Foundations of Data Analytics

Jiaqi Wang

January 20, 2024

Contents

Chapter 1

Exploratory Data Analysis

1.1 Data Types and Representations

Definition 1.1.0.1 – Data and Information In data analysis we are interested in the following two concepts:

- Data - raw numbers, facts, etc.
- Information - structured, meaningful and useful numbers, facts

Definition 1.1.0.2 – Data types We can classify data into two types:

- Categorical - data that has no intrinsic numerical value
  - Nominal - two or more outcomes that have no natural order, e.g. movie genre, hair color, etc.
  - Ordinal - two or more outcomes that have a natural order, e.g. movie ratings, level of education, etc.
- Numerical - data that has an intrinsic numerical value
  - Discrete - data that can attain certain values (typically integers). E.g. the number of days with sunshine in a certain year, the number of traffic incidents.
  - Continuous - data that can attain any value on a given measurement scale.
    - Interval data - equal intervals represent equal differences, there is *no fixed "zero point"*. e.g. temperature in Celsius, clock time, birth year, etc.
    - Ratio data - both differences and ratios make sense; there is a *fixed "zero point"*. e.g. temperature in Kelvin, height, weight, movie budget, distance, time duration, etc.

Remark 1.1.0.3 (Key points)
Some key points to remember:

- The difference between continuous and discrete data:
  Discrete data has "gaps".
- Why are distances ratio data?
  Because 0 km is an absolute minimum (and 10 km is twice as long as 5 km).

- Why is temperature in degrees Celsius interval data?
  Because 0 degrees Celsius is not an absolute minimum (and 10 degrees Celsius is not twice as hot as 5 degrees Celsius).

- Is temperature difference in Celsius not ratio data either?
  It is ratio data, since the difference of 0°C represents a fixed minimum.

- Can data represented by numbers be categorical?
  Yes, e.g. sectors of a stadium (nominal), or movie ratings (ordinal). No arithmetic operations can be performed on these numbers.

Remark 1.1.0.4 (Logarithmic scale data)
Logarithmic scale data is numerical, but it is neither ratio nor truly interval.

- Logarithmic scale data is a special case of ratio data.
- It is used to represent data that spans a large range of values.
- It is used to represent data that is highly skewed.

1.2 Tables

Definition 1.2.0.1 – Tables Tables are a common way to represent data.

- Each row represents an observation.
- Each column represents a variable.

They are good for reading off values and drawing attention to actual values.
There are two types of tables

- Reference tables - stores "all" raw data in a table so that it can be looked up easily.
- Demonstration tables - stores a subset of the data in a table to demonstrate a point.

1.3 Investigating data in tables

1.3.1 Questions you should ask

- What is the *unit of observation*?
- What is the *unit of measurement*?
- Do the values make sense when you compare columns or rows?
- What is the *data type*?
- Which column/row has the largest values?
- Which column/row has the smallest values?
- What is the *population*?
- What is the *sample*?

1.4 Going beyond plots

Plots help us to explore and give clues.

Numerical summaries like averages help us to document essential features of data sets.

One should use both plots and numerical summaries. They complement each other.

Numerical summaries are often called statistics or summary statistics (note the double meaning of the word: both a scientific field and computed numbers).

1.5 Summary Statistics

There are different types of summary statistics:

- Level - location summary statistics →what are "typical" values?
- Spread - scale summary statistics →how much do values vary?
- Relation - association summary statistics →how do values of different quantities vary simultaneously?

1.5.1 Location Summary Statistics

Definition 1.5.1.1 – **Location summary statistics** Location summary statistics are statistics that describe the location of the data.

- Mean (average) - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Median - the value separating the higher half from the lower half of a data set.
- Mode - the most frequently occurring value in a data set (it may not be unique).

Remark 1.5.1.2

The mean is sensitive to outliers, while the median and the mode are not.

The mean can be misleading / difficult to interpret for non-symmetric data sets.

Definition 1.5.1.3 – **Percentiles** Percentiles are a way to describe the location of a data point in a data set.

- The p-th percentile is the value such that p percent of the data is below it.
- The 0-th percentile is the minimum.
- The 100-th percentile is the maximum.
- For a dataset with n observations, the interval between neighboring percentiles is $\frac{100}{n-1}\%$.
- The median is the 50-th percentile.
- The first quartile is the 25-th percentile.
- The third quartile is the 75-th percentile.

Definition 1.5.1.4 – **Quartiles** Quartiles are a way to describe the location of a data point in a data set.

- The first quartile is the 25-th percentile.
- The second quartile = 20-th percentile = median.
- The third quartile is the 75-th percentile.
- The interquartile range (IQR) is the difference between the third and the first quartile.

1.5.2 Scale Summary Statistics

Definition 1.5.2.1 – **Scale summary statistics** Scale summary statistics are statistics that describe the scale of the data.

- Range - the difference between the maximum and the minimum.
- Interquartile Range (IQR) - the difference between the third and the first quartile.
- Sample Variance - $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (sometime denoted as s^2).
- Sample Standard Deviation - $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Median Absolute Deviation (MAD) - the median of the absolute deviations from the median.

The higher these statistics, the more spread/variability in the data.

Remark 1.5.2.2
- The standard deviation has right (physical) unit.
- The variance is more convenient for calculations.
- The range, variance and standard deviation are sensitive to outliers, while IQR and MAD are not.
- The standard deviation can be used as a general unit to describe variability.

Standardization

Definition 1.5.2.3 – **z-score** The z-score z (=normalized value) of values x shows is a standardized value of a data point.

- It is the number of standard deviations that a data point is above or below the mean.
- It is a dimensionless quantity.
- And it is calculated as $z = \frac{x - \bar{x}}{\sigma}$.

Remark 1.5.2.4 (Rule of thumb)
Observations with a z-score larger than 2.5 or smaller than -2.5 are considered extreme (*outliers*).

1.5.3 Association Summary Statistics

Association statistics try to quantify the strength of the relation between two variables (attributes).

Definition 1.5.3.1 – **Sample Covariance** The covariance of two variables x and y is a measure of how much they vary together.

- It is calculated as $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.
- It is sensitive to the scale of the variables.
- The sample covariance is not bounded.

In order to be useful, the sample covariance needs to be scaled / standardized.

Definition 1.5.3.2 – **Sample Correlation** The correlation of two variables x and y is a standardized measure of how much they vary together.

- It is calculated as $r = \frac{cov(x,y)}{\sigma_x \sigma_y}$.
- It is bounded between -1 and 1.

Remark 1.5.3.3 (Some remarks about correlation)

- The correlation is a measure of linear association.
- The correlation is not a measure of causation.
- The correlation is sensitive to outliers.

Remark 1.5.3.4 (Beware of spurious correlations!)
Spurious correlations are statistical phenomenon where two variables *appear to be related or correlated*, but *their relation is coincidental*, due to chance or the influence of an unaccounted-for variable.

1.6 Elementary Statistical Plots

Typical data analysis questions are:

- Are values as you expect?
- What are typical sizes?
- How much do values vary?
- How are values distributed?
- Are there exceptional values?

Graphs/Plots/Charts are useful for answering these questions.

1.6.1 Dot plots / Strip plots

Definition 1.6.1.1 – **Dot plot** A dot plot is a type of plot that displays data values as dots on a number line.

- One-dimensional numerical data
- Plotting actual values of observations
- The *jitter* option is used to avoid over-plotting.

- Tasks: find clusters, outliers, gaps, etc.
- Not suitable for large data sets.

1.6.2 Scatter plot

Definition 1.6.2.1 – **Scatter plot** A scatter plot is a type of plot that displays values of two numerical variables as points in a plane.

- Two-dimensional numerical data
- Plotting actual values of observations
- Tasks: investigate relations; find clusters, outliers, gaps, etc.
- Not suitable for large data sets.

1.6.3 Histogram

Definition 1.6.3.1 – **Histogram** A histogram is a type of plot that displays the distribution of numerical data.

- One-dimensional numerical data
- Plotting counts of observations in bins (interval of values)
- Tasks: investigate distribution; compare the counts/percentages of bins
- Suitable for large data sets.

Remark 1.6.3.2 (Bin width sensitivity)
Histograms are sensitive to the choice of bin width.

- Bin width too small →too many bins →too much detail (wiggly)
- Bin width too large →too few bins →too little detail (smooth)

Remark 1.6.3.3 (Rule of thumb)
The number of bins should be $\approx \sqrt{n}$, where n is the number of observations.

1.6.4 Cumulative histogram

Definition 1.6.4.1 – **Cumulative histogram** A cumulative histogram is a type of plot that displays the distribution of numerical data.

- One-dimensional numerical data
- Plotting counts of observations in bins (interval of values)
- Tasks: investigate distribution; explore unexpected "jumps"; to lookup percentiles or thresholds

1.6.5 Bar chart

Definition 1.6.5.1 – **Bar chart** A bar chart is a type of plot that displays the distribution of categorical data.

- One-dimensional categorical data
- Also used for two-dimensional data with one categorical variable and one numerical variable.
- Plotting counts of observations in categories
- Tasks: lookup and compare values.

1.7 Advanced Statistical Plots

1.7.1 Box plot

Definition 1.7.1.1 – **Box and whisker plot** Convenient way to display summary statistics.

- Box: IQR (interquartile range)
- Outliers: dots/crosses/diamonds... for all values above/below $1.5 \times IQR$
- Endpoints of whiskers: min and max (excluding outliers)
- Median: line in the box
- Mean: dot in the box (Optional)

Insights are:

- Is the plot symmetric with the median in the middle?
- Are there outliers?

1.7.2 Kernel density plot

Definition 1.7.2.1 – **Kernel density plot** The kernel density function shows the likelihood of finding a data point at a specific value.

- The total area under the curve is 1.
- The area under the curve between two values is the probability of finding a data point between these values.
- It is impossible to determine the minimal and maximal values or to find the mode of a dataset based on its kernel density plot!

How to generate a kernel density plot?

- Choose a kernel function and a bandwidth to be taken around each data point.
- Generate a kernel with the chose bandwidth for every data point in the data set

- For a data set with n points, the area under the curve of the kernel of a point is $\frac{1}{n}$.
- The kernel density plot is the sum of the kernels of the data points.
- Bandwidth choice is important!

1.7.3 Typical distribution shapes

- Symmetric
- Skewed (right or left skewed)
- Types of peaks: kurtosis
- Unimodal (1 peak)
- Bimodal (2 peaks)

1.7.4 Violin plot

Definition 1.7.4.1 – Violin plot A violin plot is a combination of a box plot and a kernel density plot.

- Global shape of box-and-whisker plot
- Local details of kernel density plot

1.7.5 Empirical cumulative distribution function (ECDF)

Definition 1.7.5.1 – Empirical cumulative distribution function (ECDF) The empirical cumulative distribution function (ECDF) is a function that maps a value to the fraction of values that are smaller or equal to it.

- It is a step function.
- It is a non-parametric estimator of the cumulative distribution function (CDF).
- It is a good way to visualize the distribution of a data set.
- It is a good way to compare the distribution of two data sets.

Interpreting ECDF

- For all x *smaller* than the *minimal* value of the data set, the ECDF is 0.
- For all x *larger or equal* to the *maximal* value of the data set, the ECDF is 1.

Chapter 2

Data visualization and Communication

2.1 Visualization in communication and exploration

> **Definition 2.1.0.1 – visualization** Visualization is the process that transforms (abstract) data into (interpretive) graphical representations for the purpose of exploration, confirmation or communication.

2.1.1 Purpose

- Open Exploration
- Confirmation
- Communication

2.2 Search

- Location
- Target

2.3 Human Perception

2.3.1 Visual processing

- Proximity
- Shape
- Continuity
- Colors

2.4 Encoding data

2.4.1 Marks and Channels

- Marks: points, lines, areas, bars, dots, etc.
- Channels: position, size, color, shape, etc.

2.4.2 Arrangement

How data values determine position

2.4.3 Mapping color

2.5 Idioms

Visualization idioms are common visualization techniques that are used to solve common problems.

2.5.1 Elements of a chart

2.5.2 Scatter plot

Definition 2.5.2.1 – • Two-dimensional data

- Mark:
- Channels:
- Task:

2.5.3 Bar chart

Definition 2.5.3.1 – • Two-dimensional data

- Mark: Lines (bar = a thick line)
- Channels:
  - Length to express quant value
  - Spacial regions
- Task: lookup and compare values

2.5.4 Stacked bar chart

Definition 2.5.4.1 – • Two-dimensional data

- Mark: Lines (bar = a thick line)
- Channels:
  - Length to express quant value

    - Spacial regions
    - Color to express category
- Task:
    - Compare the lowest values
    - lookup and compare categories

2.5.5 Normalized stacked bar chart

2.5.6 Line chart

2.5.7 Choosing a chart

2.5.8 Histogram

Definition 2.5.8.1 –

- Raw Data: one-dimensional numerical data
- Task: Find a distribution (shape)
- Derived Data: two-dimensional numerical data
    - Bins - key, counts - value
    - Bin size is crucial

2.5.9 Heatmap

Definition 2.5.9.1 –

- Three-dimensional data with
    - Two dimensions for the heatmap
    - One dimension for the color

2.5.10 Box plot

Definition 2.5.10.1 –

2.5.11 Violin plot

2.5.12 Box plot vs violin plot

2.5.13 Pie chart

2.5.14 Scatter plot matrix, parallel coordinates

2.6 Effectiveness, Evils, Ethics

Chapter 3

Data mining methods

3.1 What is data mining

Data mining is extracting information (knowledge) from data (observations).

3.1.1 Categories

- Do we have a predefined target (output variable)
  - Yes → **supervised learning**
  - No → **unsupervised learning**
- Do we seek information applicable to all or only some of the data?
  - All → **global**
  - Some → **local**

3.2 Data mining methods

- Linear regression
- Decision tree mining
- Clustering
- Association rule learning

3.3 Linear Regression

3.4 Clustering

We use *k-Mean* clustering method.

3.5 Distances

3.5.1 Euclidean distance

3.5.2 Network distance

3.5.3 Manhattan distance

3.6 Decision Trees

3.7 Association Rules