STAT2203 STAT7203

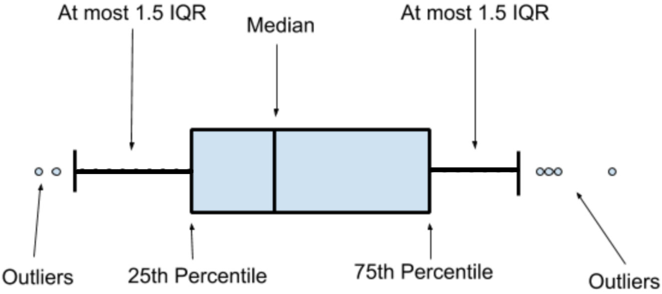
Probability models and data analysis

Lecture 1.3

Describing Data (continued...)

Visualising quantitative data (Boxplot)

The **boxplot** (or box-and-whiskers plot) is a graphical representation of the five-number summary.



Because the whiskers must end at an observed data point, the whisker lengths can look unequal, even though 1.5 IQR is the same for both sides.

Observed data points outside the boundary of the whiskers are plotted as outliers and are flagged as unusual.

Visualising quantitative data (Histogram)

A **histogram** is a common graphical representation of the distribution of a quantitative variable.

First, the range of the data is divided into a number of **bins**.

We then tally how many values fall in each bin and then make the plot by drawing rectangles whose bases are the bin intervals and whose heights are the counts.

Histograms can be sensitive to the choice of the number of bins.

Visualising quantitative data (Density plot)

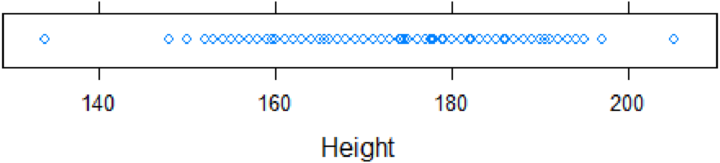
Density plots can be thought of as a smoothed version of a histogram.

The smoothness of the density plot is controlled by a bandwidth parameter (`bw`). `R` has a method for choosing the bandwidth automatically.

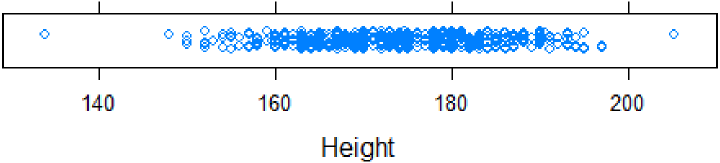
Note: Density plots are only used for continuous variables.

Visualising quantitative data (Stipplot)

A stripplot plots observations in a single line.



“Jitter” is added to show frequently occuring values.

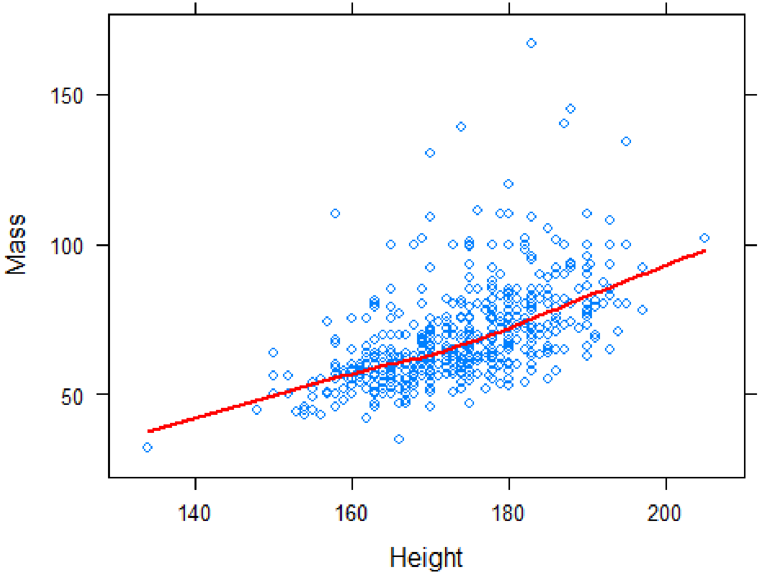


Visualising quantitative data (Scatter plot)

We can visualize patterns between two quantitative variables using a scatter plot.

We typically plot the response varaible on the vertical axis and the explanatory variable on the horizontal axis.

We often describe the relationship between two quantitative variables in terms of the direction (positive or negative), linearity, and strength (amount of variability around the global trend).


150
100
50
Mass
140
160
180
200
Height

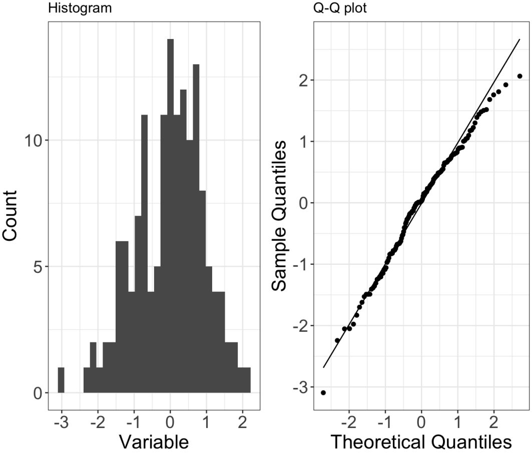
Visualising quantitative data (Quantile plot)

Most experiments resulting in the measurement of a continuous random variable provide little insight as to which probability model best describes the distribution of the measurements.
Quantile-Quantile (qq) plot is a tool to test the *goodness-of-fit* of a particular model to a random sample obtained from some population.

The basic idea of the qq-plot is to plot the quantiles of data, against the corresponding quantiles of the model distribution.

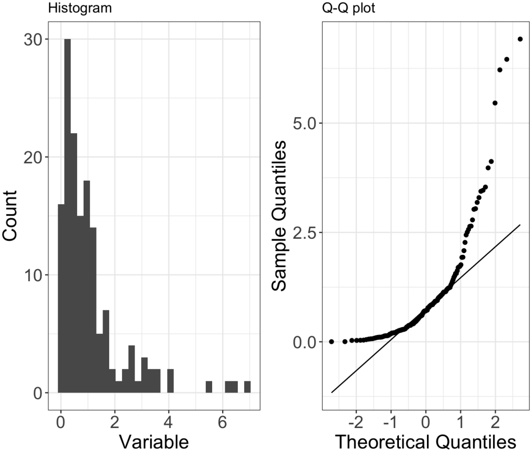
Each point in a qq-plot corresponds to the p-quantile of the variable and the model distribution for a particular value of $p \in [0, 1]$.

Visualising quantitative data (Quantile plot)



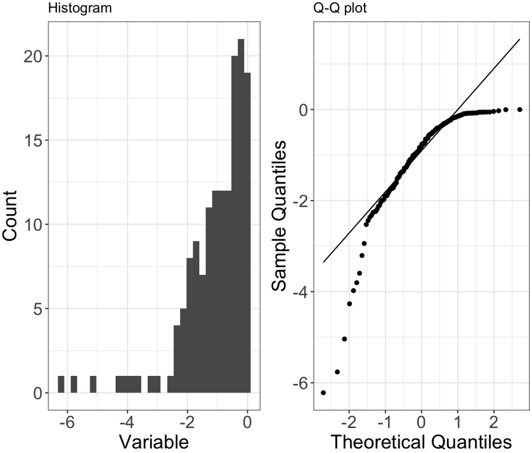
Normally distributed data

Visualising quantitative data (Quantile plot)



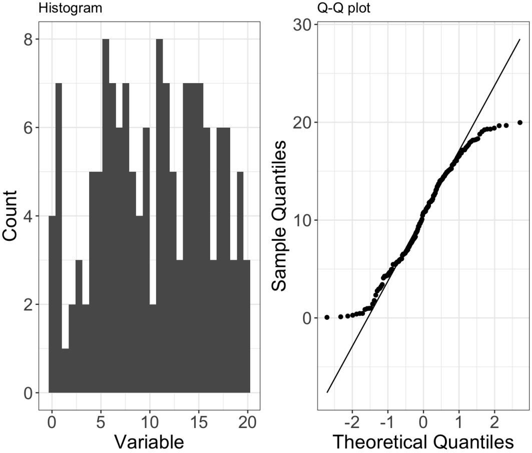
Right-skewed data

Visualising quantitative data (Quantile plot)



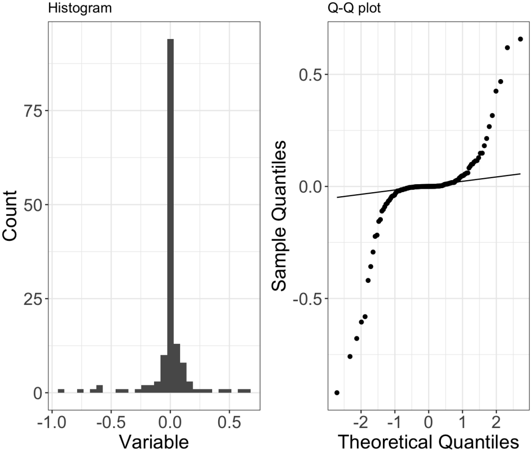
Left-skewed data

Visualising quantitative data (Quantile plot)



Under-dispersed data

Visualising quantitative data (Quantile plot)



Over-dispersed data

Visualising quantitative data (Quantile plot)

The distribution of two variables can also be compared using a qq-plot.

This kind of comparison is much more detailed than a simple comparison of means or medians.

If two variables have the same distribution, what would the quantile-quantile look like?