Basic Probability Reference Sheet

17.846, 2001

This is intended to be used in addition to, not as a substitute for, a textbook.

X is a random variable. This means that X is a variable that takes on value X with probability x. This is the density function and is written as:

$$\text{Prob(X=x)} = f(x)$$

for all

$$X \in domain$$

The cumulative probability distribution is the probability that X takes on a value less than or equal to x. This is written as:

$$\text{Prob(X} \le x) = F(x)$$

for all

$$X \in domain$$

A probability distribution is any function such that

$$f(x) > 0 \text{ and } \int_{-\infty}^{\infty} f(x)dx = 1$$

or if X is discrete

$$\sum_{x \in domain} p(x) = 1$$

These say that the probability of any event if zero or positive, and that the sum of the probabilities of all events must equal one. (In other words, you can't have a negative probability of something happening, and something must happen.)

Two important probability distributions are the standard normal and the uniform[0,N]:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

and

$$f(x) = \frac{1}{N}$$

Most probability distributions have a mean and a variance.

Mean = μ and Variance = σ^2

The standard deviation (the average distance that the sample random variable is from the true mean) is equal to the square root of the variance.

Mean is also equal to the first moment, or expected value of X.

Mean = E[X]

The mean is the average value of X, weighted by the probability that X = x, for all values of x. For a continuous distribution, this is:

$\mu = \int_{-\infty}^{\infty} x \bullet f(x) dx$ or for discrete variables $\mu = \sum_{x \in domain} x \bullet f(x)$

The expected value operator is linear. That mean:

$$E[aX+b] = aE[X]+b$$

This is true because integration and summation, the methods we use to calculate the mean, are also linear operators.

The variance is also called the second, mean-deviated moment. It is formally:

$$E[X-E[X]]^2$$

This can be reduced a bit...

$= E[E[X^2]-2XE[X]+E[X]^2] = E[X^2]-2E[X]E[X]+E[X]^2 = E[X^2]-E[X]^2$ since the expected value of an expected value is just the expected vale (E[X] is a constant, and the expected value of a constant is simply the constant).

For a continuous distribution, this is:

$$\int_{-\infty}^{\infty}(x-\mu)^2 \bullet f(x)dx$$

For a discrete distribution, this is:

$$\sum_{x \in domain}(x-\mu)^2 \bullet f(x)$$

When working with variances, most of what you need can be derived from a few simple formulas:

$Cov(aX, bY) = ab(X, Y)$ which is often used in the form $Cov(X, -Y) = -Cov(X, Y)$

and especially $Cov(X+Y) = Var(X)+Var(Y)+2Cov(X, Y)$

In the case of independence, the covariance of X and Y is zero, and we get the rule that the variance of a sum equals the sum of the variance whenever the variables are independent from each other. This can be generalized to N variables:

$Var\left(\sum_{i=1}^{N}X_i\right) = \sum_{i=1}^{N}Var(X_i)$ since we know by independence that $\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}Cov(X_i, X_j) = 0$

Here is the intuition: Even though the variance is the sum of squared deviations, the variance of a sum of increases linearly. This is because if you take the sum, the deviations *tend to cancel out.*

Thus, if the variance of a coin flip (heads = 1, tails = 0) is 0.25 (0.5 * 0.5), the variance of the sum of 100 coin flips is 25, not 0.25*(100 squared). This is because the sum centers around 50... in other words, the chance of getting 100 heads is really small. Thus, if we were to flip 100 coins and take the sum, then repeat this exercise 1000 times, we would get a bell shaped curve. If we flip two coins, take the sum, and repeat 1000 times, we do not get a bell shaped curve. We get a histogram with 0 25% of the time, 1 50% of the time, and 2 25% of the time. This yields a variance of:

$$[0.25(2-1)^2 + 0.5(1-1)^2 + 0.25(1-0)^2] = 0.5$$

If we flip just one coin, we get a 0 half the time and a 1 half the time, for a variance of 0.25.

This may not sound so important, but it becomes important when we combine it with the following rule:

$$Var(aX+b) = a^2 Var(X)$$

This equation can be derived directly from the expectation formulas, and is highly intuitive. If we go back to our single coin flip, the variance of a single coin flip is:

$$0.5(1-0.5)^2 + 0.5(0-0.5)^2 = 0.25$$

But the variance of 100 * a single coin flip is:

$$0.5(100-50)^2 + 0.5(0-50)^2 = 2500 = 100^2 \cdot 0.25$$

So you can see the difference between the variance of 100 times a single coin flip, and the variance of the sum of 100 coin flips. Mathematically:

$$Var(kX) = E[kX - E[kX]]^2 = E[k(X - E[X])]^2 = k^2 E[X - E[X]]^2 = k^2 Var(X)$$

Now we can caluculate the variance of a sample mean from the variance of the random variable itself:

$$Var(\bar{X}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} Xi\right) = \left(\frac{1}{N}\right)^2 Var\left(\sum_{i=1}^{N} Xi\right) = \frac{1}{N^2} \cdot N \cdot Var(Xi) = \frac{1}{N} \cdot Var(Xi)$$

Or, using standard deviations:

$$Stdev(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{\frac{1}{N} \cdot Var(X)} = \frac{1}{\sqrt{N}} \cdot Stdev(X)$$

So after estimating the mean and variance of X normally, you can estimate the mean and variance of the mean of X. If you combine this with the central limit theorem, this forms the entire basis of hypothesis testing and confidence interval estimation in regression analysis.

So far we have talked about the true mean and true variance. In real life, we don't know the true values. All we see is the data geenrated by some mysterious natural or human process. We assume this process can be described by a mathematical relationship (such as a probability distribution). If so, and if the probability distribution which determines how the world operates isn't too weird, we can estimate the parameters of that distribution. We do so by using the data we find in the world.

The problem is that the world is noisy - is full of error. Or perhaps our measurements are noisy. Either way, we need some way of accounting for or at least quantifying this noise. This is why we estimate the variance of our observations.

So, if X follows some distribution, and we take 100 samples of X...

{X1, X2, X3, X4, X5.... X99 X100}

We can use this sample to estimate the value of the mean of whatever distribution generated X. Our best guess at this true mean is in fact $\bar{X}$, which can be defined:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{100} Xi$$

It turns out that this is an unbiased estimated of the mean. In other words, the expected value of this sum is equal to the true mean. This is true because the expected value of any single Xi is equal to the mean, so 1/N times N such expected values of Xi is also equal to the mean. The only advantage of using more observations is that the extimate of the mean becomes more precise. We measure this using the variance.

First, we estimate the variance of Xi. We do this with the formula:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X-\bar{X})^2$$

We have an N-1 in the denominator to account for the fact that we used up a degree of freedom to estimate the sample mean. In other words, estimating the sample variance with the *same data* that we used to estimate the sample mean gives us some wiggle room, and we need to penalize ourselves for doing this. [Technically, we prove that dividing by N-1 gives us an unbiased estimator by breaking the sample variance into two pieces, and showing they are independent of each other, which is a neat little trick.]

The next useful result is the **Central Limit Theorem**, which tells us that no matter what the distribution of X (within certain limits), the distribution of $\bar{X}$ is normal as long as $\bar{X}$ uses enough observations (about 20-30, it turns out). Since we can calculate the mean and variance of X, and we know the shape of the distribution of $\bar{X}$, we can use this knowledge to make some statements about $\bar{X}$. For example, $\bar{X}$ will fall within 1.96 standard deviations of the true mean μ 95% of the time. If σ is the standard deviation of X, then $\bar{X}$ will fall within $1.96 \cdot \frac{\sigma}{\sqrt{N}}$ of μ 95% of the time. Or in other words, μ will be within $1.96 \cdot \frac{\sigma}{\sqrt{N}}$ of $\bar{X}$ 95% of the time. Thus, our 95% confidence interval is $\bar{X} \pm \left(1.96 \cdot \frac{\sigma}{\sqrt{N}}\right)$.

Now since we don't know the true σ (just as we don't know the true mean), we substitute in our estimated standard deviation $\hat{\sigma}$ for the true standard deviation. Thus we obtain our confidence intervals. Hypothesis testing is conducted by a similar procedure. We ask, if it were true that μ = K, what is the probability that a randomly generated $\bar{X}$ equals or is further away from the value of $\bar{X}$ that we observed? Well, if μ really does equal K, and $\bar{X}$ is normally distributed around μ with a standard deviation of $\frac{\sigma}{\sqrt{N}}$, then we reject the hypothesis that K = μ with 95% certainty if $|\bar{X} - K| > 1.96 \cdot \frac{\sigma}{\sqrt{N}}$. K is often zero, and the T statistic in your STATA regressions measures how many standard deviations your estimated mean is away from zero. This gives you an implicit hypothesis test against the null hypothesis than nothing is going on, or the true mean actually is zero. For different certainty levels (other than 95%) we use values other than 1.96. These values can be read off of Z tables (for the normal), or T tables if you really want to be precise.

Regression Analysis and Statistics

The above notes deal entirely with statistics. Regression itself need not be thought of a statistical operation. Instead, it can be considered a tool that uses linear algebra to calculate conditional means. We then use our knowledge of statistics to make probabilistic statements about the parameters we calculate using linear algebra.

What is a regression? Geometrically, a regression is the projection of a single vector (your dependent variable) of a dimension N (equal to your number of observations) onto a K-dimensional subspace spanned by the K vectors which we call independent variables (also of dimension N, that is, with N observations). The subspace spanned by your independent variable vectors defines a hyperplane - a K-dimensional space which can be reached by some combination of your independent variable vectors. Regression then finds the point in that hyperplane which is closest to the point reached by your dependent variable vector. In fact, it constructs a whole line of points in the independent variable hyperplane which are closest to the points in your dependent variable vector. It then tells you in what combination you should combine your independent variable vectors to reach this closet-points vector (or the best guess vector). The best-guess vector is also called the projection of Y onto the subspace spanned by the X's. Your coefficients B1, B2, etc... tell us how to create the projection line given our X1, X2, etc... vectors. The vector of error terms defines a vector perpendicular to your dependent variable hyperplane. By combining your X's in the proportions defined by your Bs, and then adding the error term, you get back to your Y vector.

Back to the real(?) world. To calculate your B coefficients, you write down the projection equation:

$$\hat{Y} = X'(X'X)^{-1}X'Y = X'B \text{ where } B = (X'X)^{-1}X'Y$$

Notice that if X and Y have only one variable, this reduces to:

$$\hat{B} = \frac{\sum XY}{\sum X^2}$$

Actually, even sinple regressions with one dependent variable usually include an extra X vector of constants, or 1's, in order to allow for an intercept. If the 1's vector is the only vector in the matrix X, then B equals the mean of Y. If there is another vector in addition to the 1's vector, then:

$$\hat{B} = \frac{\sum (X-\bar{X})\cdot(Y-\bar{Y})}{\sum (X-\bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$

If you are actually regressing X on Y, the reverse regression yields:

$$\hat{B} = \frac{\sum (X-\bar{X})\cdot(Y-\bar{Y})}{\sum (Y-\bar{Y})^2} = \frac{Cov(X,Y)}{Var(Y)}$$

The only difference is the scaling factor in the denominator. For an unbiased scaling factor, we can use:

$$Estimated\ Correlation(X, Y) = \frac{\sum (X - \bar{X}) \cdot (Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}} = \frac{Cov(X, Y)}{Stdev(X) \cdot Stdev(Y)}$$

The correlation coefficient is always bounded between -1 and 1. To grasp the intuition behind this, imagine that X and Y are plotted, and form a perfect line. Then the correlation is 1 or -1. When you introduce noise to that line, the noise terms are summed and then multiplied in the denominator, whereas they are multiplied and then summed in the numerator. So the numerator is smaller. Imagine that we have mean deviated the variables already, and thus the sample means are equal to zero. Then the equation reduces to:

$$\frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} = \sqrt{\frac{(\sum XY)^2}{\sum X^2 \sum Y^2}}$$

Since we have mean-deviated, X and Y measure the distance (positive or negative) away from 0. Thus, their product sort of measures how they move together. If you have two long X and Y vectors, and for each observation Y is always positive when X is positive, and Y is always negative when X is negative, then you will have positive correlation. If the reverse is true, you will have negative covariation. Moreover, the degree to which X and Y move the same amount in the same direction yields the magnitude of the correlation. If X = Y always, you get 1, as you can see.