

## Elementary Probability Review Continued Economics 5330, Fall 2016

This is a review of elementary probability that will be useful for our study of financial econometrics. It is based on coverage Wooldridge (2004).

The **cumulative distribution function (CDF)** of the random variable  $X$  is:

$$F(X) = P(X \leq x)$$

For discrete random variables it is obtained by summing the PDF over all values  $x_j$  such that  $x_j \leq x$ .

For a continuous random variable,  $F(X)$  is the area under the PDF,  $f(x)$  to the left of  $x$ .

Because it is a probability,  $0 \leq F(X) \leq 1$ .

If  $x_1 < x_2$  then  $P(X \leq x_1) \leq P(X \leq x_2)$ , that is  $F(x_1) \leq F(x_2)$ .

Two important properties of CDFs that are useful for computing probabilities are the following:

- For any number  $c$ ,  $P(X > c) = 1 - F(c)$
- For any numbers  $a$  and  $b$ ,  $P(a \leq X \leq b) = F(b) - F(a)$

For continuous random variables the inequalities in probability statements are not strict:

$$P(X \geq c) = P(> c)$$

$$\begin{aligned} P(a < X < b) &= P(a \leq X \leq b) \\ &= P(a \leq X < b) \\ &= P(a < X \leq b) \end{aligned}$$

Let  $X$  and  $Y$  be discrete random variables. Then for  $(X, Y)$  a **joint distribution** which is fully described by the **joint probability density function** of  $(X, Y)$ :

$$f_{XY}(x, y) = P(X = x, Y = y)$$

$X$  and  $Y$  are said to be independent if, and only if:

$$f_{XY}(x, y) = f_X(x)f_Y(y) \quad \text{for every } x \text{ and } y$$

where  $f_X$  is the PDF of the random variable  $X$ , and  $f_Y$  is the PDF of random variable  $Y$ .  $f_X$  and  $f_Y$  are referred to as the **marginal probability density functions**.

The discrete case is the easiest to grok. If  $X$  and  $Y$  are discrete and independent then

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Note: If  $X$  and  $Y$  are independent then finding the joint PDF only requires knowledge of  $P(X = x)$  and  $P(Y = y)$

Example: Consider a basketball player shooting two free throws. Let  $X$  be the Bernoulli random variable equal to 1 if he makes the first free throw, and 0 otherwise. Let  $Y$  be the Bernoulli random variable equal to 1 if he makes the second free throw. Suppose that he is an 80% free throw shooter, so that  $P(X = 1) = P(Y = 1) = 0.80$ . What is the probability of making both free throws?

If  $X$  and  $Y$  are independent:  $P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = (0.8)(0.8) = 0.64$ . Thus, a 64% chance of making both.

Independence is often reasonable in more complicated situations. In the airline example, suppose that  $n$  is the number of reservations booked. For each  $i = 1, 2, \dots, n$  let  $Y_i$  denote the Bernoulli random variable indicating whether or not customer  $i$  shows up for the flight.

Let  $\theta$  again denote the probability of success (showing up for the reservation). Each  $Y_i \sim \text{Bernoulli}(\theta)$ .

The variable of primary interest is the total number of customers showing up out of the  $n$  reservations: call this  $X$ .

$$X = Y_1 + Y_2 + \dots + Y_n$$

Assume that  $P(Y_i = 1) = \theta$  for every  $Y_i$ , and further that they  $Y_i$  are independent. Then  $X$  has a **binomial distribution**, which we write in shorthand as:  $X \sim \text{Binomial}(n, \theta)$ . The binomial PDF is the following:

$$f(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n$$

Note:  $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ , and is read as “n choose x”.

Example: If the flight has 100 seats and  $n = 120$  and  $\theta = 0.85$  then:

$$P(X > 100) = P(X = 101) + P(X = 102) + \dots + P(X = 120)$$

In econometrics we are usually interested in how one variable  $Y$  is related to one or more other variables. For now, consider only one such variable  $X$ . What we can know about how  $X$  affects  $Y$  is contained in the **conditional distribution** of  $Y$  given  $X$ . This information is summarized in the **conditional probability distribution function**:

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

In the discrete case:  $f_{Y|X}(y|x) = P(Y = y|X = x)$ , which we read as the probability that  $Y = y$  given that  $X = x$ .

If  $X$  and  $Y$  are independent, then the knowledge of  $X$  tells us nothing about  $Y$ :

$$f_{Y|X}(y|x) = f_Y(y) \quad \text{and} \\ f_{X|Y}(x|y) = f_X(x)$$

Example: Free throw shooting again. Assume the conditional PDF is given by the following:

$$f_{Y|X}(1|1) = 0.85, \text{ and } f_{Y|X}(0|1) = 0.15.$$

$$f_{Y|X}(1|0) = 0.70, \text{ and } f_{Y|X}(0|0) = 0.30.$$

These are not independent. The probability of making the second free throw depends on whether or not the first free throw was made. We can calculate  $P(X = 1, Y = 1)$  if we know  $P(X = 1)$ . Assume the probability of making the first free throw is  $P(X = 1) = 0.80$ . Then:

$$\begin{aligned} P(X = 1, Y = 1) &= P(Y = 1|X = 1) \times P(X = 1) \\ &= (0.85) \times (0.80) \\ &= 0.68 \end{aligned}$$

The **expected value** is a measure of central tendency. It is one of the most important probabilistic concepts in econometrics. If  $X$  is a random variable the **expected value** (or expectation) of  $X$ , denoted  $E(X)$  and sometimes  $\mu$ , is a weighted average of all possible values of  $X$ . The weights are determined by the PDF.

Consider the case of a discrete random variable. Let  $f(x)$  denote the PDF of  $X$ . The expected value of  $X$  is the weighted average:

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \dots + x_k f(x_k) = \sum_{j=1}^k x_j f(x_j)$$

Example: Suppose  $X$  takes on the values  $-1$ ,  $0$ , and  $2$  with probabilities  $\frac{1}{8}$ ,  $\frac{1}{2}$ ,  $\frac{3}{8}$ . Then

$$E(X) = (-1)\left(\frac{1}{8}\right) + (0)\left(\frac{1}{2}\right) + (2)\left(\frac{3}{8}\right) = \frac{5}{8}$$

Note:  $E(X)$  can take on values that are not even possible outcomes of  $X$ .

If  $X$  is a continuous random variable then

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

This is still interpreted as a weighted average.

Given a random variable  $X$  and a function  $g(\cdot)$ , we can create a new random variable  $g(X)$ . For example, if  $X$  is a random variable, then so is  $X^2$  or  $\log(X)$  (for  $x > 0$ ).

The expected value of  $g(X)$  is given by

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

or

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Example: For the random variable above let  $g(X) = X^2$ . Then

$$E(X^2) = (-1)^2\left(\frac{1}{8}\right) + (0)^2\left(\frac{1}{2}\right) + (2)^2\left(\frac{3}{8}\right) = \frac{13}{8}$$

Note:  $E[g(X)] \neq g[E(X)]$ .

Properties of Expected Values:

**Property E1:** For any constant  $c$ ,  $E(c) = c$ .

**Property E2:** For any constants  $a$  and  $b$ ,  $E(aX + b) = aE(X) + b$ .

**Property E3:** If  $a_1, a_2, \dots, a_n$  are constants and  $X_1, X_2, \dots, X_n$  are random variables then:

- $E(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n)$
- Or  $E\left(\sum_{i=1}^n a_iX_i\right) = \sum_{i=1}^n a_iE(X_i)$
- A special case is when each  $a_i = 1$  so that  $E\left(\sum_{i=1}^n E(X_i)\right) = \sum_{i=1}^n E(X_i)$ , or in other words the expected value of a sum, is the sum of the expected values.

Example: Expected revenue at a pizzeria.  $X_1$ ,  $X_2$ , and  $X_3$  are the number of small, medium, and large pizzas sold during the day. Suppose  $E(X_1) = 25$ ,  $E(X_2) = 57$ , and  $E(X_3) = 40$ . Prices are \$5.50 for a small, \$7.60 for a medium, and \$9.15 for a large. Then expected revenue is the following

$$\begin{aligned} E(5.50X_1 + 7.60X_2 + 9.15X_3) &= 5.50E(X_1) + 7.60E(X_2) + 9.15E(X_3) \\ &= 5.50(25) + 7.60(57) + 9.15(40) \\ &= 936.70 \end{aligned}$$

The outcome on any given day will differ from this, but this is the expected revenue.

If  $X \sim \text{Binomial}(n, \theta)$  then  $E(X) = n\theta$ . The expected number of successes in  $n$  Bernoulli trials is  $n\theta$ . We can see this by writing

$$X = Y_1 + Y_2 + \dots + Y_n \quad \text{where each } Y_i \sim \text{Bernoulli}(\theta)$$

Then

$$\begin{aligned} E(X) &= \sum_{i=1}^n E(Y_i) \\ &= \sum_{i=1}^n \theta \\ &= n\theta \end{aligned}$$

Example: Consider the airline problem with  $n = 120$  and  $\theta = 0.85$ . Then  $E(X) = n\theta = 120(0.85) = 102$ , which is too many.

The **median** is another measure of central tendency. If  $X$  is continuous then the median is the value  $m$  such that one-half of the area under the PDF is to the left of  $m$ , and one-half is to the right of  $m$ .

If  $X$  is discrete and takes on an odd number of finite values, the median is obtained by ordering the possible outcomes of  $X$  and selecting the middle value.

Example: For the sample  $\{-4, 0, 2, 8, 10, 13, 17\}$  the median is 8.

If  $X$  takes on an even number of values, then the median is the average of the two middle values.

Example: For the sample  $\{-5, 3, 9, 17\}$  the median is  $\frac{3+9}{2} = 6$ .

For a random variable let  $E(X) = \mu$ . There are various ways to measure how far  $X$  is from its expected value. One of the simplest is the squared distance:

$$(X - \mu)^2$$

This eliminates the sign, which corresponds with our intuitive notion of a distance measure. It treats values above and below  $\mu$  symmetrically.

The **variance** is defined as follows:

$$Var(X) = E[(X - \mu)^2]$$

The variance is sometimes denoted by  $\sigma_X^2$  or just  $\sigma^2$  when the random variable is understood to be  $X$ .

Note:

$$\begin{aligned} \sigma^2 &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

Example: If  $X \sim \text{Bernoulli}(\theta)$  we know that  $E(X) = \theta$ . Since  $X^2 = X$  it follows that  $E(X^2) = \theta$ . Then  $Var(X) = E(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$ .

Properties of variance:

**Property VAR1:**  $Var(X) = 0$  if, and only if for every  $c$  such that  $P(X = c) = 1$ , in which case  $E(X) = c$ .

**Property VAR2:** For constants  $a$  and  $b$   $Var(aX + b) = a^2 Var(X)$ .

The **standard deviation** is related to the variance as follows:  $sd(X) = \sqrt{Var(x)}$ . The standard deviation is often denoted  $\sigma_x$  or just  $\sigma$ .

Properties of the standard deviation:

**Property SD1:** For a constant  $c$ ,  $sd(c) = 0$ .

**Property SD2:** For constants  $a$  and  $b$   $sd(aX + b) = |a|sd(X)$ .

Given a random variable  $X$ , we can define a new random variable  $Z$  by

$$Z = \frac{X - \mu}{\sigma}$$

or  $Z = aX + b$  where  $a = \frac{1}{\sigma}$  and  $b = \frac{-\mu}{\sigma}$ . Then  $E(Z) = aE(X) + b = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$ .

The variance is  $Var(Z) = a^2 Var(X) = \frac{\sigma^2}{\sigma^2} = 1$ . Thus the new random variable has  $\mu = 0$  and  $\sigma^2 = 1$ . This is known as **standardizing** a random variable.

Example: Suppose  $E(X) = 2$  and  $Var(X) = 9$  then  $Z = \frac{X-2}{3}$ .

While the joint distribution completely describes the relationship between two random variables it is often useful to have a summary measure of how, on average, two random variables vary with one another.

The **covariance** is defined as follows:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

The covariance is often denoted by  $\sigma_{XY}$ . If  $\sigma_{XY} > 0$  then on average when  $X$  is above its mean  $Y$  is also above its mean. If  $\sigma_{XY} < 0$  then on average when  $X$  is above its mean  $Y$  is below its mean, and vice versa.

Note:

$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X - \mu_X)Y] \\ &= E(XY) - \mu_X \mu_Y \end{aligned}$$

Properties of covariance:

**Property COV1:** If  $X$  and  $Y$  are independent then  $Cov(X, Y) = 0$ . Note: the converse is not true. Zero  $Cov(X, Y)$  does not imply independence.

**Property COV2:** For any constants  $a_1$ ,  $b_1$ ,  $a_2$ , and  $b_2$   $Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X, Y)$ .

**Property COV3:**  $|Cov(X, Y)| \leq sd(X)sd(Y)$ .

Note: property COV2 suggests that  $Cov(X, Y)$  depends upon how the random variables are measured, not only on how strongly they are related. In other words, scale matters for  $Cov(X, Y)$ .

The **correlation coefficient** is defined as

$$Corr(X, Y) = \frac{Cov(X, Y)}{sd(X)sd(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

The correlation coefficient is sometimes denoted by  $\rho_{XY}$ .

Properties of correlation:

**Property CORR1:**  $-1 \leq Corr(X, Y) \leq 1$ .

**Property CORR2:** For constants  $a_1$ ,  $b_1$ ,  $a_2$ , and  $b_2$  with  $a_1a_2 > 0$   $Corr(a_1X + b_1, a_2Y + b_2) = Corr(X, Y)$ . If  $a_1a_2 < 0$  then  $Corr(a_1X + b_1, a_2Y + b_2) = -Corr(X, Y)$ .

With covariance and correlation defined we state further properties of the variance:

**Property VAR3:** For constants  $a$  and  $b$ ,  $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$ .

**Property VAR4:** If  $\{X_1, X_2, \dots, X_n\}$  are pairwise uncorrelated and  $\{a_i : i = 1, \dots, n\}$  are constants then  $Var(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 Var(X_i)$ .

The **conditional mean** is defined as follows:

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

Example: Let  $(X, Y)$  represent the population of all working individuals, where  $X$  is years of education and  $Y$  is hourly wages. Then  $E(Y|X = 12)$  is the average hourly wage for all the people in the population with 12 years of education (roughly high school education).  $E(Y|X = 16)$  is the average hourly wage for all people with 16 years of education.

A typical situation in econometrics will look like the following:

$$E(WAGE|EDUC) = 1.05 + 0.45EDUC$$

If this linear relationship holds then for 8 years of education the expected hourly wage is  $1.05 + 0.45(8) = 4.65$  or \$4.65 per hour.

Properties of conditional expectations:

**Property CE1:**  $E[c(X)|X] = c(X)$  for any function  $c(X)$ . In other words, functions act as constants. For example,  $E[X^2|X] = X^2$ . If we know  $X$  we also know  $X^2$ .

**Property CE2:** For functions  $a(X)$  and  $b(X)$ ,  $E[a(X)Y + b(X)|X] = a(X)E(Y|X) + b(X)$ . For example, consider the random variable  $XY + 2X^2$ .  $E(XY + 2X^2|X) = XE(Y|X) + 2X^2$ .

**Property CE3:** If  $X$  and  $Y$  are independent then  $E(Y|X) = E(Y)$ .

**Property CE4:**  $E[E(Y|X)] = E(Y)$ . This is known as the Law of Iterated Expectations.

**Property CE5:**  $E(Y|X) = E[E(Y|X, Z)|X]$ .

**Property CE6:** If  $E(Y|X) = E(Y)$  then  $Cov(X, Y) = 0$  and  $Corr(X, Y) = 0$ .

The **conditional variance** is defined as follows:

$$Var(Y|X = x) = E(Y^2|X) - [E(Y|X)]^2$$

Properties of conditional variance:

**Property CV1:** If  $X$  and  $Y$  are independent then  $Var(Y|X) = Var(Y)$ .

The **normal probability density function** is defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(X - \mu)^2}{2\sigma^2}, \quad \text{for } -\infty < x < \infty$$

where  $E(X) = \mu$  and  $Var(X) = \sigma^2$ . When a random variable is normally distributed we write  $X \sim N(\mu, \sigma^2)$ .

A special case is the **standard normal distribution**, which is defined as follows:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2}, \quad \text{for } -\infty < z < \infty$$

The standard normal cumulative distribution function is denoted by  $\Phi(z) = P(Z \leq z)$ . Using some basic facts from probability we arrive at the following helpful formulas:

$$\begin{aligned} P(Z > z) &= 1 - \Phi(z) \\ P(Z < -z) &= P(Z > z) \\ P(a \leq Z \leq b) &= \Phi(b) - \Phi(a) \end{aligned}$$



Properties of the normal distribution:

**Property NORMAL1:** If  $X \sim N(\mu, \sigma^2)$  then  $\frac{(X-\mu)}{\sigma} \sim N(0, 1)$ .

**Property NORMAL2:** If  $X \sim N(\mu, \sigma^2)$  then  $aX + b \sim N(a\mu + b, a^2\sigma^2)$ .

**Property NORMAL3:** If  $X$  and  $Y$  are jointly normally distributed, then they are independent if, and only if  $Cov(X, Y) = 0$ .

**Property NORMAL4:** Any linear combination of independent, identically distributed normal random variables has a normal distribution.

Example: Let  $X_i$  for  $i = 1, 2$ , and  $3$ , be independent random variables distributed as  $N(\mu, \sigma^2)$ . Define  $W = X_1 + 2X_2 - 3X_3$ . Then  $W$  is normally distributed. We can solve for the mean and variance as follows:

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0$$

$$Var(W) = Var(X_1) + 4Var(X_2) + 9Var(X_3) = 16\sigma^2$$

The **chi-square distribution** is obtained directly from independent, standard normal random variables. Let  $Z_i, i = 1, 2, \dots, n$ , be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the individual  $Z_i$ :

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

The new random variable  $X$  has a **chi-square distribution** with  $n$  **degrees of freedom**. This is often written as  $X \sim \chi_n^2$ .

The **t distribution** is a workhorse in classical statistics and econometrics. A  $t$  distribution is obtained from a standard normal and a chi-square random variable. Let  $Z$  have a standard normal distribution and let  $X$  have a chi-square distribution with  $n$  degrees of freedom. Also assume that  $Z$  and  $X$  are independent. Then the following random variable

$$T = \frac{Z}{\sqrt{X/n}}$$

has a  $t$  distribution with  $n$  degrees of freedom. This is denoted by  $T \sim t_n$ . The  $t$  distribution gets its degrees of freedom from the chi-square random variable.

Another important distribution for statistics and econometrics is the **F distribution**. To define an  $F$  random variable, let  $X_1 \sim \chi_{k_1}^2$  and  $X_2 \sim \chi_{k_2}^2$  and assume that  $X_1$  and  $X_2$  are independent. Then, the random variable

$$F = \frac{X_1/k_1}{X_2/k_2}$$

has an  $F$  distribution with  $(k_1, k_2)$  degrees of freedom. We denote this as  $F \sim F_{k_1, k_2}$ . The order of the degrees of freedom is important.  $k_1$  is the *numerator degrees of freedom* and  $k_2$  is the *denominator degrees of freedom*.