

Economics 406 – Fall 2009
Professor Jeffrey Smith
Lecture: Review of the fundamentals of probability
Version of September 14, 2009

Random variables and their probability distributions

Definitions

Define an experiment: a procedure that can be infinitely repeated and that yields a well-defined set of outcomes.

Random variable: a variable that takes on values determined by an experiment. It need not be a number but it is convenient to define it as such.

Bernoulli (binary, dummy, indicator) random variable: takes on values zero or one.

Discrete random variables

A discrete random variable takes on a finite or countably infinite number of values.

Explain what countably infinite means.

Example: how many cars you have, how many years of schooling you have.

The probability density function defines the probability of each possible outcome. Thus, for random variable X ,

$$f(x_j) = p(x_j) = \Pr(X = x_j)$$

Note the upper and lower case notation.

The probabilities must sum to one.

Continuous random variables

In formal terms, these variables take on any particular value with probability zero.

In practice, it is any variable with a lot of values, none of which has a very large probability by itself.

Draw a continuous pdf

While individuals do not have positive probability, intervals do. Their probability equals

$$\Pr(a < X < b) = \int_a^b f(x)dx$$

Draw the interval and the area on the graph.

Example: price of gold, even though denominated in cents so that it is in principle finite.

Example: years of schooling. Explain that sometimes you will want to treat a variable as discrete for some purposes and continuous for others.

The cumulative distribution function (cdf) gives the probability that a continuous random variable is less than or equal to some value. That is

$$\Pr(X < a) = F(a) = \int_{-\infty}^a f(x)dx$$

Draw a picture of the cdf.

Note that $0 \leq F(X) \leq 1$. Explain.

Note that $\Pr(a < X < b) = F(b) - F(a)$. Explain.

Hybrids

Some random variables have elements of both discrete and continuous.

Earnings are a good example. Lots of people earn zero earnings, but there is a more or less continuous distribution of values above zero. Thus, zero has positive probability but non-zero values do not.

The values are zero are called a mass point.

Draw a picture of a hybrid pdf.

Joint distributions, conditional distributions and independence

Joint distributions and independence

Joint distributions indicate the probability of combinations of values of more than one random variable.

Consider the example of a direct mail retailer that sends out catalogs with different prices to random households on its mailing list in order to estimate demand curves. What it collects from this exercise are pairs of (price, quantity), or possibly vectors of (price, quantity, characteristics).

The firm is explicitly interested in the relationship between price and quantity and is deliberately inducing exogenous variation in price in order to identify this relationship.

We write the joint distribution as $f_{X,Y}(x, y)$ or just as $f(x, y)$. This is true in the continuous or the discrete case.

Do the discrete example of the direct marketer with two prices, \$20 and \$25, and three quantities, 0, 1 and 2 or more. Probabilities can be 0.15, 0.20, 0.15 and 0.20, 0.20, 0.10.

Draw a picture of the cdf for two random variables.

Two random variables are independent if their joint distribution is the product of their marginal distributions, where the marginal is just the pdf for each variable considered alone.

In notation, two random variables are independent if and only if

$$f(x, y) = f(x)f(y)$$

Another way to state this is that if two variables are independent, knowing the value of one is not informative about the other.

Ex: rolling dice or flipping a fair coin.

Ex: “hot” sports figures. Is this a myth or a reality?

Ex: price and quantity pairs for our direct marketer. The marketer is interested in price and quantity precisely because they are not independent.

Conditional distributions

A conditional distribution indicates the distribution of one random variable given another.

When X and Y are discrete random variables, we write

$$f_{Y|X}(y|x) = \Pr(Y = y | X = x).$$

In general, we can define the conditional distribution using Bayes' rule, so that

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

When we estimate the parameters of a linear regression model, we will be estimating the parameters of a conditional distribution – namely the conditional mean of the dependent variable given the independent variables.

If two random variables are independent, then their conditional and unconditional (marginal) distributions are the same. That is:

$$f_{Y|X}(y|x) = f_Y(y)$$

To think about conditional distributions, return to the example of the catalog retailer. We can think about the distribution of sales conditional on price. There will be a distribution of shirt sales conditional on the low price and another conditional on the high price. Both will be of interest.

Features of probability distributions

Expected value

This is a measure of central tendency.

It equals a weighted average of all possible values of the random variable where the weights are given by the pdf.

Thus, in the case of a discrete random variable,

$$E(X) = \sum_{j=1}^J x_j f(x_j)$$

While in the case of a continuous random variable,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

We will care about expected values because we would like our estimators to be unbiased – that is, to have as their expected value the true, or population, value of the thing being estimated. That is, we want

$$E(\hat{\beta}_1) = \beta_1$$

We will talk more (well, I will talk more) about unbiasedness in the coming weeks.

Note that the expected value need not be any value that the random variable can take on. Consider a six sided die. The expected value is

$$\frac{1}{6}(1+2+3+4+5+6) = \frac{1}{6}(21) = \frac{21}{6} = \frac{7}{2} = 3\frac{1}{2}$$

Be sure to explain how to get to the first term in this equation.

The expected value of a function of a random variable is calculated in the same way, but with the value of the function replacing the value of the random variable. For example, with a discrete random variable you have

$$E(g(X)) = \sum_{j=1}^k f_j(x_j)g(x_j)$$

Note that, in general, $E(g(X)) \neq g(E(X))$. This has to do with the linearity of the expectations operator, which we will shortly discuss.

If time, give a simple example using X^2 .

Properties of expected values

The expected value of a constant is a constant: $E(c) = c$.

A constant takes on one value with probability one, so this just follows from the formula.

The expected value of a linear combination of random variables is just the linear combination of the expected values, or:

$$E(a_1x_1 + a_2x_2 + \dots + a_kx_k) = a_1E(X_1) + a_2E(X_2) + \dots + a_kE(X_k)$$

This is a very handy formula.

Medians

An alternative measure of central tendency is the median of a random variable.

The median is the value such that $F(\text{median}(X)) = 0.5$. In words, this means that half the time the random variable will be less than the median and half the time the random variable will be greater than the median.

The median is more robust than the mean in the sense that it is not influenced by extreme values with low probabilities. In some contexts this is a major advantage.

The disadvantage to the median is that it is not as easy to manipulate using simple rules like the ones we just covered for the mean.

The median and mean generally do not coincide, but do coincide for symmetric distributions.

Draw a picture to make this point.

Variance

Variance is a measure of the spread of a distribution – how much it varies around its mean.

Let $\mu = E(X)$. Then the variance is given by:

$$\text{var}(X) = \sigma_X^2 = E[(X - E(X))^2] = E[(X - \mu)^2] > 0$$

Note that the squaring means that the inside value is always positive, which is a nice feature. You do not want positive distances canceling negative distances.

Using a pair of graphs, show how distributions with different spreads will have different variances.

The variance of a constant is zero, or $\text{var}(c) = 0$. Again, this follows immediately from the formula.

The variance of a constant times a random variable is the constant squared times the value of the random variable, or

$$\text{var}(aX) = a^2 \text{var}(X).$$

If time, show how this follows from the formula with $X = aX$.

Standard deviation

The standard deviation is just the square root of the variance, or

$$\text{sd}(X) = \sigma_X = \sqrt{\text{var}(X)}.$$

The standard deviation is nice because it is in the same units as the original variable, while the units for the variance are the original units squared.

Standardizing a random variable

Sometimes it is useful to convert a random variable into a kind of common format. The common format that is typically used is a mean of zero and a variance of one.

This transformation is called standardizing the variable and can be accomplished as follows:

$$Z = \frac{X - \mu_x}{\sigma_x}.$$

In the transformation, μ is the expected value and σ is the standard deviation.

Verify that the standardized variable has mean zero.

Verify that the standardized variable has variance one.

We will use this transformation a lot when doing hypothesis testing.

Features of joint and conditional distributions

Covariance

We are interested in measures that describe the nature and strength of the inter-relationship of two variables.

The covariance between two random variables X and Y is given by:

$$\text{cov}(X, Y) = \sigma_{XY} = E[(X - \mu_x)(Y - \mu_y)] = E(XY) - \mu_x \mu_y$$

Explain the formula. If the covariance is positive, then on average, when X is above its mean, so is Y . If the covariance is negative, then on average, when X is above its mean, Y is below its mean.

Note that the variance is just a special case of the covariance.

If two random variables are independent, then their covariance equals zero.

The reverse is not true. A good counterexample is $Y = X^2$, where $X = 1$ or -1 , each with probability 0.5. These are clearly not independent and yet their covariance is zero. If time, show this using the middle term in the formula.

For linear functions of random variables, we have:

$$\text{cov}(a_1 X, a_2 Y) = a_1 a_2 \text{cov}(X, Y)$$

This is useful and also makes a point: re-scaling one variable will change the covariance, which is not a very desirable property.

Another useful property is the Cauchy-Swartz inequality, which is that

$$|\text{cov}(X, Y)| \leq \sigma_X \sigma_Y$$

Motivate this using the extreme cases where $X = X$ or $X = -X$.

Correlation coefficient

Because of the problems with scaling and the covariance, there is interest in a measure of the interrelationship between two variables without this feature. The correlation coefficient re-scales the covariance to provide such a measure.

The correlation coefficient is given by:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Note that the covariance and the correlation always have the same sign, because the denominator is always positive.

The correlation will be zero if and only if the covariance is zero, which means that correlation has the same relationship to independence that the covariance does.

The correlation is bounded, so that:

$$-1 \leq \text{corr}(X, Y) \leq 1$$

Give the intuition behind this result. The bounds are obtained at the extreme values where X equals itself or minus itself (or a linear function of itself). This result also follows from the Cauchy-Schwartz inequality with some division and some thought.

The correlation, unlike the covariance, is invariant to units of measurement. Thus, for $a_1 a_2 > 0$,

$$\text{corr}(a_1 X + b_1, a_2 Y + b_2) = \text{corr}(X, Y),$$

and for $a_1 a_2 < 0$,

$$\text{corr}(a_1 X + b_1, a_2 Y + b_2) = -\text{corr}(X, Y)$$

Variance of sums of random variables

For two random variables, we have, based on the preceding results,

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \text{cov}(X, Y).$$

When X and Y are independent (or at least uncorrelated), so that their covariance equals zero, this simplifies to:

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y).$$

Note that this means that $\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$.

Explain why you add here rather than subtracting. A simple example with

$$\text{var}(X) = \text{var}(Y) = 10$$

might do the trick.

These formulae all generalize to sums of random variables. Thus, we can derive the variance of the mean of a set of identically distributed independent random variables:

$$\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \left(\frac{1}{n}\right)^2 \text{var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}.$$

Pretty cool, eh?

Conditional expectation

The conditional expectation, or conditional mean, is written as $E(Y | X = x)$ or just $E(Y | x)$.

It indicates the expected value of Y for a given value of X .

In the catalog example, we can think about the expected number of sweaters sold for catalogs with the low prices versus the expected number of sweaters sold for sweaters with the high prices. Essentially, you simply construct the expected value by taking a weighted sum along the appropriate row of the table.

In symbols, for discrete Y we have

$$E(Y | x) = \sum_{j=1}^m y_j f_{Y|X}(y_j | x).$$

Note that this is just like the expected value, except that the conditional distribution of Y given $X = x$ is used for the weights.

The analog for continuous Y is given by:

$$E(Y | x) = \int_{-\infty}^{\infty} y f_{Y|X}(y | x) dy.$$

In the case of regression, we parameterize the conditional expectation as a linear function of X . So what we will be doing for most of the semester is thinking about conditional expectation functions of the form:

$$E(Y | X) = \beta_0 + \beta_1 X$$

Note that there is no error term here. Explain.

Properties of conditional expectation

Functions of X act like constants when computing conditional expectations with respect to X . Thus,

$$E(c(X) | X) = c(X).$$

Furthermore,

$$E(a(X)Y + b(X) | X) = a(X)E(Y | X) + b(X).$$

Explain in words what this means. It is really pretty simple; it just means you can plug in the $X = x$.

If X and Y are independent then $E(Y | X) = E(Y)$.

In words, knowing X provides no information about the expected value of Y , as you would expect if they are independent.

The law of iterated expectations states that:

$$E_X(E_Y(Y | X)) = E(Y).$$

Explain the subscript notation – it indicates the variable being summed or integrated over in computing the expectation.

Explain how this works. Give an example if time using the catalogs (or a simpler 2 x 2 example).

A more general version of the law of iterated expectations states that

$$E(Y | X) = E_Z(E_Y(Y | X, Z) | X).$$

Explain how this is really the same thing.

Like the correlation, the relationship between independence and conditional expectation is asymmetric. We saw before that independence implies that the conditional expectation equals the unconditional expectation. In contrast, the conditional expectation equaling the unconditional expectation does not imply independence, but does imply a zero covariance and correlation. Indeed, a conditional expectation equal to the unconditional expectation implies that every function of X is uncorrelated with Y .

The asymmetry has to do with the fact that the covariance and by extension the correlation measure linear association, not any association.

A useful implication of the last two properties is that if U and X are random variables such that $E(U | X) = 0$, then $E(U) = 0$ and U and X are uncorrelated.

Explain how $E(U) = 0$ follows from the law of iterated expectations.

Foreshadow the link to regression analysis, with U as the error term in the regression.

Conditional variance

The conditional variance is like the conditional mean. It gives the variance of one variable conditional on values of the other.

Thus,

$$\text{var}(Y | X = x) = E[(Y - E(Y | x))^2 | x] = E(Y^2 | x) - [E(Y | x)]^2.$$

If X and Y are independent random variables, then $\text{var}(Y | X) = \text{var}(Y)$.

Explain how this follows directly from independence.

The normal and related distributions

The normal distribution

The normal distribution plays a key role in developing test statistics and conducting hypothesis tests within a regression framework.

The density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

where $E(X) = \mu$ and $\text{var}(X) = \sigma^2$.

Go through each term. Note that the first term is positive by definition, as is the second due to the exponential function. This is good because densities should be positive.

Note that symmetry comes from the fact that the expected value is subtracted off in constructing the density and the result then squared.

Draw the picture. Lots of things in the world have normal distributions as given or after you take their natural log. Examples of the former are heights, weights and test scores. An example of the latter is income in many countries.

The notation here is $X \sim N(\mu, \sigma^2)$. Read it out. Note that this is a two-parameter distribution.

The standard normal distribution

A special case of the normal distribution is the standard normal, which has mean zero and variance one.

There is special notation for both the pdf and cdf of the standard normal. The pdf is given by:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

This formula is obtained in a trivial way by substituting in the values of the mean and variance of the standard normal into the earlier formula.

The $\Phi(x) = \Pr(Z < x)$ notation is used for the cdf of the standard normal.

Using this notation we have that $\Pr(a < Z < b) = \Phi(b) - \Phi(a)$. Draw the picture.

The cdf for the standard normal has been tabulated, which is part of the reason for all of the attention given to it. It is Table G.1 in Appendix G of Wooldridge.

The table is useful for any normally distributed random variable due to the ease of standardization. In fact, if $X \sim N(\mu, \sigma^2)$, then

$$\frac{X - \mu}{\sigma} \sim N(0, 1).$$

That is, any normal random variable, and any probability statement about a normal random variable, can be converted into a standard normal, or a probability statement about a standard normal, via standardization.

Properties of the normal distribution

Some properties that hold for any cdf are usefully noted here. First,

$$\Pr(Z > z) = 1 - \Pr(Z < z) = 1 - \Phi(z).$$

This follows from the fact that the probabilities sum (or integrate) to one. Draw the picture.

By symmetry, we have $\Pr(Z < -z) = \Pr(Z > z)$. Draw the picture.

One very useful feature of normal random variables is that linear combinations of independent normal random variables are themselves normal.

Thus, for example, if $X_1 \square N(\mu_1, \sigma_1^2)$ and $X_2 \square N(\mu_2, \sigma_2^2)$, and the two random variables are independent, then

$$a_1 X_1 + a_2 X_2 \square N(a_1 \mu_1 + a_2 \mu_2, a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2).$$

If the two variables are dependent, the resulting distribution is still normal, but the variance includes the covariance term from the standard formula.

This result does not hold for many other common distributions. For example, the sum of a $U[0,2]$ and a $U[1,3]$ variable is not a uniform random variable.

This result means, for example, that the distribution of the sample mean of a bunch of independent normal random variables with a common mean and variance is given by:

$$\bar{X} \square N(\mu, \sigma^2 / n).$$

Explain this in light of the earlier discussion of means and variances of sums of independent random variables.

Finally, two normal random variables are independent if and only if their covariance (and correlation) equal zero. This is a stronger result than the result for all random variables.

The Chi-Square distribution

The normal, t, Chi-square and F distributions all fit together, and all play important roles in inference in the linear regression model. I did not really figure this out until the first time I taught this material – you, on the other hand, get to figure it out today!

A Chi-squared distribution is what you get if you create a random variable by summing a bunch of independent standard normal random variables.

In notation,

$$X = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

The degrees of freedom refer to the number of random variables in the summation.

Note that this is a one-parameter distribution.

Foreshadow using the Chi-square distribution to look at sums of standardized residuals from a regression model.

The t distribution

The t distribution is what you get if you take the ratio of a standard normal random variable and the square root of a Chi-square random variable divided by its degrees of freedom.

That is, $T = \frac{Z}{\sqrt{X/n}} \sim t_n$.

The t distribution will come in handy for conducting tests.

The t distribution has the same basic shape as the normal distribution, but with fatter tails. As the number of degrees of freedom gets large, the t distribution gets closer and closer to the normal distribution.

The t distribution was developed by William Sealy Gossett in 1908. Because his employer, the Guinness brewery in Dublin, did not allow its workers to publish under their own names, he published as “Student”, which is why this is sometimes called the “Student’s t distribution”.

The F distribution

The F distribution is what you get if you take the ratio of two Chi-square random variables divided by their degrees of freedom.

In notation, let $X_1 \sim \chi_{k_1}^2$ and $X_2 \sim \chi_{k_2}^2$ be two independent random variables.

$$\text{Then } F = \frac{\left(\frac{X_1}{k_1}\right)}{\left(\frac{X_2}{k_2}\right)} \sim F_{k_1, k_2}.$$

This distribution will prove useful for comparing variances of residuals. Give some intuition and foreshadow what we will be doing.

Stata commands

Stata includes the following functions

`normden(z)` gives the standard normal density evaluated at z .

`normden(z, σ)` gives the density of a normal with mean zero and standard deviation σ .

`normden(z, μ , σ)` is the same but for mean μ rather than mean zero.

`norm(z)` gives the CDF of the standard normal evaluated at z .

These functions can be used with the `generate` command to create new variables, as in

```
generate normvar = norm(z)
```

Stata includes a wealth of other built-in statistical functions. Just type

```
help functions
```

in Stata for details.