**Multivariate Analysis. Dealing with more than one variable.**

In all we have done so far we have only studied the statistical analysis of one variable. The last part of the course is going to deal with the case where we have to consider more than one variable. Sometimes the additional variables will be random, sometimes they will not, but in all cases we will be concerned with the nature of the relationships between the variables. Mostly we shall concentrate on just two variables and generalize to the case of many. First we shall consider how to detect whether or not two variables are independent.

**A Goodness of Fit Test of Independence.**

Imagine two characteristics that an individual may have, say hair colour and eye colour, we may think of them as two variables and it is of interest to establish whether or not the variables are distributed independently across individuals. Suppose the entire range of hair colour is divided into c mutually exclusive and exhaustive categories numbered j = 1,..,c and similarly the entire range of eye colour is divided into r mutually exclusive and exhaustive categories numbered i = 1,..,r. The true probabilities that a randomly selected person has a particular hair colour / eye colour combination can be arranged on an r x c grid whose rows correspond to the eye colour categories and whose columns correspond to the hair colour categories as in Table 1 below.

The $p_{ij}$'s are joint probabilities of having a particular hair / eye colour combination and taken together they correspond to the joint probability distribution of the eye colour-hair colour combinations. The row sums of probabilities ($p_{i.}$'s) represent the marginal probabilities of having the i'th eye colouring regardless of which hair colour an individual has and taken together they constitute the marginal probability distribution of eye colouring. Similarly the column sums ($p_{.j}$'s) are marginal probabilities of having the j'th hair colouring regardless of which eye colour an individual has and taken together they constitute the marginal probability distribution of hair colouring. Hence all of the $p_{ij}$'s sum to one as does the sum of the $p_{i.}$'s and the sum of the $p_{.j}$'s.

**Table 1. General probability structure.**

| | hair colour j=1,..,c | | | | | | Row sums |
|---|---|---|---|---|---|---|---|
| eye colour i=1,..,r | $p_{11}$ | $p_{12}$ | | | | $p_{1c}$ | $p_{1.}$ |
| | $p_{21}$ | $p_{22}$ | | | | $p_{2c}$ | $p_{2.}$ |
| | | | | | | | |
| | | | | | | | |
| | | | | $p_{ij}$ | | | $p_{i.}$ |
| Col sums | | | | | | | |
| | | | | | | | |
| | $p_{r1}$ | $p_{r2}$ | | | | $p_{rc}$ | $p_{r.}$ |
| | $p_{.1}$ | $p_{.2}$ | | $p_{.j}$ | | $p_{.c}$ | 1 |

**Table 2. Independent Probability Structure**

| | hair colour j=1,..,c | | | | | | Row sums |
|---|---|---|---|---|---|---|---|
| eye colour i=1,..,r | $p_{1.}p_{.1}$ | $p_{1.}p_{.2}$ | | | | $p_{1.}p_{.c}$ | $p_{1.}$ |
| | $p_{2.}p_{.1}$ | $p_{2.}p_{.2}$ | | | | $p_{2.}p_{.c}$ | $p_{2.}$ |
| | | | | | | | |
| | | | | | | | |
| | | | | $p_{i.}p_{.j}$ | | | $p_{i.}$ |
| Col sums | | | | | | | |
| | | | | | | | |
| | $p_{r.}p_{.1}$ | $p_{r.}p_{.2}$ | | | | $p_{r.}p_{.c}$ | $p_{r.}$ |
| | $p_{.1}$ | $p_{.2}$ | | $p_{.j}$ | | $p_{.c}$ | 1 |

**The Theoretical Implication of Independence.**

Back in chapter 2 we observed that, if two events A and B were independent, their joint probability was equal to the product of their marginal probabilities so that $P(A \cap B) = P(A)P(B)$ under independence. If this is true for hair colour and eye colour then P(i'th eye colour and j'th hair colour) = $p_{ij}$ = P(i'th eye colour regardless of hair colour) x P( j'th hair colour regardless of eye colour) = $p_{i.}\ p_{.j}$ would have to hold for all i and j. This would change the configuration of the probabilities in Table 1, Table 2 presents how they would look.

**The sample structure.**

Suppose we have a random sample of N individuals. Each one of the individuals in the sample will have one and only one of the hair colour - eye colour category combinations. The number of people in the sample with the ijth eye-hair colour combination ($O_{ij}$) divided by the total number of people in the sample (N) can be thought of as an estimate of the true probability that a randomly selected person from the population has that particular hair colour-eye colour combination so that the estimate may be written:

$$\overline{p_{ij}} = \frac{O_{ij}}{N}$$

Similarly the sum over all possible eye colours (hair colours) of the number of people with the i'th hair colour (j'th eye colour) divided by the total number of people in the sample can be thought of as an estimate of the true probability that a randomly selected person from the population has the i'th particular hair colour (j'th particular eye colour) so that:

$$\overline{p_i} = \sum_{i=1}^{r} \frac{O_{ij}}{N}$$

$$\overline{p_j} = \sum_{j=1}^{c} \frac{O_{ij}}{N}$$

These estimates obey adding up rules in the same way that the true probabilities obey such rules namely:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\overline{p_{ij}} = \sum_{r=1}^{r}\overline{p_i} = \sum_{j=1}^{c}\overline{p_j} = 1$$

If hair colour were distributed in the population independently of eye colour one would expect it to be reflected in the random sample namely:

$$\overline{p_{ij}} \approx \overline{p_i}\,\overline{p_j} \text{ for all i=1,..,r; j=1,..,c.}$$

If both sides of this approximate inequality were multiplied by N the sample size we would have:

$$\left(O_{ij} = N\overline{p_{ij}}\right) \approx \left(N\overline{p_i}\,\overline{p_j} = E_{ij}\right) \textit{ for all } i = 1,...,r;\ j = 1,....,c$$

Where Eij represents the number expected with the i'j th eye / hair configuration if the characteristics were truly independent. This suggests a test of the form:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

which has exactly the same form as the goodness of fit test for distributions in the previous chapter except here it is used to see how well the probability structure induced by independence fits. The statistic can be shown to have (r-1)(c-1) degrees of freedom and will be used in the context of a one sided upper tailed test.

**Jointly Distributed Random Variables**

The joint probability distributions of random variables can be viewed in much the same way that the joint probabilities of events were considered. Let X and Y be two random variables, the nature of their joint density function f(x,y) will vary with the discreteness or otherwise of X and Y just as in the single or uni-variate case. Indeed the properties of the multivariate f(x,y) are very natural multi-variate extensions of their uni-variate counterparts outlined above. There is no reason why both variables should be discrete or both continuous but, for expositional convenience and because the contrary is seldom encountered in this field, that is what will be assumed here.

In the discrete case $f(x_i,y_j) = P(X=x_i,Y=y_j)$ and in the continuous case:

$$f(x,y) \geq 0$$
$$\iint_{(x,y)\in S} f(x,y)dxdy = 1$$
$$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x,y)dxdy$$

So that again Discrete Joint Probability Distributions attach probabilities to points, Continuous Joint Probability Distributions attach probabilities to intervals.

Associated with these densities are cumulated distribution functions F(x,y) which in each case yield the probability that the random variables X and Y are respectively less than some values x and y. Algebraically these may be expressed for jointly continuous and jointly discrete distributions respectively as:

$$P\left(X \leq x, Y \leq y\right) = F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(z,w)dzdw$$

$$P\left(X \leq x, Y \leq y\right) = F(x,y) = \sum_{i=1}^{k} \sum_{j=1}^{l} f(x_i, y_j);\ \textit{where}\ x_k < x < x_{k+1}, y_l < y < y_{l+1}$$

Clearly in the case of continuous distributions $\partial 2F(X,Y)/(\partial X \partial Y) = f(X,Y)$. The situation where one variable is discrete and the other continuous can be considered in an obvious fashion by the appropriate combination of summation and integration operators.

**Marginal Distributions, Conditional Distributions and Independently Distributed Random Variables**

When only one of the jointly distributed variables is of interest the marginal distribution of that variable may be obtained in the continuous case as:

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy,\ f(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

in the discrete case the analogous result is:

$$f(x_i) = \sum_{j=1}^{\infty} f(x_i, y_j), \quad f(y_j) = \sum_{i=1}^{\infty} f(x_i, y_j)$$

Casual perusal of the analogous rules for manipulating joint probabilities above will reveal an obvious correspondence with distributions of random variables which can be extended to conditional distributions and the notion of independently distributed random variables. Thus conditional distributions of x given y and y given x are for both continuous and discrete random variables:

$$f\left(x|y\right)=\frac{f\left(x,y\right)}{f\left(y\right)},\ f\left(y|x\right)=\frac{f\left(x,y\right)}{f\left(x\right)}$$

In exactly the same fashion for both continuous and discrete random variables independence between x and y implies f(x,y) = f(x)f(y). More importantly for a sequence of mutually independent random variables $x_i$, i=1,..,n $f(x_1,x_2,x_3,..,x_n) = f(x_1)f(x_2)f(x_3)....f(x_n)$.

Interested will also focus on the conditional distribution of x given that it only takes on certain values in its range, in particular we may be interested in the conditional distribution of x given that x < z. This is f(x|x<z) = f(x)/F(z), essentially the probability distribution of x below z is simply rescaled by the probability that x is less than z.

3) The Covariance of X and Y E (X-E(X))(Y-E(Y))

This function yields the covariance of X and Y which, when X and Y are independent, is equal to zero. This is easily demonstrated for continuous distributions and has a direct analogue for discrete distributions. Since X and Y are independent f(x,y) = f(x)f(y), it follows that:

$$E\left(X-E\left(X\right)\right)\left(Y-E\left(Y\right)\right)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x-E\left(x\right)\right)\left(y-E\left(y\right)\right)f\left(x\right)f\left(y\right)dxdy=$$

$$\int_{-\infty}^{\infty}\left(x-E\left(x\right)\right)f\left(x\right)dx\int_{-\infty}^{\infty}\left(y-E\left(y\right)\right)f\left(y\right)dy=E\left(x-E\left(x\right)\right)E\left(y-E\left(y\right)\right)=0$$

For discrete distributions which are independent we have:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(x_i-E(X)\right)\left(y_j-E\left(Y\right)\right)f\left(x_i,y_j\right)=\sum_{i=1}^{r}\left(x_i-E\left(X\right)\right)f\left(x_i\right)\sum_{j=1}^{c}\left(y_j-E\left(Y\right)\right)f\left(y_j\right)=0.0$$

Clearly the metric of the covariance is the product of two measures, respectively those of X and Y so that for example if X were income $US and Y were educational status measured in Education units EU's then the covariance would be measured in $US x EU's. Again we could

consider standardizing the measure to make it metric free, and 2 examples are pertinent. When standardized by the standard deviation of X times the standard deviation of Y the covariance becomes the well known correlation coefficient indicating the extent and direction of a linear relationship between X and Y.

**Correlation and Covariance.**

Looking at the definition of the covariance between X and Y (E(X-E(X))(Y-E(Y))) we see that its units of measurement are the product of the units of measurement of the individual variables. If for example we were considering the covariance between peoples height and weight the covariance would be measured in inches x lbs or perhaps in meters x kilos. Interestingly enough the value of the covariance would differ depending upon which metrics were chosen or, put another way, covariance is not a unit free measure. Correlation is a measure related to covariance which is unit free, it is a number which ranges between -1 and +1 and is of the following form:

$$-1(COR(X,Y) \leq= \frac{COV(X,Y)}{\sqrt{V(X)V(Y)}}) \leq 1$$

**Sums and differences of independent random variables.**

Returning to an issue in previous chapters we are now able to understand why it is the variance of a sum of independent random variables is equal to the sum of thier variances. Consider two independent random variables X and Y with respective expectations E(X) and E(Y) and respective variances V(X) and V(Y). Following the definition of the variance in general the variance of a sum of two random variables is given by:

$$\begin{aligned} E(X+Y-E(X+Y))^2 &= E((X-E(X)) + (Y-E(Y)))^2 \\ &= E((X-E(X))^2 + (Y-E(Y))^2 + 2(X-E(X))(Y-E(Y)) \\ &= V(X) + V(Y) + 2COV(X,Y) \end{aligned}$$

But **since the covariance of independent random variables is zero, the variance of their sum reduces to the sum of their variances**. It turns out that the variance of the difference between two random variables is also the sum of their variances; this may be seen as follows, since again in general:

$$\begin{aligned} E(X-Y-E(X-Y))^2 &= E((X-E(X)) - (Y-E(Y)))^2 \\ &= E((X-E(X))^2 + (Y-E(Y))^2 - 2(X-E(X))(Y-E(Y)) \\ &= V(X) + V(Y) - 2COV(X,Y) \end{aligned}$$

the zero covariance implied by independence results in **the variance of the difference between two independent random variables being equal to the sum of their variances**.

**Analysis of Variance.**

Suppose you own a wheat farm and wish to evaluate the effectiveness of a variety of fertilizers on the farm. Splitting your farm into a collection of homogeneous sub plots[1] you allocate one type of fertilizer to one group of plots, another to another group of plots and so on. At the end of

[1]That is they are identical in every respect, they all suffer the same amounts of sunshine and rainfall and each has exactly the same amount of nutrients.

the year you calculate the average yield of plots covered with a particular type of fertilizer (as well as the average yield off the plots with no fertilizer) and wish to establish whether fertilizing has had any effect. The statistical technique for doing this is Analysis of Variance.

**The Numerical Structure of the Problem.**

Each of the different fertilizers (including no fertilizer) is termed a treatment, let us suppose there are T treatments t = 1,..,T and that $n_t$ plots have been allocated to treatment t. This means the total number of plots (N) on the farm is given by:

$$N = \sum_{t=1}^{T} n_t$$

Let the yield from the i'th plot under the t'th treatment be $Y_{it}$. The average yield for treatment t is:

$$\overline{Y_t} = \frac{\sum_{i=1}^{n_t} Y_{it}}{n_t}$$

The average yield over all the plots of land is given by:

$$\overline{Y} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n_t} Y_{it}}{\sum_{t=1}^{T} n_t} = \frac{\sum_{t=1}^{T} n_t \overline{Y_t}}{\sum_{t=1}^{T} n_t}$$

indicating that the overall average is really a weighted sum of the individual treatment yield averages.

The question of whether the different fertilizers had different effects then becomes a question of whether all these various means are different from one another.

**The Theoretical Structure of the Problem.**

Suppose the true mean or expected yield of plots under treatment t is $\mu_t$ and that all plots are independent of each other and are subject to random shocks with the same degree of variation then the theoretical model underpinning this problem may be written as:

$$
\begin{aligned}
&Y_{it} = \mu_t + e_{it} \qquad i = 1,.....,n_t;\ t = 1,.....,T \\
&\mathrm{E}\left(e_{it}\right) = 0 \quad \textit{for all } i,t \\
&E\left(e_{it}e_{js}\right) = \sigma^2 \quad \textit{for all } i = j \wedge\ t = s \\
&\qquad\qquad = 0 \quad \textit{otherwise}
\end{aligned}
$$

The important point here is that the $\mu_t$'s and the $e_{it}$'s are never observed, all you ever measure is the $Y_{it}$'s. The variability in the yields of different plots under different treatments (the $Y_{it}$'s) comes from two sources, variability in the random component (the $e_{it}$'s) and variability in the treatment effects (the $\mu_t$'s). If the different treatments (various fertilizers) had no effect all the $\mu_t$'s would be the same and the $e_{it}$'s would be solely responsible for variability in the $Y_{it}$'s. The trick here is how do we detect this from observing just the $Y_{it}$'s!

First note that $E(Y_{it}) = \mu_t$ for all i and t so that:

$$
E\left(\overline{Y_t}\right) = E\left(\frac{\sum_{i=1}^{n_t} Y_{it}}{n_t}\right) = \mu_t \qquad \textit{for all } t = 1,....,T
$$

Suppose we further add that $e_{it}$ is normally distributed, then all the above mentioned treatment means will be normally distributed so that:

$$
\overline{Y_t} \sim N\left(\mu_t, \frac{\sigma^2}{n_t}\right)
$$

To test whether particular treatment means differ we simply apply the difference in means "t" test which was introduced in the previous chapter. However to do this for all pairs of treatments involves T(T-1)/2 comparisons.

It would be useful to have a general test for whether or not the treatments have differing effects and this is what the analysis of variance test provides. It does this by comparing an estimate of the total variability in the $Y_{it}$'s engendered by the treatment effects with an estimate of the total variability engendered by the purely random effects (the $e_{it}$'s). These two estimates are essentially the components of the total sum of squares.

**Total Sum of Squares.**

The TOtal Sum of Squares (TOSS) representing the total variability in the $Y_{it}$'s is given by:

$$TOSS = \sum_{t=1}^{T} \sum_{i=1}^{n_t} \left( Y_{it} - \overline{Y} \right)^2$$

this reflects the variability induced by variability in the unobserved $\mu_t$'s and by variability induced by the unobserved $e_{it}$'s.

**Error Sum of Squares**

The Error Sum of Squares (ERSS) representing the variability induced by the purely random effects (the $e_{it}$'s) is captured by:

$$ERSS = \sum_{t=1}^{T} \sum_{i=1}^{n_t} \left( Y_{it} - \overline{Y}_t \right)^2$$

and reflects the variability in the $Y_{it}$'s after having taken out the variability of the treatment effects.

**Treatment Sum of Squares.**

It follows that the Treatment Sum of Squares (TRSS) representing the variability due to the treatment effects is given by:

$$TRSS = TOSS - ERSS = \sum_{t=1}^{T} n_t \left( \overline{Y}_t - \overline{Y} \right)^2$$

Some simple algebra will demonstrate that this is indeed true, since:

$$TOSS = \sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(Y_{it} - \overline{Y}\right)^2 = \sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(Y_{it} - \overline{Y}_t + \overline{Y}_t - \overline{Y}\right)^2$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n_t}[\left(Y_{it} - \overline{Y}_t\right)^2 + \left(\overline{Y}_t - \overline{Y}\right)^2 + 2\left(Y_{it} - \overline{Y}_t\right)\left(\overline{Y}_t - \overline{Y}\right)]$$

$$note\ that = \sum_{t=1}^{T}\sum_{i=1}^{n_t} 2\left(Y_{it} - \overline{Y}_t\right)\left(\overline{Y}_t - \overline{Y}\right) = 0\ because$$

$$\sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(Y_{it} - \overline{Y}_t\right)\left(\overline{Y}_t\right) = 0\ and\ \sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(Y_{it} - \overline{Y}_t\right)\left(\overline{Y}\right) = 0\ so$$

$$TOSS = \sum_{t=1}^{T}\sum_{i=1}^{n_t}[\left(Y_{it} - \overline{Y}_t\right)^2 + \left(\overline{Y}_t - \overline{Y}\right)^2] = \sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(Y_{it} - \overline{Y}_t\right)^2 + \sum_{t=1}^{T} n_t\left(\overline{Y}_t - \overline{Y}\right)^2$$

Clearly the relative magnitudes of TRSS and ERSS tell us something about the relative importance of the treatment variability and the pure random variability. Before making the comparison we should rescale them by their degrees of freedom (just like we did the sum of squared deviations from mean whenever we used that in a test). This raises the question what are the degrees of freedom associated with each concept?

**The Degrees of Freedom.**

**Total Sum of Squares (TOSS)**

The degrees of freedom associated with the TOSS is straightforward it is simply a sum of N squared deviations from mean terms and as before it will have N-1 degrees of freedom (just like the sum of squared deviations from mean employed in the variance estimators earlier).

**Error Sum of Squares (ERSS)**

The degrees of freedom associated with the Error sum of squares (ERSS) is best understood by seeing it as a sum of T squared deviations from mean terms of the form:

$$\sum_{i=1}^{n_t}\left(Y_{it} - \overline{Y}_t\right)^2 \quad for\ all\ \ t = 1, ....., T$$

where each typical term has $n_t$ - 1 degrees of freedom so that the sum of them will have:

$$\sum_{t=1}^{T}\left(n_t - 1\right) = \sum_{t=1}^{T} n_t - \sum_{t=1}^{T} 1 = N - T$$

degrees of freedom.

**Treatment Sum of Squares (TRSS)**

The degrees of freedom associated with the treatment sum of squares can be understood by noting that the TRSS is really a weighted sum of T squared deviations of the treatment means from the overall mean. Noting that the overall mean is a weighted sum of the treatment means this means that this sum has only T-1 independent components since:

$$\sum_{t=1}^{T} n_t \left( Y_t - \overline{Y} \right) = \sum_{t=1}^{T} n_t Y_t - \sum_{t=1}^{T} n_t \overline{Y} = \sum_{t=1}^{T} \sum_{i=1}^{n_t} Y_{it} - \sum_{t=1}^{T} \sum_{i=1}^{n_t} Y_{it} = 0$$

that is the weighted sum of deviations of the treatment mean from the overall mean is equal to 0.

Notice that the degrees of freedom obey the same adding up rule as the concepts themselves namely:

$$\begin{aligned} \text{TOSS} &= \text{ERSS} + \text{TRSS} \\ \text{N-1} &= \text{N-T} + \text{T-1} \end{aligned}$$

**The Test Statistic.**

The ratio of the treatment sum of squares to the error sum of squares, each divided by their respective degrees of freedom, will tell us something of the relative importance of the treatment effects as opposed to the error effects in the overall variability of the $Y_{it}$'s. In fact this ratio can be shown to be an F(T-1, N-T) random variable so that:

$$\frac{\frac{TRSS}{T-1}}{\frac{ERSS}{N-T}} \sim F\left(T-1, \text{N}-\text{T}\right)$$

If TRSS is close to 0 then this ratio will be close to 0 implying that the variability of the $Y_{it}$'s is largely due to variability in the $e_{it}$'s. If, on the other hand, the ERSS is close to 0 then the ratio will be a large number implying that the variability of the $Y_{it}$'s is largely due to variability in the $\mu_t$'s. Thus to test $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_T$ against $H_1$: $\mu_j \neq \mu_k$ for some $j \neq k$, j,k = 1, 2,..,T we can use the ratio in a one-sided upper tailed test with large values of the statistic rejecting the null hypothesis and small values of the statistic failing to reject the null.

**Simple formulae for ease of calculation.**

The easiest way to do calculate the test is to calculate TOSS and ERSS from the following formulae and then calculate TRSS = TOSS – ERSS.

$$TOSS = \sum_{t=1}^{T}\sum_{i=1}^{n_t} Y_{ti}^2 - n\bar{\bar{Y}}^2$$

$$ERSS = \sum_{t=1}^{T}\sum_{i=1}^{n_t} Y_{ti}^2 - \sum_{t=1}^{T} n_t \bar{Y}_t^2$$

Notice that the first double sum term in each formula is the same so it only need be calculated once.

Finally, if it is determined that the different treatments do have different effects then we can pursue difference in means tests of the different treatment means to see which one of the treatments is best.