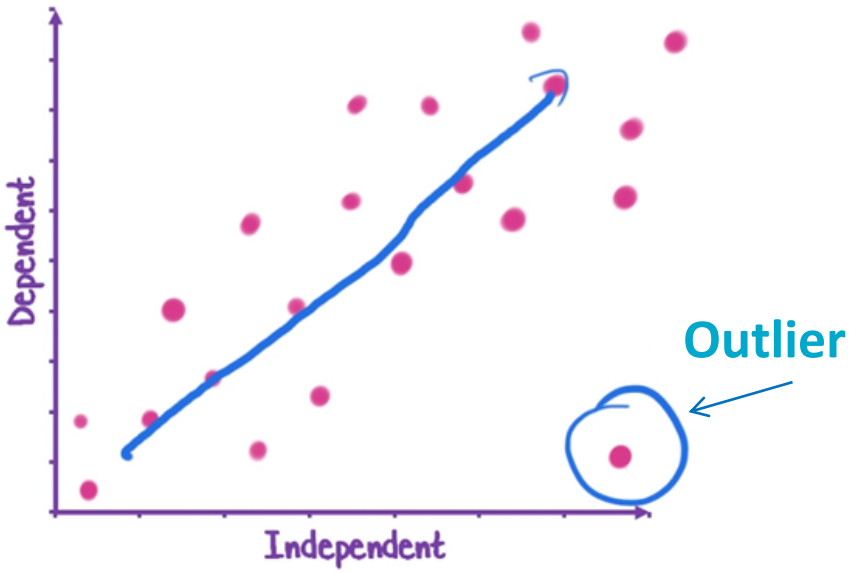


Strength-No association

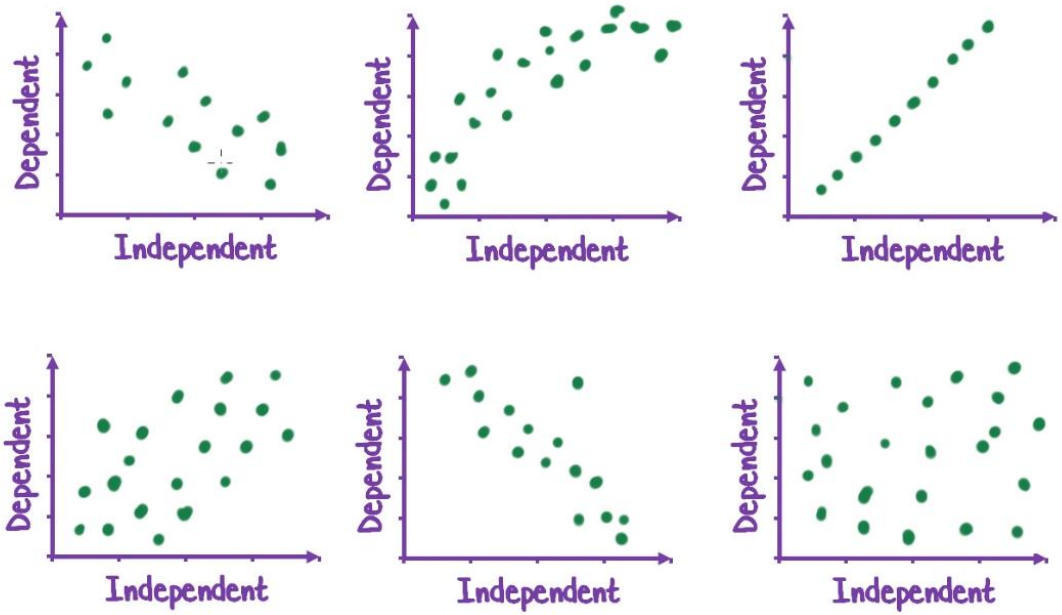


No association

Outlier



Example


Dependent
Independent
Dependent
Independent
Dependent
Independent
Dependent
Independent
Dependent
Independent
Dependent
Independent

Scatter plot Matrices

- A scatterplot matrix is a collection of scatterplots organized into a grid (or matrix).
- A scatter plot matrix can show how multiple variables are related.
- After plotting all the two-way combinations of the variables, the matrix can show relationships between variables to highlight which relationships are likely to be important.
- The matrix can also identify outliers in multiple scatter plots.

Scatter plot Matrices