STA130 Week2

TA: Gloria

Contact: shiyi.hou@mail.utoronto.ca

Vocabularies List

- Where are the data centered (approximate values if available)
- How much spread (relative to what?)
- Shape: symmetric, left-skewed, right-skewed
- The tails of the distribution (heavy-tailed or thin-tailed)
- Modes: where, how many, unimodal, bimodal, multimodal, uniform
- Outliers, extreme values
- Frequency (which category occurred the most or least often; data concentrated near a particular value or category)
- Mean, median, mode
- Standard deviation, interquartile range (IQR)

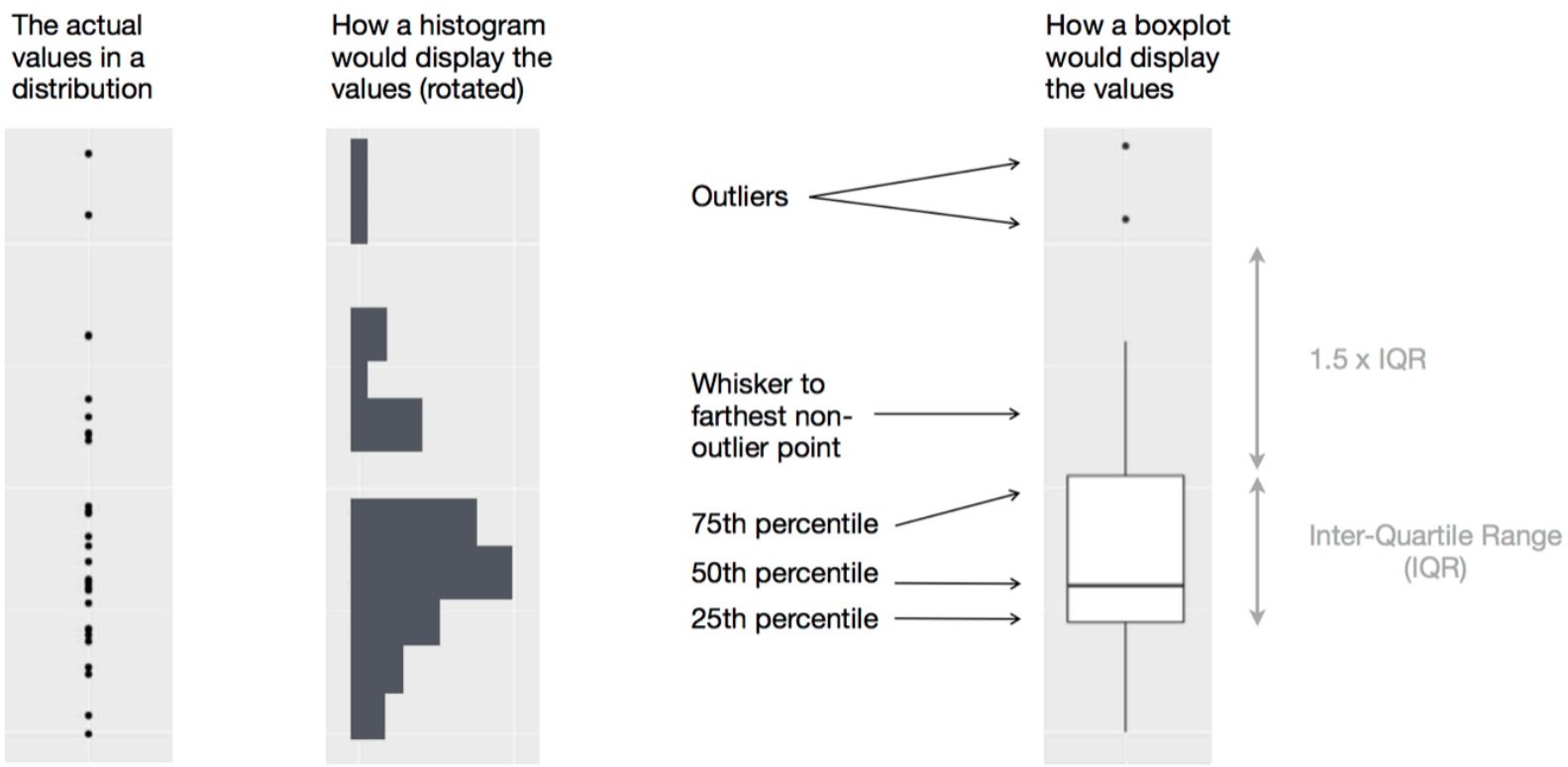
Vocabularies - Boxplot

- To describe the distribution of the data:
  - Mean: the average
  - Median: the number x such that 50% of data is greater than x and the other 50% of data is smaller than x
  - Standard deviation: to measure the spread of data
  - Variance: also to measure the spread of data
- To visualize the distribution of the data using plots:
  - Boxplot
    - Interquartile range: IQR = Q3 – Q1, IQR is such that 50% data fall within this range
    - Quartile: including Q1, Q2, Q3, and Q4.
    - Outlier: data points that are far away from the majority of the data points.

Vocabularies – R Language Terminologies

- Types of variables
  - Character variable: a word with quotation marks
  - Numeric(quantitative) variable: variables represented in integers and decimal numbers
  - Logical variable: TRUE and FALSE
- More:
  - R object: the variable name in R in which input values are stored
  - (atomic) Vector: think of it as a list, a collection of values,
  - Data frame: used for storing datasets, with column being variables and rows being the observations
  - Summary table: displays summary statistics, constructed using R function: summarize( )
  - Summary statistics: e.g. mean, min, standard deviation, etc.
  - Proportion: refers to the fraction of the total that possesses a certain attribute

Boxplot

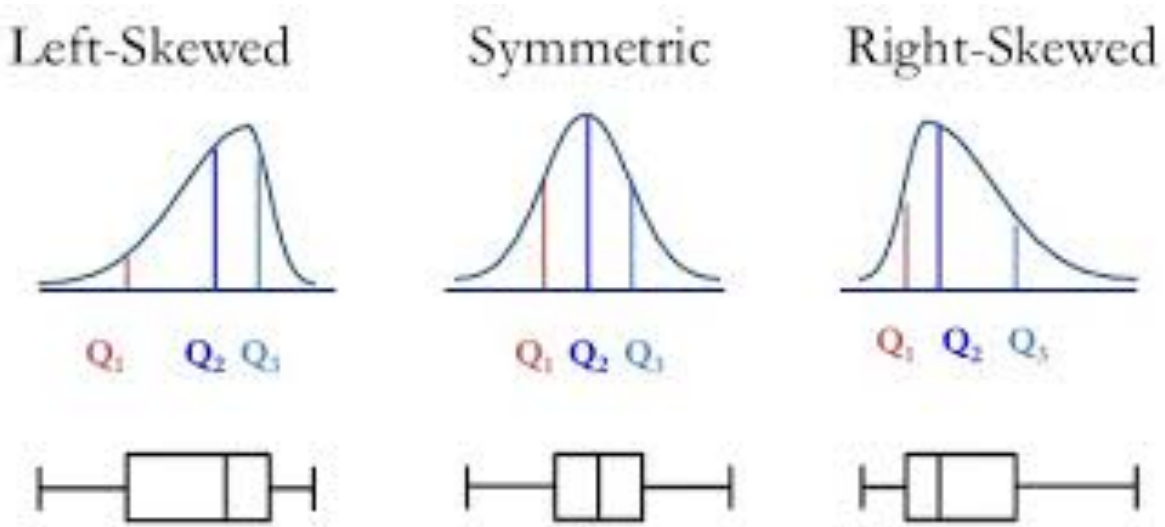


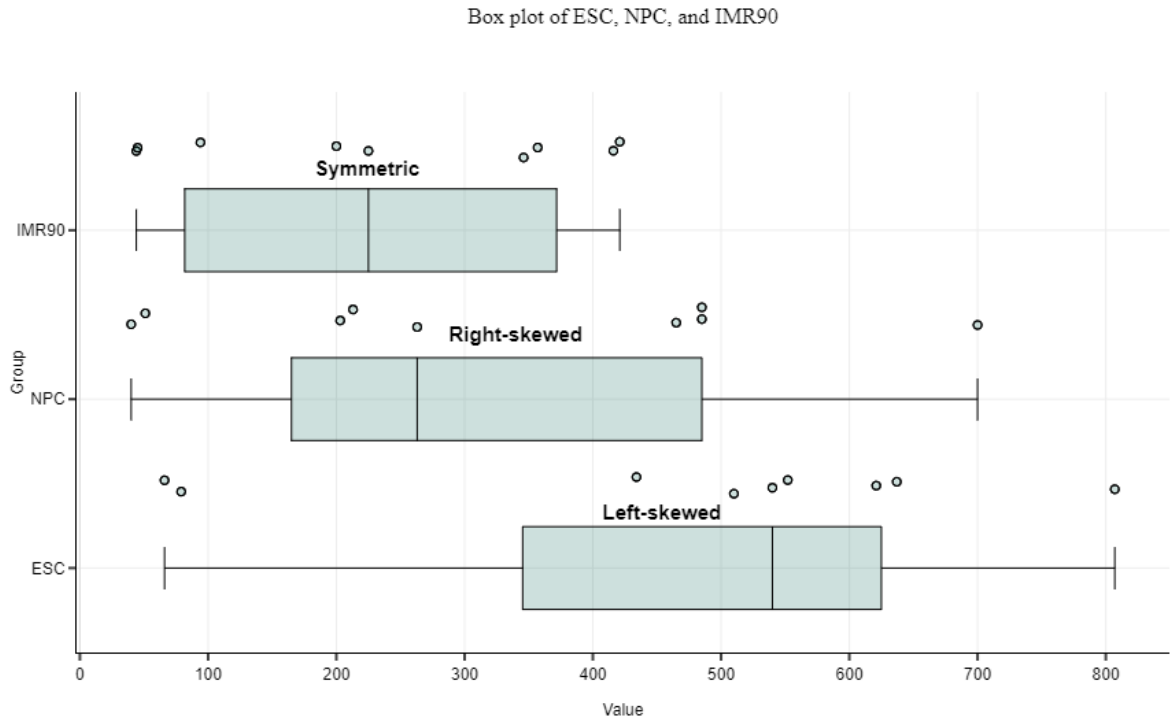
Whisker need not to have a length of 1.5*IQR(as long as it stays within the 1.5*IQR range) , for examples when all the data points are condensed together; However, the whisker can sometimes have length up to 1.5*IQR, when the data points are more dispersed.

More examples

What does boxplot do

- To summarize the distribution of a quantitative variables
  - Five statistics:
    - Minimum
    - Maximum
    - Medium (NOT mean)
    - 1st quartile
    - 3rd quartile
  - And outliers
- To summarize these vales according to a categorical variable of interest
- Example:
  - Amount of rainfall, a continuous variable
  - How this distribution varies by important categorical variables, such as: month, region, year, etc.

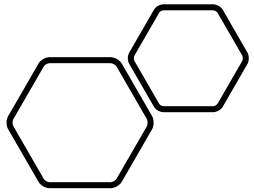

Left-Skewed
Symmetric
Right-Skewed
Q1 Q2 Q3
Q1 Q2 Q3
Q1 Q2 Q3

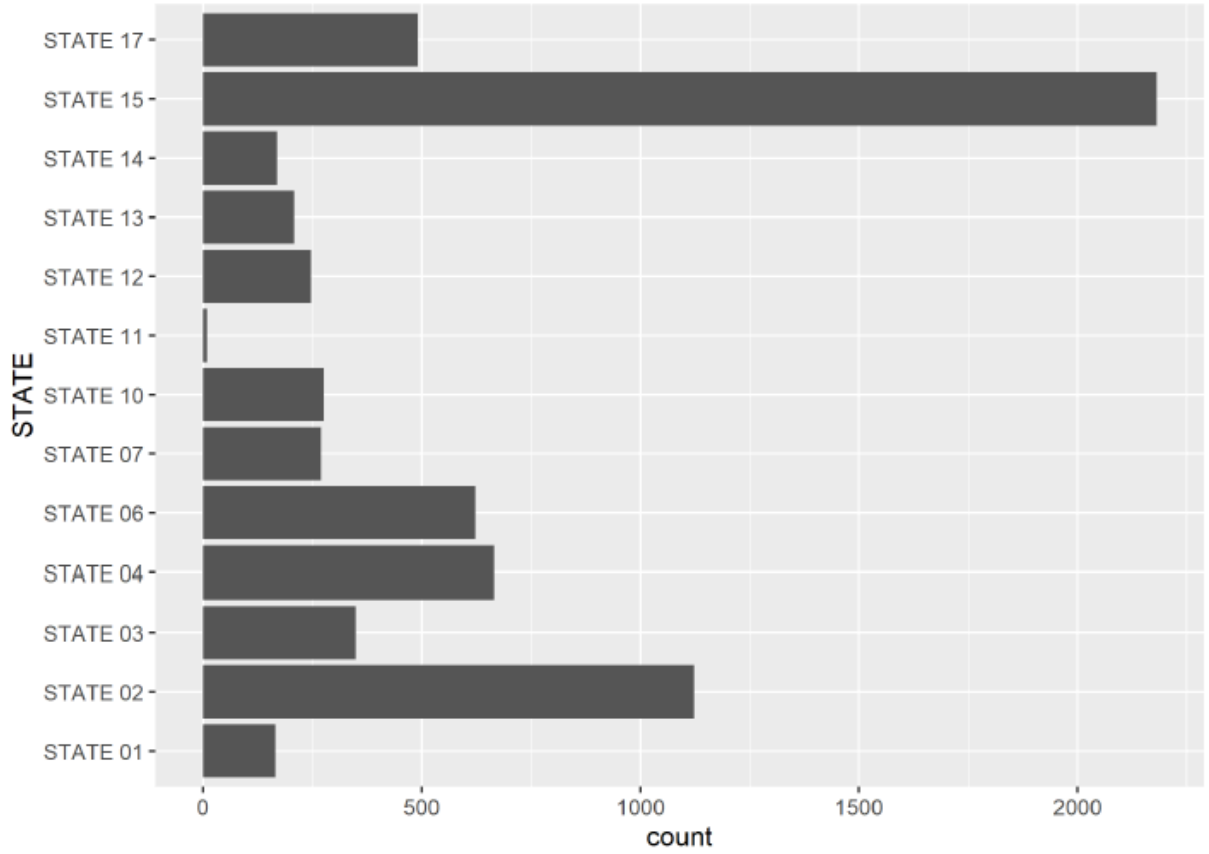

Box plot of ESC, NPC, and IMR90
Symmetric
Right-skewed
Left-skewed
IMR90
NPC
ESC
Group
0
100
200
300
400
500
600
700
800
Value

Group discussion

- *Why might it be useful to provide a summary table? (You made one in Question 2h)*

city_name <chr>	n <int>	mean <dbl>	median <dbl>	sd <dbl>	min <dbl>	max <dbl>	x1 <int>
Adelaide	110	0.9372727	0	5.0707989	0	48.0	4
Brisbane	52	4.2192308	0	14.9225062	0	97.0	6
Canberra	11	0.1636364	0	0.5427204	0	1.8	0
Melbourne	50	1.5520000	0	4.4289051	0	19.0	3
Perth	51	0.2176471	0	0.6704345	0	2.8	0
Sydney	110	5.0036364	0	19.1484731	0	191.0	14

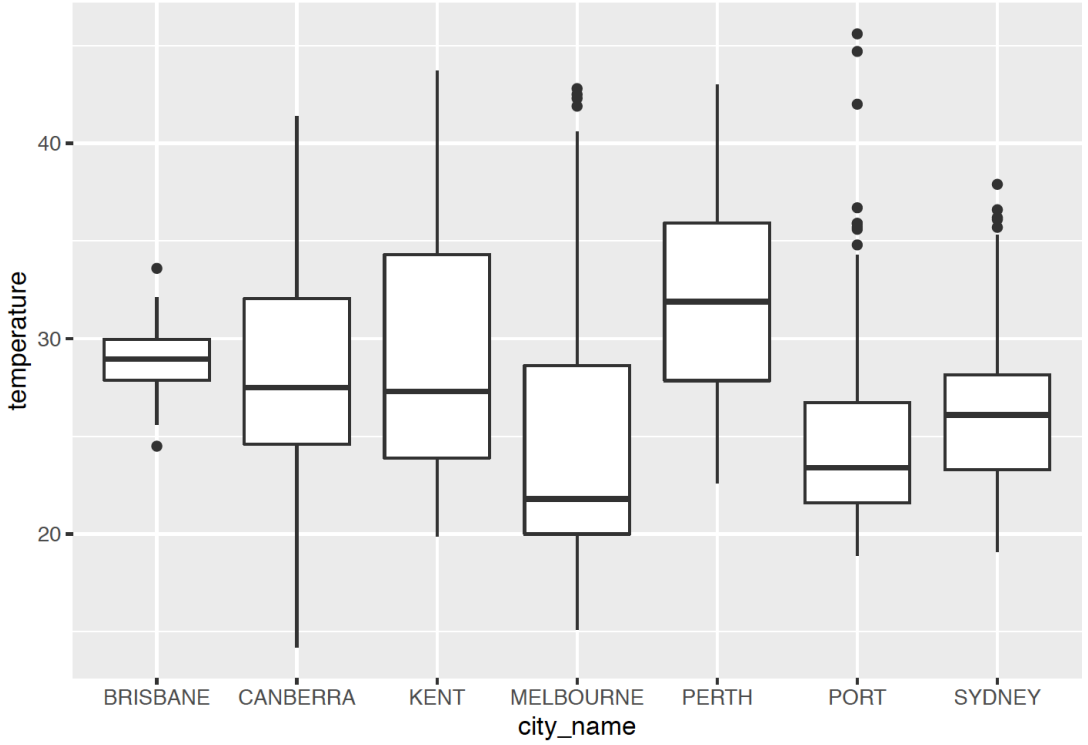


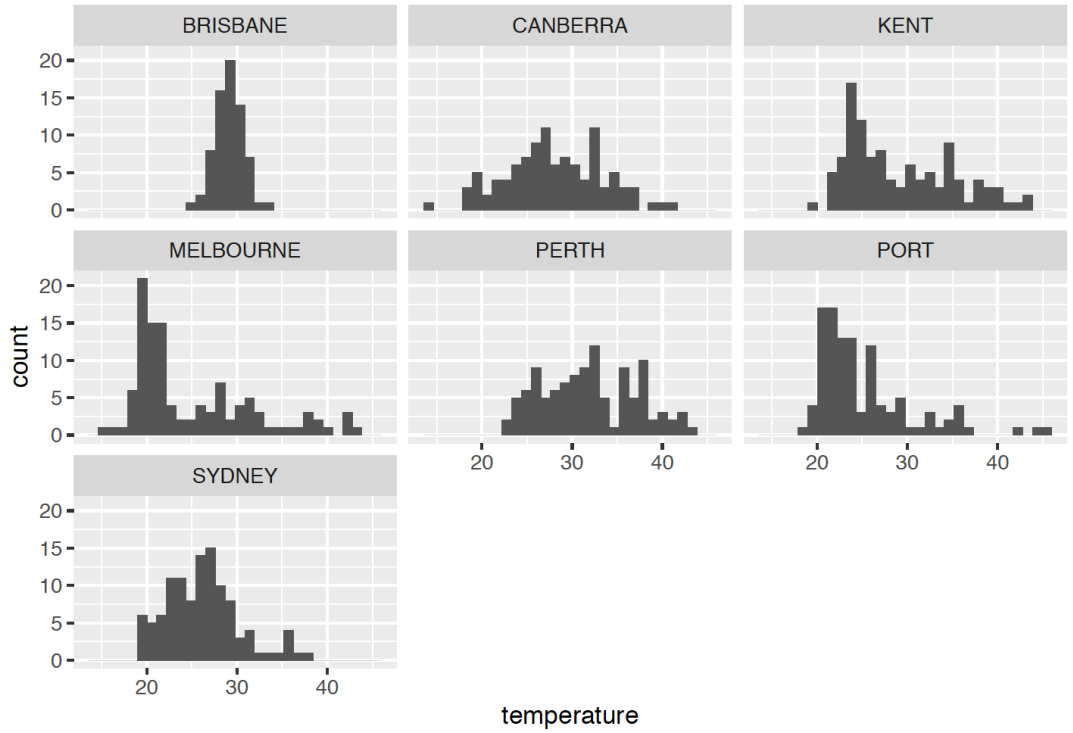



```
## # A tibble: 13 x 2
##    STATE         n
##    <fct>     <int>
##  1 STATE 01    166
##  2 STATE 02   1122
##  3 STATE 03    348
##  4 STATE 04    666
##  5 STATE 06    622
##  6 STATE 07    269
##  7 STATE 10    276
##  8 STATE 11      9
##  9 STATE 12    247
## 10 STATE 13    208
## 11 STATE 14    169
## 12 STATE 15   2180
## 13 STATE 17    491
```

Would you prefer the bar chart or the table?

Which state has the most insurance claims in our sample?





Group discussion

- *For Question 1, you used both histograms and boxplots to visualize your data.*
- *Which features were easier/harder to observe from each of the visualizations?*
- *In what situations may you want to choose a boxplot over a histogram, or vice versa? Explain.*

Histogram v.s. Boxplot

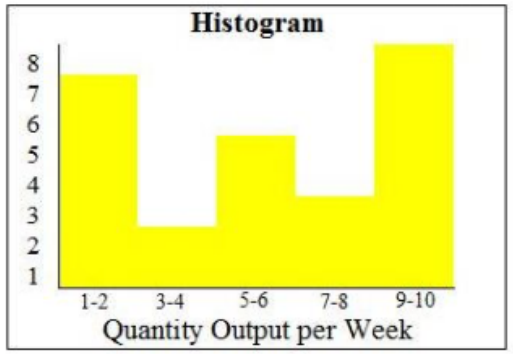
- Histogram:
  - Underlying distribution of the data.
  - Histograms are normally used to represent moderate to large amount of continuous data and we need at least 25 data points to determine if a histogram follows a particular distribution. If the data size is too small or the measurement system has a low resolution, the histogram may show very few columns and may not accurately display the shape of the distribution.
- Boxplot:
  - Comparing several distributions against each other.
  - Summarize key statistics from the data, and the outliers.
  - They provide a quick way for examining the variation present in the data.
    - A wider range boxplot indicates more variability.
  - Example: Sometimes when we're comparing distributions we don't care about overall shape, but rather where the distributions lie with regard to one another. Plotting the quantiles side by side can be a useful way of doing this without distracting us with other details that we may not care about.

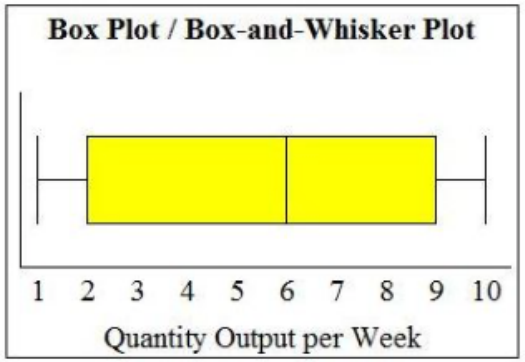
More on Histogram v.s. Boxplot

- Both charts effectively represent different data sets; however, in certain situations, one chart may be superior to the other in achieving the goal of identifying variances among data.
- The type of chart chosen depends on
  - the type of data collected
  - rough analysis of data trends
  - and project goals.
- Let’s look at some case-by-case comparisons:

Scenario 1: Histogram or Boxplot

- A histogram is highly useful when wide variances exist among the observed frequencies for a particular data set:

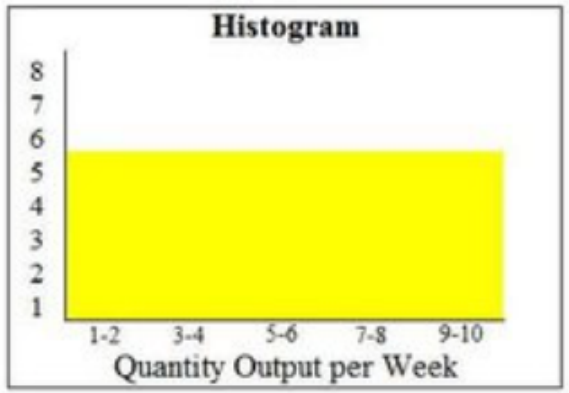


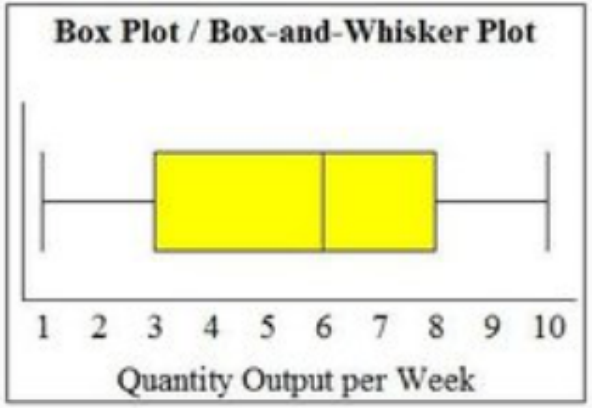


- This histogram shows that there are three peaks within the data, indicating it is tri-modal (three commonly recurring groups of numbers). Had this data simply been graphed using a box plot, the values would average one another out, causing the distribution to look roughly normal.

Scenario 2: Histogram or Boxplot

- A histogram is preferable over a box plot is when there is very little variance among the observed frequencies:

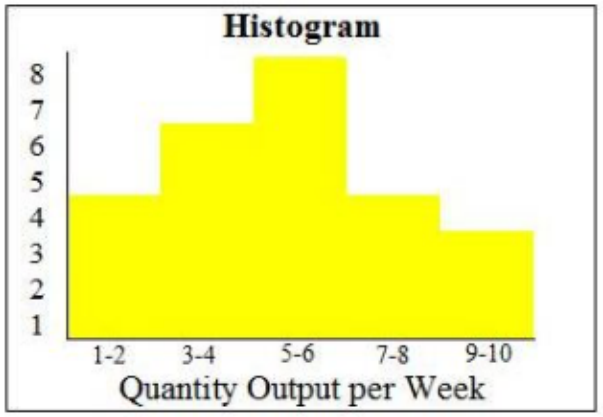


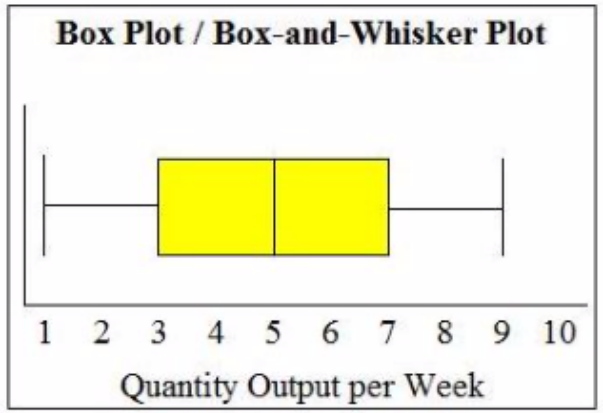


- This histogram shows that there is little variance across the groups of data; however, when the same data points are graphed on a box plot, the distribution looks roughly normal with a high portion of the values falling below six.

Scenario 3: Histogram or Boxplot

- This is an instance when a box plot can be more useful than a histogram.





- This occurs when there is moderate variation among the observed frequencies, which causes the histogram to look ragged and non-symmetrical due to the way the data is grouped. This may lead one to assume the data is slightly skewed. However, when a box plot is used to graph the same data points, the chart indicates a perfect normal distribution.

Summary:

- **Histograms**: The paired histograms give us a good overall impression of the distribution of ticket prices among survivors and non-survivors, but it is difficult to extract estimates of the mean, median, and quantiles, as well as the bounds of each bin.
- **Boxplots**: The boxplots make it easy for us to compare the medians, quartiles, IQR (and outliers) of ticket prices across the two groups, although we cannot easily extract exact values for these. Also, since boxplots only display a small number of summary statistics, we lose information about the shape of the distributions.
- **Summary table:** The summary table makes it easy to compare numerical values of key statistics. It is only from the summary table that we noticed that some passengers were recorded to have paid 0 pounds for their tickets. However, it is more difficult to get a quick sense of the overall shape of the distributions from these summary statistics alone, although these could be used to sketch a pair of boxplots.