



EBU5601

Data Design

Descriptive Statistics

Dr Chao Shu, Dr Xiaolan Liu

School of Electronic Engineering and Computer Science
Queen Mary University of London

Sep. 2024

Learning Outcomes

- The main outcomes are:
 - **[LO2.1]** Identify data types and variable types
 - **[LO2.2]** Understand the concept of random variables and implement basic statistic notation
 - **[LO2.3]** Analyse measures of centre and spread
 - **[LO2.4]** Understand how the histogram and box plot is constructed and use them to visualise quantitative data
 - **[LO2.5]** Understand Common shapes that data takes on and how to handle outliers
 - **[LO2.6]** Understand how to use the bar chart to visualise and analyse categorical data

INTRODUCTION

Introduction

- **Descriptive Statistics** is useful in many different jobs, and activities. Having a good understanding of descriptive statistics will help anyone working in:
 - Business Analytics
 - Data Analysis
 - Data Engineering
 - Product Management
- Data is defined as distinct pieces of information and it can come in many forms.
 - numbers in a spreadsheet, text, video, databases, images, audio recording

Data Types

- **Quantitative**

- Quantitative data takes on numeric values that allow us to perform mathematical operations.
- **Continuous Data:** Quantitative Values that can be split into smaller values. e.g., Age, train speed, time
- **Discrete Data:** Quantitative Values that are countable. The number of dogs is an example of a discrete data type. e.g. number of pets, number of packages delivered by Amazon/JD/Taobao

- **Categorical**

- Categorical is used to label a group or set of items.
- **Ordinal:** categorical values that are ranked. e.g., Grades (A, B, C, D, E, Fail), Day of week
- **Nominal:** categorical values that do not have ranked order. e.g., Marital Status: (Single/Married/Divorced), Country of Residence.

Data Types

Exercise

- All of the variables below are quantitative. Please select all variables that are continuous.
 1. Travel Distance from home to work
 2. Number of pages in a book
 3. Amount of rain in a year
 4. Time to run a kilometre
 5. Number of movies watched a week
 6. Amount of water consumed in a day
 7. Number of phones per household

Go to
www.menti.com

Enter the code

6195 8382



Or use QR code

Data Types

Exercise

- All of the variables below are categorical. Please select all variables that are nominal.
 1. Letter Grades (A, B, C, etc.)
 2. Types of Fruit (Apple, Banana, etc.)
 3. Ratings on a Survey (Poor, OK, Great)
 4. Types of Dog Breeds
 5. Genres of Movies (Horror, Comedy, etc.)
 6. Gender
 7. Nationality
 8. Education (Bachelor, Master, PhD)

Go to
www.menti.com

Enter the code

6195 8382









Or use QR code

ANALYSING QUANTITATIVE DATA

Random Variables

- Consider I run a website to sell snacks. In order to know how this website contributes to my snack retail business, questions I might have are:
 - How many people visit the site?
 - How much time do visitors spend on the site?
 - Are there differences in traffic depending on the day of the week?
 - How many visitors purchase an item through the website?

					
Walkers Classic Variety Multipack Crisps 12 per pack	Walkers Cheese & Onion Multipack Crisps 12 per pack	Walkers Meaty Variety Crisps 12 per pack	Naturya Organic Overnight Break... 300g	Perkier Crunchy Peanut Protein bars 3 x 35g	Perkier Salted Caramel & Dark... 3 x 37g
Price Lock	Price Lock	Price Lock			
★★★★★ (25)	★★★★★ (34)	★★★★★ (4)	★★★★★ (15)	★★★★★ (58)	★★★★★ (68)
£3.25 £	£3.25	£3.25 £	£5	£2.20	£2.95
27.1p each	27.1p each	27.1p each	£1.67 per 100g	£2.10 per 100g	£2.66 per 100g
Add to trolley	Add to trolley	Add to trolley	Add to trolley	Add to trolley	Add to trolley

Random Variables

- We can gather data in a spreadsheet.
- A column of our dataset is associated with a **random variable**. In a spreadsheet, each column commonly holds a specific variable, while each row is commonly called an instance or individual.

Date	Day of Week	Time	Buy

Random Variables

Quick Quiz

- What type of variable is the random variable X and Y?

Date	Day of Week	Time Spent on Site (X)	Buy (Y)
June 15	Thursday	5	No
June 15	Thursday	10	Yes
June 16	Friday	20	Yes

Go to
www.menti.com

Enter the code

6195 8382



Or use QR code

Random Variables

- Random Variables and Observations
 - A random variable is a placeholder for the possible values of some process
 - the amount of time someone spends on our site is a **random variable**, noted by capital letters, e.g., X
 - Observed values of the random variables are noted by lowercase letters with subscriptions, e.g., x_n

X

Date	Day of Week	Time	Buy
		5	
		10	
		...	
		...	
		m	

Observations

The diagram illustrates the relationship between a random variable X and its observed values x_1, x_2, \dots, x_n . A table with four columns (Date, Day of Week, Time, Buy) shows several rows. The 'Time' column is highlighted in blue. Arrows point from the values in the 'Time' column (5, 10, ..., m) to labels x_1, x_2, \dots, x_n on the right. A bracket groups these labels under the heading 'Observations'.

Random Variables

Example

- If 5 individuals visit our website, the first spend 10 minutes, the second spends 20 minutes, the third spend 45 mins, the fourth spends 12 minutes, and the fifth spends 8 minutes; we can notate this problem in the following way:

X is the amount of time an individual spends on the website.

$$x_1=10, x_2=20, x_3=45, x_4=12, x_5=8$$

The capital X is associated with this idea of a random variable, while the observations of the random variable take on lowercase x values.

Random Variables

Example

- If 5 individuals visit our website, the first spend 10 minutes, the second spends 20 minutes, the third spend 45 mins, the fourth spends 12 minutes, and the fifth spends 8 minutes

X is the amount of time an individual spends on the website.

$$x_1=10, x_2=20, x_3=45, x_4=12, x_5=8$$

What is the probability someone spends more than 20 minutes in our website?

In notation, we would write: $P(X > 20) = 1/5 = 20\%$

What is the probability of an individual spending 20 or more minutes on our website?

Answer:

$$P(X \geq 20) = 2/5 = 40\%$$

Four Aspects for Quantitative Data

- There are four main aspects to analyzing Quantitative data.
 - **Centre**
 - **Spread**
 - **Shape**
 - **Outliers**

Measures of Centre

- **Mean:** The mean is often called the average or the expected value in mathematics.
 - We calculate the mean by adding all of our values together and dividing by the number of values in our dataset.

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

Example

- Consider the amount of time each of the 5 users spent on our website:

$$x_1=10, x_2=20, x_3=45, x_4=12, x_5=8$$

$$\text{The mean is } \bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = \frac{10+20+45+12+8}{5} = 19$$

Measures of Centre

- **Median:** The median is a value that divides our dataset such that 50% of the values are larger while the remaining 50% are smaller.
 - Sort the values first
 - Odd number of observations:
Median = the number in the direct middle
 - Even number of observations:
Median = the average of the two values in the middle

Measures of Centre

Example

- Consider the amount of time each of the 5 users spent on our website. What is the median duration users spent on our website?

$$x_1=10, x_2=20, x_3=45, x_4=12, x_5=8$$

1. Sort the values first

8, 10, 12, 20, 45

2. Odd number of observations (5 observations):

Median = the number in the direct middle = 12

Measures of Centre

Example

- Consider the amount of time each of the 6 users spent on our website. What is the median duration users spent on our website?

$$x_1=10, x_2=20, x_3=45, x_4=12, x_5=8, x_6=2$$

1. Sort the values first

2, 8, 10, 12, 20, 45

2. Even number of observations (6 observations):

Median = the average of the two values in the middle

$$= (10+12) / 2 = 11$$

Measures of Centre

- **Mode:** The mode is the most frequently observed value in our dataset.
 - There might be multiple modes for a particular dataset, or no mode at all.
 - If all observations in our dataset are observed with the same frequency, there is no mode.

Example

- If we have the dataset: {1, 1, 2, 2, 3, 3, 4, 4}
 - There is no mode because all observations occur the same number of times.
- If we have the dataset: {1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9}
 - There are two modes 3 and 6, because these values share the maximum frequencies at 3 times, while all other values only appear once.

Measures of Centre

Exercise

- If we have the data: {5, 8, 15, 7, 10, 22, 3, 1, 15}. What is the mean, median and mode?

Go to
www.menti.com

Enter the code

6195 8382



Or use QR code

Measures of Centre

Case Study

- Suppose some of my friends have dogs at home. Ashley has 1 dog, Steve has 1 dog, Jeff has 2 dogs, Kylie has 3 dogs, and Lisa has 8 dogs. We want to summarise the number of dogs our friends have into a single number. We will use the measures of centre for this problem.
- What is the mean, median, and mode for the number of dogs our friends have?

Answer: Mean=3, Median=2, Mode=1

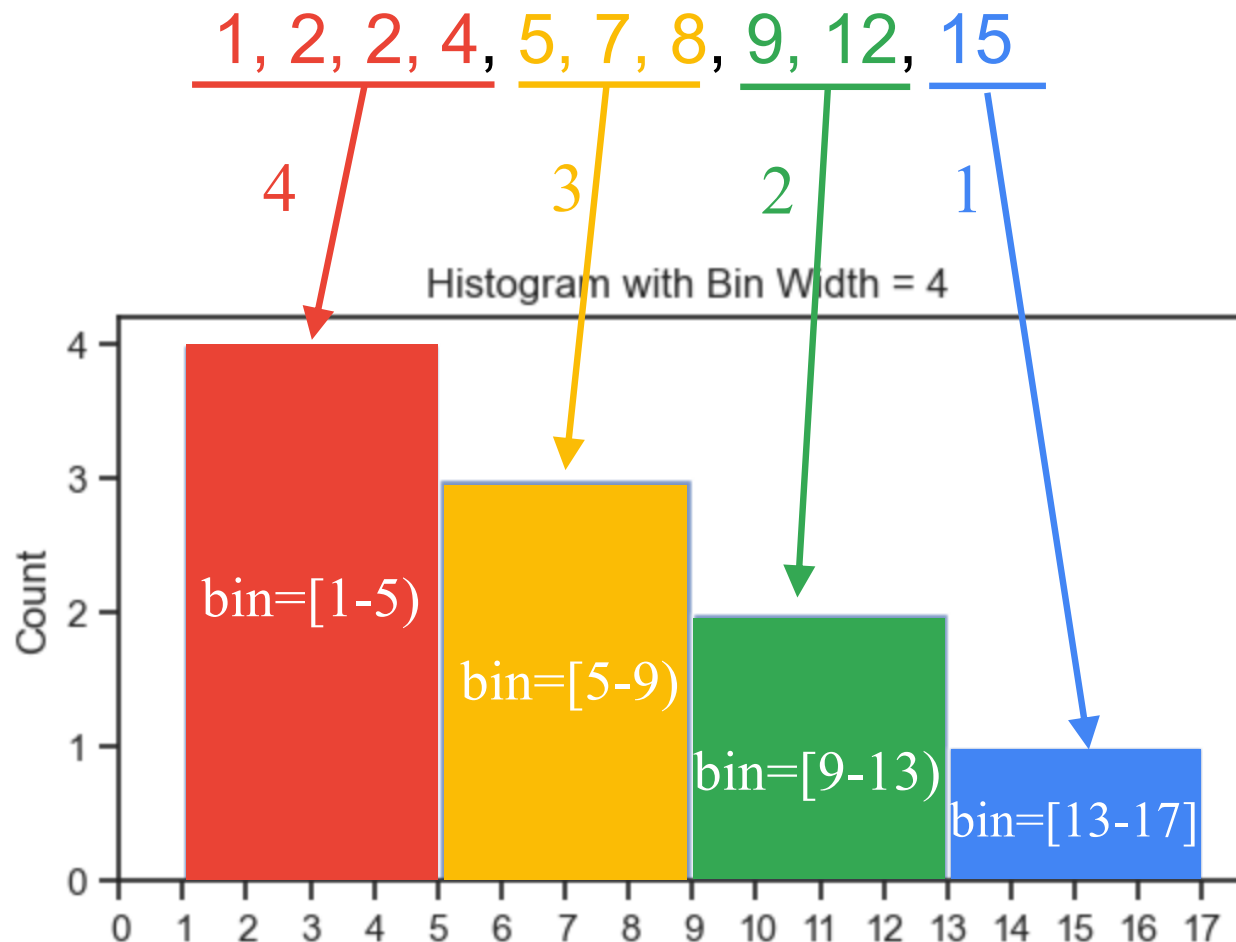
Discussion: There is no measure of centre that is always best, so we need to try all three to see what makes sense in this situation.

Four Aspects for Quantitative Data

- There are four main aspects to analyzing Quantitative data.
 - **Centre**
 - **Spread**
 - **Shape**
 - **Outliers**

Histograms

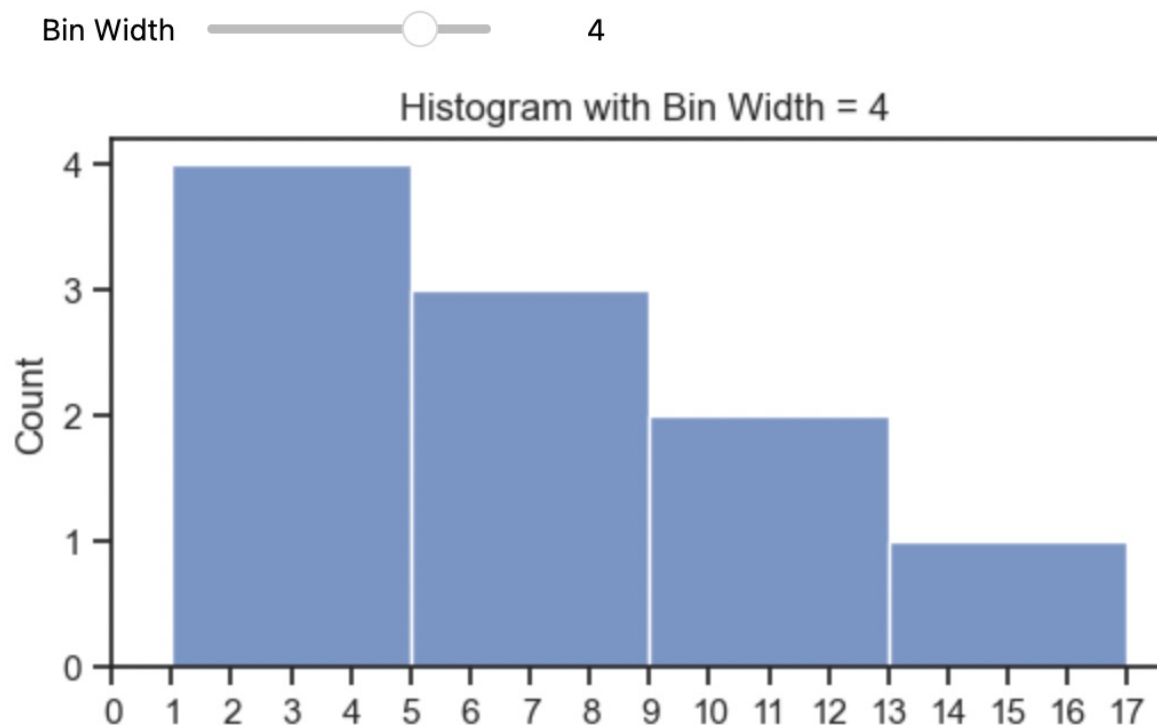
- The histogram provides a visual representation of the distribution of the data, showing the number of observations that fall within each bin.



Histograms

Demo

- Suppose we have discrete data $\{1, 2, 2, 4, 5, 7, 8, 9, 12, 15\}$. Let's plot a histogram to show the distribution of the data. Then we'll change the bin width and find if there is any change in the shape of the histogram.



Quantiles

- In statistics and probability, **quantiles** are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.
 - Common quantiles have special names, such as **quartiles** (four groups), **deciles** (ten groups), and **percentiles** (100 groups).
- In a dataset, the P th **percentile** is a value such that P percent of the values fall below this point.
 - The percentile is a special case of a quantile, with quantiles indexed by fractions. For example, the 0.8 quantile is the same as the 80th percentile.
- **Quartiles** are actually the 25th, 50th, and 75th percentiles.
- **Deciles** are actually the 10th, 20th, ..., 90th percentiles.

Quantiles

Example

- 0.5 quantile = 50th percentile = median
- 0.25 quantile = 25th percentile = first quartile

Measures of Spread

- **The Five-Number Summary**

- **Minimum:** The smallest number in the dataset
- **Q_1 :** The value such that 25% of the data falls below, which is the median of the data on the left side of Q_2 (median)
- **Q_2 :** The value such that 50% of the data falls below, which is the median of the dataset
- **Q_3 :** The value such that 75% of the data falls below, which is the median of the data on the right side of Q_2 (median)
- **Maximum:** The largest value in the dataset.

Measures of Spread

- **Range**

- The **range** is calculated as the difference between the maximum and the minimum.

- **Interquartile Range (IQR)**

- The **interquartile range** is calculated as the difference between Q_3 and Q_1
 - Q_3 and Q_1 can be considered as the medians of the data on either side of Q_2

Measures of Spread

Example

- Find the five-number summary, range and IQR for the following dataset:

{5, 8, 3, 2, 1, 3, 10}

- Answer:

1. Order the data: 1, 2, 3, 3, 5, 8, 10

2. Min=1, Max=10

3. Q2 (Median) = 3

4. Q1 is the median of 1, 2, 3, which is 2

5. Q3 is the median of 5, 8, 10, which is 8

6. Summary (Min, Q1, Q2, Q3, Max) = (1, 2, 3, 8, 10)

7. Range = $10 - 1 = 9$

8. IQR = $Q3 - Q1 = 6$

Measures of Spread

Exercise

- Find the five-number summary, range and IQR for the following dataset:
 $\{5, 8, 3, 2, 1, 3, 10, 105\}$

Go to
www.menti.com

Enter the code

6195 8382



Or use QR code

Box Plots

- A **box plot**, also known as a **box-and-whisker plot**, is a graphical representation of the distribution of a dataset through its quartiles.
- A boxplot is a standardized way of displaying the dataset based on the five-number summary: the “minimum”, the “maximum”, the sample median, and the first and third quartiles.

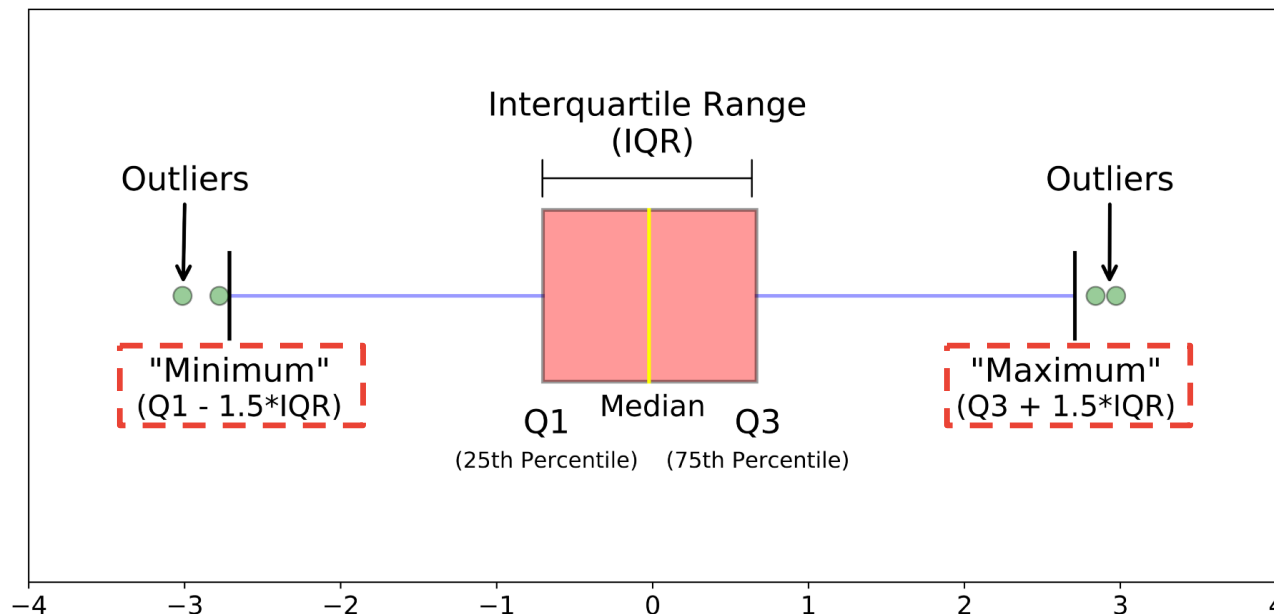


Image source: <https://www.simplypsychology.org/boxplots.html>

Box Plots

- Commonly used in statistical analysis and data visualization to gain insights into the distribution and spread of the data.
- Provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers. It is easy to see where the main bulk of the data is, and to make the comparison between different groups easier than using histograms.

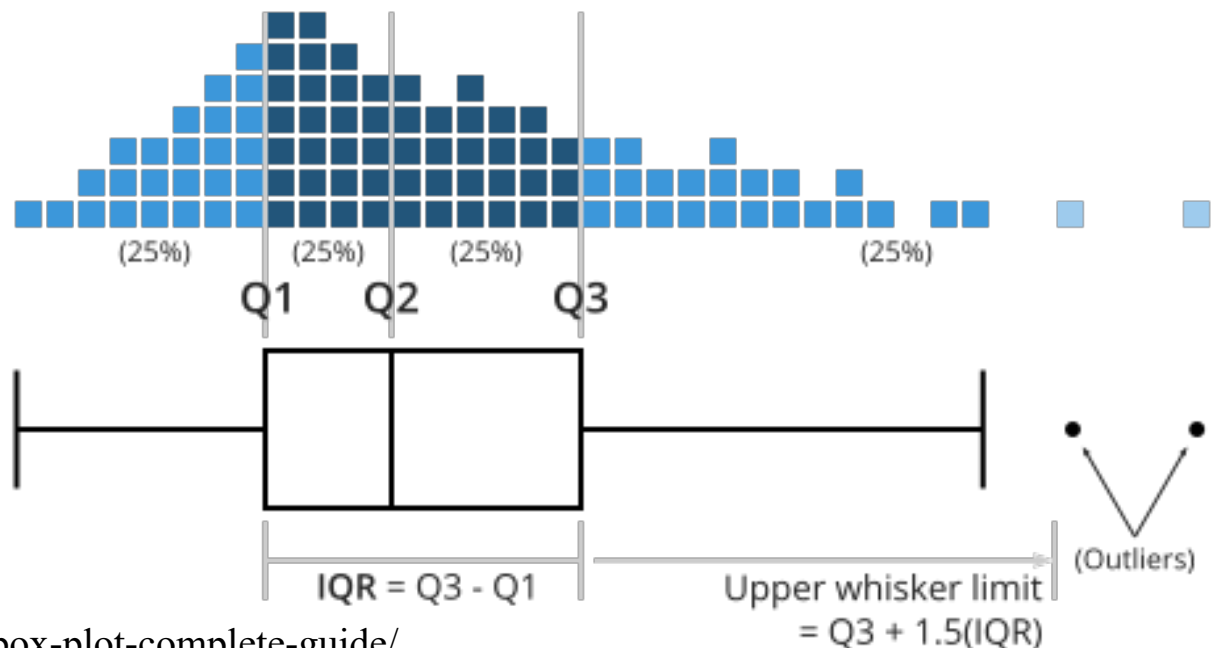
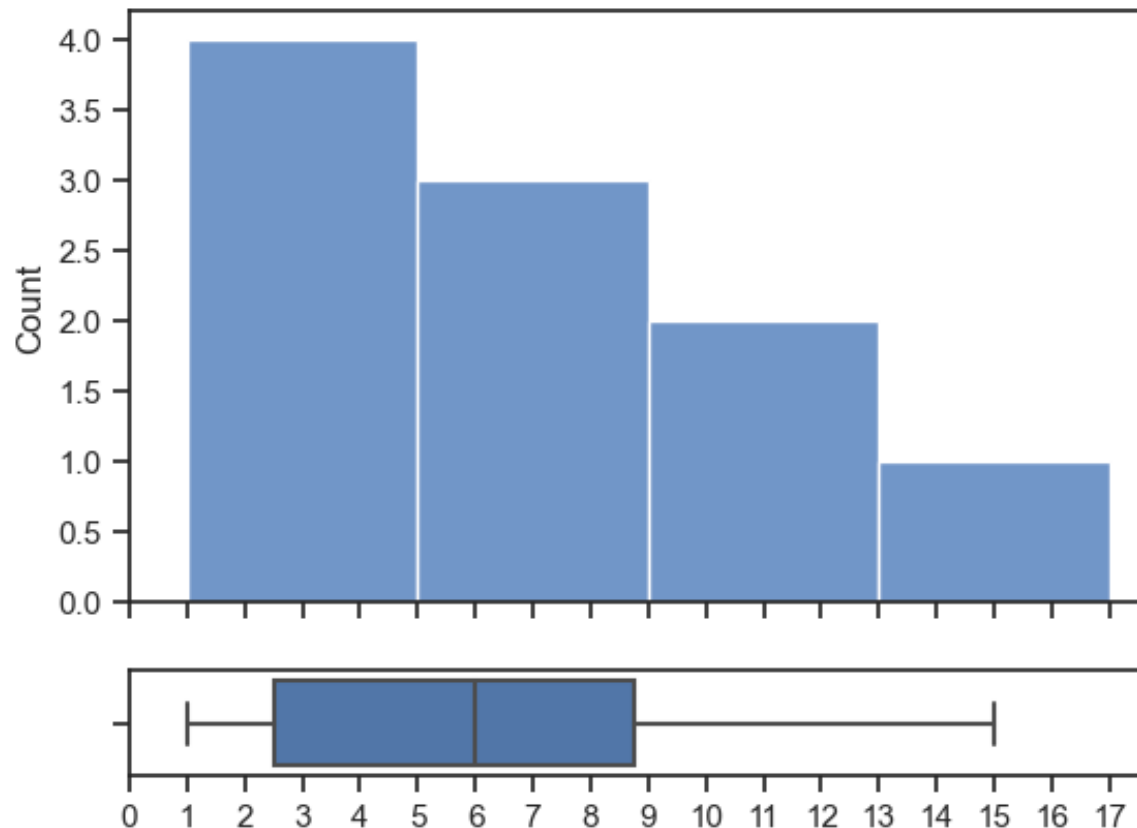


Image source: <https://chartio.com/learn/charts/box-plot-complete-guide/>

Box Plots

Demo

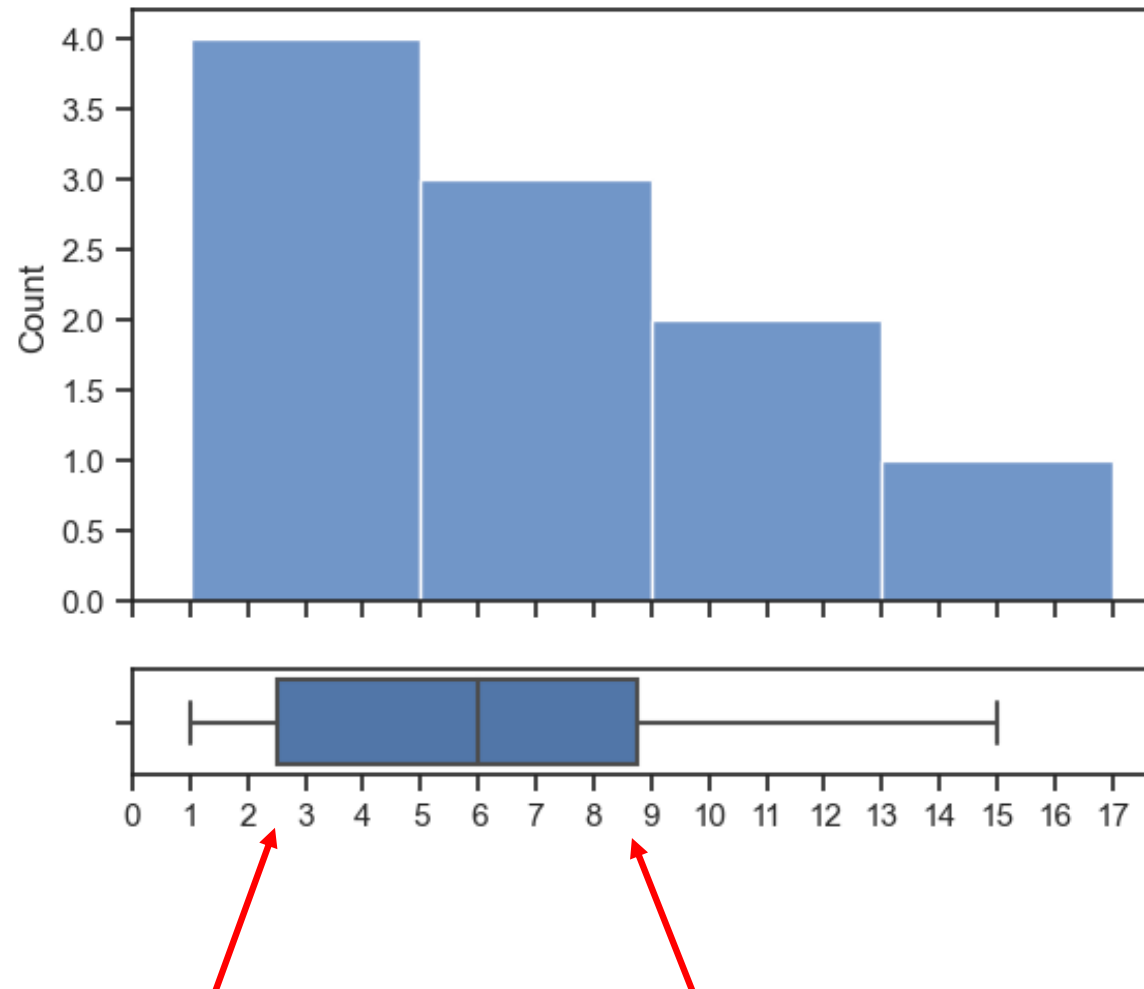
- Suppose we have discrete data $\{1, 2, 2, 4, 5, 7, 8, 9, 12, 15\}$. It is the same dataset we used to plot the histogram in the histogram demo. Let's plot a box plot along with the histogram.



Box Plots

Discussion

- Are the values of Q_1 and Q_3 the same as you expect?



Measures of Spread

- The **standard deviation** or **variance** is one of the most common ways to measure the spread of data with **a single value**.

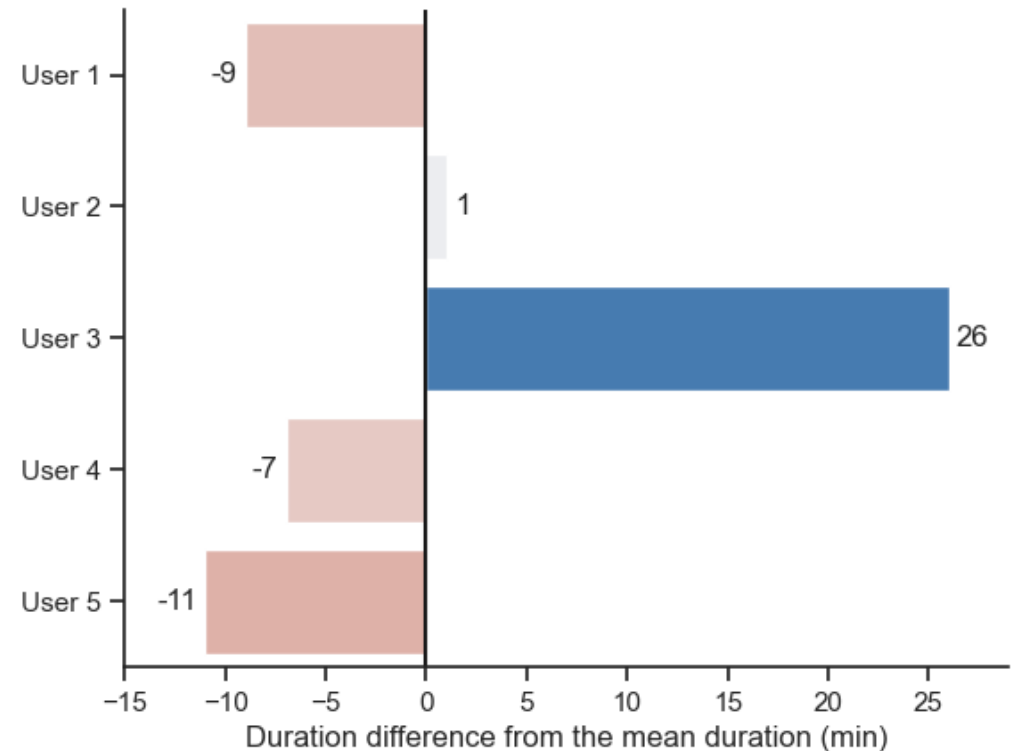
- Standard Deviation**

- The **standard deviation** is defined as the average distance of each observation from the mean.

- Variance**

- The **variance** is the average squared difference of each observation from the mean.

The amount of time each user spent on the website. (mean = 19 min)



Measures of Spread

- **Variance**

- The **variance** is the average squared difference of each observation from the mean.

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- **Standard Deviation**

- The **standard deviation** is defined as the average distance of each observation from the mean.

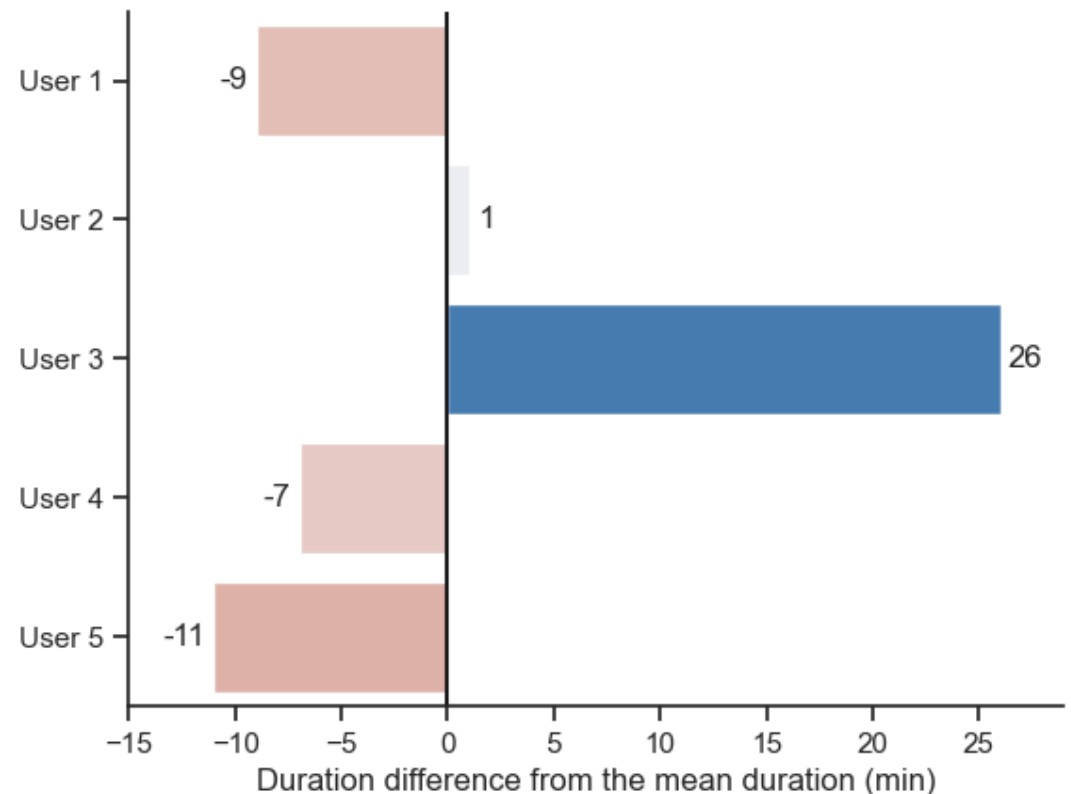
$$s = \sqrt{\text{Variance}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Measures of Spread

Example

- Consider the amount of time each of the 5 users spent on our website. Find the standard deviation and variance.

$$x_1=10, x_2=20, x_3=45, x_4=12, x_5=8$$



Answer:

The variance and standard deviation can be calculated as

$$s^2 = \frac{1}{5} [(-9)^2 + 1^2 + 26^2 + (-7)^2 + (-11)^2] = 185.6$$

$$s = \sqrt{185.6} = 13.62$$

Measures of Spread

- **Important Notes**

- The standard deviation or variance is used to compare the spread of two different groups. A set of data with a higher standard deviation/variance is more spread out than a dataset with a lower standard deviation/variance.
- When comparing the spread between two datasets, the units of each must be the same.
- When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.
- The standard deviation is used more often in practice than the variance, because it shares the units of the original dataset.

Standard Deviation & Variance

Case Study: Applied Standard Deviation & Variance

- Investment Data: Consider we have two investment opportunities. The returns for 6 consecutive years for each investment are shown above. Use this information to answer the questions below.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Investment 1	5%	5%	5%	5%	5%	5%
Investment 2	12%	-2%	10%	0%	7%	3%

- Use the information above to find the mean/expected return for each investment.

Standard Deviation & Variance

Case Study: Applied Standard Deviation & Variance

- Investment Data: Consider we have two investment opportunities. The returns for 6 consecutive years for each investment are shown above. Use this information to answer the questions below.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Investment 1	5%	5%	5%	5%	5%	5%
Investment 2	12%	-2%	10%	0%	7%	3%

2. Which statement(s) is true?

- (1) The risk associated with Investment 1 is lower than the risk associated with Investment 2
- (2) The standard deviation associated with Investment 1 is smaller than the standard deviation associated with Investment 2
- (3) Knowing the mean return across all the years for each investment provides us with all of the information necessary to understand which investment we should choose

Standard Deviation & Variance

Case Study: Applied Standard Deviation & Variance

- Investment Data: Consider we have two investment opportunities. The returns for 6 consecutive years for each investment are shown above. Use this information to answer the questions below.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Investment 1	5%	5%	5%	5%	5%	5%
Investment 2	12%	-2%	10%	0%	7%	3%

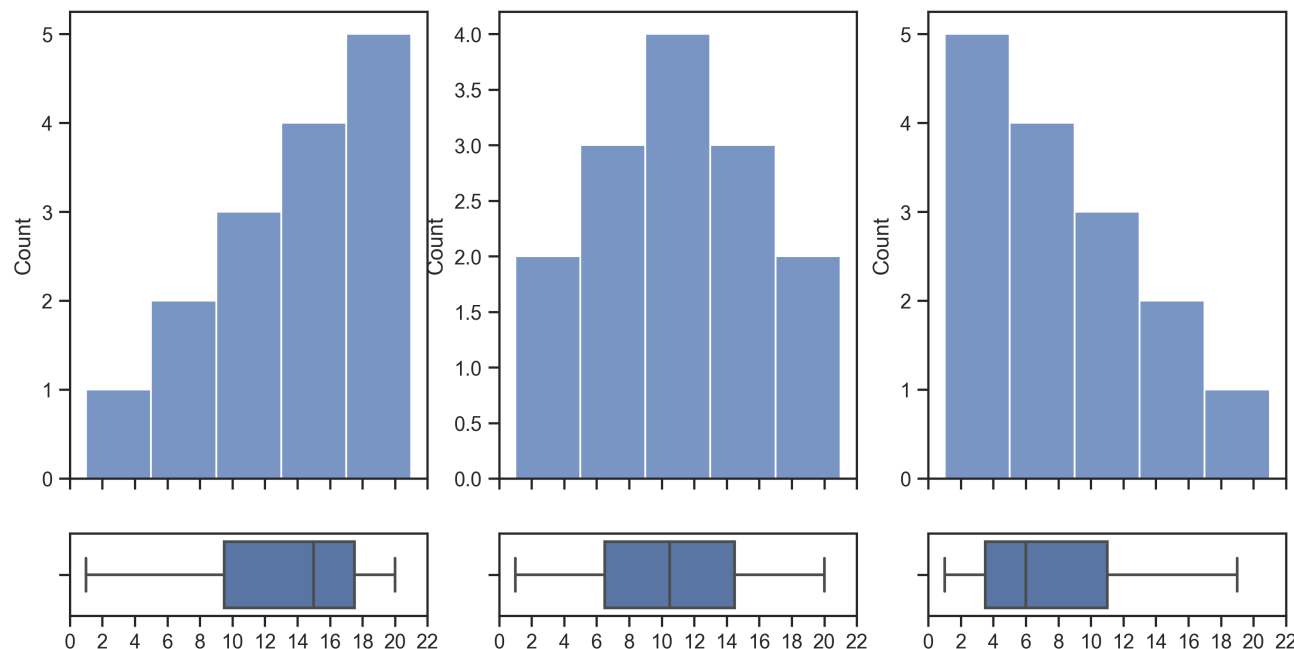
3. Based on the observed data, which of the above two investments has the best opportunity of earning more than 7%?

Four Aspects for Quantitative Data

- There are four main aspects to analyzing Quantitative data.
 - **Centre**
 - **Spread**
 - **Shape**
 - **Outliers**

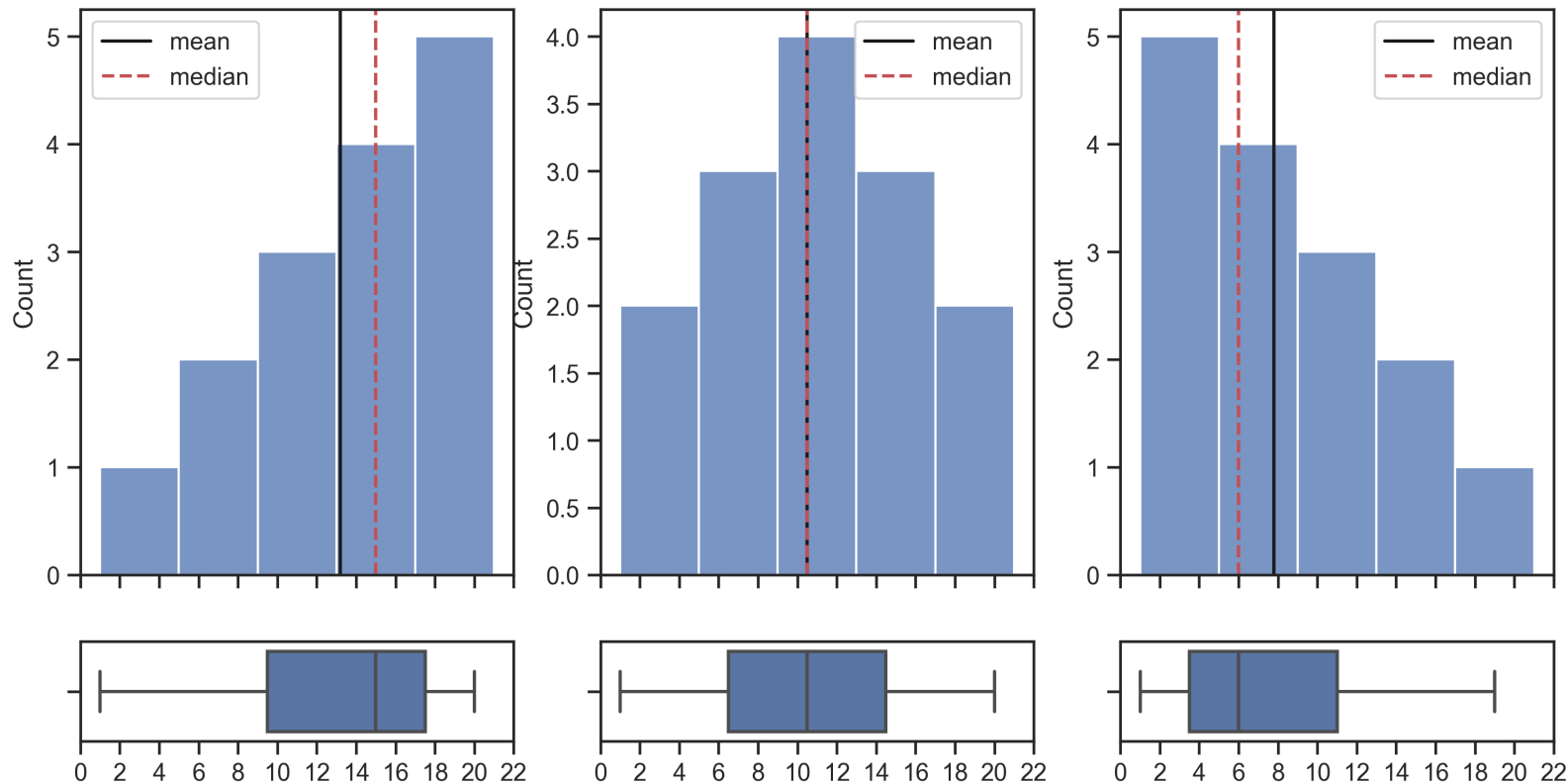
Shape

- The distribution of our data is frequently associated with one of the three shapes:
 - **Right-skewed**: The distribution has a longer **right** tail. The longer whisker is on the right
 - **Left-skewed**: The distribution has a longer **left** tail. The longer whisker is on the left
 - **Symmetric** (frequently normally distributed)



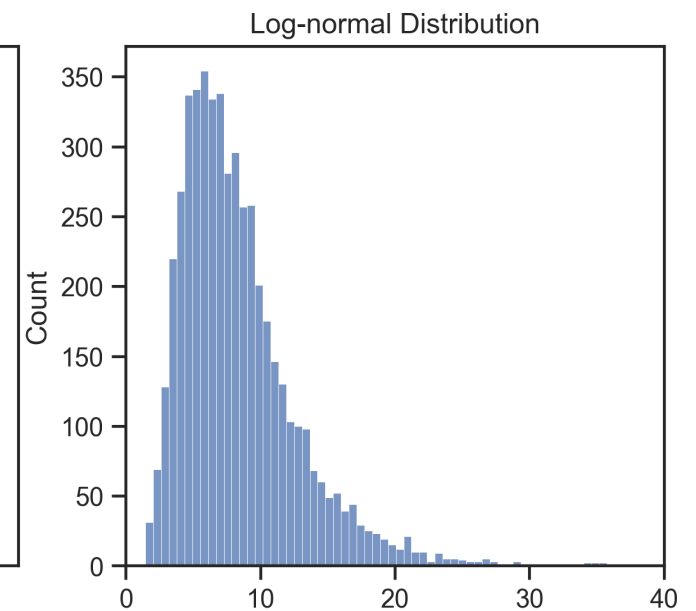
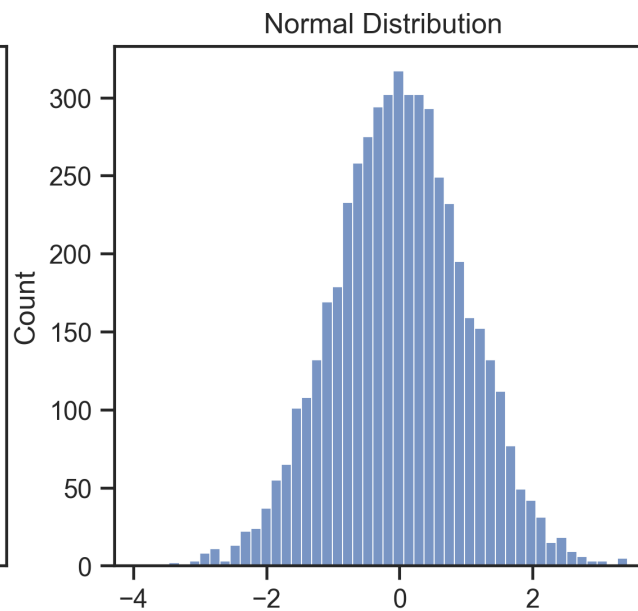
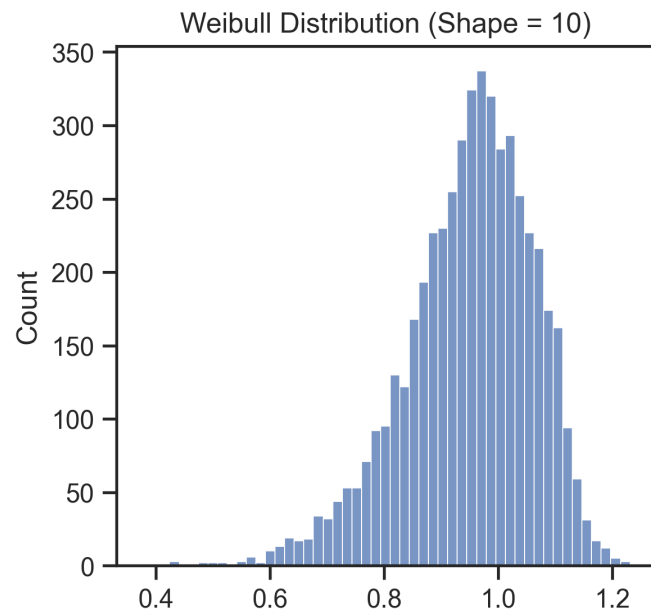
Shape

- Centres of the three shapes:
 - Right-skewed: Mean $>$ Median
 - Left-skewed: Mean $<$ Median
 - Symmetric (frequently normally distributed): Mean \approx Median



Shape

- Some distributions that are:
 - Right-skewed: Log-normal Distribution
 - Left-skewed: Weibull Distribution (when shape parameter > 3.7)
 - Symmetric: Normal Distribution



Shape

- **Summary**

- Right-skewed, Left-skewed and Symmetric distributions are the most common distributions.
- Data in the real world can be messy and it might not follow any of these distributions.

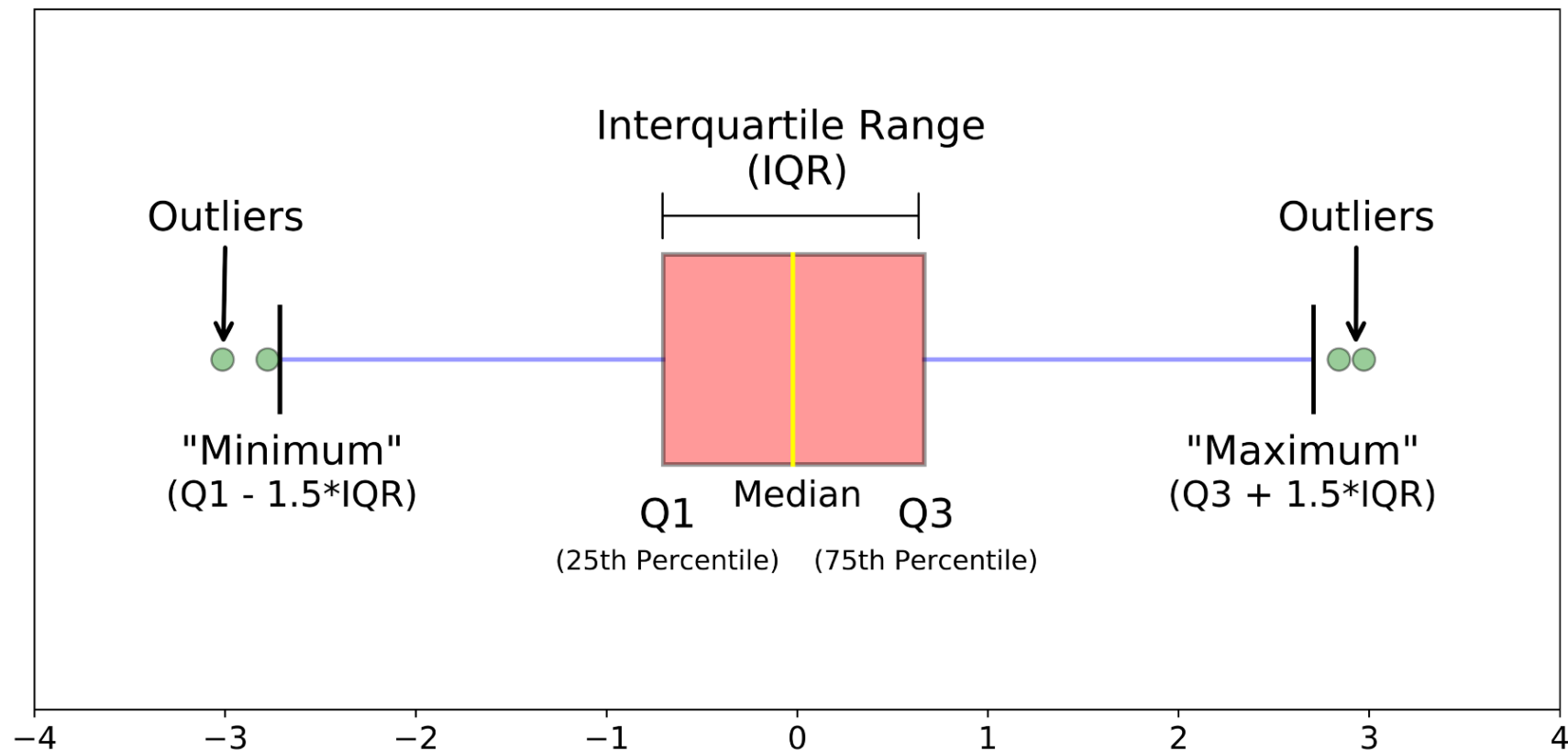
Shape	Mean vs. Median	Examples
Symmetric (Normal)	Mean \approx Median	Height, Weight, Errors, Precipitation
Right-Skewed	Mean $>$ Median	Amount of drug remaining in a bloodstream
Left-Skewed	Mean $<$ Median	Age of Death

Four Aspects for Quantitative Data

- There are four main aspects to analyzing Quantitative data.
 - **Centre**
 - **Spread**
 - **Shape**
 - **Outliers**

Outliers

- An outlier is any value that falls very far from the rest of the values in the dataset.
- The exact definition of an outlier is somewhat subjective, although certain conventions are used in various data summaries and plots, e.g., the box plot.

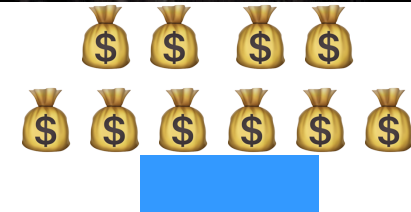
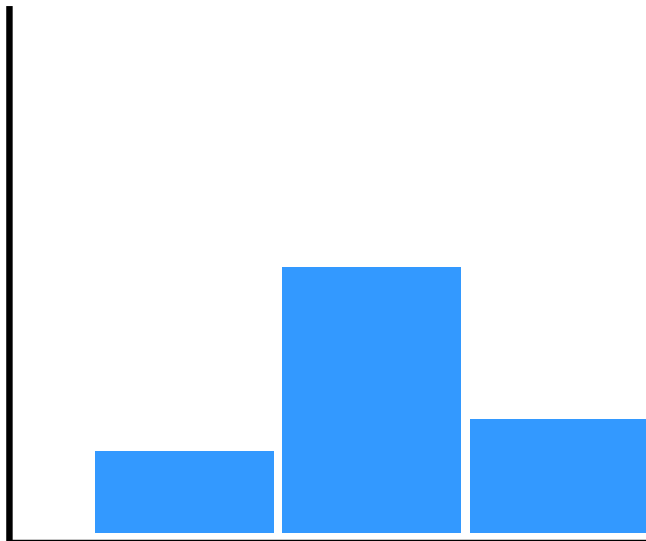


Source: <https://www.simplypsychology.org/boxplots.html>

Outliers

Case Study: Annual Earnings of Entrepreneurs

- Suppose we want to research the annual earnings of entrepreneurs, and we are going to select ten entrepreneur earnings to form our dataset. Nine values of earnings in thousands of dollars are: {45, 68, 92, 53, 105, 56, 24, 15, 155}
- The tenth value we select is the earning of Elon Musk. According to Forbes, he gained 68 billion dollars (68000000k dollars) in 2022, which is considered an outlier in the dataset.



Source: <https://www.forbes.com/profile/elon-musk/?sh=1a4486ca7999>

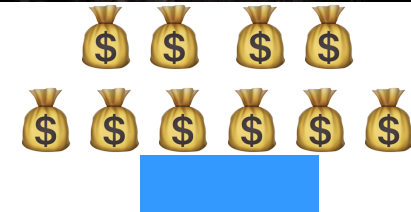
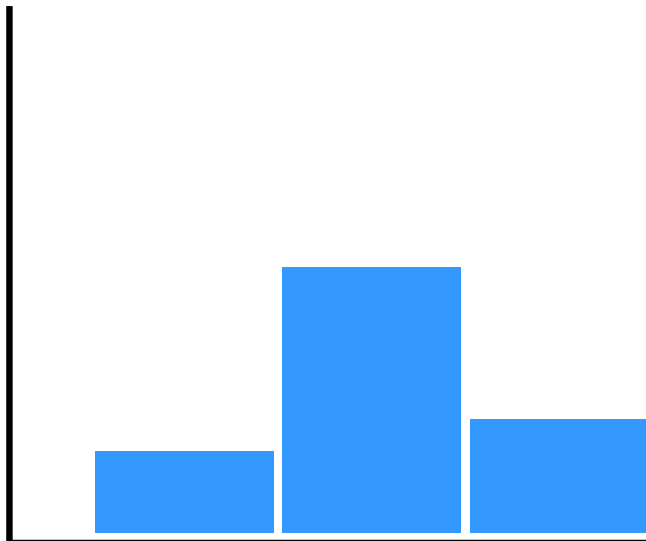
Image Source:

<https://www.gizchina.com/2022/04/05/forbes-elon-musk-is-now-the-richest-man-in-the-world/>

Outliers

Case Study: Annual Earnings of Entrepreneurs

- If we use the mean (6.8 billion) to describe the centre of the dataset, it can be misleading as no entrepreneur earned this or is close to it.
- A better measure of centre in this case is the median (62k dollars), it is a better indication of how much an entrepreneur is likely to earn based on our data.



Source: <https://www.forbes.com/profile/elon-musk/?sh=1a4486ca7999>

Image Source:
<https://www.gizchina.com/2022/04/05/forbes-elon-musk-is-now-the-richest-man-in-the-world/>

Outliers

Work with Outliers

- When outliers are present we should consider the following points.
 - Noting they exist and the impact on summary statistics.
 - If they are the result of data errors such as mixing data of different units (kilometers versus meters) or bad readings from a sensor - fix or remove them.
 - Understanding why they exist, and the impact on questions we are trying to answer about our data. When outliers are the result of bad data, the mean will result in a poor estimate of centre, while the median will still be valid.
 - Be careful in reporting. Reporting the five number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.
 - 1. If no outliers and your data follow a normal distribution - use the mean and standard deviation to describe your dataset, and report that the data are normally distributed.
 - 2. If you have skewed data or outliers, use the five-number summary to summarize your data and report the outliers.

Outliers

| Work with Outliers

- In any case, outliers should be identified and are usually worthy of further investigation.
- There's an entire field about identifying and measuring outliers called **anomaly detection**

ANALYSING CATEGORICAL DATA

Analysing Categorical Data

- **Categorical data** is usually analysed by looking at the counts or proportion of individuals that fall into each group.
 - For example, if we look at the menti multiple choice questions, we would care about how many students chose each option, or what proportion of students chose each option.

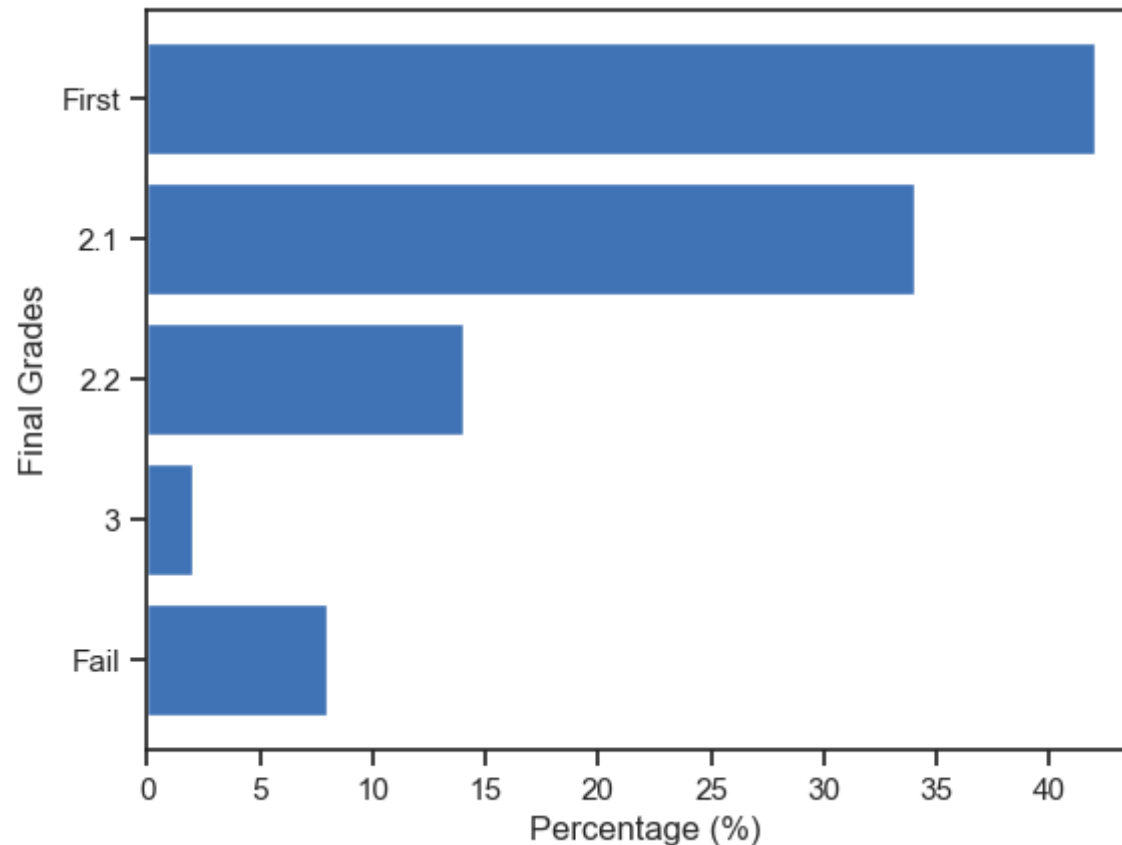
Example

- The table below shows the percentage of the final grades for JP 2018 cohort (graduated in 2022).

First	2.1	2.	3	Fail
42%	34%	14%	2%	8%

Bar Charts

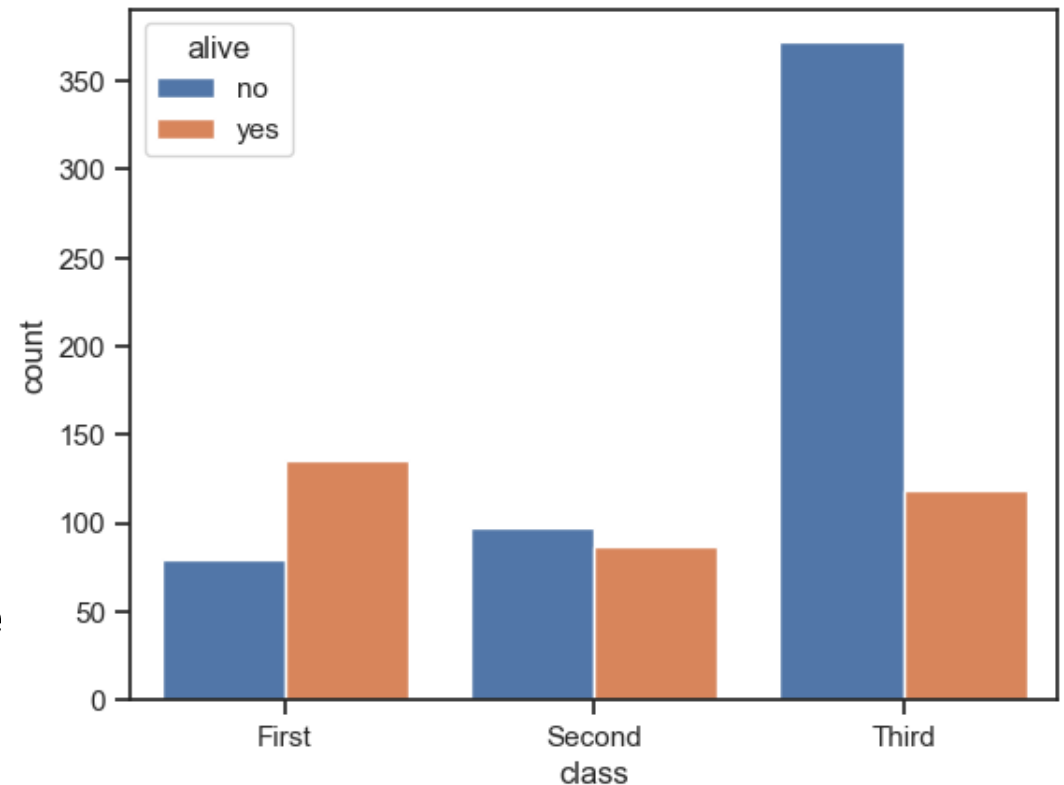
- **Bar charts** are a common visual tool for displaying a single categorical variable.
 - Categories are listed on the x-axis, and frequencies or proportions on the y-axis.



Bar Charts

Demo

- Let's use the famous Titanic dataset to demonstrate a variant of the bar chart: the **grouped bar chart**, also known as a **clustered bar chart**
- A **grouped bar chart** is a type of bar chart that displays multiple bars side by side within each category or group. Each group represents a distinct category, and the bars within the group represent different sub-categories or data series.



Bar Charts

- **Differences between a bar chart and a histogram**
 - In a bar chart, the x-axis represents different categories of a factor variable, while in a histogram the x-axis represents values of a single variable on a numeric scale.
 - In a histogram, the bars are typically shown touching each other, with gaps indicating values that did not occur in the data. In a bar chart, the bars are separated from one another.

RECAP

Descriptive Statistics

Recap

- **Data/Variable Types**
 - Quantitative
 - Continuous
 - Discrete
 - Categorical
 - Ordinal
 - Nominal

Recap

- **Quantitative Data Analysis**

Four main aspects used to describe **quantitative** variables:

- Measures of Centre

- Mean
- Median
- Mode

- Measures of Spread

- Five-number Summary
- Range
- Interquartile Range (IQR)
- Variance
- Standard Deviation

- Shape

- Mean
- Median
- Mode

- Outliers

- Visualisation

- Histograms
- Box Plots

Recap

- **Categorical Data Analysis**
 - Count/Proportion/Percentage of each category
 - Mode
 - Visualisation
 - Bar charts

Questions

Use student forum on QM+

chao.shu@qmul.ac.uk

xiaolanliu@qmul.ac.uk