Chapter 2: A Review of Probability*

Manuel V. Montesinos†

Statistics Brush-Up Course
Competition and EPP Master Programs

Fall 2020

I. Random Variables and Probability Distributions

A. Probabilities, sample space and random variables

In Chapter 1, we started from a data sample and we learned how to describe it. Now we will review some of the most important concepts in ***probability theory***, which would allow us to explain how these data are generated from a population by means of statistical models. A way of describing the process that generated the data that we observe is a ***random experiment*** or ***trial***. We call it random because, even though the process can be replicated under similar conditions, the result is not known with certainty, since there is more than one possible outcome. This is unlike a ***deterministic experiment***, which has only one possible outcome.

Most aspects of the world around us have an element of randomness. The result of rolling a dice or tossing a coin, the gender of the next person you meet, your grade on an exam, and the number of times your computer will crash while writing a term paper all have an element of chance or randomness, because if you repeat them, you will not necessarily obtain the same result. The mutually exclusive potential results of a random experiment are called the ***outcomes***. For example, your computer might never crash, it might crash once, it might crash twice, and so on. The outcomes are mutually exclusive, because only one of them will actually occur, and they do not need to be equally likely. The set of all possible outcomes is called

*These notes are partially based on James H. Stock and Mark W. Watson's textbook *Introduction to Econometrics*; Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer's textbook *Introduction to Econometrics with R*, and the Probability and Statistics courses taught by Joan Llull, Jordi Caballé and Anna Houštecká at the Universitat Autònoma de Barcelona and the Barcelona GSE. Typos, misprints, misconceptions and other errors are all mine.

†Departament d'Economia i d'Història Econòmica. Universitat Autònoma de Barcelona. Campus de Bellaterra – Edifici B, 08193, Bellaterra, Cerdanyola del Vallès, Barcelona (Spain). E-mail: manuelvicentemontesinos@gmail.com.

the ***sample space***, and an ***event*** is a subset of the sample space, that is, a set of one or more outcomes. The event "my computer will crash no more than once" is the set consisting of two outcomes: "no crashes" and "one crash".

We refer to the ***probability*** of an outcome as the proportion of the time that the outcome occurs in the long run, that is, if the experiment is repeated many times. If the probability of your computer not crashing while you are writing a term paper is 80%, then over the course of writing many term papers you complete 80% of them without a crash.

All these ideas are unified in the concept of ***random variable***, which is a numerical summary of a random outcome. The number of times your computer crashes while you are writing a term paper is random and takes on a numerical value, so it is a random variable. Random variables can be discrete or continuous. As their names suggest, a ***discrete random variable*** takes on only a discrete set of values, e.g., 0 and 1, whereas a ***continuous random variable*** takes on a continuum of possible values.

B. Probability distributions of discrete random variables

A typical example for a discrete random variable D is the result of a dice roll: in terms of a random experiment this is nothing but randomly selecting a sample of size 1 from a set of numbers which are mutually exclusive outcomes. In this case, the sample space is $\{1, 2, 3, 4, 5, 6\}$ and we can think of many different events, such as "the observed outcome lies between 2 and 5".

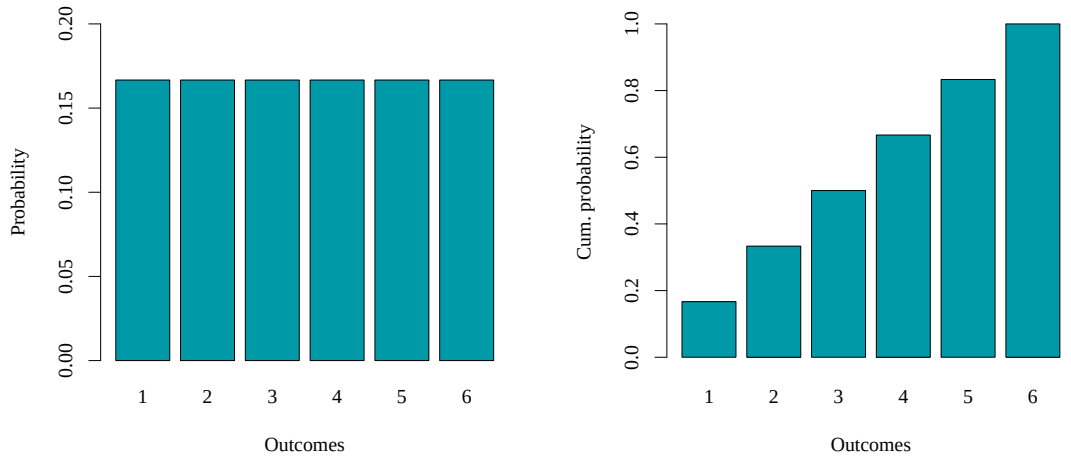
The ***probability distribution*** of a discrete random variable is the list of all possible values of the variable and the probability that each value will occur. These probabilities sum to 1. Meanwhile, the ***cumulative probability distribution*** function gives the probability that the random variable is less than or equal to a particular value. For the dice roll, the probability distribution and the cumulative probability distribution are summarized in Table 1 and represented in Figure 1.

Table 1: Probability Distribution and Cumulative Probability Distribution of a Dice Roll

Outcome	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6
Cumulative Probability	1/6	2/6	3/6	4/6	5/6	1

Figure 1: Probability Distribution and Cumulative Probability Distribution of a Dice Roll

(a) Probability Distribution (b) Cum. Probability Distribution



The probability of an event can be computed from the probability distribution. For example, the probability of the event "get 1 or 2" in a dice roll is the sum of the probabilities of the constituent outcomes:

$$\Pr(D = 1 \text{ or } D = 2) = \Pr(D = 1) + \Pr(D = 2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3},$$

which coincides with the probability that we get a value smaller or equal than 2, as given by the cumulative probability distribution.

An important special case of a discrete random variable is when the random variable is binary, that is, the outcomes are 0 or 1. A binary random variable is called a ***Bernoulli random variable***, and its probability distribution is called the ***Bernoulli distribution***.[1] The result of a single coin toss or the gender of the next new person you meet are examples of Bernoulli random variables, since they are variables with two possible distinct outcomes (head/tails, male/female). If we let G be the gender of the next new person you meet, where $G = 0$ indicates that the person is male and $G = 1$ that she is female, the outcomes of G and their probabilities are

$$G = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p. \end{cases} \tag{1}$$

Now imagine that you are about to meet 10 new persons in a row, and wonder

[1] In honor of the 17th-century Swiss mathematician Jacob Bernoulli.

how likely it is to end up meeting five girls. This is an example of what we call a ***Bernoulli experiment***, as it consists of $n = 10$ Bernoulli trials that are independent of each other and we are interested in the likelihood of observing $k = 5$ successes (meeting a girl). Let's assume that the probability of meeting a girl is $p = 0.5$ in each trial, and note that the order of the outcomes does not matter here. It is a well-known result that the number of successes k in a Bernoulli experiment follows a ***Binomial distribution***. We denote this as

$$k \sim B(n, p). \tag{2}$$

The probability of observing k successes in the experiment $B(n, p)$ is given by

$$f(k) = \Pr(k) = \begin{pmatrix} n \\ k \end{pmatrix} \times p^k \times (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \times p^k \times (1-p)^{n-k}, \tag{3}$$

where $n!/k!(n-k)!$ is the *binomial coefficient*. Then, the probability that we meet five girls out of ten new persons in a row is

$$\Pr(k = 5) = \frac{10!}{5! \times 5!} \times 0.5^5 \times 0.5^5 = 0.246.$$

Now assume that we are interested in the probability of meeting 4, 5, 6 or 7 girls out of 10 new persons. This probability can be computed as

$$\Pr(4 \leq k \leq 7) = \Pr(k \leq 7) - \Pr(k \leq 3),$$

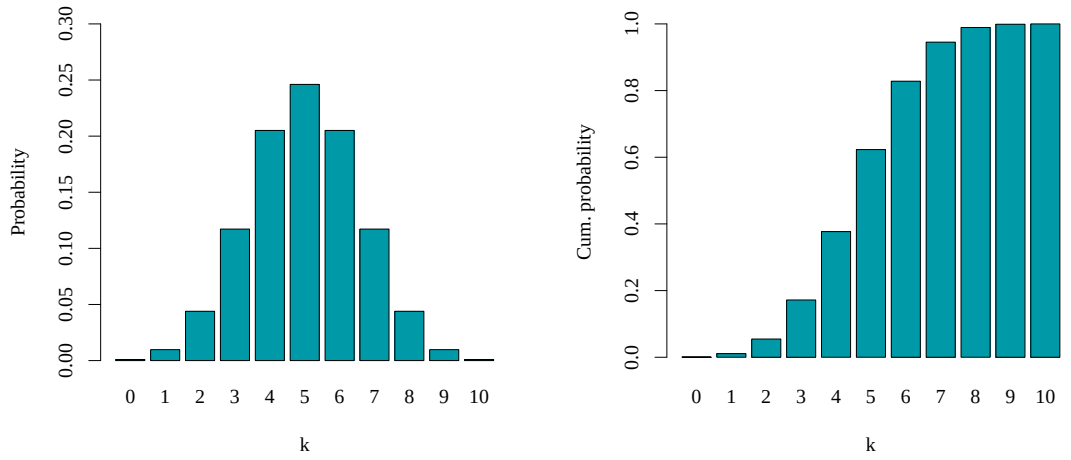
which is approximately equal to 0.773. Figure 2 represents the probability distribution and cumulative probability distribution of k.

C. Probability distributions of continuous random variables

Since a continuous random variable takes on a continuum of possible values, we cannot use the concept of a probability distribution as used for discrete random variables, which lists the probability of each possible value of the random variable. Instead, the probability of a continuous random variable is summarized by the ***probability density function*** (a.k.a. PDF, density function, or simply, density). The area under the probability density function between any two points is the probability that the random variable falls between these two points. Formally, let $f_Y(y)$ denote the probability density function of random variable Y. Then, the probability that Y falls between a and b, provided that $a < b$, is given by the integral of $f_Y(y)$

Figure 2: Probability Distribution and Cumulative Probability Distribution of $k \sim B(10, k)$

(a) Probability Distribution (b) Cum. Probability Distribution



between a and b:

$$\Pr(a \leq Y \leq b) = \int_a^b f_Y(y)\mathrm{d}y. \tag{4}$$

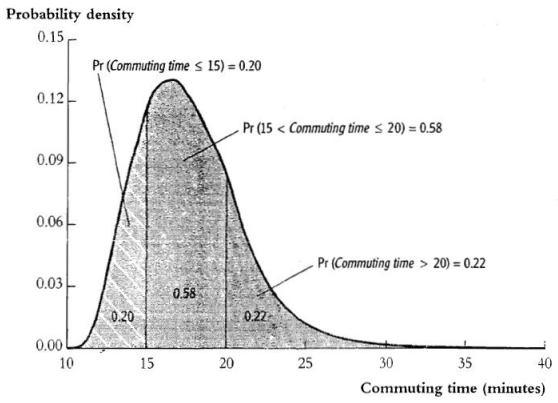
We further have that $\Pr(-\infty \leq Y \leq \infty) = \int_{-\infty}^{\infty} f_Y(y)\mathrm{d}y = 1$.

As an example, consider a student who drives from home to school. This student's commuting time can take on a continuum of values and, because it depends on random factors such as the weather and traffic conditions, it is natural to treat it as random. The left panel of Figure 3 plots a hypothetical probability density function of commuting times. The probability that the commute takes between 15 and 20 minutes is given by the area under the PDF between 15 and 20 minutes, which is 0.58 or 58%.

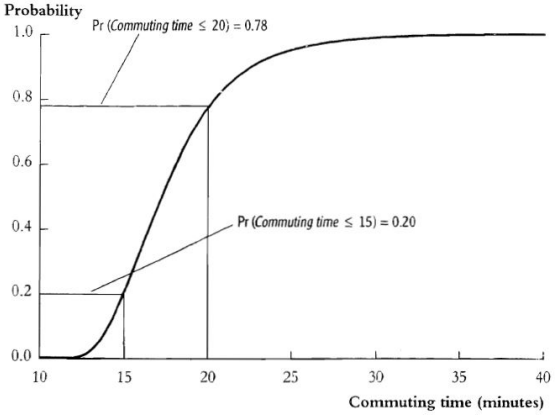
The ***cumulative probability distribution function*** (a.k.a. CDF or distribution function) is defined just as it is for a discrete random variable. Therefore, the CDF of a continuous random variable gives the probability that the random variable is less than or equal to a particular value. The right panel of Figure 3 represents the cumulative distribution of commuting times in our example. We can see that the probability that the commute takes less than 15 minutes is 20%, and the probability that it takes less than 20 minutes is 78%. Note that the probability density function and the cumulative probability distribution show the same information in different formats. For instance, the probability that the commute takes between 15 and 20 minutes can also be seen on the CDF as the difference between the probability that the commute takes less than 20 minutes (78%) and the probability that it takes less

FIGURE 3: PROBABILITY DENSITY AND CUMULATIVE DISTRIBUTION FUNCTIONS OF COMMUTING TIME

(A) PDF



(B) CDF



than 15 minutes (20%).

II. Expected Value and Variance

A. Expected value of a random variable

The ***expected value*** of a random variable Y (a.k.a. the expectation of Y or the mean of Y), denoted by $\mathbb{E}[Y]$ or μ_Y, is the long-run average value of the random variable over many repeated trials or occurrences. For a discrete random variable, the expected value is computed as a weighted average of the possible outcomes of that random variable, where the weights are the probabilities of each outcome. Formally, assume the random variable Y takes on k possible values $y_1, y_2, ..., y_k$, where y_1 denotes the first value, y_2 denotes the second value, and so forth, and the probability that Y takes on y_1 is p_1, the probability that Y takes on y_2 is p_2, and so on. Then, the expected value of Y is defined as

$$\mathbb{E}[y] = y_1 p_1 + y_2 p_2 + ... + y_k p_k = \sum_{i=1}^{k} y_i p_i. \tag{5}$$

For example, suppose you lend someone else \$100 at 10% interest. If the loan is repaid, you will get \$110, but there is a risk of %1 that this person will default and you will get nothing at all. Thus, the amount you will be repaid is a random variable that equals \$110 with probability 0.99 and \$0 with probability 0.01. Over many such loans, 99% of the time you would be paid back \$110, but 1% of the time you would get nothing, so on average you would be repaid $\$110 \times 0.99 + \$0 \times 0.01 = \$108.90$,

which is the expected value of repayment.

In the dice example, the random variable D takes on 6 possible values $d_1 = 1, d_2 = 2, ..., d_6 = 6$. Assuming a fair dice, each of the 6 outcomes occurs with probability 1/6. Therefore, the expected value is

$$\mathbb{E}[D] = \frac{1}{6}\sum_{i=1}^{6} d_i = 3.5. \tag{6}$$

As another example, consider the number of computer crashes M while writing a term paper with the probability distribution in Table 2. The expected number of computer crahses while writing a term paper is

$$\mathbb{E}[M] = 0 \times 0.8 + 1 \times 0.1 + 2 \times 0.06 + 3 \times 0.03 + 4 \times 0.01 = 0.35.$$

Table 2: Probability of Your Computer Crashing M Times

Outcome	0	1	2	3	4
Probability	0.8	0.1	0.06	0.03	0.01
Cumulative Probability	0.8	0.9	0.96	0.99	1

An important special case of (5) is the mean of a Bernoulli random variable. Let G be a Bernoulli random variable with the probability distribution in (1). The expected value of G is

$$\mathbb{E}[G] = 1 \times p + 0 \times (1 - p) = p. \tag{7}$$

Thus, the expected value of Bernoulli random variable is p, the probability that it takes on value 1.

For a continuous random variable, the expected value is also the probability-weighted average of the possible outcomes of the random variable. Due to continuity, we use integrals instead of sums. Then, the expected value of a continuous random variable Y with PDF $f_Y(y)$ is defined as

$$\mathbb{E}[Y] = \int y f_Y(y) \mathrm{d}y. \tag{8}$$

B. Variance and standard deviation

Other frequently encountered concepts are the variance and the standard deviation, which measure the dispersion or the *spread* of a probability distribution. The ***variance*** of a random variable Y, denoted by $\text{Var}[Y]$ or σ_Y^2, is the expected value of the square of the deviation of Y from its mean. For a discrete random variable, this is

$$\text{Var}[Y] = \mathbb{E}\left[(Y - \mu_Y)^2\right] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 p_i. \tag{9}$$

The standard deviation of Y is σ_Y, the square root of the variance, which is easier to interpret than the variance, because it is in the same units as Y.

In the dice example, the variance of D is

$$\text{Var}[D] = \frac{1}{6}\sum_{i=1}^{6} (d_i - 3.5)^2 = 2.92. \tag{10}$$

In the computer crashes example, the variance of M is

$$\text{Var}[M] = \sum_{i=1}^{5} (m_i - \mu_M)^2 p_i = (0 - 0.35)^2 \times 0.8 + (1 - 0.35)^2 \times 0.1 \tag{11}$$

$$+ (2 - 0.35)^2 \times 0.06 + (3 - 0.35)^2 \times 0.03 + (4 - 0.35)^2 \times 0.01 = 0.6475, \tag{12}$$

and the standard deviation is $\sigma_M = \sqrt{0.6475} \approxeq 0.8$.

For a Bernoulli random variable G, the variance is

$$\text{Var}[G] = \sigma_G^2 = (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p = p(1 - p), \tag{13}$$

and the standard deviation is $\sigma_G = \sqrt{p(1-p)}$.

For a continuous random variable Y with density $f_Y(y)$, the variance is defined as

$$\text{Var}[Y] = \int (y - \mu_Y)^2 f_Y(y) \mathrm{d}y. \tag{14}$$

C. Expected value and variance of a linear function of a random variable

Now let's consider random variables X and Y that are related by a linear function. For example, consider an income tax scheme under which a worker is taxed at a rate of 20% on his or her earnings and then given a (tax-free) grant of \$2000. Under this

tax scheme, after-tax earnings Y are related to pre-tax earnings X by

$$Y = 2000 + 0.8X. \tag{15}$$

That is, after-tax earnings Y is 80% of pre-tax earnings X, plus \$2000.

Let's assume an individual's pre-tax earnings next year are a random variable with mean μ_X and variance σ_X^2. Pre-tax earnings are random, and so are after-tax earnings. What are the mean and standard deviation of her after-tax earnings under this tax? After taxes, her earnings are 80% of the original pre-tax earnings, plus \$2000. Thus the expected value of ther after-tax earnings is

$$\mathbb{E}[Y] = \mu_Y = 2000 + 0.8\mu_X. \tag{16}$$

The variance of after-tax earnings is the expected value of $(Y - \mu_Y)^2$. Note that $Y = 2000 + 0.8X$, so $Y - \mu_Y = 2000 + 0.8X - (2000 + 0.8\mu_X) = 0.8\,(X - \mu_X)$. Thus,

$$\text{Var}[Y] = \mathbb{E}\left[(Y - \mu_Y)^2\right] = \mathbb{E}\{[0.8\,(X - \mu_X)]^2\} = 0.64\mathbb{E}\left[(X - \mu_X)^2\right] = 0.64\text{Var}[X], \tag{17}$$

and the standard deviation of the distribution of Y is

$$\sigma_Y = 0.8\sigma_X, \tag{18}$$

that is, the standard deviation of the distribution of after-tax earnings is 80% of the standard deviation of the distribution of pre-tax earnings.

In general, for $Y = a + bX$, the expected value, variance and standard deviation are

$$\mathbb{E}[Y] = a + b\mu_X, \tag{19}$$

$$\sigma_Y^2 = b^2\sigma_X^2, \tag{20}$$

$$\sigma_Y = b\sigma_X. \tag{21}$$

D. Other measures of the shape of a distribution

The expected value and standard deviation measure two important features of a distribution: its center (the expected value) and its spread (the standard deviation). Now we are going to discuss measures of two other features of a distribution: the *skewness*, which measures the lack of symmetry of a distribution, and the *kurtosis*, which measures how thick or "heavy" are its tails. The expected value, variance, skewness, and kurtosis are all based on what are called the ***moments of a distri-***

bution.[2]

Figure 4 plots four distributions: two which are symmetric (a and b) and two which are not (c and d). Visually, the distribution in (d) appears to deviate more from symmetry than does the distribution in (c). The skewness of a distribution provides a mathematical way to describe how much a distribution deviates from symmetry. Formally, the ***skewness*** of the distribution of a random variable Y is defined as

$$\text{Skewness} = \frac{\mathbb{E}\left[(Y - \mu_Y)^3\right]}{\sigma_Y^3}, \tag{22}$$

where σ_Y is the standard deviation of Y. For a symmetric distribution, the likelihood of a value of Y that is a given amount above its mean is the same as the likelihood of a value of Y that is the same amount below its mean. In that case, positive values of $(Y - \mu_Y)^3$ will be offset on average (in expectation) by equally likely negative values. Therefore, the skewness for a symmetric distribution is zero, because $\mathbb{E}\left[(Y - \mu_Y)^3\right] = 0$. By contrast, if a distribution is not symmetric, a positive value of $(Y - \mu_Y)^3$ will not be offset on average by an equally likely negative value, so the skewness will not be zero. We divide by σ_Y^3 in Equation (22) to cancel the units of Y^3 in the numerator, so the skewness is unit free; in other words, changing the units of Y does not change its skewness.

Below each of the four distributions in Figure 4 its skewness is reported. If a distribution has a long right tail, positive values of $(Y - \mu_Y)^3$ will not be fully offset by negative values, and the skewness will be positive. If a distribution has a long left tail, its skewness will be negative.

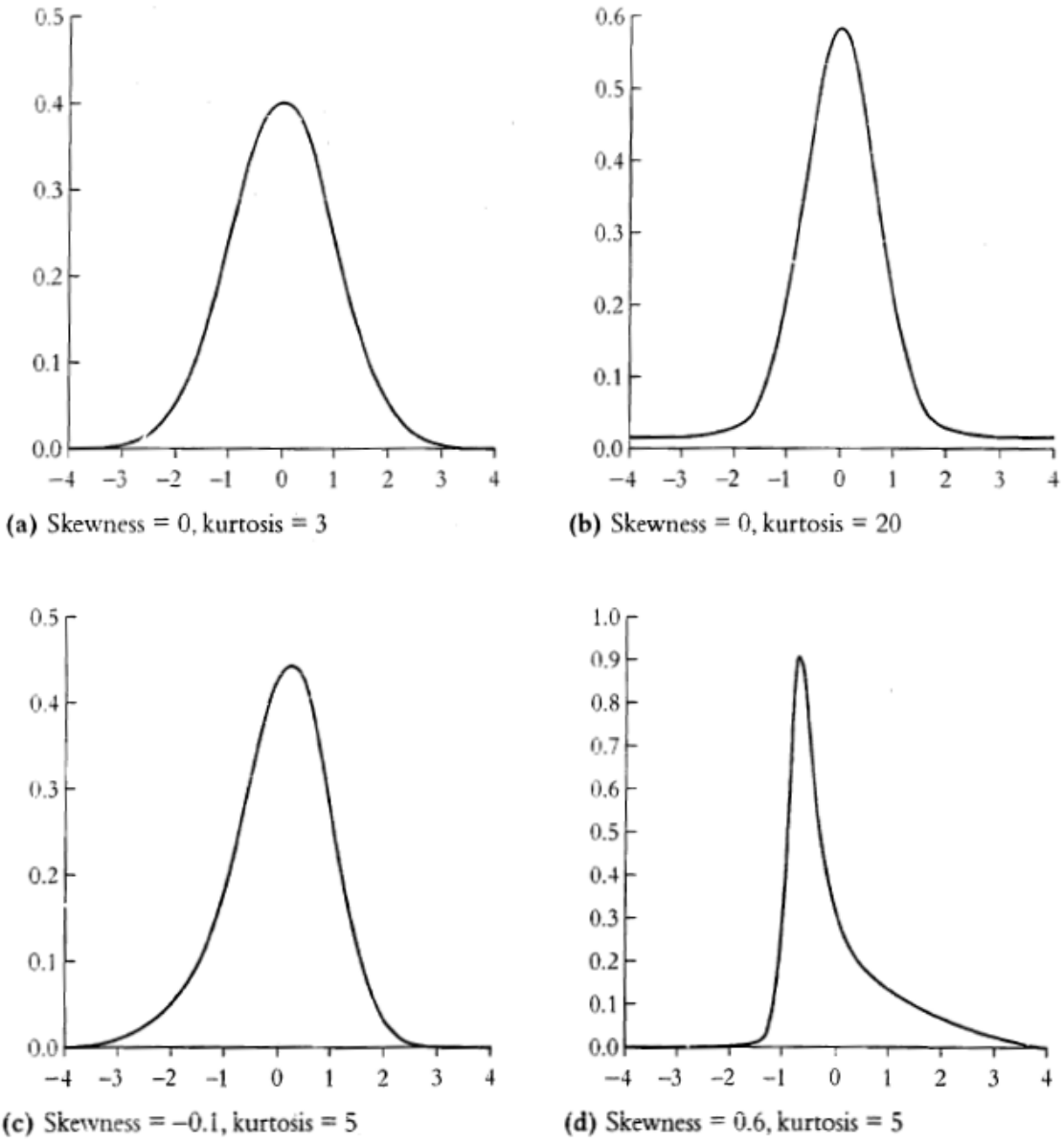
The ***kurtosis*** of a distribution is a measure of how much mass is in its tails and, therefore, is a measure of how much of the variance of Y arises from extreme values. An extreme value of Y is called an ***outlier***. The greater the kurtosis of a distribution, the more likely are outliers. The kurtosis of the distribution of Y is defined as

$$K = \frac{\mathbb{E}\left[(Y - \mu_Y)^4\right]}{\sigma_Y^4}. \tag{23}$$

If a distribution has a large amount of mass in its tails, then some extreme departures of Y from its mean are likely, and these very large values will lead to large values of $(Y - \mu_Y)^4$ on average (in expectation). Thus, for a distribution with a large amount of mass in its tails, the kurtosis will be large. Because $(Y - \mu_Y)^4$

[2]The expected value of Y, $\mathbb{E}[Y]$ is also called the first moment of Y, and the expected value of the square of Y, $\mathbb{E}[Y^2]$, is called the second moment of Y. In general, the expected value of Y^r is called the r^{th} *moment* of Y, $\mathbb{E}[Y^r]$. The skewness is a function of the first, second, and the third moments of Y, and the kurtosis is a function of the first through fourth moments of Y.

FIGURE 4: FOUR DISTRIBUTIONS WITH DIFFERENT SKEWNESS AND KURTOSIS



Note: The four distributions have a mean of 0 and a variance of 1. The distributions with skewness of 0 (a and b) are symmetric; the distributions with nonzero skewness (c and d) are not symmetric. The distributions with kurtosis exceeding 3 (b-d) have heavy tails.

cannot be negative, the kurtosis cannot be negative.

The kurtosis of a normally distributed random variable is 3, so a random variable with kurtosis exceeding 3 has more mass in its tails than a normal random variable. The kurtosis is unit free, so changing the units of Y does not change its kurtosis.

Below each of the four distributions in Figure 4 the kurtosis is reported. The distributions in b-d are heavy-tailed.

III. Two Random Variables

Most of the interesting questions in economics involve two or more variables. Are college graduates more likely to have a job than nongraduates? How does the distribution of income for women compare to that for men? These questions concern the distribution of two random variables, considered together (education and employment status in the first example, income and gender in the second). Answering such questions requires an understanding of the concepts of *joint, marginal* and *conditional probability distributions*.

A. Joint and marginal distributions

The ***joint probability distribution*** of two discrete random variables, say X and Y, gives the probability that the random variables simultaneously take on certain values, such as x and y, respectively. The probabilities of all possible (x, y) combinations must add up to 1. The joint probability distribution can be written as the function $\Pr(X = x, Y = y)$.

For example, weather conditions affect the commuting time of the student in one of our previous examples. Let Y be a binary random variable that equals 1 if the commute is short (less than 20 minutes) and equals 0 otherwise, and let X be a binary random variable that equals 0 if it is raining and 1 if not. Between these two random variables, there are four possible outcomes: it rains and the commute is long ($X = 0, Y = 0$); rain and short commute ($X = 0, Y = 1$); no rain and long commute ($X = 1, Y = 0$); and no rain and short commute ($X = 1, Y = 1$). The joint probability distribution is the frequency with which each of these four outcomes occurs over many repeated commutes. A hypothetical joint distribution of these two variables is given in Table 3. According to this distribution, over many commutes, 15% of the days have rain and long commute ($X = 0, Y = 0$); that is, the probability of a long, rainy commute is 15%, or $\Pr(X = 0, Y = 0) = 0.15$. Also, $\Pr(X = 0, Y = 1) = 0.15$, $\Pr(X = 1, Y = 0) = 0.07$, and $\Pr(X = 1, Y = 1) = 0.63$. These four possible outcomes are mutually exclusive and constitute the sample space,

so the four probabilities sum to 1.

Table 3: Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.3	0.7	1

The ***marginal probability distribution*** of a random variable Y is the name for its probability distribution when we consider more than one random variables together. This term is used to distinguish the distribution of Y alone (*marginal*) from the *joint* distribution of Y and another random variable. The marginal distribution of Y can be computed from the joint distribution of X and Y by adding up the probabilities of all possible outcomes for which X takes on a specified value. If X can take on l different values $x_1, ..., x_l$, then the marginal probability that Y takes on the value y is

$$\Pr(Y = y) = \sum_{i=1}^{l} \Pr(X = x_i, Y = y). \tag{24}$$

For example, in Table 3, the probability of a long rainy commute is 15% and the probability of a long commute with no rain is 7%, so the probability of a long commute (rainy or not) is 22%. The marginal distribution of commuting times is given in the final column of the table. Similarly, the marginal probability that it will rain is 30%, as shown in the final row of the table.

B. Conditional distributions

The distribution of a random variable Y conditional on another random variable X taking on a specific value is called the ***conditional distribution*** of Y given X. The conditional probability that Y takes on the value y when X takes on the value x is written $\Pr(Y = y|X = x)$.

For example, what is the probability of a long commute ($Y = 0$) if you know it is raining ($X = 0$)? From Table 3, the joint probability of a rainy short commute is 15% and the joint probability of a rainy long commute is 15%, so if it is raining, a long commute and a short commute are equally likely. Thus the probability of a long commute ($Y = 0$) conditional on it being rainy ($X = 0$) is 50%, or $\Pr(Y = 0|X = 0) = 0.5$. Equivalently, the marginal probability of rain is 30%, that

is, over many commutes it rains 30% of the time. Of this 30% of commutes, 50% of the time the commute is long (0.15/0.3).

In general, the conditional distribution of Y given $X = x$ is

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}. \tag{25}$$

For instance, the conditional probability of a long commute given that it is rainy is $\Pr(Y = 0|X = 0) = \Pr(X = 0, Y = 0) / \Pr(X = 0) = 0.15/0.3 = 0.5$.

As a second example, consider a modification of the crashing computer example. Assume that you borrow a computer in the library to type your term paper and the librarian randomly assigns you a computer from those available, half of which are new and half of which are old. Because you are randomly assigned to a computer, the age of the computer you use is a random variable. Let's call it A (= 1 if the computer is new, = 0 if it is old). A hypothetical joint distribution of the number of computer crashes M and A is given in Panel A of Table 4. The conditional distribution of computer crashes, given the age of the computer, is reported in Panel B. For example, the joint probability $M = 0$ and $A = 0$ is 0.35. Because half the computers are old, the conditional probability of no crashes, given that you are using an old computer, is $\Pr(M = 0|A = 0) = \Pr(M = 0, A = 0) / \Pr(A = 0) = 0.35/0.5 = 0.7$, or 70%. In contrast, the conditional probability of no crashes given that you are assigned a new computer is 90%. According to the conditional distribution in Panel B of Table 4, the newer computers are less likely to crash than the old ones. For instance, the probability of three crashes is 5% with an old computer, but 1% with a new computer.

Table 4: Joint and Conditional Distributions of Computer Crashes (M) and Computer Age (A)

A. Joint distribution	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
Old computer ($A = 0$)	0.35	0.065	0.05	0.025	0.01	0.5
New computer ($A = 1$)	0.45	0.035	0.01	0.005	0	0.5
Total	0.8	0.1	0.06	0.03	0.01	1
B. Conditional distribution of M given A	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
$\Pr(M\|A = 0)$	0.7	0.13	0.1	0.05	0.02	1
$\Pr(M\|A = 1)$	0.9	0.07	0.02	0.01	0	1

Given the conditional distribution of Y given X, it is possible to compute the ***conditional expectation*** (or *conditional mean*) of Y given X, which is the mean of the conditional distribution of Y given X, i.e., the expected value of Y computed

from the conditional distribution of Y given X. If Y takes on k values $y_1, ..., y_k$, then the conditional mean of Y given $X = x$ is

$$\mathbb{E}(Y|X = x) = \sum_{i=1}^{k} y_i \Pr(Y = y_i|X = x). \tag{26}$$

Based on the conditional distribution in Table 4, the expected number of computer crashes, given that the computer is old, is $\mathbb{E}(M|A = 0) = 0 \times 0.7 + 1 \times 0.13 + 2 \times 0.1 + 3 \times 0.05 + 4 \times 0.02 = 0.56$. The expected number of computer crashes, given that the computer is new, is $\mathbb{E}(M|A = 1) = 0.14$, less than for the old computers.

From the conditional expectation of Y given X, it is possible to recover the mean of Y as the weighted average of this conditional expectation, using the probability distribution of X as weights. For example, the mean height of adults is the weighted average of the mean height of men and the mean height of women, weighted by the proportions of men and women. Formally, if X takes on the l values $x_1, ..., x_l$, then

$$\mathbb{E}[Y] = \sum_{i=1}^{l} \mathbb{E}(Y|X = x_i) \Pr(X = x_i). \tag{27}$$

Stated differently, the expectation of Y is the expectation of the conditional expectation of Y given X,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]], \tag{28}$$

where the inner expectation on the right-hand side is computed using the conditional distribution of Y given X, and the outer expectation is computed using the marginal distribution of X. This result is known as the ***law of iterated expectations***.

For example, the mean number of crashes M is the weighted average of the conditional expectation of M given that the computer is old, and the conditional expectation of M given that the computer is new, so $\mathbb{E}[M] = \mathbb{E}[M|A = 0] \times \Pr(A = 0) + \mathbb{E}[M|A = 1] \times \Pr(A = 1) = 0.56 \times 0.5 + 0.14 \times 0.5 = 0.35$. This is the mean of the marginal distribution of M, as calculated in Equation (28).

The law of iterated expectations implies that if the conditional mean of Y given X is zero, then the mean of Y is zero: if $\mathbb{E}[Y|X] = 0$, then $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[0] = 0$. In other words, if the mean of Y given X is zero, then it must be that the probability-weighted average of these conditional means is zero, that is, the mean of Y must be zero.

The law of iterated expectations also applies to expectations that are conditional on multiple random variables. For example, let X, Y and Z be random variables that are jointly distributed. Then the law of iterated expectations says that $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X, Z]]$, where $\mathbb{E}[Y|X, Z]$ is the conditional expectation of Y given both X

and Z. For example, in the computer crash illustration of Table 4, let P denote the number of programs installed on the computer. Then $\mathbb{E}[M|A,P]$ is the expected number of crashes for a computer with age A that has P programs installed. The expected number of crashes overall, $\mathbb{E}[M]$, is the weighted average of the expected number of crashes for a computer with age A and number of programs P, weighted by the proportion of computers with that value of both A and P.

Similarly, we can define the variance of the conditional distribution of Y given X. Formally, the ***conditional variance*** of Y given X is

$$\text{Var}[Y|X=x] = \sum_{i=1}^{k} [y_i - \mathbb{E}[Y|X=x]]^2 \Pr(Y=y_i|X=x). \tag{29}$$

For example, the conditional variance of the number of crashes given that the computer is old is $\text{Var}[M|A=0] = (0-0.56)^2 \times 0.7 + (1-0.56)^2 \times 0.13 + (2-0.56)^2 \times 0.1 + (3-0.56)^2 \times 0.05 + (4-0.56)^2 \times 0.02 \approxeq 0.99$. The standard deviation of the conditional distribution of M given that $A=0$ is thus $\sqrt{0.99} = 0.99$. The conditional variance of M given that $A=1$ is the variance of the distribution in the second row of Panel B of Table 4, which is 0.22, so the standard deviation of M for new computers is $\sqrt{0.22} = 0.47$. For the conditional distribution in Table 4, the expected number of crashes for new computers (0.14) is less than that for old computers (0.56), and the spread of the distribution of the number of crashes, as measured by the conditional standard deviation, is smaller for new computers (0.47) than for old (0.99).

C. Independence

We say that two random variables X and Y are ***independently distributed***, or ***independent***, if knowledge of one of the variables provides no information about the other. Specifically, X and Y are independent if the conditional distribution of Y given X equals the marginal distribution of Y. That is, X and Y are independently distributed if, for all values x and y,

$$\Pr(Y=y|X=x) = \Pr(Y=y). \tag{30}$$

Substituting (30) into (25) gives an alternative expression for independent random variables in terms of their joint distribution. If X and Y are independent, then their joint distribution is equal to the product of their marginal distributions:

$$\Pr(X=x, Y=y) = \Pr(X=x)\Pr(Y=y). \tag{31}$$

D. Covariance and correlation

One measure of the extent to which two random variables move together is their covariance. The ***covariance*** between X and Y is the expected value

$$\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right], \tag{32}$$

where μ_X is the mean of X and μ_Y is the mean of Y. If X can take on l values and Y can take on k values, then the covariance is given by

$$\begin{aligned}\mathrm{Cov}(X,Y) = \sigma_{XY} &= \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_j-\mu_X)(y_i-\mu_Y)\Pr(X=x_j, Y=y_i).\end{aligned} \tag{33}$$

To interpret the covariance, suppose that when X is greater than its mean (so that $X-\mu_X$ is positive), then Y tends to be greater than its mean (so that $Y-\mu_Y$ is positive), and when X is less than its mean (so that $X-\mu_X<0$), then Y tends to be less than its mean (so that $Y-\mu_Y<0$). In both cases, the product $(X-\mu_X)\times(Y-\mu_Y)$ tends to be positive, so the covariance is positive. In contrast, if X and Y tend to move in opposite directions (so that X is large when Y is small, and vice versa), then the covariance is negative. Finally, if X and Y are independent, then the covariance is zero.

However, since the covariance is the product of X and Y, deviated from their means, its units are, awkwardly, the units of X multiplied by the units of Y. This "units" problem can make numerical values of the covariance difficult to interpret.

The correlation is an alternative measure of dependence between X and Y that solves the "units" problem of the covariance. Specifically, the ***correlation*** between X and Y is the covariance between X and Y divided by the product of their standard deviations:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}. \tag{34}$$

Because the units of the numerator are the same as those of the denominator, the units cancel and the correlation is unitless. Its values are always between -1 and 1, that is, $|\mathrm{Corr}(X,Y)\leq 1|$.

We say that the random variables X and Y are ***uncorrelated*** if $\mathrm{Corr}(X,Y)=0$. This happens if the conditional mean of Y does not depend on X (or vice versa):

$$\mathbb{E}[Y|X]=\mu_Y \Rightarrow \mathrm{Cov}(X,Y)=0 \iff \mathrm{Corr}(Y,X)=0. \tag{35}$$

Proof. Assume that Y and X are zero-mean,[3] so that

$$\text{Cov}(Y, X) = \mathbb{E}\left[(Y - \mu_Y)(X - \mu_X)\right] = \mathbb{E}[YX].$$

By the law of iterated expectations,

$$\mathbb{E}[YX] = \mathbb{E}\left[\mathbb{E}\left[YX|X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[Y|X\right]X\right] = 0,$$

because $\mathbb{E}[Y|X] = 0$, so $\text{Cov}(Y, X) = 0$. Equation (35) follows by substituting $\text{Cov}(X, Y) = 0$ into (34). □

It is not necessarily true, however, that if X and Y are uncorrelated, then the conditional mean of Y given X does not depend on X. It is possible for the conditional mean of Y to be a function of X (Y and X are not independent) while Y and X are uncorrelated.

IV. Normal, Chi-Squared, Student t and F Distributions

In this section we are going to discuss some of the most popular probability distributions in economics: the normal, chi-squared, Student t and F distributions.

A. The normal distribution

A continuous random variable with a ***normal distribution*** has the familiar symmetric, bell-shaped probability density shown in Figure 5. A normal distribution is characterized by its mean μ and its standard deviation σ, concisely expressed by $\mathcal{N}(\mu, \sigma^2)$. The normal distribution has the PDF

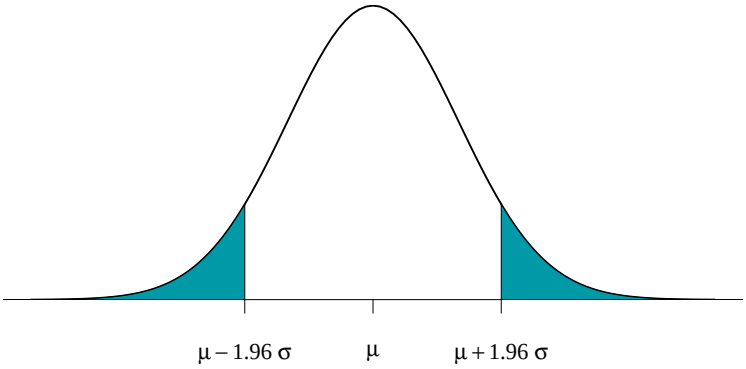
$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{36}$$

The normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is the ***standard normal distribution***, denoted by $\mathcal{N}(0, 1)$. Random variables that have a $\mathcal{N}(0, 1)$ distribution are often denoted Z, and the standard normal cumulative distribution function is denoted by $\Phi(z)$, such that $\Pr(Z \leq c) = \Phi(c)$, where c is a constant. The standard normal CDF is plotted in Figure 6.

To look up probabilities for a normal variable with a general mean and variance, we must ***standardize*** the variable by first substracting the mean, and then by dividing the result by the standard deviation. For example, suppose Y is distributed

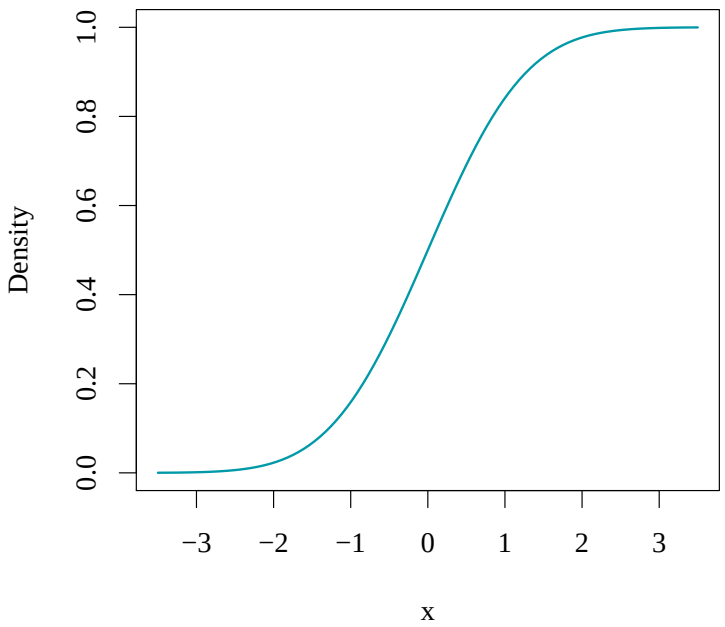
[3] If Y and X are not zero-mean, we can subtract off their means and the proof still applies.

Figure 5: Normal Probability Density Function



Note: The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal PDF between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is equal to 0.95.

Figure 6: Standard Normal Cumulative Distribution Function



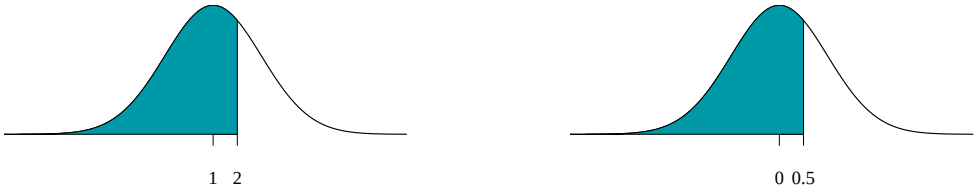
as $\mathcal{N}(1,4)$, that is, Y is normally distributed with a mean of 1 and a variance of 4. What is the probability that $Y \leq 2$? That is, what is the shaded area in Figure 7a? The standardized version of Y is $(Y-1)/\sqrt{4} = (Y-1)/2$. Accordingly, the random variable $(Y-1)/2$, plotted in Figure 7b, is normally distributed with mean 0 and variance 1. Now $Y \leq 2$ is equivalent to $(Y-1)/2 \leq (2-1)/2 = 1/2$. Thus,

$$\Pr(Y \leq 2) = \Pr\left[\frac{(Y-1)}{2} \leq \frac{1}{2}\right] = \Pr\left(Z \leq \frac{1}{2}\right) = \Phi(0.5) = 0.691, \tag{37}$$

where the value 0.691 is taken from the tables of the standard normal distribution.

FIGURE 7: CALCULATING THE PROBABILITY THAT $Y \leq 2$ WHEN Y IS DISTRIBUTED $\mathcal{N}(1,4)$

(A) $Y \sim \mathcal{N}(1,4)$ (B) $Z \sim \mathcal{N}(0,1)$



In general, we can standardize $Y \sim \mathcal{N}(\mu, \sigma^2)$ by subtracting its mean and dividing by its standard deviation, and use the resulting $Z = (Y-\mu)/\sigma$ to compute probabilities. Then, if we have two numbers, c_1 and c_2, such that $c_1 < c_2$, we can compute $d_1 = (c_1 - \mu)/\sigma$ and $d_2 = (c_2 - \mu)/\sigma$ and write

$$\Pr(Y \leq c_2) = \Pr(Z \leq d_2) = \Phi(d_2) \tag{38}$$

$$\Pr(Y \geq c_1) = \Pr(Z \geq d_1) = 1 - \Phi(d_1) \tag{39}$$

$$\Pr(c_1 \leq Y \leq c_2) = \Pr(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1). \tag{40}$$

The normal distribution can be generalized to describe the joint distribution of a set of random variables. In this case, the distribution is called the ***multivariate normal distribution***, or, if only two variables are being considered, the ***bivariate***

normal distribution. The bivariate normal PDF is given by

$$g_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \tag{41}$$
$$\times \exp\left\{\frac{1}{-2\left(1-\rho_{XY}^2\right)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{XY}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

where ρ_{XY} is the correlation between X and Y. In general, we say that the $m \times 1$ vector random variable $\boldsymbol{V}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu_V}$ and covariance matrix $\boldsymbol{\Sigma_V}$ if it has the joint probability density function

$$g(\boldsymbol{V}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma_V})}} \exp\left[-\frac{1}{2}\left(\boldsymbol{V}-\boldsymbol{\mu_V}\right)'\boldsymbol{\Sigma_V}^{-1}\left(\boldsymbol{V}-\boldsymbol{\mu_V}\right)\right], \tag{42}$$

where $\det(\boldsymbol{\Sigma_V})$ is the determinant of the matrix $\boldsymbol{\Sigma_V}$. The multivariate normal distribution is denoted $\mathcal{N}(\mu_{\boldsymbol{V}}, \boldsymbol{\Sigma_V})$, and it has four important properties:

1. If X and Y have a bivariate normal distribution with covariance σ_{XY} and if a and b are two constants, then $aX + bY$ has the normal distribution

$$\mathcal{N}\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\right). \tag{43}$$

   More generally, if n random variables have a multivariate normal distribution, then any linear combination of these variables (such as their sum) is normally distributed.

2. If a set of variables has a multivariate normal distribution, then the marginal distribution of each of the variables is normal (this follows from (43) by setting $a = 1$ and $b = 0$).

3. If variables with a multivariate normal distribution have covariances that equal zero, then the variables are independent. Thus, if X and Y have a bivariate normal distribution and $\sigma_{XY} = 0$, then X and Y are independent. This result, that zero covariance implies independence, is a special property of the multivariate normal distribution that is not true in general.

4. If X and Y have a bivariate normal distribution, then the conditional expectation of Y given X is linear in X. That is, $\mathbb{E}[Y|X = x] = a + bx$, where a and b are constants. Joint normality implies linearity of conditional expectations, but linearity of conditional expectations does not imply joint normality.

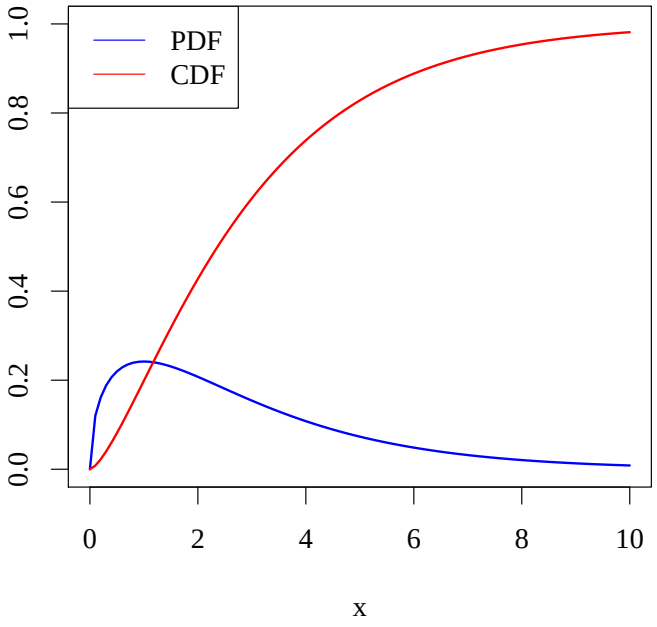
B. The chi-squared distribution

The ***chi-squared distribution*** is the distribution of the sum of m squared independent standard normal random variables. Frequently used when testing certain types of hypothesis in statistics and econometrics, this distribution depends on m, which is the number of degrees of freedom:

$$Z_1^2 + ... + Z_M^2 = \sum_{m=1}^{M} Z_m^2 \sim \chi_M^2 \text{ with } Z_m \overset{i.i.d.}{\sim} \mathcal{N}(0,1). \tag{44}$$

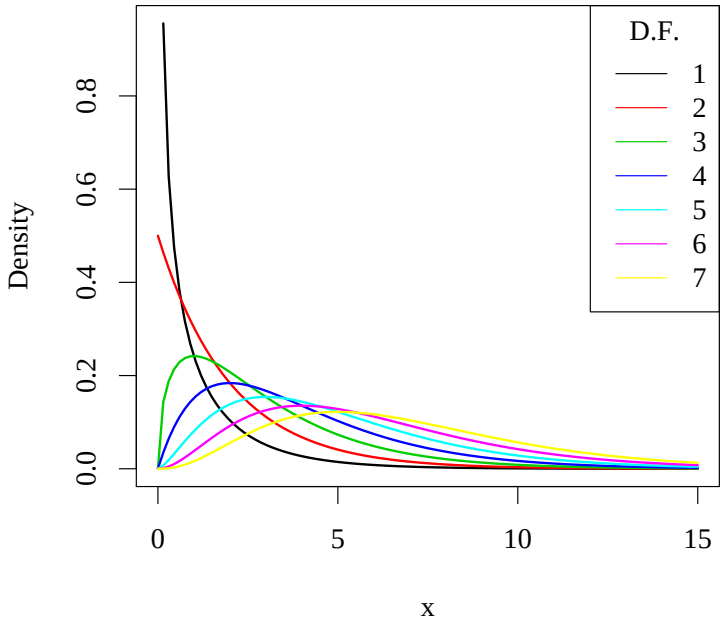
Since the outcomes of a χ_m^2 distributed random variable are always positive, the support of the related PDF and CDF is $\mathbb{R}_+$. As an example, the PDF and the CDF of a χ_3^2 random variable are plotted in Figure 8.

FIGURE 8: PDF AND CDF OF CHI-SQUARED DISTRIBUTION WITH $m = 3$



A χ_m^2 distributed random variable has expectation m, mode $m-2$ if $m \geq 2$ and variance $2m$. As the expectation and the variance depend solely on the degrees of freedom, the shape of the distribution changes drastically with the degrees of freedom. This is depicted by Figure 9.

FIGURE 9: CHI-SQUARED DISTRIBUTED RANDOM VARIABLES (DIFFERENT NUMBER OF DEGREES OF FREEDOM)



C. The Student t distribution

The ***Student t distribution*** with m degrees of freedom is defined as the distribution of the ratio of a standard normal random variable to the square root of an independently distributed chi-squared random variable with m degrees of freedom divided by m.[4] Formally, let Z be a standard normal random variable, let W be a random variable with a chi-squared distribution with m degrees of freedom, and let Z and W be independently distributed. Then the resulting random variable has a Student t distribution with m degrees of freedom:

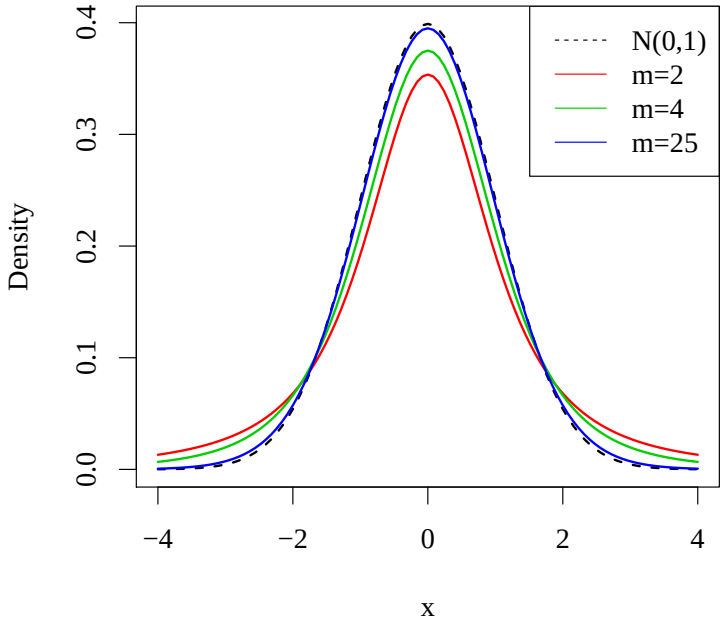
$$\frac{Z}{\sqrt{W/M}} = X \sim t_m. \tag{45}$$

A t_m distributed random variable X has expectation $\mathbb{E}[X] = 0$ if $m > 1$, and

[4]The Student t distribution takes its name from William Sealy Gosset's 1908 paper in *Biometrika* under the pseudonym "Student". Gosset worked at the Guiness Brewery in Dublin, Ireland, and was interested in the problems of small samples –for example, the chemical properties of barley where sample sizes might be as few as 3. One version of the origin of the pseudonym is that Gosset's employer preferred staff to use pen names when publishing scientific papers instead of their real name, so he used the name "Student" to hide his identity. Another version is that Guiness did not want their competitors to know that they were using the t-test to determine the quality of raw material.

variance $\text{Var}[X] = m/(m-2)$ if $m > 2$. Similar to the χ^2_m distribution, the shape of a t_m distribution depends on m. This kind of distribution is symmetric, bell-shaped and looks similar to a normal distribution, especially when m is large. This is not a coincidence: for a sufficiently large m, the t_m distribution can be approximated by the standard normal distribution. This approximation works reasonably well for $m \geq 30$, and actually, in the limit when m goes to ∞, the t distribution becomes the standard normal distribution. We illustrate this point in Figure 10.

FIGURE 10: DENSITIES OF t DISTRIBUTIONS COMPARED TO $\mathcal{N}(0,1)$



D. The F distribution

The ***F distribution*** with m and n degrees of freedom, denoted by $F_{m,n}$, is defined as the distribution of the ratio of a chi-squared random variable with m degrees of freedom, divided by m, to an independently distributed chi-squared random variable with n degrees of freedom, divided by n.[5] Formally, let W be a chi-squared random variable with m degrees of freedom and let V be a chi-squared random variable with n degrees of freedom, where W and V are independently distributed. Then,

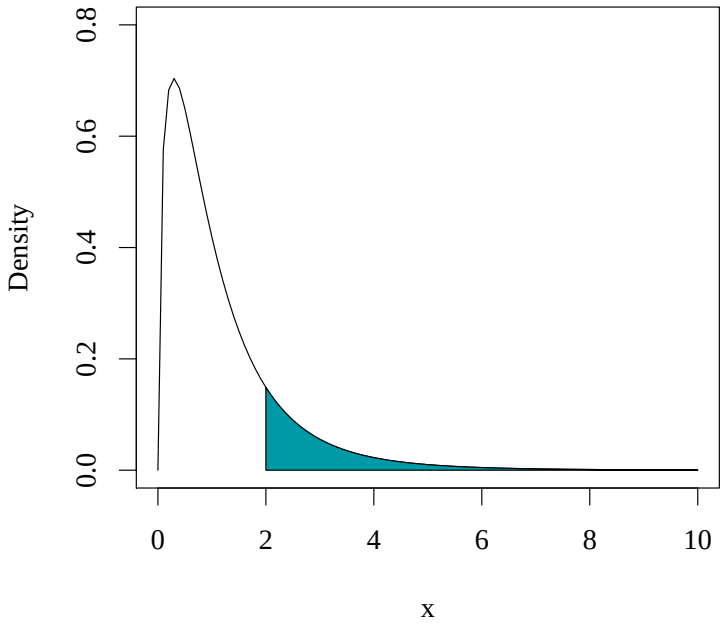
$$\frac{W/m}{V/n} \sim F_{m,n} \text{ with } W \sim \chi^2_m, V \sim \chi^2_n. \tag{46}$$

[5]The F distribution was first derived by George Snedecor, but was named in honor of Sir Ronald Fisher.

By definition, the support of both the PDF and the CDF of an $F_{m,n}$ distributed random variable is $\mathbb{R}_+$. As an example, Figure 11 plots the PDF of an $F_{3,14}$ distributed random variable.

In statistics and econometrics, an important special case of the F distribution arises when the number of degrees of freedom of the denominator is large enough that the $F_{m,n}$ distribution can be approximated by the $F_{m,\infty}$ distribution. In this limiting case, the denominator variable V is the mean of infinitely many chi-squared random variables, and that mean is 1 because the mean of a squared standard normal random variable is 1. Thus the $F_{m,\infty}$ distribution is the distribution of a chi-squared random variable with m degrees of freedom, divided by m: $W/m \sim F_{m,\infty}$.

FIGURE 11: DENSITY OF $x \sim F_{3,14}$



Appendix A: Means, Variances, and Covariances of Sums of Random Variables

Here is a number of results concerning the expected value, variance and covariance of sums of random variables:

$$\mathbb{E}\left[a + bX + cY\right] = a + b\mu_X + c\mu_Y$$

Proof. This result follows from the definition of expected value. □

$$\text{Var}\left[a + bY\right] = b^2\sigma_Y^2,$$

Proof. Using the definition of variance, we can write

$$\begin{aligned}\text{Var}[a + bY] &= \mathbb{E}\{[a + bY - \mathbb{E}(a + bY)]^2\} = \mathbb{E}\{[b\,(Y - \mu_Y)]^2\} \\ &= b^2\mathbb{E}\left[(Y - \mu_Y)^2\right] = b^2\sigma_Y^2.\end{aligned}$$

□

$$\text{Var}\left[aX + bY\right] = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2$$

Proof. Using the definition of the variance, we can write,

$$\begin{aligned}\text{Var}[aX + bY] &= \mathbb{E}\{[(aX + bY) - (a\mu_X + b\mu_Y)]^2\} \\ &= \mathbb{E}\{[a\,(X - \mu_X) + b\,(Y - \mu_Y)]^2\} \\ &= \mathbb{E}\left[a^2\,(X - \mu_X)^2\right] + 2\mathbb{E}\left[ab\,(X - \mu_X)\,(Y - \mu_Y)\right] + \mathbb{E}\left[b^2\,(Y - \mu_Y)^2\right] \\ &= a^2\text{Var}[X] + 2ab\text{Cov}(X, Y) + b^2\text{Var}[Y] \\ &= a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2,\end{aligned}$$

where the second equality follows by collecting terms, the third equality follows by expanding the quadratic, and the fourth equality follows by the definition of the variance and covariance. □

$$\mathbb{E}[Y^2] = \sigma_Y^2 + \mu_Y^2$$

Proof. This result can be obtained from

$$\begin{aligned}\mathbb{E}[Y^2] &= \mathbb{E}\{[(Y-\mu_Y)+\mu_Y]^2\} \\ &= \mathbb{E}\left[(Y-\mu_Y)^2\right] + 2\mu_Y\mathbb{E}\left(Y-\mu_Y\right) + \mu_Y^2 \\ &= \sigma_Y^2 + \mu_Y^2,\end{aligned}$$

because $\mathbb{E}\left[Y-\mu_Y\right] = 0$. □

$$\mathrm{Cov}\left(a+bX+cZ, Y\right) = b\sigma_{XY} + c\sigma_{ZY}$$

Proof. Using the definition of covariance, we can write

$$\begin{aligned}\mathrm{Cov}\left(a+bX+cZ, Y\right) &= \mathbb{E}\{\left[a+bX+cZ-\mathbb{E}\left[a+bX+cZ\right]\right]\left[Y-\mu_Y\right]\} \\ &= \mathbb{E}\{\left[b\left(X-\mu_X\right)+c\left(Z-\mu_Z\right)\right]\left[Y-\mu_Y\right]\} \\ &= \mathbb{E}\{b(X-\mu_X)\left[Y-\mu_Y\right]\} + \mathbb{E}\{\left[c(Z-\mu_Z)\right]\left[Y-\mu_Y\right]\} \\ &= b\sigma_{XY} + c\sigma_{ZY}.\end{aligned}$$

□

$$\mathbb{E}[XY] = \sigma_{XY} + \mu_X\mu_Y$$

Proof. We can write:

$$\begin{aligned}\mathbb{E}[XY] &= \mathbb{E}\{\left[(X-\mu_X)+\mu_X\right]\left[(Y-\mu_Y)+\mu_Y\right]\} \\ &= \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] + \mu_X\mathbb{E}[Y-\mu_Y] + \mu_Y\mathbb{E}[X-\mu_X] + \mu_X\mu_Y \\ &= \sigma_{XY} + \mu_X\mu_Y.\end{aligned}$$

□

$$|\text{Corr}(X, Y)| \leq 1 \text{ and } |\sigma_{XY}^2| \leq \sqrt{\sigma_X^2 \sigma_Y^2}$$

Proof. Let $a = -\sigma_{XY}/\sigma_X^2$ and $b = 1$. Applying the above result concerning the variance of a linear combination of two random variables, we have

$$\begin{aligned}
\text{Var}[aX + Y] &= a^2\sigma_X^2 + \sigma_Y^2 + 2a\sigma_{XY} \\
&= \left(-\frac{\sigma_{XY}}{\sigma_X^2}\right)^2 \sigma_X^2 + \sigma_Y^2 + 2\left(-\frac{\sigma_{XY}}{\sigma_X^2}\right)\sigma_{XY} \\
&= \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}.
\end{aligned}$$

The variance cannot be negative, so the result in the final line above must be greater or equal than 0. Rearranging, we obtain the covariance inequality

$$\sigma_{XY}^2 \leq \sigma_X^2 \sigma_Y^2,$$

which implies that $\sigma_{XY}^2/(\sigma_X^2\sigma_Y^2) \leq 1$ or, equivalently, $|\sigma_{XY}/(\sigma_X\sigma_Y)| \leq 1$, which by the definition of the correlation coefficient, proves the correlation inequality. □