**Univariate analysis**

Measures of central tendency

mean, median, mode

most robust: median – less sensitive to extreme values

*\*mode: unimodal*

Measures of spread

variance, standard deviation, IQR, range (min. – max.)

most robust: IQR – less sensitive to extreme outliers

Robustness = few changes between repeated samples collected from a population.

If you collect repeated samples from a population, the minimum, maximum and range tend to change drastically from sample to sample, while the variance and standard deviation change less, and the IQR least of all.

The minimum and maximum of a sample may be useful for detecting outliers, especially if you know something about the possible reasonable values for your variable. They often (but certainly not always) can detect data entry errors such as typing a digit twice or transposing digits (e.g., entering 211 instead of 21 and entering 19 instead of 91 for data that represents ages of senior citizens.) The IQR has one more property worth knowing: for normally distributed data only, the IQR approximately equals 4/3 times the standard deviation. This means that for Gaussian distributions, you can approximate the sd from the IQR by calculating 3/4 of the IQR

Skewness and kurtosis

Skewness is a measure of asymmetry. Kurtosis is a more subtle measure of peakedness compared to a Gaussian distribution.

**Skewness (e) or kurtosis (u) Conclusion**

−2SE(e) < e < 2SE(e) not skewed

e ≤ −2SE(e) negative skew

e ≥ 2SE(e) positive skew

−2SE(u) < u < 2SE(u) not kurtotic

u ≤ −2SE(u) negative kurtosis

u ≥ 2SE(u) positive kurtosis

e: skewness, u: kurtosis

For a positive skew, values far above the mode are more common than values far below, and the reverse is true for a negative skew. When a sample (or distribution) has positive kurtosis, then compared to a Gaussian distribution with the same variance or standard deviation, values far from the mean (or median or mode) are more likely, and the shape of the histogram is peaked in the middle, but with fatter tails. For a negative kurtosis, the peak is sometimes described has having “broader shoulders” than a Gaussian shape, and the tails are thinner, so that extreme values are less likely

**Multivariate analysis**

Covariance

* Technically, independence implies zero correlation, but the reverse is not necessarily true.

Correlation

**Degrees of freedom**

Used for samples.

Consider 5 numbers with a mean of 10. To calculate the variance of these numbers we need to sum the squared deviations (from the mean). It really doesn’t matter whether the mean is 10 or any other number: as long as all five deviations are the same, the variance will be the same. This make sense because variance is a pure measure of spread, not affected by central tendency. But by mathematically rearranging the definition of mean, it is not too hard to show that the sum of the deviations (not squared) is always zero. Therefore, the first four deviations can (freely) be any numbers, but then the last one is forced to be the number that makes the deviations add to zero, and we are not free to choose it. It is in this sense that five numbers used for calculating a variance or standard deviation have only four degrees of freedom (or independent useful pieces of information). In general, a variance or standard deviation calculated from n data values and one mean has n − 1 df.

**Data**

* plot residuals
* correlation matrix