COC131 Data Mining - Brief Statistics Revision

Martin D. Sykora m.d.sykora@lboro.ac.uk
Tutorial 01, Friday 15th February 2008

1 Statistical Modelling

We wish to collect data from the real-world and draw conclusions. In order to use this data in a meaningful way we need to build a model that well-represents the data. If a model is a poor fit, then the observations made will be poor as well.

1.1 Model building

1.2 Population and samples

A sample is a small subset of a population of interest. It is used to infer things about the population as a whole. In terms of data-mining, the inferences can be either descriptive or predictive. The larger the sample, the more likely that it will reflect the whole population. While random samples from the population will vary, on average for large samples they will remain fairly similar.

1.3 Simple Models

1.3.1 Mean, median, mode

The mean is a simple summary of data.

$$\mu = \frac{\sum_{i=0}^{n} x_i}{n}$$

The median is the middle value. If there is an odd number of observations, it is the middle value. If there is an even number of observations, it is the

mean of the two middle values. It is less sensitive to extreme scores (outliers) than the mean and makes a better measure for highly skewed data.
The mode is the most frequently occuring score in a distribution and is greatly subject to sample fluctuation. Many distributions have more than one mode (multi-modal).

1.3.2 Variance and Standard Deviation

We could calculate the error of an observation by subtracting the mean as $\epsilon = x_i - \mu$. We could then say that the total error is the sum of the observation errors.

$$\epsilon = \sum_{i=0}^{n} (x_i - \mu)$$

This is not a good way to calculate the error as it depends on the direction of the error. For example, a combination of negative errors and positive errors can lead to a 0 error which would be an entirely false observation about the data. One way to get around this problem is to square each of the errors.

$$ss\epsilon = \sum_{i=0}^{n} (x_i - \mu)^2$$

This is a good model of accuracy of the model but is dependent on the amount of data collected. We can overcome this by dividing by the number of observations and finding the average error in the sample. More interesting is using the error in the sample to estimate the error in the population, so we divide instead by the number of observations minus one. This is called the variance.

$$variance = \sigma^2 = \frac{\sum_{i=0}^{n} (x_i - \mu)^2}{(n-1)}$$

It is the average error between the mean and the observations made, but is in units squared which is inconvenient. The square root of the variance is known as the standard deviation.

$$standarddeviation = \sigma = \sqrt{\frac{\sum_{i=0}^{n} (x_i - \mu)^2}{(n-1)}}$$

A small standard deviation indicates that the data is close to the mean. Conversely, a large standard deviation indicates that the data is far from the mean. This will indicate if the mean is a good or poor fit to the data.

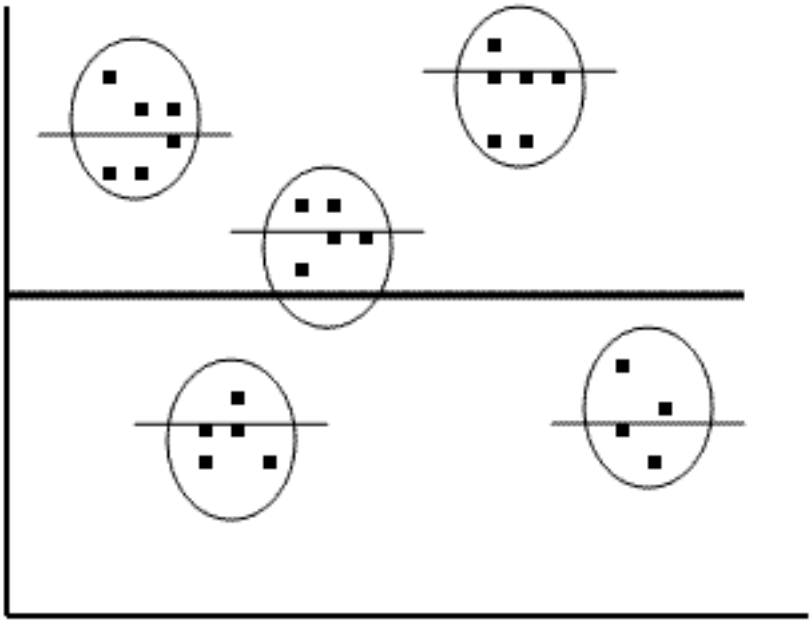

Figure 1: Standard Error (not to accurate scale)

1.3.3 Standard Error

The standard error is not the standard deviation. It is best explained with an example. Take 5 samples from a population with each sample containing 5 observations. In Figure 1, the samples are represented by the ovals and means of the samples are represented by the lines that pass through the ovals. The global mean is represented by the thick black line passing through the space. The standard error is the standard deviation of the difference between the sample means and the overall mean. A large standard error indicates that there is a lot of variability, whereas a small standard eror indicates that samples are more similar to the population.

2 Exploring Data

It is important to understand our dataset and to check properties of our data so that it meets the criteria necessary for the statistical procedures we wish to use. The easiest way to see trends in our data is to plot a graph. There are 6 basic things that should interest us about a dataset.

1. Shape of a dataset will be the main factor in determining which set of summary statistics best summarises our dataset. It should hence be

the first characteristic to be noted. Shape is commonly categorised as symmetric, left-skewed or right-skewed, and as uni-modal, bi-modal or multi-modal.

2. Measure the central tendency, common measures of central tendency are the mean and median. Less common are the mode (the most frequent value), the mid-range (the value midway between the minimum and maximum values) and the truncated mean (where a fixed percentage of the largest and smallest scores are deleted from the dataset and the mean of the remaining data is calculated).

3. Spread is a measure of variation in the data. Common measures of spread are variance, standard deviation and the interquartile range. Less commonly used is the range, as it is strongly skewed by outliers and not very robust.

4. Outliers are data values that lie away from the general cluster of other data values. Each outlier needs to be examined to determine if it represents a possible value from the population being studied, in which case it should be retained, or if it is non-representative (or an error) in which case it can be excluded. If distribution of the dataset is a normal distribution, which is the most common case with majority of data, then 68% of the observations are within $\mu \pm \sigma$ and 95% of the observations are within $\mu \pm 2\sigma$.

5. Clustering implies that the data tends to bunch up around certain values, eg. annual wages for a factory may cluster around \$20 000 for unskilled factory workers, \$35 000 for tradespersons and \$50 000 for management.

6. Granularity implies that only certain discrete values are allowed, eg. a commodities future may only be traded in multiples of 100. Discrete data has some granularity as only certain values are possible. Continuous data can show granularity if the data is rounded.

3 Correlation

Correlation is the linear relationship between two or more variables. A positive correlation means that as one variable increases, the other increases as well. A negative correlation means that as one variable increases, the other decreases.

3.1 Covariance

Variance is the average amount data varies from the mean. If there is a correlation between two variables, then as one deviates from the mean, we expect the other to have similar deviations.

$$cov(x, y) = \frac{\sum_{i=0}^{n}(x_i - \mu_x)(y_i - \mu_y)}{(n-1)}$$

This is the average sum of the combined differences. It is dependent on the scale of measurement and changing the units effect the covariance. It needs to be standardised. In order to standardise measurements, we normally subtract the mean and divide by the standard deviation. In the case of covariance, the mean is already subtracted, so we simply divide by the standard deviation.

$$r = \frac{cov(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=0}^{n}(x_i - \mu_x)(y_i - \mu_y)}{(n-1)\sigma_x \sigma_y}$$

This is the Pearson product moment correlation coefficient and lies between the values of -1 and 1. A value of 1 indicates perfect correlation which means that as one variable increases, the other increases by a proportional amount.
In interpreting correlation, we can conclude that with a strong correlation that as one variable increases, the other increases as well. We cannot, however, say that one variable increasing caused the other to increase. This is because in any bivariate correlation, there may be other measured or unmeasured variables effecting the results. Furthermore, the correlation coefficient does not say anything about what variable caused another to change.
The correlation coefficient can be squared producing the measure of the amount of variability in one variable that is explained by the other. If, say, two variables had a correlation of -0.431, then the r^2 value will be 0.1858, so we can say that one variable accounts for 18.58% of the variability in the other variable. Still this cannot be used to infer causal effects.