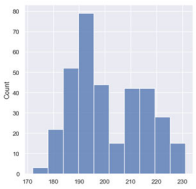
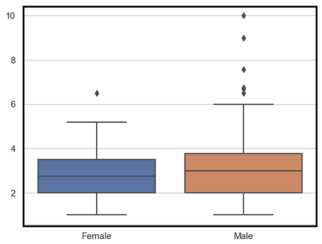
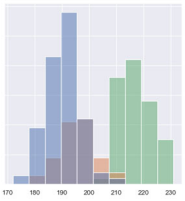
Chapter 5.4 – Visualising Distribution in Data

Contents

- Basic Distribution Plots
- Histogram
- Box Plots
- Violin Plots

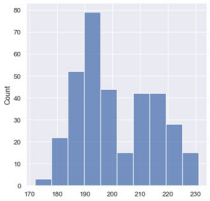


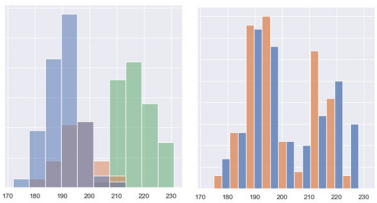

1

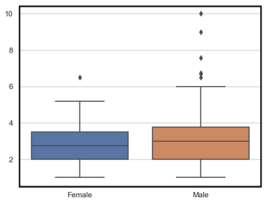
Basic Distribution Plots

Visualising Distribution in the Data

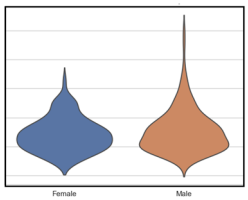
- Distribution plots gives a visual insight into how the data **values are distributed** within the dataset.
- **Histograms** – are often used to visualise the distribution of a **single variable**. Multiple variables[1] are possible but may suffer from occlusion and clutter.
- **Box & Violin Plots** – are useful for exploring the distribution in **multiple variables** simultaneously. The violin plot visualises the **density curve** of the variable, whereas the box plot visualises the **median** and **interquartile range** of each variable.


Standard Histogram


Histogram with Multiple Categories



Box Plot



Violin Plot

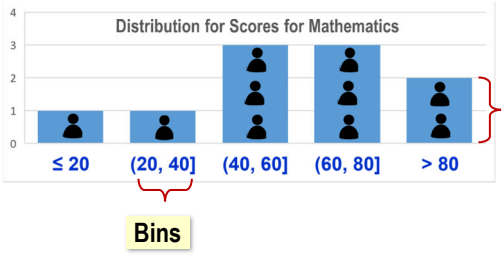
[1] Visualizing distributions of data (2019), https://seaborn.pydata.org/tutorial/distributions.html

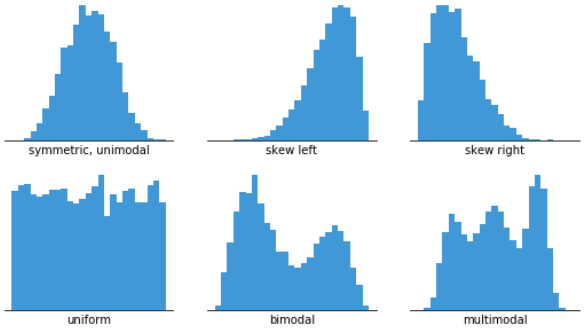
2

Histograms

Visualising Data Distribution

- Histograms plot the distribution of a variable's numeric values as a series of bars. Each bar covers a range of numeric values (**bin**) and its **height** encodes the **frequency** of data points that have values within the corresponding bin.
- Histograms reveals the general distributional features of variables, their positions of peaks, whether the spread is skewed, symmetric, and if there are any outliers[2].





[2] Mike Yi, A Complete Guide to Histograms (2019), https://chartio.com/learn/charts/histogram-complete-guide/

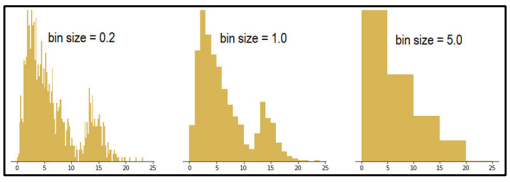
3

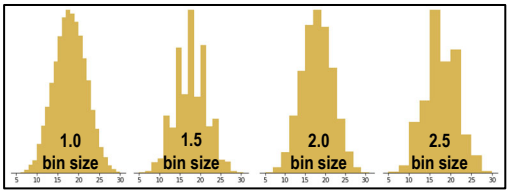
3

Histograms

Selecting Number of Bins and Boundaries

- The bin count and their boundaries for tallying data points must be chosen properly as it can have an effect of the **shape** of the plot and how it is interpreted[2].
- **Too many** bins may result in a **noisy** plot that is difficult to discern general data distribution. **Too few** bins may **reduce details** needed to discern patterns in the data (e.g. a double peak distribution).
- **Bin sizes** should match the **type of values** the data can have. Fractional bin sizes (e.g. 1.5) may not be suitable for integer variables as not all bins can take the same numbers of valid data values, leading to "bumpy" looking plots[2].





NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[2] Mike Yi, A Complete Guide to Histograms (2019), https://chartio.com/learn/charts/histogram-complete-guide/

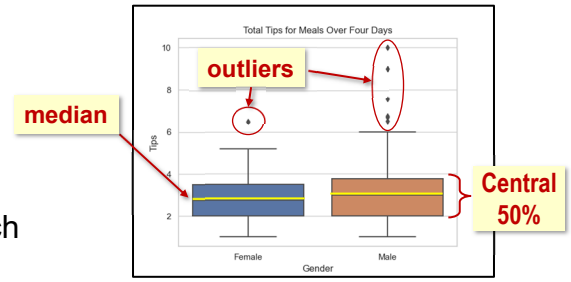
4

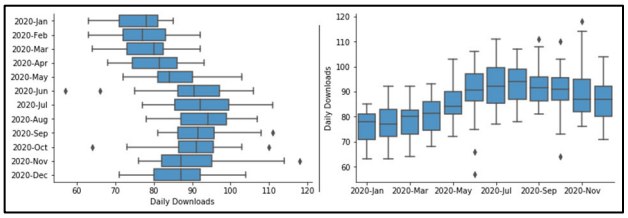
4

Box Plots

Visualising Statistical Properties of Multiple Variables

- Box plots (whisker plot) uses boxes and lines to depict the distributions of one or more numeric variables.
- The **box limits** show the range of the **central 50%** of the data, with a central line marking its **median**. Lines extend from each box to capture the **range** of the remaining data, with dots placed past the line edges to indicate outliers.
- Box plots allow multiple variables to be visualised either vertically or horizontally[3].





[3] Mike Yi, A Complete Guide to Box Plot (2019), https://chartio.com/learn/charts/box-plot-complete-guide/

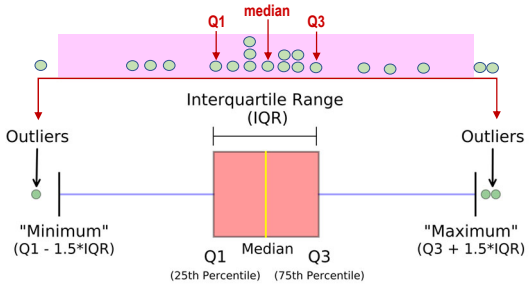
5

5

Box Plots

Constructing the Box and Whiskers

- Box plots display data variable's quartiles (values that divide the dataset into equal fourths).
- The 1st quartile (**Q1**) is greater than 25% of the data. The 2nd quartile (**Q2**), the **median**, sits in the middle, dividing the data in half. The 3rd quartile (**Q3**) is larger than 75% of the data. The box boundaries and its centre line mark the locations of Q1, Q2 and Q3[4].



- The left and right whiskers are **1.5 times** the interquartile range (IQR) from Q1 and Q3 respectively, but not exceeding data points at the extremities. All points outside these limits are considered **outliers** and are plotted individually horizontally[4].

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[4] M. Galarnyk, Understanding Boxplots (2018), https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

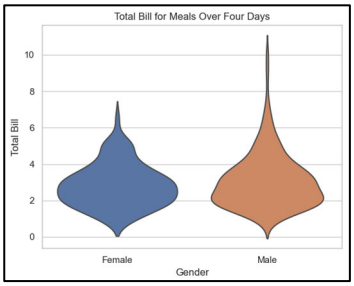
6

6

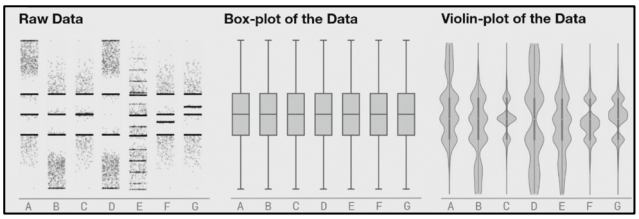
Violin Plots

Visualising Kernel Density Curves of Multiple Variables

- Violin plots visualise distributions of numeric data values using kernel density curves, whose widths encode the approximate frequency of data points in each region[5].
- Distribution comparison between multiple groups can be done as the peaks, valleys, and tails in of their respective density curves can be contrasted by the side-by-side plots.



- Datasets can be modified in a way that the quartiles do not change, but the shape of the distribution differs dramatically[6]. In such cases, violin plots give **more insights** into the **data distribution** than box plots.



[5] Mike Yi, A Complete Guide to Violin Plots (2019), https://chartio.com/learn/charts/violin-plot-complete-guide/

[6] J. Matejka, G. Fitzmaurice, Same Stats, Different Graphs (CHI'17), https://www.autodesk.com/research/publications/same-stats-different-graphs

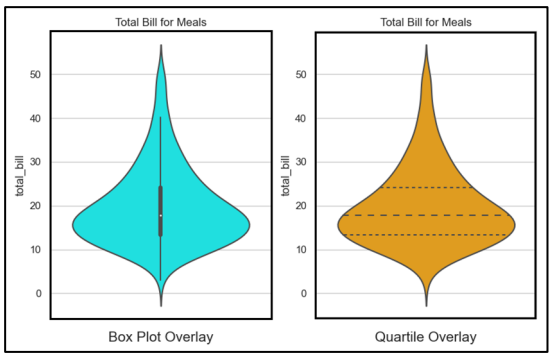
7

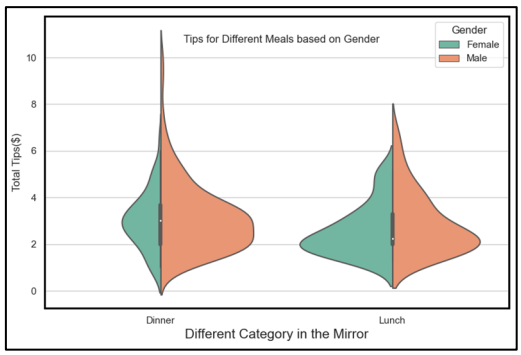
7

Violin Plots

Getting the Best of Two Worlds

- Violin plots have additional flexibility compared to box plots. They can also be overlayed with additional information such as the **box plot** or **quartile lines**.
- Each **mirrored half** of the violin plot can be used to encode the distribution of another **complementary** data category (e.g. females on the left and males on the right)[7].

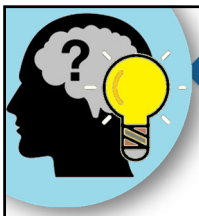




[7] Seaborn, Violin Plots, https://seaborn.pydata.org/generated/seaborn.violinplot.html

8

8

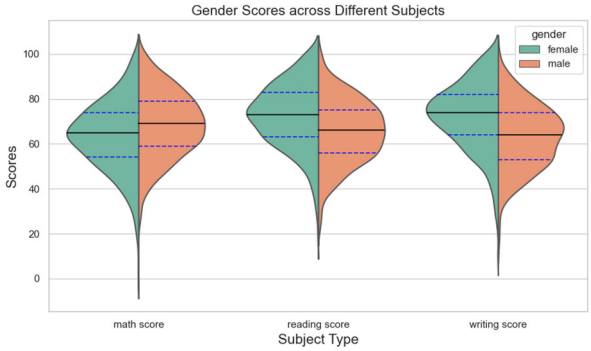


Think and Apply

What Affects Academic Performance?

- Analyse the math, reading and writing scores of 1000 students with different backgrounds, gender and preparedness. Looking at various score distributions, answer these questions:

  - What subjects do the male students do better in?
  - Which subject was the most difficult?
  - Did taking the prep course help improve scores?
  - Do ethnicity and parent's educational background influence performance?



Summary

Distribution Plots

- Distribution plots allow us to visualise the way the data values in the dataset is **distributed** across the permissible range.
- The **histogram** is a bar chart that shows the distribution of data values. Its shape can **reveal much more** (e.g. skewed, bimodal, etc) than what typical statistical measures like mean and variance can inform.
- **Box plots** is useful for visualising the **statistical** properties of **multiple variables** simultaneously.
- **Violin plots** allow the **kernel density curves** of multiple variables to be visualised and compared in a **pairwise** manner.

References for Distribution Plots

[1] Visualizing distributions of data (2019), https://seaborn.pydata.org/tutorial/distributions.html

[2] Mike Yi, A Complete Guide to Histograms (2019), https://chartio.com/learn/charts/histogram-complete-guide/

[3] Mike Yi, A Complete Guide to Box Plot (2019), https://chartio.com/learn/charts/box-plot-complete-guide/

[4] M. Galarnyk, Understanding Boxplots (2018), https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

[5] Mike Yi, A Complete Guide to Violin Plots (2019), https://chartio.com/learn/charts/violin-plot-complete-guide/

[6] J. Matejka, G. Fitzmaurice, Same Stats, Different Graphs (CHI'17), https://www.autodesk.com/research/publications/same-stats-different-graphs

[7] Seaborn, Violin Plots, https://seaborn.pydata.org/generated/seaborn.violinplot.html

Note: All online articles were accessed between May to June 2021