GSÜSEM

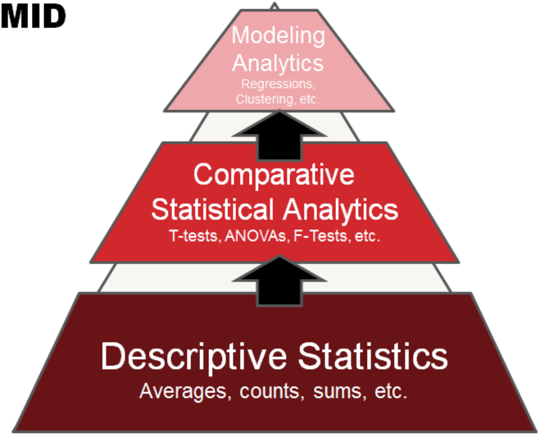
GALATASARAY ÜNİVERSİTESİ
Sürekli Eğitim Uygulama ve Araştırma Merkezi

Descriptive Data Analysis and Preprocessing

Ar. Gör. Pınar Uluer & Dr. Öğr. Üyesi Günce Keziban Orman

20/12/2019

ORGANIZATIONAL DATA CONSUMPTION PYRAMID



Modeling Analytics
Regressions, Clustering, etc.

Comparative Statistical Analytics
T-tests, ANOVAs, F-Tests, etc.

Descriptive Statistics
Averages, counts, sums, etc.

Outline

1. Data Definition
2. Descriptive Data Analysis
3. Quantitative Data Description
   - “Middle”
   - “Variation”
   - “Frequency”
   - Normal Distribution
   - Standard Score
4. Qualitative Data Description
   - Nominal Data Description
   - Ordinal Data Description
5. Summary
6. What’s Next?

Part 1 - Descriptive Data Analysis

Outline

1. Data Definition
2. Descriptive Data Analysis
3. Quantitative Data Description
4. Qualitative Data Description
5. What's Next?

Keywords

- Data:

  "*Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer*" - Cambridge English Dictionary

  Example: What do you think about the following statements?

  1. Temperature readings all over the world for the past 100 years
  2. Global temperature is rising

- Data set:

  A collection of data

Keywords

- Sample:
  A set of data collected and/or selected from a statistical population by a defined procedure
- Instance:
  One object in the data set, an independent example $\rightsquigarrow$ rows
- Attribute/Variable/Feature:
  Value used to describe one aspect of an instance $\rightsquigarrow$ columns
- Dimension:
  Number of features used for describing an instance

Assumptions

Type of data set: Collection of records with fixed set of features

If instances have the same fixed set of features:
→ they can be thought of as points in a multi-dimensional space
→ each dimension represent a distinct feature

Such data set can be represented by an *m* by *n* data matrix where:
– *m* rows → one for each instance
– *n* columns → one for each feature

For now:

Focus on the individual feature vectors ✓

Relation and/or dependence between them ✗

Data Matrix - Example

ID	Drink	Type	Calories	Sugars (gr)	Caffeine (mg)
13424	Brewed Coffee	Hot	4	0	260
13425	Cafe Latte	Hot	100	14	75
13426	Cafe Mocha	Hot	170	27	95
13427	Cappuccino	Hot	60	15	120
13428	Iced Coffee	Cold	60	15	120

Table: Nutritional data on some popular drinks

Outline

1. Data Definition
2. Descriptive Data Analysis
3. Quantitative Data Description
4. Qualitative Data Description
5. What's Next?

What is Descriptive Data Analysis?

→ Usually the first step of the analysis process

→ Objective: To have a general idea about data

Implemented before applying any algorithms to make a prediction:
"prediction" → inferential statistics

→ Answer to the questions:

- "What happened?"
- "What is happening?"

→ Provide a quick and simple description: quantitative, visual, etc.

How to Describe What We Have?

Description based on:

- a measure of central tendency
  → e.g. arithmetic mean
- a measure of spread
  → e.g. standard deviation
- a measure of distribution shape
  → e.g. skewness, kurtosis
- (*if more than one feature, a measure of dependence*
  → *e.g. correlation coefficient*)

Data Description: Types of Features

- Qualitative or symbolic:
  - Dichotomous: presence/absence
  - Nominal or categorical: several unordered symbols → e.g. blue, red, white, etc.
  - Ordinal: several ordered symbol → e.g. cold, warm, hot

- Quantitative or numerical:
  - Binary: 0/1
  - Ordinal or rank: 1st, 2nd, 3rd, 4th, etc.
  - Interval or scaled: continuous value expressed on an scale whose zero is an arbitrary value → e.g. Celsius degrees
  - Ratio: like interval, except the zero is not arbitrary → e.g. Kelvin degrees

The difference between interval and ratio scales is that, while interval scales are void of absolute or true zero for example temperature can be below 0 degree Celsius (-10 or -20), ratio scales have a true zero value, for example, height or weight it will always be measured between 0 to maximum but never below 0.

Data Types - Example

Name	Gender	Language	Education	Income
Susan	F	English	University	40000
Jason	M	English	High School	30000
Michael	M	French	University	45000
John	M	German	High School	35000
Emily	F	French	High School	45000

Table: Information about people and their incomes

Outline

1. Data Definition
2. Descriptive Data Analysis
3. Quantitative Data Description
   - “Middle”
   - “Variation”
   - “Frequency”
   - Normal Distribution
   - Standard Score
4. Qualitative Data Description
5. What’s Next?

Which Group is Smarter?

IQ Scores of Group 1	IQ Scores of Group 2
102	127
115	162
128	131
109	103
131	96
89	111
98	80
106	109
140	93
119	87
93	120
97	105
110	109

Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.

Quantitative Data Description

Focus on the numerical data, analysis based on 3 key measures:

- Central Tendency $\rightsquigarrow$ “Middle”
  To find one number summarizing the entire data set:
  $\rightarrow$ a central number to represent the data
- Spread/Dispersion $\rightsquigarrow$ “Variation”
  To check the variability within the data:
  $\rightarrow$ spread of the values around the central tendency
- Distribution $\rightsquigarrow$ “Frequency”
  To find the number of times it appears in a sample:
  $\rightarrow$ probability of the occurrence of an event

Mean

- The simplest and most intuitive way
- A number around which the data is spread out
- **mean** is the statistical jargon for **average**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example:

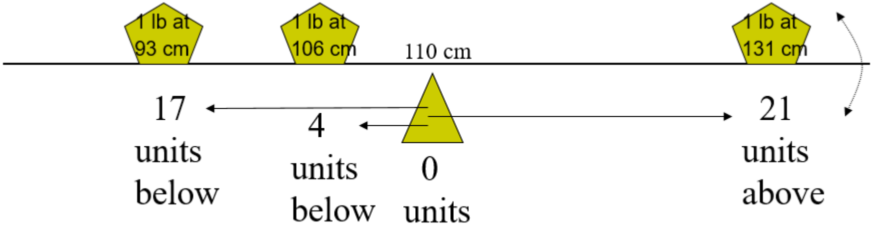
$$x_{group1}^T = [102, 115, 128, 109, 131, 89, 98, 106, 140, 119, 93, 97, 110]$$
$$x_{group2}^T = [127, 162, 131, 103, 96, 111, 80, 109, 93, 87, 120, 105, 109]$$
$$\bar{x}_{group1} = 110.5385$$
$$\bar{x}_{group2} = 110.2308$$

They're roughly the same!

Mean

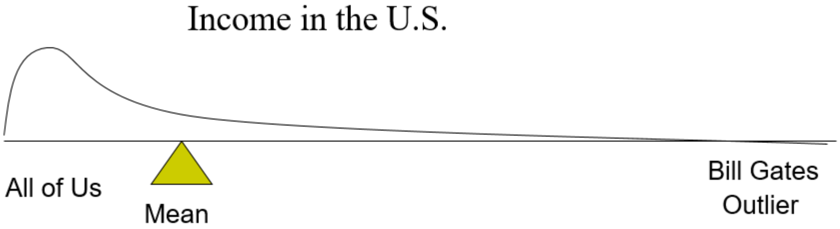
The mean is the *balance point.*

Each person's score is like 1 pound placed at the score's position on a see-saw. Below, on a 200 cm see-saw, the mean equals 110, the place on the see-saw where a fulcrum finds balance:



Mean

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency or common experience



Median

- The middle term of the **sorted** data
- The value which divides the data in 2 equal parts

Example:

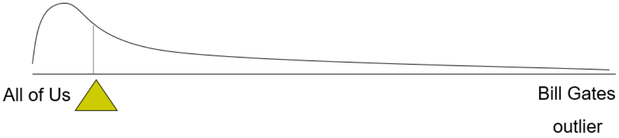
$$x_{group1}^{T} = [102, 115, 128, 109, 131, 89, 98, 106, 140, 119, 93, 97, 110]$$
$$\Rightarrow x_{sorted}^{T} = [89, 93, 97, 98, 102, 106, \mathbf{109}, 110, 115, 119, 128, 131, 140]$$
$$\Rightarrow x_{sorted}^{T} = [140, 131, 128, 119, 115, 110, \mathbf{109}, 106, 102, 98, 97, 93, 89]$$
$$median(x) = 109$$

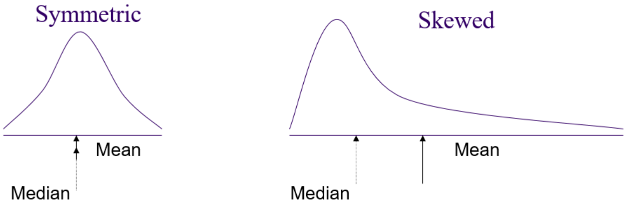
Question: What if we had the following data set?

$$x^{T} = [15, 20, 21, 18, 36, 15, 25, 15]$$

Median

1. The median is unaffected by outliers, making it a better measure of central tendency, better describing the ”typical person” than the mean when data are skewed.
2. If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
3. In skewed data, the mean lies further toward the skew than the median.





Median

1. The middle score or measurement in a set of ranked scores or measurements; the point that divides a distribution into two equal halves.
2. Data are listed in order—the median is the point at which 50% of the cases are above and 50% below.
3. The 50th percentile.

Mode

- The term having the **highest frequency** of occurrence
- You can either sort the data and count the number of occurrence or just find the frequency without any sorting operation

Example: The combined IQ scores for Group 1 and Group 2

$$x^T = [80, 87, 89, 93, 93, 96, 97, 98, 102, 103, 105, 106, \mathbf{109}, \mathbf{109}, \mathbf{109}, 110, 111, 115, 119, 120, 127, 128, 131, 131, 140, 162]$$

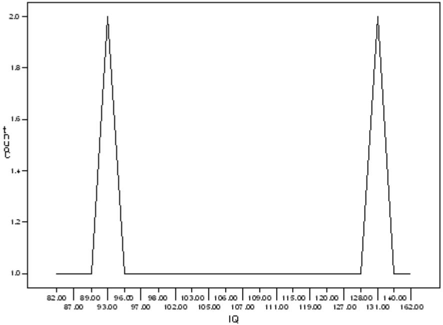
$$mode(x) = 109$$

Question: What if we had the following data sets?

$$x_1^T = [15, 16, 17, 18, 19, 20, 21, 22]$$
$$x_2^T = [15, 20, 21, 18, 36, 15, 25, 25]$$

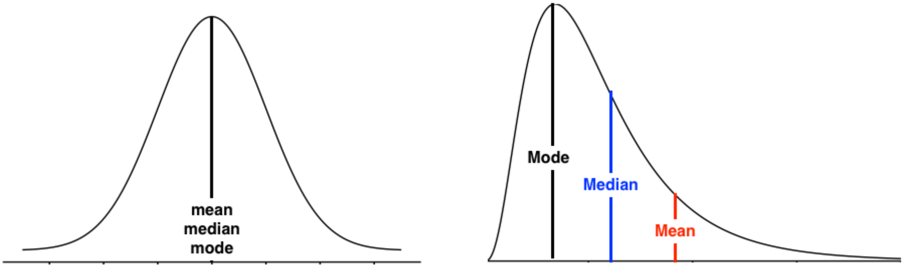
Mode

1. It may mot be at the center of a distribution.
2. Data distribution on the bottom is ”bimodal” (even statistics can be open-minded)
3. It may give you the most likely experience rather than the “typical” or “central” experience.
4. In symmetric distributions, the mean, median, and mode are the same.
5. In skewed data, the mean and median lie further toward the skew than the mode.



Mean, Median & Mode

Assume that the yearly income is distributed as follows for 2 different countries: In which country would you like to live? Why?



(a) symmetric distribution (b) skewed distribution

Fig: $x \rightarrow$ Income per citizen, $y \rightarrow$ Number of citizen (Image courtesy: `http://statisticshelper.com/`)

Variation

Now we have some measures to represent the data:
mean, median or mode

What if two feature vectors have the same central tendency?

→ We need another measure:

- To examine further the similarities and/or differences
- To check the **variability** within the data

Variability:
“Spread of the values around the central tendency”
“A measure of the extent to which individual values differ from the mean”

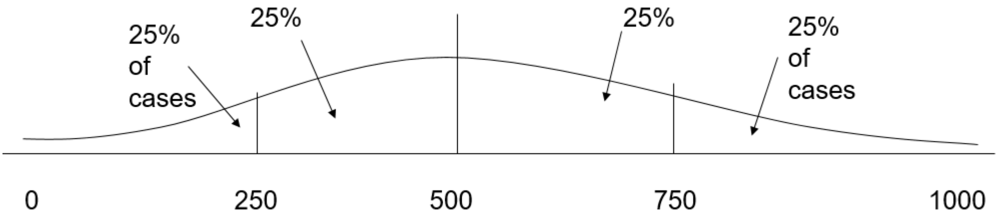
Range

The spread, or the distance, between the lowest and highest values of a variable.
To get the range for a variable, you subtract its lowest value from its highest value.
The difference between the maximum and minimum

$$
\begin{aligned}
x_{group1}^{T} &= [102, 115, 128, 109, 131, 89, 98, 106, 140, 119, 93, 97, 110] \\
x_{group2}^{T} &= [127, 162, 131, 103, 96, 111, 80, 109, 93, 87, 120, 105, 109] \\
range_{group1} &= 140 - 89 = 51 \\
range_{group2} &= 162 - 80 = 82
\end{aligned}
$$

Interquartile Range

- A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.
- The median is a quartile and divides the cases in half.
- 25th percentile is a quartile that divides the first 1/4 of cases from the latter 3/4.
- 75th percentile is a quartile that divides the first 3/4 of cases from the latter 1/4.
- The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



Percentile

- A way to represent the position of a value in the data set
  → **The spread of the sorted data**

- Similar to median, the data set should be ordered, preferably in ascending order

- The value such that:
  P percent of the values take on this value or less
  $(100 - P)$ percent take on this value or more

Percentile

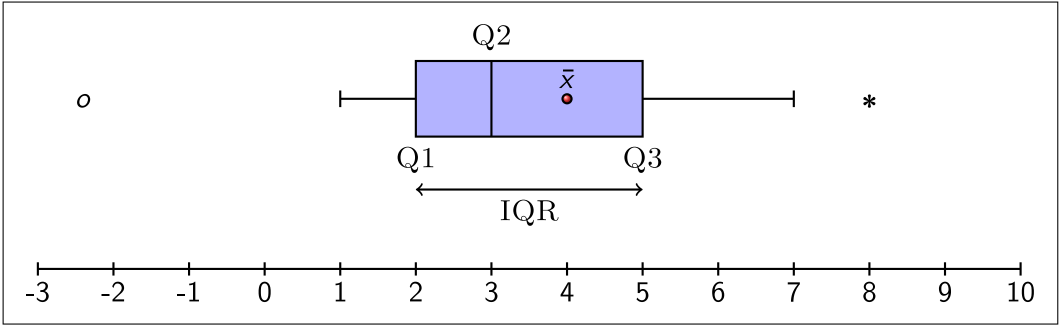
- In general, if k is n^{th} percentile → n% of the total terms are less than k:

$$\begin{array}{llcl} \rightarrow & \text{The } 0^{th} \text{ percentile} & : & \left.\begin{array}{l}\text{Minimum} \\ \text{Maximum}\end{array}\right] \text{Range} \\ \rightarrow & \text{The } 100^{th} \text{ percentile} & : & \\ \rightarrow & \text{The } 50^{th} \text{ percentile} & : & \text{Median} \end{array}$$

- The data can be divided into “quartiles”:
  Cutpoints @ 25% (Q1), 50% (Q2) and 75% (Q3)

- Extra vocabulary: Q1 ↔ Q3 : “interquartile range”

Box and Whisker Plot



What does the boxplot tell you?

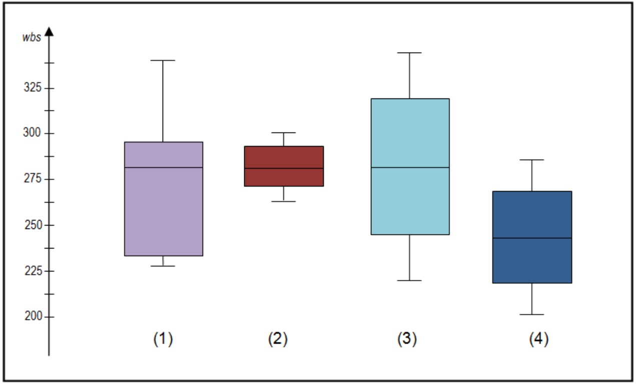
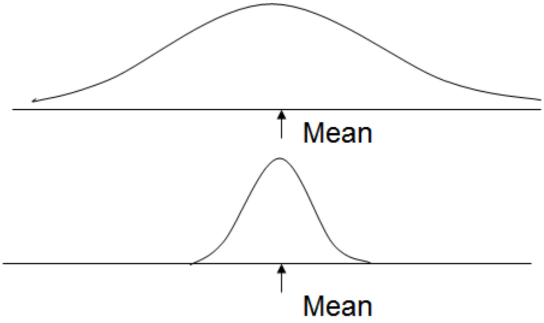


Fig: Different boxplot examples drawn for different groups based on wellbeing at school scale scores (Image courtesy: `https://www.wellbeingatschool.org.nz/`)

Variance

- A measure of the spread of the recorded values on a variable. A measure of dispersion. → variance (σ^2)
- The larger the variance, the further the individual cases are from the mean.
- The smaller the variance, the closer the individual scores are to the mean.



Variance

Variance is a number that at first seems complex to calculate.

- Calculating variance starts with a "deviation."
- A deviation is the distance away from the mean of a case's score.

$$x_{group1} = [102, 115, 128, 109, 131, 89, 98, 106, 140, 119, 93, 97, 110]$$
$$\bar{x}_{group1} = 110.54$$

Question: The deviation $(x_i - \bar{x})$ of 102 (an IQ score from Group 1) from 110.54 (mean IQ score of Group 1) is?
102 - 110.54 = -8.54

Question: Deviation of 115?
115 - 110.54 = 4.46

Variance

- We want to add these to get total deviations, but if we were to do that, we would get zero every time. Why?
- We need a way to eliminate negative signs.
- Squaring the deviations will eliminate negative signs..
- A Deviation Squared: $(x_i - \bar{x})^2$

Back to the IQ example, A deviation squared for 102 and 115 are:
$(102 - 110.54)^2 = (-8.54)^2 = 72.93$
$(115 - 110.54)^2 = (4.46)^2 = 19.89$

Variance

If you were to add all the squared deviations together, you'd get what we call the **Sum of Squares**.

Sum of Squares (SS) $= \sum (x_i - \bar{x})^2$

$$SS = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$$

Variance

$$x_{group1} = [102, 115, 128, 109, 131, 89, 98, 106, 140, 119, 93, 97, 110]$$
$$\bar{x}_{group1} = 110.54$$

$$SS = (102-110.54)^2 + (115-110.54)^2 + (126-110.54)^2 + (109-110.54)^2$$
$$+(131-110.54)^2 + (89-110.54)^2 + (98-110.54)^2 + (106-110.54)^2$$
$$+(140-110.54)^2 + (119-110.54)^2 + (93-110.54)^2 + (97-110.54)^2$$
$$+(110-110.54) = 2825.39$$

Variance

The last step

- The approximate average sum of squares is the variance.
- SS/N = Variance for a population.
- SS/n-1 = Variance for a sample.

$$variance = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

For Group1 ;

$$Variance = \frac{2825.39}{n-1} = \frac{2825.39}{12} = 235.45$$

Standard Deviation

- Another way to determine how data is spread out from the mean:
  To measure the average distance between a single observation and the mean
  $\rightarrow$ standard deviation (σ)
- More accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range
- Low σ → data points tend to be close to the mean
  High σ → data points are spread out over a wider range

Standard Deviation

To convert variance into something of meaning, let's create standard deviation. How to find standard deviation σ?

- The Square root of the variance reveals the average deviation of the observations from the mean

$$\sigma = \sqrt{\frac{\text{sum of (individual value - mean value)}^2}{\text{number of values}}}$$

- Standard deviation of a population

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

- Standard deviation of a sample

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

Standard Deviation

For Group 1, the standard deviation is:

$$\sqrt{235.45} = 15.34$$

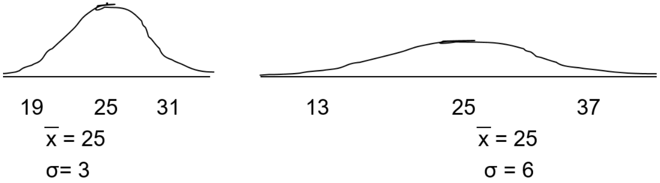
The average of persons' deviation from the mean IQ of 110.54 is 15.34 IQ points.

Review:

1. Deviation
2. Deviation squared
3. Sum of squares
4. Variance
5. Standard deviation

Standard Deviation

- Larger σ means greater amounts of variation around the mean.
- For example:



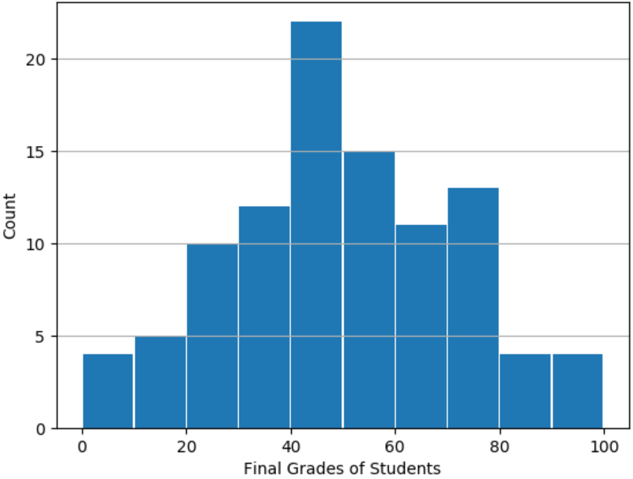
- $\sigma = 0$ only when all values are the same (only when you have a constant and not a ”variable”).
- If you were to ”rescale” a variable, the σ would change by the same magnitude (e.g. if we changed units above so the mean equaled 250, the σ on the left would be 30, and on the right, 60).
- Like the mean, the σ will be inflated by an outlier case value.

Frequency

- The data is distributed in some manner throughout the various possible values
- One of the most common ways to describe a single variable is to find the number of times it appears in a sample
  → Frequency
- To reflect the probability of the occurrence of an event
  → Frequency distribution
- Another convenient way to summarize data

Histogram

- A graph displaying visually the relation between values and their frequencies
  → A mapping from values to frequencies
- To plot quantitative data with ranges of the data grouped into bins or intervals
- To show distributions of variables
  → It does not make sense to rearrange the bars of a histogram



Probability Density Function

Probability: A frequency expressed as a fraction of the sample size

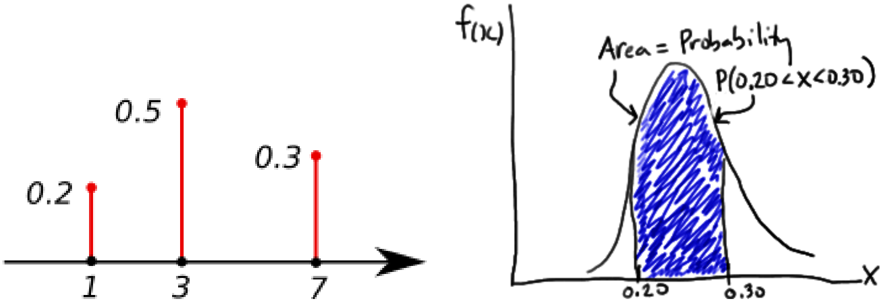
Reminder:

- Definition of Probability: "the percent chance that some event will occur"
- Commonly quantified in the range of 0 to 1:
  0 means we are certain this will not occur
  1 means we are certain it will occur
- A probability distribution is :
  → A function which represents the probabilities of all possible outcomes in a statistical experiment
  → A table or an equation that links each outcome with its probability of occurrence

Probability Density Function

- Discrete versus Continuous → Probability Mass Function (PMF) versus Probability Density Function (PDF)
- PMF: A representation of a distribution as a function that maps from values to probabilities
- PDF: A representation of a distribution as a function that describes the relative likelihood for the random variable to take on a given value → integration

Mass versus Density



(a) PMF

(b) PDF

Fig: PMF vs PDF (Image courtesy: `https://www.medium.com/`)

Normal (Gaussian) Distribution

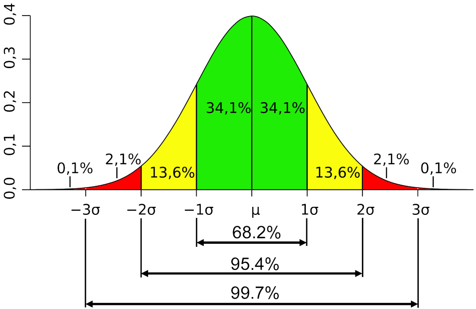
Different types of distribution: Uniform, Poisson, Power Law, Normal (Gaussian), etc.

- Uniform → rolling a dice, random number generation
- Poisson → number of automobiles arriving at a traffic light within the hour
- Power law → Google page rank, “rich gets richer”
- Normal distribution → birth weights of babies for the last 10 years, petal lengths of iris flower, serum level measurements of healthy individuals, etc.

Normal (Gaussian) Distribution

Normal distribution is popular because:

- it is symmetric: mean $=$ median $=$ mode $\rightarrow$ symmetry about the center
- it can be fully characterized by just two parameters:
  $\rightarrow$ mean & standard deviation (or variance)
  $\rightarrow$ often referred to as $N(\mu, \sigma^2)$
- the probability of any value occurring can be obtained simply by knowing how many σ separate the value from μ:
  - likely to be within 1 σ
  - very likely to be within 2 σ
  - almost certainly within 3 σ



Normal (Gaussian) Distribution

Example: 95% of employees have salaries ranging from \$1100 to \$1700. Assuming this data is normally distributed, what can we find out about salaries?

The mean is halfway between \$1100 and \$1700 :

$$\mu = (1100 + 1700)/2 = \$1400$$

95% is 2 standard deviations either side of the mean :

$$\begin{aligned}\sigma &= (1700 - 1100)/4 \\ &= 600/4 \\ &= \$150\end{aligned}$$

Normal (Gaussian) Distribution

- It is best suited for data that meets the following conditions:
  - a strong tendency to take on a central value
  - positive and negative deviations from this central value are equally likely
  - the frequency of the deviations falls off rapidly as we move further away from the central value (zero skewness & no kurtosis)
- What about the situation when we use a normal distribution to characterize data that is non-normal?
  → Then there is a cost we pay

Standard Normal Distribution

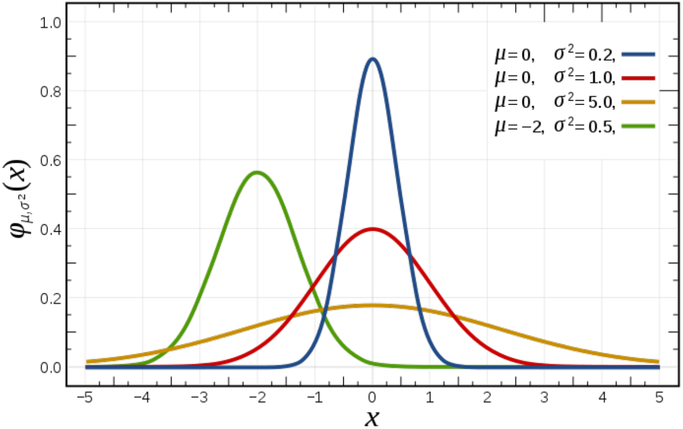


Fig: PDF of Normal Distribution (Image courtesy of Wikimedia Commons)

Standard Normal Distribution

- Normal distributions do not necessarily have the same means and standard deviations
- Hence the need to standardizing the distributions: Standard Normal distribution with $\mu = 0$ and $\sigma = 1$
- It is often used to:
  - compare two or more distributions of data
  - estimate or to compute probabilities of events involving normal distributions

What does "Standard Score" mean?

$$\begin{aligned}\text{Standard score} &= \text{normal score} = \text{standardized variable}\\ &= \text{z-score} = \text{z-value}\end{aligned}$$

To convert a value to standard score:

- first subtract the mean
- then divide by the standard deviation

$$\text{standard score} = \frac{\text{(value to be standardized - mean)}}{\text{standard deviation}}$$

$$z = \frac{(x_i - \mu)}{\sigma}$$

Standard Score

Example: Assume that you have to compare 2 candidates in terms of their English speaking skills for an open position:
– One of them has a score of 1100 points from TOEFL
– The other has 25 points from IELTS.

We know that:
– For TOEFL: $\mu = 1000$, $\sigma = 100$
– For IELTS: $\mu = 22$, $\sigma = 2$

Question: Who should get the job?

Outline

1. Data Definition
2. Descriptive Data Analysis
3. Quantitative Data Description
4. Qualitative Data Description
   - Nominal Data Description
   - Ordinal Data Description
5. What's Next?

Qualitative Data Description

So far, we have dealed with numerical values...

But not all data is quantitative...

What was the qualitative data?

Data that can take on only a specific set of values representing a set of possible categories:

- Nominal
- Ordinal

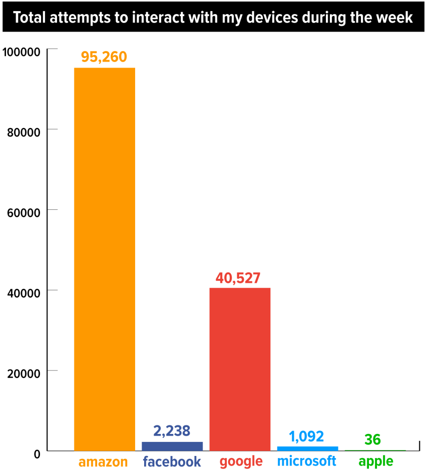
Nominal Data: What is it?

- No hierarchical relation among categories

  Example:
  – Animal species: Pig is not higher than bird, lower than fish
  – Nationality: Being Turkish, Icelandic, or Japanese does not imply an ordered relationship

- No numerical relationship between the different categories:
  $\rightarrow$ mean ✗, median ✗, mode ✓
- The best we can do is to indicate which category was most frequently reported
  $\Rightarrow$ Check the frequency: How many categories? How many instances?

Example - ‘I Cut the Big 5 Tech Giants from my Life’ (https://gizmodo.com/)



Example

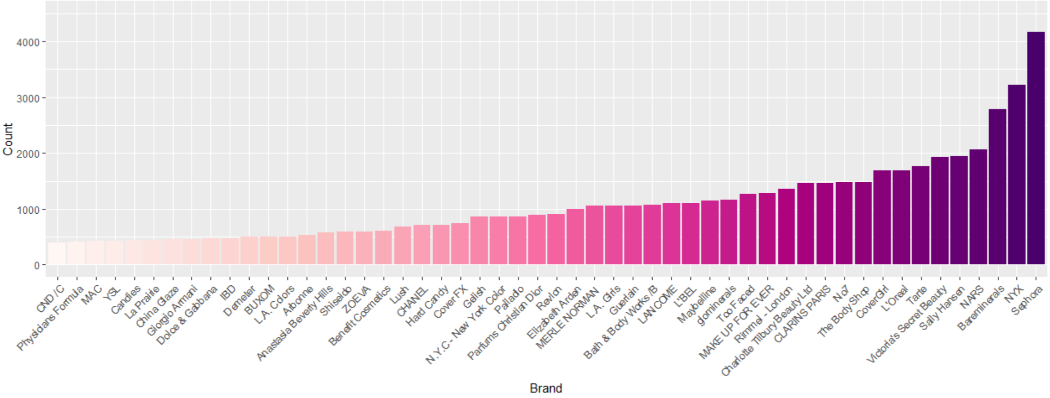
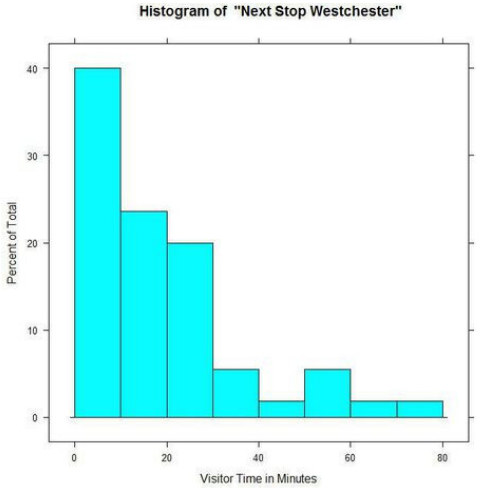
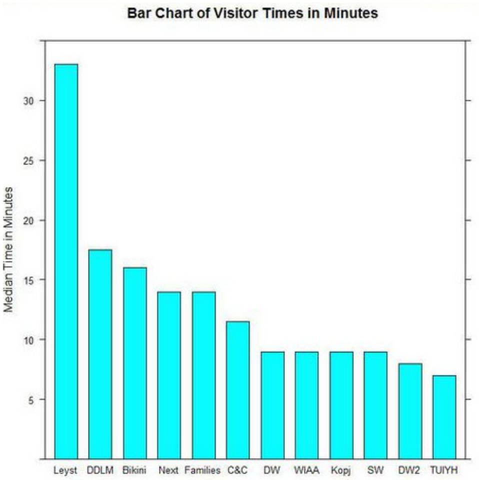


Fig: Bar chart of brand name

Just a Reminder ('Paying Attention: Visitors and Museum Exhibitions')





Ordinal Data: What is it?

- An explicit hierarchy among the categories → It can be ranked or sorted BUT no numerical relationship between categories
- We know that it can be rank ordered but we don't know anything about the space/distance between the rankings
  → mean is (mostly) meaningless
  → median should be used with discretion
  → long live the mode!
- The best we can do is to indicate which category was most frequently reported
  ⇒ Check the frequency: How many categories? How many instances?

Ordinal Data - Example

Case 1: No idea about the relative distance between the ranks

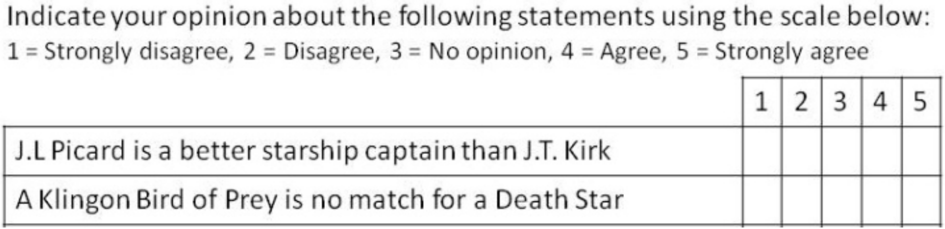
Ranking	Time (minutes)
1^{st}	30:00
2^{nd}	30:01
3^{rd}	400:00

Table: Rankings of a 10K run

→ Without the time data, we don't have any idea about the distance between the ranks, median seems to be significant but with the timing of runner we can say that the median is not a good measure to represent this data set

Ordinal Data - Example

Case 2: Assumption of equal magnitudes between the categories

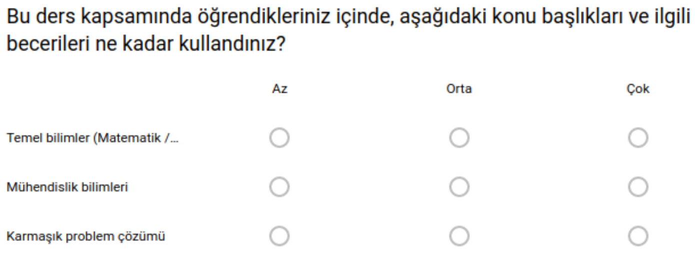


Indicate your opinion about the following statements using the scale below:
1 = Strongly disagree, 2 = Disagree, 3 = No opinion, 4 = Agree, 5 = Strongly agree

	1	2	3	4	5
J.L Picard is a better starship captain than J.T. Kirk					
A Klingon Bird of Prey is no match for a Death Star					

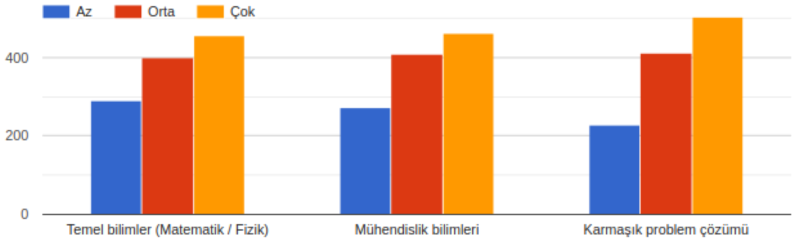
Fig: Likert scale in social sciences

→ the difference between 2 and 3 & the difference between 4 and 5 can be reasonably assumed to be similar



Bu ders kapsamında öğrendikleriniz içinde, aşağıdaki konu başlıkları ve ilgili becerileri ne kadar kullandınız?

	Az	Orta	Çok
Temel bilimler (Matematik /...	○	○	○
Mühendislik bilimleri	○	○	○
Karmaşık problem çözümü	○	○	○



Summary

- Usually, but not necessarily, the first step in the analysis
- How to define & describe data:
  - Central tendency
  - Dispersion
  - Distribution
- Quantitative vs Qualitative:
  Different data types $\Leftrightarrow$ different methodologies
- Importance of standardization
- It is a never-ever-ending dynamic process

What about the Next Step?

- Coming up next: Python practice with zomato data set:
  - How to import a data file?
  - What libraries and methods to use?
  - How to describe the data we have?
  - How to visualize it?
- Part 2:
  - Data preprocessing
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction

Resources

- Textbooks:
  - Thomas Haslwanter, An Introduction to Statistics with Python with Applications in the Life Sciences, Springer (2016)
  - Joel Grus, Data Science from Scratch: First Principles with Python, O'Reilly (2015)
  - Peter C. Bruce & Andrew Bruce, Practical Statistics for Data Scientists: 50 Essential Concepts, O'Reilly (2017)
- Online:
  - `https://towardsdatascience.com/`
  - `https://stats.stackexchange.com/`
  - `https://realpython.com/`
  - `http://pandas.pydata.org/`
- Slides:
  - http://www.sjsu.edu/people/james.lee/courses/102/s1/asDescriptive_Statistics2.ppt

Any questions so far?