Introduction to Econometrics
ECO4421 Spring & Fall

Department *of* Economics
Florida State University

Review of Probability & Statistics

Lectures:	Check Syllabus	Instructor:	Farasat A.S. Bokhari
Internet:	http://mailer.fsu.edu/~fbokhari/eco4421	Email:	fbokhari@fsu.edu
Office:	Bellamy 284	Tel:	(850) 644-7098
Hours:	Check Syllabus	Fax:	(850) 644-4535

Introduction to Econometrics - ECO4421

Lecture Notes on Review of Probability & Statistics

Last updated on January 24, 2006

Farasat A.S. Bokhari ©

fbokhari@fsu.edu

Sources

These lecture notes are to be used as a supplement to Appendix A of your textbook (Gujarati $4th.$ edition). These notes are based on a number of sources. Primary among these are,

(1) Introduction to Econometrics, Stock and Watson. Addison-Wesley, 2003.
(2) Understandable Statistics (6th Ed.), Brase & Brase. Houghton-Mifflin, 1999.
(3) Mathematical Statistics with Applications (4th Ed.), Mendenhall, Wackerley & Scheaffer. PWS-KENT, 1990.
(4) Econometric Methods (4th. Ed.), Johnston & Dinardo. McGraw-Hill, 1997.
(5) Econometric Analysis (3rd Ed.), Greene. Prentice Hall, 1997
(6) An Introduction to Classical Econometric Theory, Rudd. Oxford, 2000.

The Stock and Watson book is new. It is an excellent text book that covers all relevant topics that we will be covering during the semester. If possible, you should get a copy of this book and use it as a supplement to your own textbook. (Note: You are not required to purchase this book, nor will I expect you to read it. Getting a copy of the book is only a suggestion.)

1. Sample Space, Probabilities & Random Variables

Definition 1 (**Sample Space**)**.** The set of all possible outcomes is called a sample space.

Definition 2 (**Event**)**.** An event is a subset of the sample space.

Suppose you were to toss two coins. Then the sample space would consist of the set of all possible outcomes, E1 = HH, E2 = HT, E3= TH, E4 = TT. Of these four possible outcomes, E1,E2, E3, E4 etc. are events. To each of these events, we can assign a probability.

Probability. The probability of an event is the proportion of the time the event occurs in the long run. Let A be an event in a sample space. Then $P(A)$, the probability of event A is the proportion of times the event A will occur in repeated trials of an experiment. $P(A)$ is a real valued function and has the following properties

- $0 \leq P(A) \leq 1$ for every A.
- If $A, B, C, \ldots$ constitute an exhaustive set of events, then $P(A + B + C + \ldots) = 1$ where $A + B + C$ means A or B or C.
- If $A, B, C, \ldots$ are mutually exclusive events, then

$$P(A + B + C + \ldots) = P(A) + P(B) + P(C) + \ldots \tag{1}$$

Definition 3 (**Random Variable**)**.** A Random Variable is a real valued function for which the domain is a sample space.

Simply put, this means that if we have a sample space and we can find a mapping between all the elements in the sample set and the number line, then the value assigned to any element on the number line is by definition a random number.

Example 1 (**Toss Two Coins**)**.** Let Y be the number of heads observed when tossing two coins. Then the sample space consists of the set E1 = HH, E2 = HT, E3= TH, E4 = TT (See Figure 1).

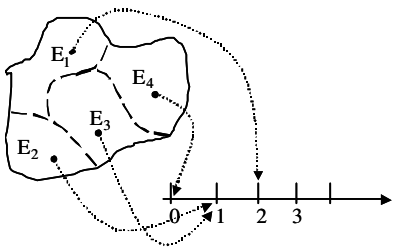


FIGURE 1. Tossing Two Coins

Now, if we let Y be a real valued function such that

$$Y(E1) = 2; Y(E2) = 1; Y(E3) = 1; Y(E4) = 0,$$

then, Y is a random variable. Note again that Y is a number on the number line which takes the same value as the number of heads observed in the toss of the two coins. If the two coins were fair and we distinguished between the events E2 (HT) and E3 (TH) then the probabilities of each of the four events would be 1/4.

Example 2 (**Map to Unity**)**.** Let S be the set of integers from 1 to 10 inclusive and let f a real valued function such that $Y = f(s) = 100$ for all $s \in S$. Then Y is a random variable.

Even a dumb function, like the one in example 2, where all numbers from 1 to 10 are being mapped to 1 is a random variable, even though nothing about it seems random. This is because the probability of an outcome Y = 100 is 1. In general, you observe the following properties about random variables:

- Every point in the sample space maps to a point on the real number line. Thus, in example of a toss of two coins above, E1 maps to 2, E2 and E3 map to 1 and E4 maps to zero.
- The converse is not true. For instance, in the example of the sample space for the toss of 2 coins, there is no point in the sample space that maps to the number 3 or 3.5 or 4.
- More than one point in the sample space may map to the same point on the number line. In the example of the toss of two coins, E2 and E3 and both map to the value 1 on the number line.
- Again, the converse is not true. A point in the sample space cannot map to more than one point on the real number line.

There are two types of random variables: discrete and continuous. Loosely speaking, the difference between a continuous and a discrete random variable is weather the random variable Y can only take certain values on the number line such as the integers or if the Y can even take fraction values on the number line.

Discrete random variables. Variables that can only assume a countable number of values (e.g. positive or negative whole numbers)

- Number of people at the symphony
- Number of children in poverty
- Number of Heads observed in the toss of 2 coins

Continuous random variables. Variables that can take any numerical value-including fractions (i.e, when the values that can be assumed by the variable are not countable).

- Time it takes to fly to Chicago
- Weight of a new-born baby
- Amount (in volume) of milk I drink with my chocolate chip cookies before going to bed

2. Probability Distributions

2.1. Cumulative Probability Distribution

The cumulative probability distribution is the probability that the random variable is less than or equal to a particular value. Thus, for a random variable Y, the cumulative density $F(\ )$ at a specific value y is $F(y) = P(Y \leq y)$. The cumulative probability distribution is always between 0 and 1. It is also known as the cumulative density function (CDF) or just the *distribution function*.

Definition 4 (**Distribution Function**)**.** Let Y denote any random variable. The distribution function of Y, denoted by $F(y)$, is given by $F(y) = P(Y \leq y)$, $-\infty < y < \infty$.

Theorem 1 (**Properties of a Distribution Function**)**.** If $F(y)$ is a distribution function, then the following are true:

(1) $\lim_{y \to -\infty} F(y) = F(-\infty) = 0$
(2) $\lim_{y \to \infty} F(y) = F(\infty) = 1$
(3) $0 \leq F(y) \leq 1$
(4) $F(y_c) \geq F(y_a)$ if $y_c > y_a$
(5) $Prob(y_a < y \leq y_c) = F(y_c) - F(y_a)$

2.2. Probability Distribution of a Discrete Random Variable

The probability distribution for a random variable Y is a mapping from the possible values of Y to the probability that Y takes on each of those values. Thus, it is the list of all possible values of the variable along the probability assigned to each value of the random variable. The sum of the probabilities is always equal to 1: $\sum P(y) = 1$. The probability distribution is also known as the Probability Density Function (PDF).

Definition 5 (Discrete Probability Density Function). Let Y be a discrete rv taking distinct values $y_1, y_2, \ldots, y_n, \ldots$. Then the function

$$f(y) = \begin{cases} P(Y = y_i) & for\ i = 1, 2, 3, \ldots, n, \ldots \\ 0 & for\ y \neq y_i \end{cases} \tag{2}$$

is called the probability density function (PDF) of Y, where $P(Y = y_i)$ means the probability that the discrete rv Y takes the value of y_i.

Note the relationship between PDF and CDF: if $f(y)$ is the PDF of a discrete rv Y, then the CDF of Y (given by $F(y)$) is

$$\begin{aligned} F(y) &= \sum_{Y \leq y} f(y) \\ &= Prob(Y \leq y). \end{aligned} \tag{3}$$

Lets look at some specific (discrete) random variables and their distributions.

General Distribution with known probabilities for every outcome: Suppose that Dr. Ahmed can perform upto 3 cardiac surgeries in a 24 hr. period (his malpractice insurance company does not allow him to perform more than 3 a day). Analysis of his historic data suggest that he will perform between 0 and 3 surgeries with probabilities as given in the table below. The tabulation of all possible outcomes is the PDF. The table also provides the CDF.

Outcome	0	1	2	3
PDF $(P(y))$	0.125	0.375	0.375	0.125
CDF $(F(y < Y)$	0.125	0.500	0.875	1.000

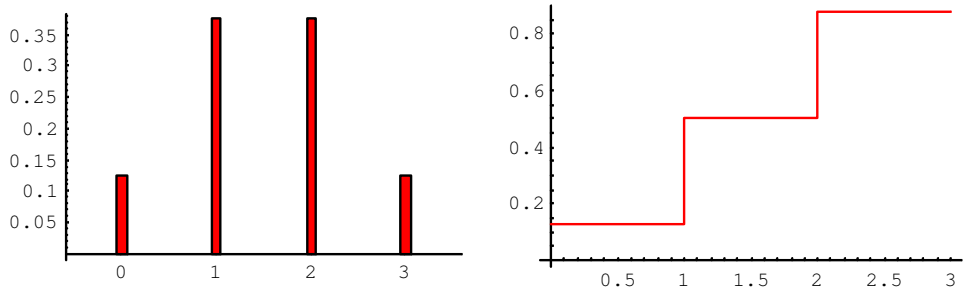


FIGURE 2. PDF and CDF of a Discrete Random Variable

2.3. **Discrete Uniform Random Variable**

When a random variable Y can take n discrete values with equal probabilities, we say that Y has a uniform distribution. In this case the probability of any one out come is just $1/n$, i.e., the PDF of Y given by $f(y)$, is $1/n$. Thus, A random variable Y has a discrete uniform distribution if it has a finite number of possible outcomes with values $y_1, y_2, \ldots, y_n$ and

$$f(y) = P(Y = y_i) = \frac{1}{n} \tag{4}$$

As an example, let y be the number of dots showing when you roll a dice once. The probability of any of the six outcomes (1,2,...,6) is 1/6. The probability density function and the CDF are given in the table below.

y	1	2	3	4	5	6
$f(y)$(PDF)	1/6	1/6	1/6	1/6	1/6	1/6
$F(y)$(CDF)	1/6	2/6	3/6	4/6	5/6	6/6

2.4. **Bernoulli Random Variable**

An important case of the discrete random variable is when it is a binary variable (i.e., takes on only two values). Without loss of any generality, let these be 0 and 1. For instance, Let Y be the pass or fail outcome of a driving test where $Y = 0$ indicates that the person failed and $Y = 1$ indicates that the person passed. The outcomes of Y and the probabilities are

$$Y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

where p is the probability of success (i.e., passing the test) and $1 - p$ is the probability of failure. The PDF of the Bernoulli random variable is

$$f(y) = p^y (1 - p)^{(1-y)} \tag{5}$$

Parameters of a Distribution. It is often easy to summarize the distribution (discrete or continuous) in terms of a few constants. These constants are known as the **parameters** of the distribution. For instance, in the case of the Bernoulli distribution above, the parameter of the distribution is just the probability of success p.

2.5. **Binomial Random Variable**

Imagine a process in which there are two or more consecutive Bernoulli trials, each of which has just two possible outcomes (success or failure), and the probability of success remains the same from one trial to the next (the trials are independent). The binomial random variable Y is the number of successes (y) in n trials. Note the following features of the Binomial Random Variables.

- There are a fixed number of trials, n.
- The n trials are independent and repeated under identical conditions
- Each trial has only two outcomes: success (S) and failure (F) with probabilities p and $q = 1 - p$.
- For each individual trial, the probability of success is the same.
- The central problem is to find the probability of y successes out of n trials. Thus, the random variable of interest Y is the number of successes observed during n trials.
- The parameters of the distribution are n and p and the PDF of a binomial random variable is given by

$$P(Y = y) = C_{n,y} p^y q^{(n-y)} \tag{6}$$

where $P(Y = y)$ is the probability of y successes in n trials when the probability of a single success is p and $C_{n,y}$ is the factorial combination given by

$$C_{n,y} = \frac{n!}{y!(n-y)!}. \tag{7}$$

and where the symbol $n!$ stands for factorial [1].

Some examples of binomial random variables are: (1) Number of heads in n tosses of a coin, (2) Number of odd-numbered faces in n throws of die, (3) Number of intoxicated drivers in a random stop of 100 cars, (4) Number of bad debts in an audit of 50 credit accounts, and (5) Number of defective pistons in a quality check of 40 engines.

Example 3. Suppose that we wanted to compute the probability that we will get exactly 2 heads if we flip a fair coin three times. Then, observing heads (successes) can be modeled as a binomial random variable where $n = 3, p = 0.5, q = 0.5$ and $y = 2$. Then by eqn. 6

$$P(Y = 2) = C_{3,2}.5^2.5^{(3-2)} \quad = 3 \times .25 \times .5 = .375$$

The PDF and CDFs of a binomial random variable when the parameters are $n = 10$ and (i)p = .2,(ii) p = .5. and (iii) p=.8 respectively, are given in Figure 3.

[1] For instance 10! would be $10 \times 9 \ldots \times 2 \times 1$

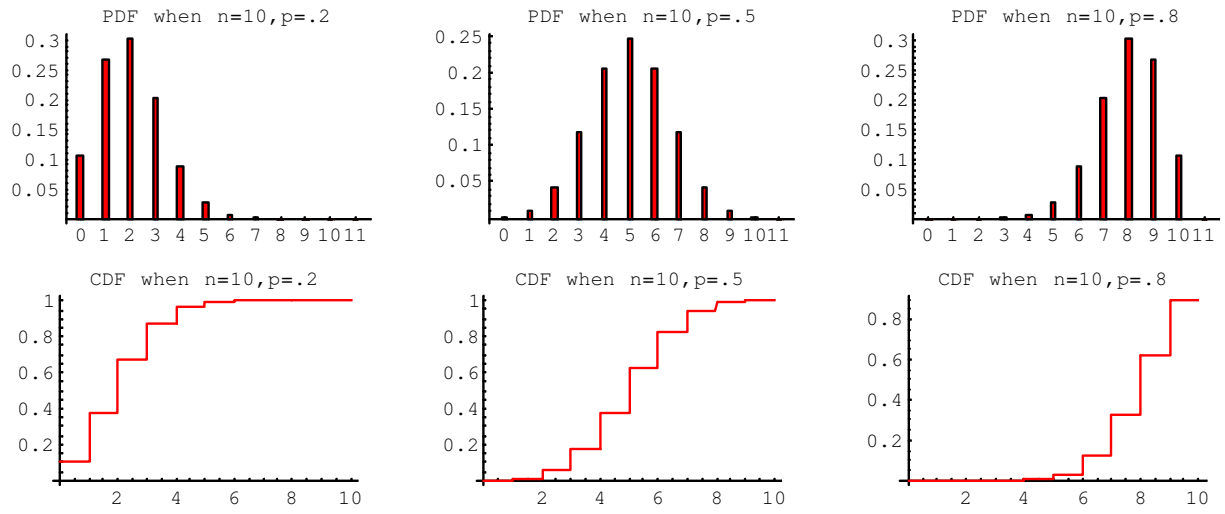


FIGURE 3. Binomial PDFs and CDFs

2.6. **Geometric Random Variable**

Imagine a process in which there are two or more consecutive trials, each of which has just two possible outcomes (success or failure), and the probability of success remains the same (say p) from one trial to the next (the trials are independent). However, the process continues until you see the first success. Then the sample space for this process consists of the points

$$\begin{aligned}
&E_1\text{: S} && \text{(First success on first trial)}\\
&E_2\text{: FS} && \text{(First success on second trial)}\\
&E_3\text{: FFS} && \text{(First success on third trial)}\\
&E_4\text{: FFFS} && \text{(First success on fourth trial)}\\
&\qquad\vdots\\
&E_5\text{: } \underbrace{\text{FFFF}\ldots\text{FS}}_{k-1} && \text{(First success on kth. trial)}
\end{aligned}$$

Note that the size of the sample space is countably infinite (the process may continue for ever if you never get any success since the process only stops when you get the first success). Suppose that we define Y to be a random variable that maps the points $E_1, E_2, \ldots, E_k$ to the real number line such that

$$Y(E_1) = 1,\ Y(E_2) = 2, \quad Y(E_3) = 3,\ \ldots, \quad Y(E_k) = K$$

i.e., the random variable of interest Y, is the number of trial on which the first success occurs. Then, Y is said to be a random variable with a **geometric** probability distribution. In general, since the experiment will continue until you observe a success, say on the yth trial, the sample space will consist of only $E_1, E_2, \ldots, E_y$ and the probability of observing success on the yth trial, i.e., the

PDF of Y is

$$f(y) = P(Y = y) = P(E_y) \quad = P\big(\underbrace{\text{FFFF}\ldots\text{FS}}_{y-1}\big) = q^{(y-1)}p$$

where as before $q = 1 - p$. Note the following features of the geometric random variable (and compare them to the case of the binomial random variable)

- Sequence of independent bernoulli processes and the number of trials is not fixed.
- Again, the trials must be identical and independent.
- Each trial has only two outcomes: success (S) and failure (F) with probabilities p and $q = 1 - p$.
- For each individual trial, the probability of success is the same.
- However, the geometric random variable Y is the number of the trail on which the first success occurs (as opposed to the binomial case where Y is the number of successes observed in n trials).
- The parameter of the distribution is p and the PDF is as given earlier but reproduced here for completeness.

$$\begin{aligned} f(y) = P(Y = y) = P(E_y) \quad &= P\big(\underbrace{\text{FFFF}\ldots\text{FS}}_{y-1}\big) \\ &= q^{(y-1)}p \end{aligned} \tag{8}$$

The PDF and CDF for a random variable Y when $p = .2, .5$ and $.8$ respectively is given in Figure 4.

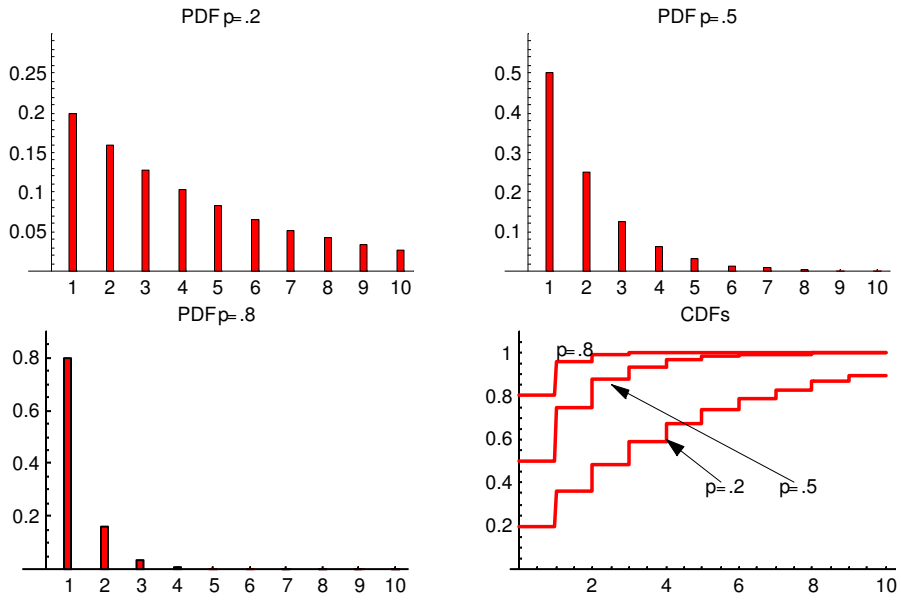


FIGURE 4. Geometric PDFs and CDFs

Example 4. Suppose that we play a game of flipping a coin and the game ends the first time that we get a head. The the probability probability that the game will end on, say the 12 $th.$ flip can be computed using the PDF give in equation 8 and noting that $p = .5, y = 12$. Then,

$$P(Y = 12) = .5^{12-1} \times .5 = .000244$$

2.7. Negative Binomial Random Variable

Imagine now a process in which again there are two or more consecutive trials, each of which has just two possible outcomes (success or failure), and the probability of success remains the same (say p) from one trial to the next (the trials are independent). However, this time the process continues not until you see the first success but rather the second success (or third, or fourth, or $\ldots, kth.$ success where $k = 2, 3, 4, \ldots$). Then the sample space consists of the set of elements E_y where $k-1$ successes occurred on the first $y-1$ trials and the $kth.$ success occurred on the $yth.$ trail (Note (1) that we are not interested in which of the $y-1$ trials the $k-1$ successes occurred, i.e, we are not concerned about the order of success or failure in the first $y-1$ trials as long as $k-1$ successes did occur and $y \geq k$, i.e., the number of trial is greater than or equal to the number of preselected successes).

Suppose that we define Y to be a random variable that maps the points E_y the real number line such that $Y(E_y) = y$ where $y \geq k$. Then the random variable of interest Y, is the number of the trial on which the $kth.$ success occurs. The random variable Y has what is called the **negative binomial** probability distribution. Note the following about the negative binomial distribution (and compare it to the features of the binomial and geometric distributions).

- Similar to geometric and binomial (but not the same). The number of trials is not fixed.
- Again, the trials must be identical and independent.
- Each trial has only two outcomes: success (S) and failure (F) with probabilities p and $q = 1 - p$.
- However, the negative binomial random variable Y is the number of the trail on which the $kth.$ success occurs (compare to geometric where the random variable Y is the number of the trial on which the first success occurs)
- The parameter of the distribution is p and k and the PDF of a negative binomial random variable is given by

$$f(y) = P(Y = y) = C_{(y-1),(k-1)} p^k q^{(y-k)} \text{ where } y \geq k \tag{9}$$

Example 5. Suppose that we play a game of flipping a coin and the game ends the third time that we get a head. Then the probability that the game will end on, say the 12 *th*. flip can be computed using the PDF above (eqn. 9) and noting that $p = 0.5, q = 0.5$ and $k = 3$ and $y = 12$. Then,

$$P(Y = 12) = C_{(12-1),(3-1)} \times .5^3 \times .5^{12-3} = .0134$$

The PDFs and CDFs for a negative binomial random variable Y when $k = 3$, and $p = .2, .5$ and $.8$ and when $k = 8$, and $p = .2, .5$ and $.8$ respectively are given in Figure 5 and Figure 6.

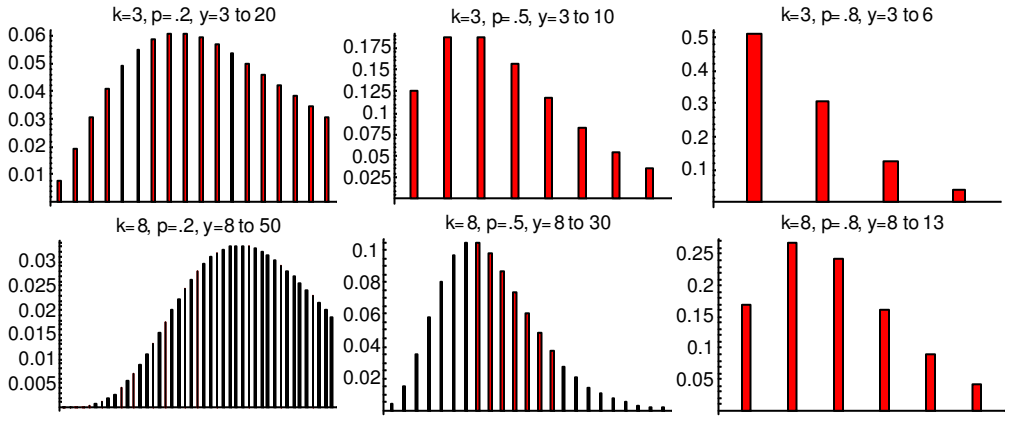


FIGURE 5. Negative Binomial PDFs

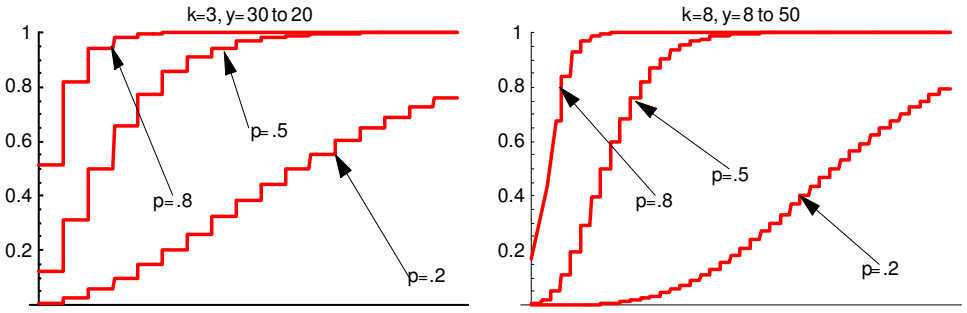


FIGURE 6. Negative Binomial CDFs

2.8. **Poisson Probability Distribution**

One last discrete distribution that we should review is the poisson distribution, which is also related to the binomial distribution. Recall that in the binomial case, the random variable Y is the number of successes/events observed when total trials are n and the probability of success/event on any one trial is p. Consider a *continuous* interval or region (day, week, month, year whatever), which can be

broken up into n subintervals, each of which is so small that at most only one trial can take place in each of the subintervals (and the probability of a success in a single subinterval is p and that of failure is $1-p$). If we were to add up the total number of subintervals in which a success occurred, we would just get back the total number of successes in the entire interval/region. Further, if the successes/events in one interval are independent from another interval, then the total number of successes will have a binomial distribution. If we let np be constant (say $np = \lambda$) and choose smaller and smaller subintervals such that n increases, (i.e., $n \to \infty$) then it can be shown that the limit of the binomial pdf given in 6 is

$$f(y) = P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}.$$

Thus, the poisson distribution can be defined as the limit of the binomial distribution when $n \to \infty$ and np is constant (let the constant number be called λ)[2]. It describes the behavior of a large number n of independent experiments of which only a very small fraction $pn = \lambda$ is expected to yield successes.

A random variable Y given as above has a poisson distribution and it tells us the probability of observing $Y = y$ successes in the given period. Note that it is not written as a function of p (an individual success) but rather of λ which is to be interpreted as the *average number of successes* in the period and is the parameter of this distribution. The poisson distribution is often used to model events such as number of large earth quakes in say a 10 year period, or the number of patents a company may file for in a year, etc. Note the following about the Poisson distribution.

- We count the number of successes, or occurrences of events, in a specified interval or region (time, space etc.)
- The number of events occurring in one specific interval are independent of the number of events in a different interval.
- Each event occurs at a single point on the interval (again, time, space etc.)
- The probability of more than one event at a single point in the interval is zero.
- The Poisson random variable Y is the number of successes/events in the interval/region and for the parameter λ its PDF is given by

$$f(y) = P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}. \tag{10}$$

Example 6. Suppose that a random system of police patrol is devised so that a patrol officer may visit a given beat location $Y = 0, 1, 2, 3, \ldots$ times per half-hour period and that the system is arranged so that each location is visited on an average of once per half-hr. Then,

[2] note: $e = 2.718\ldots$

- the probability that a given location will be missed in a half hr. is $P(Y = 0) = \frac{1^0}{0!}e^{-1} = 1/e = .368$
- the probability that it will be visited once is $P(Y = 1) = \frac{1^1}{1!}e^{-1} = 1/e = .368$
- the probability it will be visited twice is $P(Y = 2) = \frac{1^2}{2!}e^{-1} = 1/2e = .184$
- the probability it will be visited at least once is $P(Y \geq 1) = 1 - P(Y = 0) = 1 - .368 = .632$

The PDFs for a poisson random variable Y when $\lambda = 3, 10, 20$ and 30 respectively are given in figure Figure 7.

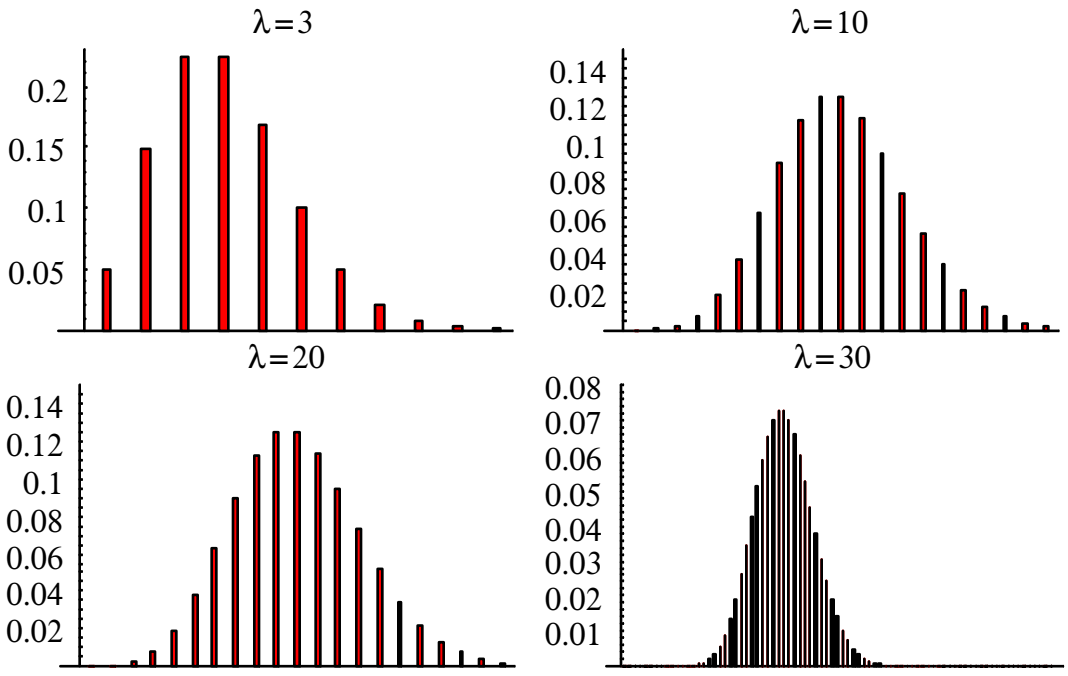


FIGURE 7. Poisson Distributions

2.9. **Probability Distribution of a Continuous Random Variable**

Since a continuous random variable takes on a continuum of possible values, we define the probability of an outcome y in a small interval around it. Specifically, we talk about the probability $P(a \leq y \leq c)$ and summarize the probability of an outcome y (again between small intervals around it) using the *probability density function*, also known as the PDF.

Definition 6 (Continuous Probability Density Function). Let Y be a continuous random variable. Then $f(y)$ is said to be the PDF of Y if the following conditions are satisfied:

$$f(y) \geq 0 \tag{11a}$$

$$\int_{-\infty}^{+\infty} f(y)dy = 1 \tag{11b}$$

$$\int_{a}^{c} f(y)dy = P(a \leq y \leq c) \tag{11c}$$

where $f(y)$ is known as the probability element (the probability associated with a small interval of a continuous variable) and where $P(a \leq y \leq c)$ means the probability that Y lies in the interval a to c.

Graphically, the area under the PDF between any two points (say a and c) is the probability that the continuous random variable is between these two values. Note that for a continuous rv, the probability that it takes on a specific value is zero. Finally, note the relationship between PDF and

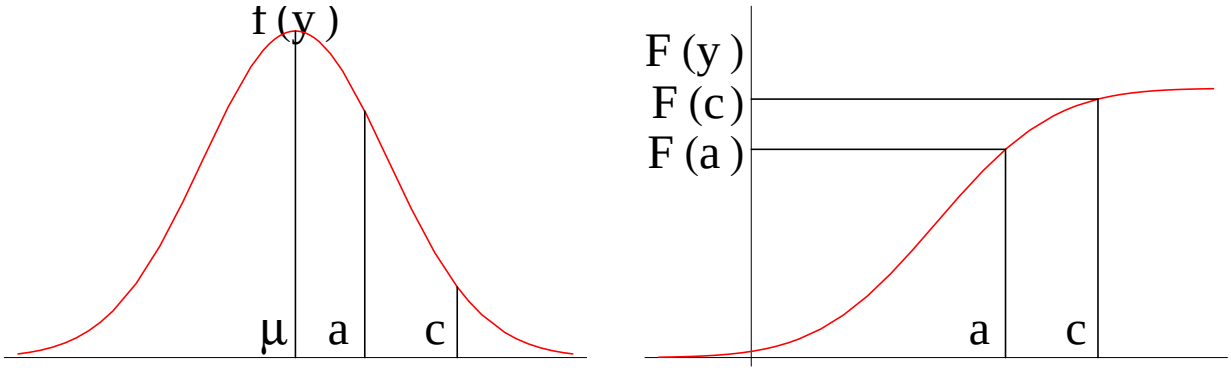


FIGURE 8. PDF and CDF of a Continuous Random Variable

CDF of a continuous random variable (infact,some authors often define PDF via this relationship): For a continuous random variable Y

$$F(y) = \int_{-\infty}^{y} f(t)dt \quad \text{and } f(y) \quad = \frac{dF(y)}{dy} \tag{12}$$

The CDF of the continuous rv Y is given as shown in the right hand side panel in figure 8, and where as in the PDF curve the probability $P(a \leq y \leq c)$ is given by the area under the curve between a and c, in the CDF curve (see Item 5 of Theorem 1 again) it is given by

$$P(a \leq y \leq c) = F(c) - F(a)$$

i.e., by the difference in the value of the function as on shown on the vertical axis.

There are a number of important continuous distributions (uniform, normal, beta, etc.). I will delay the discussion on these for the moment and will later return to four continuous distributions: (1) Normal distribution, (2) Students t^2 distribution, (3) F-distribution and (4) Chi Sq (χ^2) distribution.

3. Moments of a Distribution

3.1. Mean and Variance of a Random Variable

The expected value of a random variable Y, denoted $E(Y)$, is the long-run average, or mean value of the outcome repeated over many trials. It is a measure of the central tendency of the probability distribution of the random variable. For a discrete rv, it is equal the the weighted average of all the possible outcomes, where the weights are the probabilities of each of the outcomes. Thus, to compute the expected value of a random variable Y with k possible outcomes, we multiple the value of each outcome with the probability of the outcome and then sum over all the possible outcomes. For continuous random variable, the expected values is essentially the same thing except that instead of using a specific outcome, we consider the value of the random variable within a small interval and then weight it by the probability element of that interval before "summing" over all the possible outcomes.

Definition 7 (**Expectation**)**.** Let Y be a random variable with probability distribution $f(y)$. Then, the expected value of Y, denoted $\mathrm{E}(Y)$ is

$$\mathrm{E}(Y) = \begin{cases} \sum_y yf(y) & \text{if Y is a discrete random variable} \\ \int_{-\infty}^{+\infty} yf(y)dy & \text{if Y is a continuous random variable} \end{cases} \tag{13}$$

In the definition above, (i) for the discrete case the random variable Y can take k discrete values, $\sum_y$ means the sum over all the k discrete values and $f(y)$ is the discrete PDF of Y and (ii) for the continuous case, where say Y can take any value between $-\infty$ and $+\infty$, then $\int_{-\infty}^{+\infty}$ means the 'sum' (i.e. integral) over the values of Y and $f(y)$ is the continuous PDF of Y. Note that the expected value of a constant is the same constant. Draw the number $Y = 4$, say 1000 times. The average or the expected value will also be 4. Also, the expected value of a constant multiplied by a random variable is equal to the constant multiplied by the expected value of the random variable. Thus, we have the following theorem.

Theorem 2 (**Expectation of a constant and a RV**)**.** Let Y be a random variable and c be a constant. Then,

(1) $E(c) = c$
(2) $E(c + Y) = c + E(Y)$
(3) $E(cY) = cE(Y)$.

Proof.

For (1), let X be a random variable such that $\mathrm{P}(X = c) = 1$. Thus, $X = c$ then by definition 7,

$$E(c) = E(X) = \sum_x xf(x) = \sum_x cf(x) = c\sum_x f(x) = c.1 = c.$$

For (2), observe that

$$E(c + Y) = \sum_y (c + y)f(y) = \sum_y cf(y) + \sum_y yf(y) = c + E(Y).$$

For (3), observe that

$$E(cY) = \sum_y cyf(y) = c\sum_y yf(y) = cE(Y)$$

□

The variance of a random variable Y, denoted $\mathrm{var}(Y)$ is a measure of the spread of the probability distribution and is the expected value of the square of the deviation of Y from its mean.

Definition 8 (**Variance**)**.** Let Y be a random variable with the expected value equal to μ_Y, i.e., $\mathrm{E}(Y) = \mu_Y$. Then the variance of Y, denoted $\mathrm{var}(Y)$ is defined as the expectation of $(Y - \mu_Y)^2$. Thus

$$\mathrm{var}(Y) = \mathrm{E}[(Y - \mu_Y)^2]. \tag{14}$$

Note that, denoting $\mathrm{var}(Y)$ by $\sigma_Y^2, \mathrm{E}(Y)$ by μ_Y, and the PDF by $f(y)$, then definitions 7 and 8 imply that variance of Y is

$$\mathrm{var}(Y) = \sigma^2 = \begin{cases} \sum_y (Y - \mu_Y)^2 f(y) & \text{if Y is a discrete random variable} \\ \int_{-\infty}^{+\infty} (Y - \mu_Y)^2 f(y)dy & \text{if Y is a continuous random variable} \end{cases} \tag{15}$$

Often, a useful way of computing the variance of a random variable is to compute the difference between expectation of the square of the variable and the square of the expectation.

Theorem 3 (**Variance as expectation of square minus square of expectation**).

$$\text{var}(Y) = E[Y^2] - (E[Y])^2 \tag{16}$$

Proof. Left as exercise □

Since the units of the variance are the units of square of the variable Y, we often measure the spread by the standard deviation, which is the square root of the variance (denoted std(Y)).

Example 7 (General Discrete Distribution)**.** Let Y be the number of surgeries performed by Dr. Ahmed in a day (see table on p. 5), then

$$\begin{aligned}
\text{E}(Y) &= 0 \times .125 + 1 \times .375 + 2 \times .375 + 3 \times .125 \\
&= 1.5 \\
\text{var}(Y) &= (0 - 1.5)^2 \times .125 + (1 - 1.5)^2 \times .375 + (2 - 1.5)^2 \times .375 + (3 - 1.5)^2 \times .125 \\
&= .75 \\
\text{std}(Y) &= \sqrt{.75} = .866
\end{aligned}$$

Example 8 (Bernoulli Random Variable)**.** Consider a Bernoulli random variable Y equal to your income in a month which takes a value of \$100 with probability .1 if you fall sick and equal to \$600 with probability .9 if you do not fall sick. Then,

$$\begin{aligned}
\text{E}(Y) &= 100 \times .1 + 600 \times .9 && = \$550 \\
\text{var}(Y) &= (100 - 550)^2 \times .1 + (600 - 550)^2 \times .9 && = 205,000 \\
\text{std}(Y) &= \sqrt{205000} = \$452.77
\end{aligned}$$

Example 9 (Discrete Uniform Distribution)**.** Consider a random variable Y with values equal to y which are equal to the number of dots on a throw of a dice. Then, Y has a discrete uniform distribution (each outcome with probability 1/6) and

$$\begin{aligned}
\text{E}(Y) &= 1 \times (1/6) + 2 \times (1/6) + \ldots + 6 \times (1/6) && = 3.5 \\
\text{var}(Y) &= (1 - 3.5)^2 \times (1/6) + \ldots + (6 - 3.5)^2 \times (1/6) && = 2.9166 \\
\text{std}(Y) &= \sqrt{2.9166} = 1.7078
\end{aligned}$$

Example 10 (Binomial Distribution)**.** Suppose I flip a coin 3 times and call it a success if I observe heads. If we let Y be the number of heads that we observe out of these 3 trials, and if $p = .5$ is the probability of a success on any one trial, then Y is a binomial random variable. To compute

the expected number of successes, their variance and standard deviation, first observe that Y can take on 4 values (0,1,2,3) and the probabilities of each of these outcomes is (.125,.375,.375,.125) (see equations 6 & 7 on p. 7) i.e., hence, as in example 7

Outcome	0	1	2	3
PDF $(P(y))$	0.125	0.375	0.375	0.125
CDF $(F(y < Y)$	0.125	0.500	0.875	1.000

$$\begin{aligned}
\mathrm{E}(Y) &= 0 \times .125 + 1 \times .375 + 2 \times .375 + 3 \times .125 \\
&= 1.5 \\
\mathrm{var}(Y) &= (0 - 1.5)^2 \times .125 + (1 - 1.5)^2 \times .375 + (2 - 1.5)^2 \times .375 + (3 - 1.5)^2 \times .125 \\
&= .75 \\
\mathrm{std}(Y) &= \sqrt{.75} = .866
\end{aligned}$$

For many distributions (both discrete and continuous) we can often derive a formula in closed form the expectation and variance of the distribution. For some of these, see Table 1

TABLE 1. Moments of Discrete Distributions

Distribution	E(y)	var(y)
*Uniform (Discrete)	$\frac{b+a}{2}$	$\frac{(b-a)(b-a+2)}{12}$
Bernoulli	p	$p\,(1-p)$
Binomial	np	$np\,(1-p)$
Geometric	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Poisson	λ	λ

* where Y takes values on n integers $a, a+1, \ldots, b$

3.2. Linear Function of a Random Variable

Say X is a random variable with mean μ_X and standard deviation σ_X. Then a linear transformation of X to Y, given by $Y = a+bX$ where a and b are some constants, also results in a random variable, i.e., Y is also a random variable. Further, since Y is a random variable, we can compute its mean and standard deviation from the mean and standard deviation of X. Thus, we have the following theorem.

Theorem 4 (**Linear function of a random variable**). Let X be a random variable where the expectation and variance of X are $\mathrm{E}(X) = \mu_X$ and $\mathrm{var}(X) = \sigma_X^2$. Then a linear transformation of X, given by

$$Y = a + bX$$

where a and b are any constants, gives another random variable Y and

$$E(Y) = \mu_Y = a + b\mu_X \tag{17a}$$

$$\mathrm{var}(Y) = \sigma_Y^2 = b^2\sigma_X^2 \tag{17b}$$

Proof. By Theorem 2, observe that

$$E(Y) = E(a + bX) = E(a) + E(bX) = a + bE(X) = a + b\mu_X$$

To prove the second part, observe that

$$\begin{aligned}
\sigma_Y^2 = var(Y) &= E\big[(Y - E(Y))^2\big] && \text{by definition of variance (see def.8)} \\
&= E\big[(a + bX - a - b\mu_X)^2\big] && \text{by what we just proved above} \\
&= E\big[b^2(X - \mu_X)^2\big] && \text{algebraic simplification} \\
&= b^2E\big[(X - \mu_X)^2\big] && \text{by theorem 2} \\
&= b^2\sigma_X^2 && \text{\& by definition of variance of X (see def.8)}
\end{aligned}$$

□

Example 11. Suppose that your health insurance policy costs you \$200 and it stipulates that you will be responsible 5% of any hospital bills incurred during the coming year. If the mean and variance of the your total hospital bills is \$1500 and \$1000 respectively, we can compute the mean and variance of the total cost to you (i.e., cost of insurance plus the portion of medical bills that you will be paying out of pocket) by using the linear transformation given above. Let X be

the rv denoting the total hospital bills incurred during the year such that μ_X =E(X)=1500 and σ_X^2 =var(X) = 100. Now let Y be the total cost to you, given by

$$Y = 200 + .05X$$

then

$$\mu_Y = \mathrm{E}(Y) = 200 + .05\mu_X = \$275$$
$$\sigma_Y^2 = \mathrm{var}(Y) = .05^2\sigma_X^2 = 2.5$$
$$\sigma_Y = \mathrm{std}(Y) = .05\sigma_X = \$1.58$$

4. **Joint, Marginal & Conditional Distributions**

So far we have considered the PDF, CDF and moments (mean, variance) of a single random variable. Often, in econometric work we are more interested in the joint distributions of two or more random variables and the mean and variance of the joint distributions.

4.1. **Joint Distribution**

The joint probability distribution of two discrete random variables, say X and Y, denoted $f(x, y)$ is the tabulation of probabilities of all possible combinations of outcomes for the two variables. Thus, for example, if X can take on 4 values, say 1,2,3 and 4 and Y can take on only 2 value, say A or B, then listing out the probabilities of all possible combinations $\{(1, A), (1, B), (2, A), (2, B) \ldots (4, A), (4, B)\}$ in either a table or a graph would be specifying the joint probability distribution. The joint probability distribution itself would be the frequency of each of these eight possible outcomes over repeated trails. More generally, for two discrete random variables X and Y the joint probability distribution would be

$$f(x, y) = \mathrm{P}(X = x, Y = y).$$

Similarly, if X and Y were two continuous random variables then the joint probability distribution would tell us the probability of X and Y within some intervals around specific values x and y, i.e.,

$$f(x, y) = \mathrm{P}(a \leq x \leq b, c \leq y \leq d).$$

Definition 9 (**Discrete Joint PDF**)**.** Let X and Y be two discrete random variables. Then the function $f(x, y)$ is known as the discrete joint density function and is defined as

$$\begin{aligned} f(x, y) &= \mathrm{P}(X = x \text{ and } Y = y) \\ &= 0 \text{ when } X \neq x \text{ and } Y \neq y \\ \text{and } \sum_X \sum_Y f(x, y) &= 1 \end{aligned} \tag{18}$$

where $\sum_X \sum_Y$ means summation over all possible values of X and Y.

Definition 10 (**Continuous Joint PDF**)**.** Let X and Y be two continuous random variables. Then the function $f(x, y)$ is known as the continuous joint density function and is defined such that

$$Prob(a \leq x \leq, c \leq y \leq d) = \int_a^b \int_c^d f(x, y)\, dydx \tag{19}$$

where $f(x, y) \geq 0$ and $\int_x \int_y f(x, y)\, dydx = 1$.

Just as we defined the expectation of random variable in the univariate case, we can define the expectation of any function of two random variables.

Definition 11 (**Expected value of a function of random variables**)**.** Let X and Y be two random variable with joint distribution given by $f(x, y)$. Then, for any function $g(x, y)$ of X and Y, the expected value of $g(x, y)$, denoted $E[g(x, y)]$, is defined as

$$E[g(x, y)] = \begin{cases} \sum_y \sum_x g(x, y) f(x, y) & \text{if X and Y are discrete random variables} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dxdy & \text{if X and Y are continuous random variables} \end{cases} \tag{20}$$

Definition 12 (**Joint Cumulative Distribution**)**.** If X and Y are two random variables, then the joint cumulative distribution, or the bivariate CDF, $F(x, y) \equiv \mathrm{P}(X \leq x, Y \leq y)$ is given by

$$F(x, y) = \begin{cases} \sum_{s \leq x} \sum_{t \leq y} f(s, t) & \text{if } X, Y \text{ discrete} \\ \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, dsdt & \text{if } X, Y \text{ continuous} \end{cases} \tag{21}$$

Example 12 (Two Dice - Joint Discrete PDF)**.** Suppose you were to throw two dice and you were interested in knowing joint probabilities of pairs of outcomes such as (6,1), (1,6), (3,4) etc. Then we can summarize the probability of all possible outcomes using the joint PDF. Let X be number of dots on dice 1 and and Y be the number of dots on dice 2. The the joint PDF is as given in Table 2.

TABLE 2. Joint Distribution (PDF) of Rolling Two Dice

	X=1	X=2	X=3	X=4	X=5	X=6	Marginal Probability $f_Y(y)$ =P$(Y=y)$
Y=1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Y=2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Y=3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Y=4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Y=5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Y=6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Marginal Probability $f_X(x)$ =P$(X=x)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Example 13 (Joint PDF of Difficulty on Exam and Passing the Exam)**.** Next, consider the case when a student (not fully prepared) faces two random events, (i) if they will pass the exam ($Y=1$) or not ($Y=0$) and (ii) if the exam will be easy ($X=1$), moderate ($X=2$) or tough ($X=3$). The frequency of outcomes from past exams, i.e., the joint probability distribution (for a hypothetical class with the same instructor and same syllabus), are summarized in Table 3 below.

TABLE 3. Joint Distribution (PDF) of Passing an Exam and Level of Difficulty

	X=1	X=2	X=3	Marginal Probability $f_Y(y)$ =P$(Y=y)$
Y=1	.20	.40	.20	.80
Y=0	.05	.10	.05	.20
Marginal Probability $f_X(x)$ =P$(X=x)$	.25	.50	.25	1

In this example, the probability of the joint event that the exam will be moderate and that the student will fail the exam P$(X=2, Y=0)$ is .10 while the probability of the joint event that the exam will be difficult and that the student will pass P$(X=3, Y=1)$ is .20.

4.2. **Marginal Distribution**

In both the examples above, note the numbers appearing in the last columns and the last rows. These are equal to the row sums and row columns. For instance, in Table 2 the number 1/6 in the first row last column is the sum of all the numbers in the first row. Same goes for the sum of numbers in the second row. Similarly, the number 1/6 in last row first column is the sum of

numbers in the first column. These numbers in the last column and the last row are called the marginal probabilities. Notice that these are just the probabilities of observing a 1,2,3,4,5,or 6 on a roll of single dice. Thus, the marginal probability distribution of a random variable is just another name for its probability distribution ... something we have already seen in the previous pages. Consequently, the sum of the last row itself, or the last column, is 1. This is also equal to the sum of all the numbers in the table excluding the last row and the last column.

This can be seen more easily in the next table (Table 3). The first three numbers in the first, .20,.40 and .20 give the probability of the joint events that the student will pass and the exam will be easy, moderate and tough respectively. But the sum of these three numbers, .80, gives you the probability that the student will pass the exam. Similarly, the first three numbers in the second row .05,.10 and .05 give you the probabilities of the joint events that the student will fail the exam and that the exam will be easy, moderate and tough respectively. The sum of these, .20, is the probability that the student will fail the exam. These two numbers, .80 and .20 are the marginal probability distribution (or the probability distribution) and represented by $f_Y(y)$, for the random variable Y representing the outcomes that the student will pass ($Y = 1$) or fail ($Y = 0$). Likewise, the marginal probability distribution of the random variable X that the exam will be easy ($X = 1$), moderate ($X = 2$) or tough ($X = 3$) is given by $f_X(x)$ where the probability of receiving an easy exam is .25 (sum of first column), that of a moderate exam is .50 (sum of second column) and that of a tough exam is .25 (sum of the third column). The formal definitions follow.

Definition 13 (**Marginal Probability Distribution**)**.** If X and Y are two random variables jointly distributed, then the marginal probability distributions (or the marginal PDFs) of X and Y are

$$f_X(x) = \begin{cases} \sum_y f(x,y) & \text{for the discrete case} \\ \int_y f(x,t)\,dt & \text{for the continuous case} \end{cases} \tag{22a}$$

$$\text{and, } f_Y(y) = \begin{cases} \sum_x f(x,y) & \text{for the discrete case} \\ \int_x f(s,y)\,ds & \text{for the continuous case.} \end{cases} \tag{22b}$$

A useful result to consider at this point is the expected value of the sum of two random variables.

Theorem 5 (**Expected value of sum of two random variables**)**.** Let X and Y be two random variables. Then

$$E(X+Y) = E(X) + E(Y). \tag{23}$$

Proof. To prove this result, we start with the definition of expected value of a function of random variables, i.e. in definition 11 let $g(x,y) = X + Y$. Then, by definition 11 the expected value of $(X+Y)$ is $E(X+Y) = \sum_x \sum_y (x+y) f(x,y)$. Next, observe the following:

$$\begin{aligned}
E(X+Y) &= \sum_x \sum_y (x+y) f(x,y) \\
&= \sum_x \sum_y x f(x,y) + \sum_x \sum_y y f(x,y) \\
&= \sum_x x \Big(\sum_y f(x,y) \Big) + \sum_y y \Big(\sum_x f(x,y) \Big) \\
&= \sum_x x f_x(x) + \sum_y y f_y(y) \text{ (by definition of marginals above)} \\
&= E(X) + E(Y).
\end{aligned}$$

□

Example 14 (General Discrete Bivariate Probability Distribution)**.** In the table below, the cell entries give the probability of the joint occurrence of the associated $1, 2, \ldots, m$ values of X and $1, 2, \ldots, p$ values of Y, where p_{ij} = probability that $X = x_i$ and $Y = y_j$. The column and row totals give the marginal probabilities for X and Y respectively and the period indicates the subscript over which summation has taken place. Thus, $f_x(x) = (p_{1\cdot}, \ldots, p_{i\cdot}, \ldots p_{m\cdot})$

TABLE 4. Bivariate Probability Distribution

	X_1	$\ldots$	X_i	$\ldots$	X_m	Marginal Probability $f_Y(y) = \mathrm{P}(Y = y)$
Y_1	p_{11}	$\ldots$	p_{i1}	$\ldots$	p_{m1}	$p_{\cdot 1}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
Y_j	p_{1j}	$\ldots$	p_{ij}	$\ldots$	p_{mj}	$p_{\cdot j}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
Y_p	p_{1p}	$\ldots$	p_{ip}	$\ldots$	p_{mp}	$p_{\cdot p}$
Marginal Probability $f_X(x) = \mathrm{P}(X = x)$	$p_{1\cdot}$	$\ldots$	$p_{i\cdot}$	$\ldots$	$p_{m\cdot}$	1

4.3. **Conditional Distribution**

In a bivariate distribution, there is a conditioning distribution over y for each value of x (and vice versa). For instance, the distribution of the random variable Y conditional on *a specific value* of the random variable X, say x, is called the conditional distribution of Y given X and is denoted by $f(y|x)$. Thus, the conditional distribution tells us the probability that Y takes on a value of y when X is held at the value x, i.e., $\mathrm{P}(Y = y|X = x)$. In table 4, the probability that $Y = Y_1$ when given that $X = X_1$ is

$$P(Y = Y_1|X = X_1) = \frac{p_{11}}{p_{1.}}$$

and similarly the probability that $Y = Y_2$ when given that $X = X_1$ is

$$P(Y = Y_2|X = X_1) = \frac{p_{12}}{p_1}.$$

If we enumerated the probabilities of all possible outcomes for Y for a given value of $X = X_1$, we would have listed the conditional distribution of Y given $X = X_1$, i.e., $f(y|X_1)$. Clearly, we can do this (i) for other values of X and (2) by reversing the process, i.e., compute probabilities of X for given values of Y. Finally, note that to compute the conditional probabilities, all that we are doing is that we are dividing the joint probability by the marginal probability.

Definition 14 (**Conditional Distribution**)**.** If X and Y are two random variables with the joint distribution $f(x, y)$ and marginal distributions $f_X(x)$ and $f_Y(y)$, then the conditional marginal probability distributions are

$$f(y|x) = \frac{f(x,y)}{f_X(x)} \text{ and } \quad f(x|y) = \frac{f(x,y)}{f_Y(y)}. \tag{24}$$

These definitions provide a useful way of computing the joint distribution

$$f(x,y) = f(y|x) \cdot f_X(x) \quad = f(x|y) \cdot f_Y(y) \tag{25}$$

Lets make all these ideas concrete by looking at a specific numerical example.

Example 15 (Income and Vacation Expenditure (Hypothetical Population))**.** Suppose that every one in a hypothetical population earns either $20, 30$ or 40 thousand dollars in a given year (random variable X) and gets to spend $1, 2, 3, 4, 5$ or 6 thousand on vacations (random variable Y). Further, suppose that the joint probabilities of earing X and spending Y are as given in the upper left part of table 5. Thus, the probability that someone earns \$20,000 and spends \$1,000 on vacations is 0.28 while the probability that someone earns \$40,000 and spend \$4,000 is 0.15. Given this joint distribution we can compute the marginal distributions as well as conditional distributions.

TABLE 5. Joint, Marginal, & Conditional Distributions

	Col1	Col2	Col3	Col4	Col5		Col6	Col7	Col8	Col9
		X($'000)								
	Y($'000)	20	30	40	$f_Y(y)$					
Row1	1	.28	.03	0	.31					
Row2	2	.08	.15	.03	.26					
Row3	3	.04	.06	.06	.16					
Row4	4	0	.06	.15	.21					
Row5	5	0	0	.03	.03					
Row6	6	0	0	.03	.03					
Row7	$f_X(x)$	.40	.30	.30	1					
		Conditional Probabilities								
		$f(y\|x)$								
		20	30	40			20	30	40	
Row8	1	.7	.1	0			.9	.1	0	1
Row9	2	.2	.5	.1			.31	.58	.12	1
Row10	3	.1	.2	.2		$f(x\|y)$	.25	.38	.38	1
Row11	4	0	.2	.5			0	.29	.71	1
Row12	5	0	0	.1			0	0	1	1
Row13	6	0	0	.1			0	0	1	1
Row14		1	1	1						

Example from Johnston & Dinardo, p.14

First, the marginal distributions can be computed by by taking the sum of the rows or columns. For $f_Y(y)$, given in column 5, .31 is .28+.03 (sum of row 1); .26 is .08+.15+.03 (sum of row 2) and so on. Similarly for $f_X(x)$ given in row 7, take the sum of rows 1 through 6 in column 2 to get .40 and take the sum of rows 1-6 in column 3 to get .30 and so on. Next, to compute the conditional probabilities, $f(y|x)$, first note that there are three different conditional distributions, one for each value of X. The conditional distribution $f(y|X = 20)$ is given is column 2 (rows 8-13), that of $f(y|X = 30)$ is in column 3 (rows 8-13) and $f(y|X = 40)$ is in column 4 (rows 8-13). To compute the actual numbers in the conditional distributions (say $f(y|X = 20)$), all we need to do is divide the joint probability with the value of the marginal distribution $f_X(x)$ at that value of X (in this case 20). Thus, to get the values for the conditional distribution $f(y|X = 20)$, divide .28 by .40 to get .7 (given in col2, row8), divide .08 by .40 to get .2 (in col2 row9) and divide .04 by .40 to get .10 (in col2 row10). Similarly, to compute $f(y|X = 30)$, you would divide each of the joint probabilities in column 3, rows 1-6 by .30 to get the numbers in col3, rows 8-13. Finally, to compute the conditional probabilities, $f(x|y)$, note that there are five different conditional distributions, one for each value of Y. These can be computed in the analogous way by dividing the joint probabilities by a specific value of the

the marginal distribution $f_Y(y)$. Thus, dividing the joint probabilities in row 1, cols 2-4 by .31 gives the conditional distribution $f(x|Y=1)$ and is given in cols 6-8 of row 8.

4.4. Conditional Mean and Variance

Eqns. 13 through 15 give the mean and variance of a distribution of a random variable. These were the unconditional mean and variance. We can do the same for the conditional distributions.

Definition 15 (Conditional Mean). Conditional mean is the mean of the conditional distribution. Thus, the conditional mean of Y for a specific value of $X = x$, denoted $\mathrm{E}(Y|x)$, is

$$\mathrm{E}(Y|x) = \begin{cases} \int_y yf(y|x)\,dy & \text{if } y \text{ is continuous} \\ \sum_y yf(y|x) & \text{if } y \text{ is discrete.} \end{cases} \tag{26}$$

Definition 16 (Conditional Variance). Conditional variance of Y for a specific value of $X = x$, denoted $\mathrm{var}(Y|x)$, is the variance of the conditional distribution

$$\begin{aligned} \mathrm{Var}(Y|x) =& \mathrm{E}\big((Y - \mathrm{E}(Y|x))^2|x\big) \\ =& \mathrm{E}(Y^2|x) - (\mathrm{E}(Y|x))^2 \\ =& \begin{cases} \int_y (y - \mathrm{E}(Y|x))^2 f(y|x)\,dy & \text{if } y \text{ is continuous} \\ \sum_y (y - \mathrm{E}(Y|x))^2 f(y|x) & \text{if } y \text{ is discrete.} \end{cases} \end{aligned} \tag{27}$$

The conditional mean function, $E(Y|x)$ is called the **regression** of y on x. The conditional variance is called the **scedastic function** and is generally a function of x. Typically, the conditional variance does not vary with x. This does not imply that that $\mathrm{var}(Y|x)$ is the same as $\mathrm{var}(y)$. All it means is that the conditional variance is a constant. The cases were the conditional variance does not vary with x is called **homoscedasticity**.

In Equation 26 we are computing the mean of Y for a specific value of $X = x$. If we repeated this calculation for all specific values of X and then took the mean of all those values, it would give us back just the mean value of Y, i.e., the unconditional mean of Y. This is known as the law of iterated expectations.

Theorem 6 (Law of Iterated Expectations). The expectation of Y is the expectation of the conditional expectation of Y given $X = x$.

$$\mathrm{E}(Y) = \mathrm{E}_X[E(Y|x)]$$

where $\mathrm{E}_X[E(Y|x)]$ means $\sum_x \Big[\sum_y yf(y|x)\Big] f_X(x)$.

In the equation above, note that on the right hand side, we are first computing the inner conditional expectation of Y given $X = x$ and then taking the outer expectation $E_X[.]$ over the values of X. For the first (inner) expectation, we use the conditional distribution of Y, i.e. $f(y|x)$, and for the second (outer) expectation, we use the marginal distribution of X, i.e., $f_X(x)$.

Proof.

$$\begin{aligned}
\mathrm{E}_X[E(Y|x)] &= \sum_x \Big[\sum_y y f(y|x)\Big] f_X(x) = \sum_x \sum_y y f(y|x) f_X(x) \\
&= \sum_x \sum_y y f(y,x) \text{ by equation 25 and definition 14} \\
&= \sum_y y \sum_x f(y,x) \text{ by switching the order of summation} \\
&= \sum_y y f_Y(y) \text{ by equation 22b in definition 13} \\
&= \mathrm{E}(Y) \text{ by definition of expected value}
\end{aligned}$$

□

Theorem 7 (**Decomposition of Variance**)**.** In a joint distribution,

$$\mathrm{var}(Y) = \mathrm{var}_X[E(Y|x)] + \mathrm{E}_X[var(Y|x)] \tag{28}$$

Thus, the unconditional variance of Y is equal to the variance of the conditional expectation plus the expectation of the conditional variance.

Proof. The proof follows by substituting in the definitions of conditional mean (definition 15) and of conditional variance (definition 16) on the right hand side of the relationship above and then applying the results of theorem 3 and the law of iterated expectations (already proved above). To proceed, observe the following.

First, by theorem 3, for any random variable a we can write $\mathrm{var}(a) = E[a^2] - (E[a])^2$. Hence, for the expression $var_X[E(Y|x)]$ we can write $E_X\big((E(Y|x))^2\big) - \big(E_X[E(Y|x)]\big)^2$. But note that by the law of iterated expectations, the second expression within the squared sign, $E_X[E(Y|x)]$, is just $E(Y)$ and hence $(E_X[E(Y|x)])^2$ can be written as $(E(Y))^2$.

Second, from the definition of conditional variance (definition 15, eqn. 27) $E_X[Var(Y|x)]$ can be written as $E_X\Big(E(Y^2|x) - (E(Y|x))^2\Big)$. Expanding this out, it becomes, $E_X\Big(E(Y^2|x)\Big) - E_X\Big((E(Y|x))^2\Big)$ (by theorem 5). But once again, by law of iterated expectations, the fist of these terms, $E_X\Big(E(Y^2|x)\Big)$ is just the expectation of Y^2, i.e., $E_X\Big(E(Y^2|x)\Big) = E(Y^2)$ (where we are using the fact that since the law of iterated expectations holds for *any* random variable, it must

also be true for the random variable Y^2. Thats it. Now we just put these terms together and simplify. Thus,

$$\begin{aligned} var_X[E(Y|x)] + E_X[var(Y|x)] &= E_X\big((E(Y|x))^2\big) - \big(E_X[E(Y|x)]\big)^2 + E_X\big(E(Y^2|x) - (E(Y|x))^2\big) \\ &= E_X\big((E(Y|x))^2\big) - (E(Y))^2 + E_X\big(E(Y^2|x)\big) - E_X\big((E(Y|x))^2\big) \\ &= E_X\big((E(Y|x))^2\big) - (E(Y))^2 + E(Y^2) - E_X\big((E(Y|x))^2\big) \\ &= E(Y^2) - (E(Y))^2 \\ &= var(Y). \end{aligned}$$

□

Example 16 (Example 15 continued)**.** For the data given in example 15, lets first compute (1) E(Y), then (2) E$(Y|x)$, then (3) var$(Y|x)$ and finally (4) $\mathrm{E}_X[\mathrm{E}(Y|x)]$ where, the last numeric calculation is to confirm the law of the iterated expectations, since we already know that answer to (4) should be the same as that for (1).

(1) The expectation E(Y) can be computed using just the values of Y and the marginal probabilities for Y, i.e., $\sum_y y f_Y(y)$
So E$(Y) = \sum_y y f_Y(y) = 2.48$

Y	1	2	3	4	5	6
$f_Y(y)$	0.31	0.26	0.16	0.21	0.03	0.03
$y \times f_Y(y)$	0.31	0.52	0.48	0.84	0.15	0.18

(2) Next, to compute the expected value of Y given $X = 20, 30$ *or* 40, we will use the values of Y and the distribution of Y conditional on these three values of X.

	$X=20$		$X=30$		$X=40$	
Y	$f(y\|x)$	$yf(y\|x)$	$f(y\|x)$	$yf(y\|x)$	$f(y\|x)$	$yf(y\|x)$
1	0.7	0.7	0.1	0.1	0	0
2	0.2	0.4	0.5	1	0.1	0.2
3	0.1	0.3	0.2	0.6	0.2	0.6
4	0	0	0.2	0.8	0.5	2
5	0	0	0	0	0.1	0.5
6	0	0	0	0	0.1	0.6
E $(Y\|x) = \sum_y f(y\|x)$	1.4		2.5		3.9	

(3) Next, to compute the variance of Y conditional on $X = 20, 30,$ or 40, once again we use the conditional distribution $f(y|x)$ but this time multiply them with the square of the deviation of Y from its conditional mean values (calculated in item 2 above).

		$X = 20$			$X = 30$			$X = 40$	
	$(Y-1.4)^2$	$f(y\|x)$	$f(y\|x)\times$ $(Y-1.4)^2$	$(Y-2.5)^2$	$f(y\|x)$	$f(y\|x)\times$ $(Y-2.5)^2$	$(Y-3.9)^2$	$f(y\|x)$	$f(y\|x)\times$ $(Y-3.9)^2$
Y=1	0.16	0.7	0.112	2.25	0.1	0.225	8.41	0	0
Y=2	0.36	0.2	0.072	0.25	0.5	0.125	3.61	0.1	0.361
Y=3	2.56	0.1	0.256	0.25	0.2	0.05	0.81	0.2	0.162
Y=4	6.76	0	0	2.25	0.2	0.45	0.01	0.5	0.005
Y=5	12.96	0	0	6.25	0	0	1.21	0.1	0.121
Y=6	21.16	0	0	12.25	0	0	4.41	0.1	0.441
$\text{var}(Y\|x) = \sum_y (y - \text{E}(Y\|x))^2 f(y\|x)$			0.44			0.85			1.09

(4) To do the final computation, multiply $\text{E}(Y|x)$ in item 2 with the marginal distribution of X, i.e., with $f_X(x)$ and then add up the answers:

X	20	30	40
$f_X(x) =$	0.4	0.3	0.3
$\text{E}(Y\|x) =$	1.4	2.5	3.9
$f_X(x) \times \text{E}(Y\|x) =$	.56	.75	1.17

and so, the sum of the last row of the table above is $\sum_x f_X(x) \times \text{E}(Y|x) = 2.48$, i.e., $\text{E}(Y) = \text{E}_X[E(Y|x)] = 2.48$ which is the same answer that we got in item 1.

The main thing to note in (1) vs. (4) is that in (1) we got the expectation of Y by multiplying the values of Y with the marginal distribution of Y (and then adding up these values) where as in (4), we multiplied the conditional expected values of Y with the marginal distribution of X (and them added these numbers up) to get the same answer.

4.5. **Independence, Covariance & Correlation**

We use the concept of independence of events to describe the independence of random variables. For instance, in your earlier probability classes you may have seen that the probability of event A and B is equal to the probability of event A multiplied with the probability of event B if the events were independent (say rolling a dice and tossing a coin at the same time, then the probability of observing 4 dots on the dice and a head on the coin is just $P(4, H) = P(4) \times P(H) = 1/6 \times 1/2$.

In the case of random variables, we do the same by using the PDFs. Specifically, they are said to be independent if the conditional distribution is the same as the marginal distribution.

Definition 17 (**Independence of random variables**). Two random variables are statistically independent if and only if their joint density is the product of marginal densities.

$$f(x,y) = f_X(x)f_Y(y) \Leftrightarrow x \, and \, y \, \text{are statistically independent.} \tag{29}$$

Intuitively, it is easier to understand independence as that two random variables X and Y are independently distributed, if information about the value of X does not provide any information about the value of Y, for instance, $P(Y|x) = P(y)$. In terms of the distributions, this can be written as $f(y|x) = f_Y(y)$. In fact, some authors define independence using this relationship, i.e.,

$$f(y|x) = f_Y(y) \Leftrightarrow x \, and \, y \, \text{are statistically independent.} \tag{30}$$

Which ever of these you start with as the definition of independence, you can always derive the other condition from it.

Example 17. Prove that if $f(x,y) = f_X(x)f_Y(y)$ then it implies that that $f(y|x) = f_Y(y)$

Proof. From Equation 24 and Equation 25 we know that $f(x,y) = f(y|x)f_X(x)$. Thus,

$$\begin{aligned} f(y|x) &= \frac{f(x,y)}{f_X(x)} && \text{from eqns. 24, 25} \\ &= \frac{f_X(x)f_Y(y)}{f_X(x)} && \text{given hypothesis} \\ &= f_Y(y) \end{aligned}$$

□

Example 18. Prove that if $f(y|x) = f_Y(y)$ then it implies that $f(x,y) = f_X(x)f_Y(y)$.

Proof. Left as exercise □

Theorem 8.

$$f(x,y) = f_X(x)f_Y(y) \Leftrightarrow f(y|x) = f_Y(y). \tag{31}$$

Proof. Proof given in the previous two examples. □

When considering two random variables, X and Y, we can also measure how much they move together. For instance, if we draw a large value of Y, do we also typically draw a large value of X? One way to measure this is by measuring the covariance between X and Y.

Definition 18 (**Covariance**)**.** Let X and Y be two random variables. Then, the covariance, $\sigma_{XY} = \text{cov}(X, Y)$ is the expected value of $(X - \mu_X)(Y - \mu_Y)$.Thus,

$$\begin{aligned} \text{cov}(X, Y) = \sigma_{XY} &= \text{E}\big((X - \mu_X)(Y - \mu_Y)\big) \\ &= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) & \text{if discrete} \\ \int_x \int_y (x - \mu_X)(y - \mu_Y) f(x, y) dxdy & \text{if continuous.} \end{cases} \end{aligned} \tag{32}$$

When σ_{XY} is positive it means that X and Y move in the same direction, i.e., on average if X is larger than its mean value μ_X then so is Y larger than its mean value of μ_Y and conversely, when X is small (say below its mean value) then on average Y is also small (below its own mean value). On the other hand, when σ_{XY} is negative, it means that X and Y move in opposite directions, i.e., on average large values of X are associated with small values of Y and vice versa. When the covariance is zero, it means that there is no such association between X and Y.

It is hard to compare association between two pairs of random variables, say X and Y and between two other random variables, say W and V because the units are in terms of products of X and Y (deviated from there means) etc. To overcome this difficulty, we can normalize the covariance by dividing it by the variance of each of the two variables. This is called correlation, denoted ρ_{XY}, and it measures a degree of *linear* association between X and Y. Thus,

Definition 19 (**Correlation**)**.** Let X and Y be two random variables with covariance $\sigma_{XY} = \text{cov}(X, Y)$. The correlation $\text{corr}(X, Y) = \rho_{XY}$ is a measure of linear association between X and Y and is given by

$$\begin{aligned} \rho_{XY} &= \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \\ &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \end{aligned} \tag{33}$$

Note that the correlation is always between -1 and +1. Let us now review some rules about correlations, independence and variances of functions of random variables.

Theorem 9 (**Properties of independently distributed random variables**)**.** If X and Y are two independently distributed random variables, i.e. if $f(x, y) = f_X(x) f_Y(y)$ then

(1) $f(y|x) = f_Y(y)$
(2) $f(x|y) = f_X(x)$
(3) $cov(X, Y) = \sigma_{XY} = 0$
(4) $cor(X, Y) = \rho_{XY} = 0$
(5) $E(Y|x) = E(Y) = \mu_Y$

(6) $var(Y|x) = var(Y) = \sigma_Y^2$

While it is true that if X and Y are independent, $\text{cov}(X, Y) = 0$, it is not necessarily true that if $\text{cov}(X, Y) = 0$ then X and Y are independent.[3]

Proof. For proof of (1) and (2), see Theorem 8. For proof of (3) observe the following. Since X and Y are independent then $f(x, y) = f_X(x)f_Y(y)$. Hence,

$$\begin{aligned}\sigma_{XY} &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\ &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f_X(x) f_Y(y) \\ &= \sum_x (x - \mu_X) f_X(x) \sum_y (y - \mu_Y) f_Y(y) \quad = 0 \times 0 = 0.\end{aligned}$$

Proof of (4) follows from the definition of correlation (see Equation 33), i.e., if covariance is zero, then correlation is zero as well. For proof of (5), observe that

$$\begin{aligned}E(Y|x) &= \sum_y y f(y|x) \\ &= \sum_y y f_Y(y) = E(Y).\end{aligned}$$

Proof of (6) is left as an exercise. (Hint: To prove (6), start with the definition of conditional variance in the form that (conditional) variance is equal to (conditional) expectation of square minus the square of the (conditional) expectation and then use the result of (5) above.) □

Theorem 10 (**Conditional mean and correlation**)**.** If conditional mean of Y does not depend on X, then X and Y are uncorrelated. Thus,

$$\text{If } \mathrm{E}(Y|x) = \mathrm{E}(Y) = \mu_X \text{ then } \sigma_{XY} = \rho_{XY} = 0$$

Proof. (Sketch) First create two new variables $Y^* = Y - \mu_Y$ and $X^* = X - \mu_X$ and show that $cov(Y^*X^*) = cov(YX)$. Next, prove the statement above for X^* and Y^*. Lets prove the second part first: Observe that $cov(Y^*X^*) = E[(Y^* - \mu_Y^*)(X^* - \mu_X^*)] = E(X^*Y^*)$ because $\mu_Y^* = \mu_X^* = 0$. Next, by law of iterated expectations $E(Y^*X^*) = E_{X^*}[E(Y^*|X^*)X^*]$. But $E(Y^*|X^*) = E(Y^*)$ by hypothesis and $E(Y^*) = 0$. Hence, $E_{X^*}(0.X^*) = E_{X^*}(0) = 0$ and so $cov(Y^*X^*) = 0$. Since covariance is zero, hence correlation is also zero. Finally, the first part can be proved by direct

[3] One exception is the case when X and Y are normally distributed. In that case, $\text{cov}(X, Y) = 0$ does imply independence of X and Y. We will see this later when we review the normal and other continuous distributions.

substitution: $cov(Y^*X^*) = E[(Y^* - \mu_Y^*)(X^* - \mu_X^*)] = E(Y^*X^*) = E[(Y - \mu_Y)(X - \mu_X)] = cov(YX)$ □

Once again, the converse is not true, i.e., it is not necessarily true that if $\text{cov}(X, Y) = 0$ then $\text{E}(Y|x) = \text{E}(Y)$.

Finally, we can collect various rules so far in one theorem.

Theorem 11. Let X,Y and Z be three random variables with expectations, μ_X, μ_Y and μ_Z. Similarly, let σ_X^2, σ_Y^2 and σ_Z^2 be the variances of the these random variables and $\sigma_{XY}, \sigma_{XZ}$and σ_{YZ} be the pairwise covariances among them. Finally, let a, b and c be three constants. Then,

(1) $E(a + bX + cY) = a + b\mu_X + c\mu_Y$
(2) $\text{var}(a + bY) = b^2\sigma_Y^2$
(3) $\text{var}(aX + bY) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$
(4) $\text{E}(Y^2) = \sigma_Y^2 + \mu_Y^2$
(5) $\text{cov}(a + bX + cZ, Y) = b\sigma_{XY} + c\sigma_{ZY}$
(6) $\text{E}(XY) = \sigma_{XY} + \mu_X\mu_Y$

Proof. Proofs given here extensively make use of Theorem 2, 3, 4 and 5.

Proof of (1): Use the definition of expectation and theorems 2, 3, 4 and 5. Thus,

$$\begin{aligned} E(a + bX + cY) &= E(a) + E(bX) + E(cY) \\ &= a + bE(X) + cE(Y) = a + b\mu_X + c\mu_Y. \end{aligned}$$

Proof of (2): Use the definition of variance, theorems 2, 3, 4, 5 and result (1) above.

$$\begin{aligned} var(a + bY) &= E[(a + bY - E(a + bY))^2] \\ &= E[(a + bY - a - bE(Y))^2] = E[b^2(Y - E(Y))^2] = b^2E[(Y - E(Y))^2] \\ &= b^2\sigma_Y^2. \end{aligned}$$

Proof of (3): Use the definition of variance, theorems theorems 2, 3, 4, 5 and the two results above.

$$\begin{aligned} var(aX + bY) &= E\Big[\big((aX + bY) - E(aX + bY)\big)^2\Big] \\ &= E\Big[\big((aX + bY) - a\mu_X + b\mu_Y\big)^2\Big] = E\Big[\big(a(X - \mu_X) + b(Y - \mu_Y)\big)^2\Big] \\ &= E[a^2(X - \mu_X)^2] + E[b^2(Y - \mu_Y)^2] + 2E[ab(X - \mu_X)(Y - \mu_Y)] \\ &= a^2var(X) + b^2var(Y) + 2abcov(X, Y) \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}. \end{aligned}$$

Proof of (4): Start by adding and subtracting μ_Y from Y^2 and then use the definition of expectation. Expand the squares and apply theorems 2, 3, 4, 5 as well as the results above. In the last step, use the definition of variance.

$$\begin{aligned} E(Y^2) &= E\big[((Y-\mu_Y)+\mu_Y)^2\big] = E\big[((Y-\mu_Y)^2+\mu_Y^2+2\mu_Y(Y-\mu_Y)\big] \\ &= E[(Y-\mu_Y)^2]+\mu_Y^2+2\mu_Y E[Y-\mu_Y] \\ &= \sigma_Y^2+\mu_Y^2 \text{ where, note that } E[Y-\mu_Y]=0. \end{aligned}$$

Proof of (5): Start with the definition of covariance, use theorems 2, 3, 4, 5 and the previous results.

$$\begin{aligned} cov(a+bX+cZ,Y) &= E\Big[\big(a+bX+cZ-E(a+bX+cZ)\big)\big(Y-\mu_Y\big)\Big] \\ &= E\Big[\big(b(X-\mu_X)+c(Z-\mu_Z)\big)\big(Y-\mu_Y\big)\Big] \\ &= E\big[b(X-\mu_X)(Y-\mu_Y)+c(Z-\mu_Z)(Y-\mu_Y)\big] \\ &= bE[(X-\mu_X)(Y-\mu_Y)]+cE[(Z-\mu_Z)(Y-\mu_Y)] \\ &= bcov(XY)+ccov(ZY) \\ &= b\sigma_{XY}+c\sigma_{ZY}. \end{aligned}$$

Proof of (6): Similarly, first subtract and add the means of the X and Y, and then apply the definition of expectation. Use theorems 2, 3, 4, 5 and the previous results as necessary and in the last step, apply the definition of covariance and the use the fact that $E[Y-\mu_Y]=E[X-\mu_X]=0$.

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

□

5. Continuous Distributions

5.1. Normal, Standard and Bi-variate Normal Distributions

Normal Distribution. A continuous random variable X with the **normal distribution** is the usual familiar bell shaped curve. Its parameters are the mean (μ_X) and variance (σ_X^2) of the random variable and the PDF is given by

$$N(\mu_X,\sigma_X^2)=f(x)=\frac{1}{\sigma_X\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\Big) \tag{34}$$

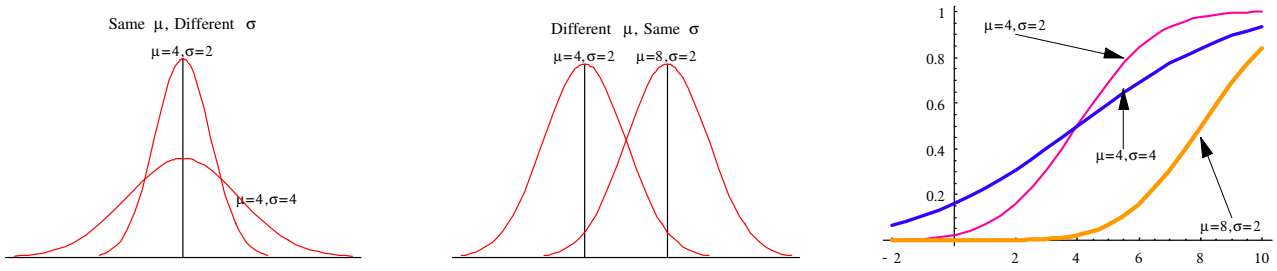


FIGURE 9. Normal Distributions, PDFs and CDFs

Because of the importance and use of this distribution, special notation is often used for this distribution. The normal distribution with parameters μ and σ^2 (where I am now suppressing the subscript X for shorthand) is written as $N(\mu, \sigma^2)$. Thus, if a random variable Y had a normal distribution with $\mu = 1$ and $\sigma^2 = 4$, we could convey this information by simply writing $Y \sim N(1, 4)$.

A normal density function with parameter μ (mean) and σ^2 (variance) is symmetric around μ and has 95% of its probability mass between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$. In general, however, it is also true that distributions that are 'approximately' bell shaped, the following empirical rule applies:

(1) Approximately 68.2% of the area under the curve lies between $\mu \pm \sigma$
(2) Approximately 95.4% of the area under the curve lies between $\mu \pm 2\sigma$
(3) Approximately 99.7% of the area under the curve lies between $\mu \pm 3\sigma$

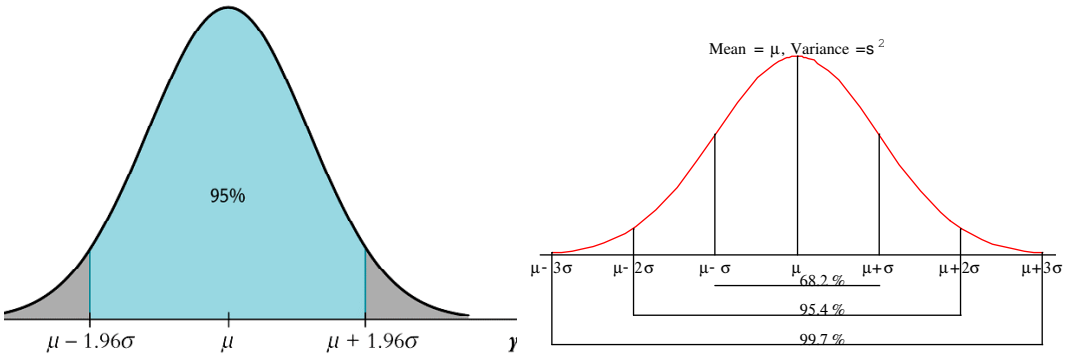


FIGURE 10. Area under a normal curve

Theorem 12 (Linear Combination of Two Normal Distributions). If X and Y are two normal distributions such that $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ and that they are independent, then for any two constants a and b if $W = aX + bY$ then

$$W \sim N[(a\mu_X + b\mu_Y), (a^2\sigma_X^2 + b^2\sigma_Y^2).$$

Thus, a linear combination of normally distributed variables is itself normally distributed and this result can be generalized to a linear combination of more than just two random variables.

Standard Normal Distribution. A special case of the normal distribution is the case when $\mu = 0$ and $\sigma = 1$. In this case it is called the **standard normal distribution**, or $N(0, 1)$. Most authors use the letter Z to denote the random variable with the standard normal distribution $N(0, 1)$. Thus,

$$N(0,1) = f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{35}$$

Similarly, the greek letter Φ is reserved for the cumulative distribution (CDF) of Z. Thus, for some number $Z = c$, $\Phi(c)$ is the probability that the standard normal variable Z is less than or equal to c. The standardized distribution can be used to compute probabilities that a normal random variable is between certain values of is less than some specific value. For instance, say Y is a random variable with normal distribution such that $Y \sim N(1, 4)$ and we want to know the probability that $Y \leq 2$. Then, we can compute this easily by first covering Y to a standard normal variable Z and then looking up the appropriate probability in the tables for the standard normal distribution (Table D.1. on page 960 in your text book). Here is how we would compute it: Let

$$Z = \frac{Y - \mu}{\sigma} \tag{36}$$

Then $Y = 2$ corresponds to $Z = \frac{2-1}{2} = .5$ To compute $\mathrm{P}(Y \leq 2)$ we need to compute $\mathrm{P}(Z \leq .5)$ i.e., $\Phi(.5)$ Now, if you look at table D.1. on page 960 of your text book, this can be read off from the table as $0.5 + .1915 = .6915$ Thus, we have the following rule.

If $Y \sim N(\mu, \sigma^2)$ and c_1 and c_2 are any two numbers such that $c_1 \leq c_2$ then

(1) $\mathrm{P}(Y \leq c_2) = \mathrm{P}(Z \leq d_2) = \Phi(d_2)$

(2) $\mathrm{P}(Y \geq c_1) = \mathrm{P}(Z \geq d_2) = 1 - \Phi(d_2)$

(3) $\mathrm{P}(c_1 \leq Y \leq c_2) = \mathrm{P}(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1)$

where $d_1 = (c_1 - \mu)/\sigma$ and $d_2 = (c_2 - \mu)/\sigma$

Bivariate Normal Distribution. The normal distribution can be extended to the joint distributions as well. In the case of two random variables, X and Y, if they have a joint distribution which is also normal, it is called the **bivariate normal distribution**. Where as a univariate normal distribution has just two parameters (μ and σ), a bivariate distribution has five parameters. Thus, if X and Y have a bivariate normal distribution then the parameters of the distribution are

$\mu_X, \sigma_X, \mu_Y, \sigma_Y$ and ρ_{XY}. These parameters correspond to the mean and standard deviations of the two random variables and the correlation between them.

Theorem 13 (**Bivariate Distributions**). If X and Y have a bivariate normal distribution with parameters $\mu_X, \sigma_X, \mu_Y, \sigma_Y$ and ρ_{XY} then

(1) for any two constants a and b, W = aX + bY has a distribution given by

$$W \sim N[(a\mu_X + b\mu_Y), (a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})]$$

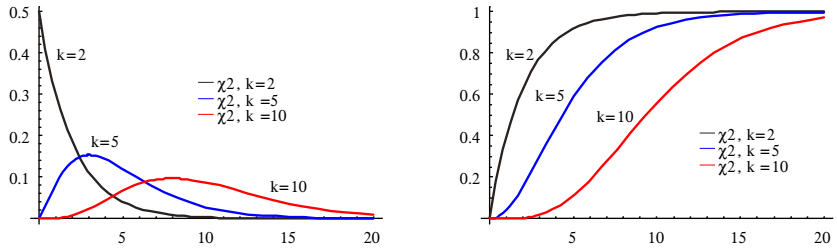
(2) $f_X(x) \sim N(\mu_X, \sigma_X)$ and $f_Y(y) \sim N(\mu_Y, \sigma_Y)$, i.e., if the joint distribution is bivariate normal, then the marginal distributions will also be normal.

(3) If $\rho_{XY} = 0$ then the variables X and Y are independently distributed

Note that in theorem 9 we noted that if two random variables X and Y are independent then, $\rho_{XY} = 0$ but that in general, if $\rho_{XY} = 0$ then it does not imply that the random variables are independently distributed. However, in the case of bivariate normal distributions, it is true that if $\rho_{XY} = 0$ then the random variables are independently distributed.

5.2. The χ^2, F and t Distributions

Three other important continuous distributions that we will be using extensively in the rest of the course are the χ^2, F and t distributions and will mostly be used for the purpose of hypotheses testing.



χ^2 PDF χ^2 CDF

FIGURE 11. χ^2 Distributions with k degrees of freedom

χ^2 Distribution. A χ^2 distribution (pronounced chi-square) is the sum of k squared *independent* standard normal distributions. The parameter of the distribution is k, i.e., the number of Z^2 distributions added up and is called the degrees of freedom of the chi-squared distribution. Thus, if

$Z_1, Z_2, \ldots, Z_k$ are k independent standardized normal variables (i.e., $Z_i \sim N(0,1)$ for $i = 1, 2, \ldots k$) and

$$Y = \sum_{i=1}^{k} Z_i^2 \quad \text{then} \quad Y \sim \chi^2(k). \tag{37}$$

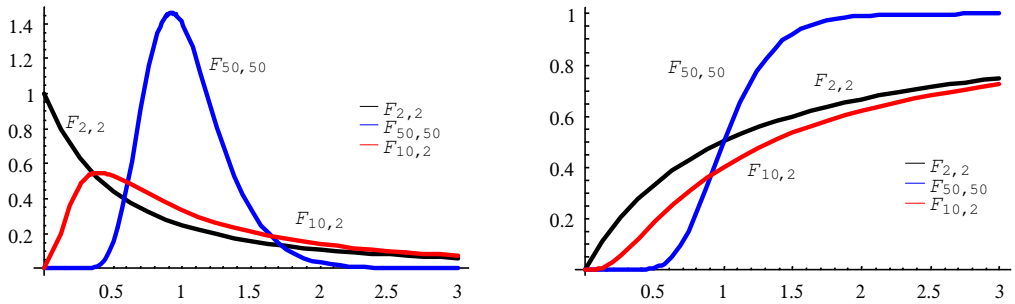
i.e, Y has a χ^2 distribution with k degrees of freedom and is written as $Y \sim \chi^2(k)$. Figure 11 shows the PDF and CDF for the case when $k = 2, 5,$ and 10. Note the following properties.

(1) A random variable with a χ^2 distribution only takes on positive values.
(2) A χ^2 distribution is skewed to the right but for large values of k it becomes more symmetrical.
(3) The mean and variance of the χ^2 distribution are k and $2k$ respectively.
(4) If Y_1 and Y_2 are two independent χ^2 variables with k_1 and k_2 degrees of freedom then $Y_1 + Y_2$ is also a χ^2 variable with $k_1 + k_2$ degrees of freedom.

Example 19. Suppose that a random variable Y is such that $Y \sim \chi^2(20)$, i.e. it has a χ^2 distribution with 20 degrees of freedom. Then, as table D.4 in your text book shows, the probability that $Y > 10.85$ is 0.950. Similarly, the probability that $Y > 39.997$ is 0.005.

F Distribution. If Y_1 and Y_2 are two independent χ^2 variables with k_1 and k_2 degrees of freedom respectively, and you define a new variable X such that

$$X = \frac{Y_1/k_1}{Y_2/k_2} \quad \text{then} \quad X \sim F_{k_1 k_2} \tag{38}$$



F Distribution PDF F Distribution CDF

FIGURE 12. F Distributions with k1, k2 degrees of freedom

i.e., X has what is called the (Fisher's) F-distribution with parameters k_1 and k_2 (and written as $F_{k_1 k_2}$). These parameters are called the numerator and denominator degrees of freedom respectively. Figure 12 shows the PDF and CDF for the case when (1) $k_1 = 2, k_2 = 2$; (2) $k_1 = 50, k_2 = 50$; and (3) $k_1 = 10, k_2 = 2$. Note the following properties.

(1) A random variable with a $F_{k_1k_2}$ distribution only takes on positive values.

(2) Like the χ^2 distribution, $F_{k_1k_2}$ is also right skewed but as k_1 and k_2 become large, the $F_{k_1k_2}$ distribution approaches the normal distribution.

(3) The mean of a F distributed variable is $k_2/(k_2-2)$ and is defined for $k_2 > 2$ and the variance is

$$\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$$

and is defined for $k_2 > 4$.

(4) If $Z_1, Z_2, \ldots, Z_{k_1}$ are k_1 standard normal variables, then $(Z_1^2 + Z_2^2 + \ldots + Z_{k_1}^2)/k_1$ has a $F_{k_1\infty}$ distribution.

(5) Equivalently, a χ^2 random variable with k_1 degrees of freedom divided by k_1 has a $F_{k_1\infty}$ distribution, i.e., if $Y \sim \chi^2(k_1)$ then $Y/k_1 \sim F_{k_1\infty}$.

(6) Alternatively, we can reexpress the last two items as saying that if the denominator degrees of freedom in k_2 is fairly large, then the following relationship holds between the χ^2 and F distributions:

$$k_1 F_{k_1\infty} = \chi^2(k_1). \tag{39}$$

Example 20. If $Y \sim F_{6,9}$ i.e. the random variable Y has a F distribution with $k_1 = 6$ and $k_2 = 9$ degrees of freedom, then what is the probability that we will observe a value of Y such that (i) $Y \geq 1.61$, (ii) $Y \geq 3.37$ and (iii) $Y \geq 5.80$? To compute the answers, we can look at table D.3 (p.962) of your text book. These correspond to probabilities of .25, .05 and .01 respectively.

Example 21. (1) In table D.3 of your text book, look up the critical F value for probability p= .01 when $k_1 = 10$ and $k_2 = \infty$. (2) Next, look up the critical chi^2 value corresponding to p = .01 and degrees of freedom equal to 10. (3) Finally, compare the two answers in light of the last property stated for the F distributions. Answer: (1) 2.32; (2) 23.2093; (3) per the last property given for the F distributions, we know that if k_2 is large (here ∞), then the F value times k_1 should be about the same as the value from the χ^2 with k_1 degrees of freedom. Thus, answer in (1) multiplied by 10 should be about the same as the answer in (2), which it is!!

t Distribution. If Z and X are two independent random numbers such that Z has a standard normal distribution (N(0,1)) and X is a Chi-square random variable with k degrees of freedom and is independently distributed from Z (ie., $Z \sim N[0,1]$ and $X \sim \chi^2(k)$ and independent from Z),

then the random variable Y defined as

$$Y = \frac{Z}{\sqrt{X/k}} \sim t_k \tag{40}$$

has a students t-distribution with k degrees of freedom.

The PDF of the student t distribution has a bell shape curve (like the normal distribution) but has more mass in the tails (ie., it is a fatter bell shaped curve). The difference in t distribution and the standard normal distribution is for low values of k (the lower the value of k the fatter the t-distribution), and disappears for large values of k. Infact, t_∞ is equal to the normal distribution. Figure 13 shows the PDF for the case when (1) $k = 5$(2) $k = 120$ which can be regarded as the normal distribution. Note the following properties:

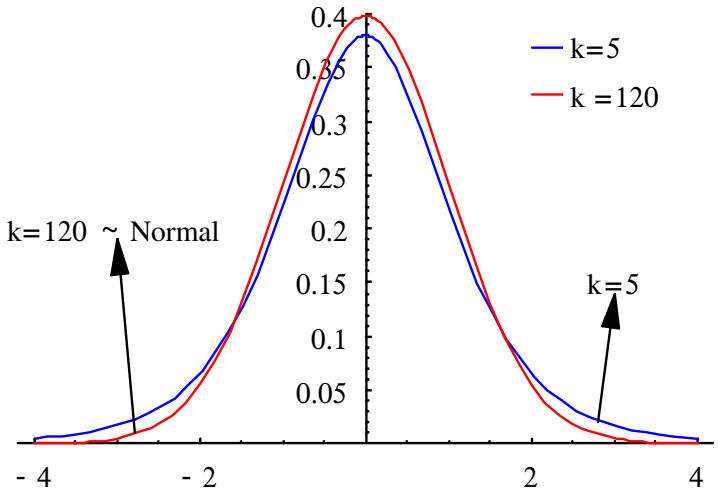


FIGURE 13. t Distributions with k=5, k=120 degrees of freedom

(1) The t-distribution is symmetrical like the normal distribution but has more mass in the tails.
(2) A k becomes large, it approximates the standard normal distribution
(3) The mean is zero and variance is $k/(k-2)$ and in the limit is equal to 1.

Example 22. If Y is a random variable with t_{10} distribution, what is the probability that we will observe a value of Y which is equal to or larger than (i) 1.812,(ii) 2.764? The answers are .05 and .01, which can be found by looking up in table D.2 in the text book.

Example 23. Suppose that Y had a t_{500} distribution. What is the probability that we will observe a value of Y that is 1.96 or larger? From table D2., the answer is .025. Since the degrees of freedom are large, we can also look up the answer in the Z tables (table D1). For Y to be as large as 1.96, the probability is 1-(.5+.4750)=.025.

6. Sampling and Sampling Distributions

If we know the underlying probability distribution function, along with the parameters, then we know the mean and variance etc. Eg. Income in a hypothetical population is distributed as $N[20K, 5K]$ then we know that the *population* mean and variance are, $\mu_Y = 20K, \sigma_Y^2 = 5K$, where Y is the value of the random variable income.

Suppose instead that we took 1 *sample* of 20 people from this population and found the sample average to be $\overline{Y} = 19.5K$. Can be be confident that the number 19.5 is 'pretty close' to the true population mean?

Next, suppose that we repeated the process of sampling, say 15 times, and each time computed the sample average. How would the average of the sample averages compare to the true underling population mean?

To answer these, we need to distinguish between a population and a sample and develop the following concepts:

(1) Random sampling and IID draws.
(2) Sample average itself is a random variable.
(3) Distribution function of the sample average.
(4) Mean and variance of the sample average.
(5) Law of large numbers.
(6) Central limit theorem.

First some terminology ...

Simple Random Sampling. A sampling procedure that assures that each element in the *population* has the same probability of being selected in the *sample* is referred to as simple random sampling.

Random Sample. Suppose that a population consists of exactly 1000 values of Y, $(Y_1, Y_2, \ldots, Y_{1000})$ and you randomly select 20 of these values, $Y_1, Y_2, \ldots, Y_{20}$. Because you selected them randomly, the values of these 20 observations can change from one sample to the next one. Hence, each of these Y_i is a random variable, i.e., $Y_1, Y_2, \ldots, Y_{20}$ are all random variables.

Identically Distributed. Since $Y_1, Y_2, \ldots, Y_{20}$ are drawn from the same population, the marginal distribution of Y_i is the same for each value of $i = 1, \ldots, 20$ and is equal to the marginal distribution

of Y in the population. When Y_i has the same marginal distribution for each value of $i = 1, \ldots, 20$, then $Y_1, Y_2, \ldots, Y_{20}$ are said to be identically distributed.

Independently and Identically Distributed. If knowing the value of Y_1 provides no information about the value of Y_2 then the conditional distribution of Y_2 given Y_1 is the same as the marginal distribution of Y_2, i.e., Y_2 is independently distributed of Y_1. Thus, if Y_i has the same marginal distribution for each value of $i = 1, \ldots, 20$, and each of these is independently distributed then $Y_1, Y_2, \ldots, Y_{20}$ independently and identically distributed (called iid).
Simple random sampling results is iid draws of $Y_1, Y_2, \ldots, Y_{20}$

Definition 20 (**Sample Average**)**.** If a sample consists of n observations, $Y_1, Y_2, \ldots, Y_n$, then the **Simple Average**, denoted $\overline{Y}$ is

$$\overline{Y} = \frac{1}{n}\sum_{i}^{n} Y_i. \tag{41}$$

Similarly,

Definition 21 (**Sample Variance**)**.** If a sample consists of n observations, $Y_1, Y_2, \ldots, Y_n$, then the **Simple Variance** denoted s^2 is,

$$s^2 = \frac{1}{n-1}\sum_{i}^{n}(Y_i - \overline{Y})^2. \tag{42}$$

Note that both, the sample average and the sample variance, are themselves random variables. Each will have its own PDF and CDF as well as the expected values and variance. The next theorem provide the expected value and variance of the the sample average when the expected value and variance of the underling population is known and the draws are iid.

Theorem 14 (Mean and Variance of $\overline{Y}$)**.** Let $Y_1, Y_2, \ldots, Y_n$, be iid draws from a population such that the mean and variance of Y_i are μ_Y and σ_Y^2. Then,

(1) $E(\overline{Y}) = \mu_Y$
(2) $var(\overline{Y}) = \frac{\sigma_Y^2}{n}$ and $sd.(\overline{Y}) = \frac{\sigma_Y}{\sqrt{n}}$

Proof. (for 1) Start with the definition of a sample average, take the sum of expected values of random variables and then use the fact the these are iid draws and hence $E(Y_i) = \mu_Y$

$$E(\overline{Y}) = E(\frac{Y_1 + Y_2 + \ldots + Y_n}{n}) = \frac{1}{n}E(Y_1 + Y_2 + \ldots + Y_n)$$
$$= \frac{1}{n}(E(Y_1) + E(Y_2) + \ldots + E(Y_n)) = \frac{1}{n}\sum_{i}^{n} E(Y_i) = \mu_Y.$$

For (2): Observe that since Y_i are iid, then $cov(Y_i, Y_j)$ is zero between any two pairs. Next use item (3) in Theorem 11 where $a = b = 1/n$. Then,

$$var(\overline{Y}) = var(\frac{1}{n}\sum_i Y_i) = var(Y_1/n + Y_2/n + \ldots + Y_i/n)$$
$$= var(aY_1 + aY_2 + \ldots + aY_i) = a^2 var(Y_1 + Y_2 + \ldots + Y_n) = a^2(n\sigma_Y^2) = \sigma_Y^2/n.$$

□

Note two things about this theorem. First, it tells us that on average, the sample average will be equal to the average of the population average (i.e., the expected value of the sample average is equal to the population average) and two, that the variance of the sample average is equal to the variance of the population divided by the sample size. Thus, any one value of the sample average will not be equal to the population average, (only its expectation is equal to the population average) but very importantly, the variance of the sample average will decrease as the sample size increases (ie., as sample size increases). This begs the question as to what happens to the sample average if the sample size becomes indefinitely large? Intuitively, if n gets very large, the variance of the sample average will go to zero and the sample average itself will get very close to the population average. These idea are made more precise by the concept of consistency and the law of large numbers.

Definition 22 (**Convergence in Probability**)**.** Let $S_1, S_2, \ldots, S_n, \ldots$ be a sequence of random variables indexed by the sample size. Then, the sequence $\{S_n\}$ is said to **converge in probability** to limit μ (written as $S_n \xrightarrow{p} \mu$) if the probability that S_n is within $\pm\delta$ of μ tends to one as $n \to \infty$ for every $\delta > 0$. That is

$$S_n \xrightarrow{p} \mu \text{ if and only if } \lim_{n\to\infty} Pr[|S_n - \mu| < \delta] = 1. \tag{43}$$

Definition 23 (**Consistency**)**.** If $S_n \xrightarrow{p} \mu$ then S_n is a consistent estimator of μ.

In the definitions above, think of $S_1, S_2 \ldots$ as sample averages $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_n$ indexed by the sample size.

Theorem 15 (**Law of Large Numbers**)**.** Let $Y_1, Y_2, \ldots, Y_n$ be iid draws and $E(Y_i) = \mu_Y$ and $var(Y_i) < \infty$ then

$$\overline{Y}_n \xrightarrow{p} \mu_Y. \tag{44}$$

The law says that as the sample size increases, the sampling distribution of $\overline{Y}$ concentrates around the population means μ_Y. Further, as the sample size increases, the variance of $\overline{Y}$ decreases and

that the probability that $\overline{Y}$ falls outside of a small interval $\pm\delta$ of the population mean goes to zero. This is most easily seen by observing that in theorem 14, the variance of $\overline{Y}$ is $\frac{\sigma_Y^2}{n}$ which goes to zero as n becomes large.

Theorem 16 (Distribution of $\overline{Y}$). Let $Y_1, Y_2, \ldots, Y_n$ be iid draws from $N[\mu_Y, \sigma_Y^2]$. Then,

$$\overline{Y} \sim N[\mu_Y, \sigma_Y^2/n]. \tag{45}$$

Proof. Per Theorem 12, sum of independent and normal distributions is itself a normal distribution. Hence, sum of $Y_1, Y_2, \ldots, Y_n$ will be normally distributed. Further, since $Y_1, Y_2, \ldots, Y_n$ are iid draws from $\sim N[\mu_Y, \sigma_Y^2]$ then per theorem 14, the mean and variance of $(\overline{Y})$ is μ_Y and σ_Y^2/n.

□

Theorem 16 states that if $Y_1, Y_2, \ldots, Y_n$ were iid draws from $N[\mu_Y, \sigma_Y^2]$. then the exact distribution of $\overline{Y}$ is $N[\mu_Y, \sigma_Y^2/n]$. However, the central limit theorem (stated formally later) states that even if $Y_1, Y_2, \ldots, Y_n$ were not drawn from a normal distribution, then as the sample size gets large, the distribution of $\overline{Y}$ approximates a normal distribution. Specifically, under very general conditions, the standardized sample average converges in distribution to a normal random variable. To formalize this notion, we need to develop the concept of *convergence in distribution* (which is different from the concept of *convergence in probability*)

Convergence in Distribution. The basic idea of convergence in distribution is when in some limit the CDF of a sequence of random numbers converges to some CDF. For instance, recall that

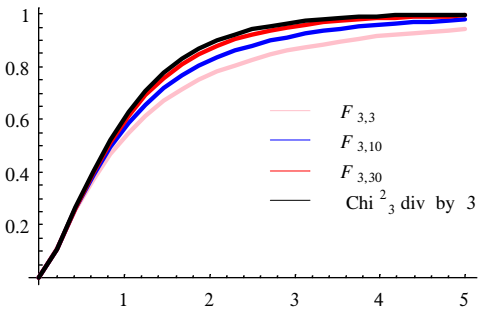


Figure 14. CDF of $\chi_3^2/3$ and of F_3, k where $k = 5, 10, 30$

earlier when discussing the properties of a F distribution, we said the if the denominator degrees of freedom were vary large (or ∞) while the numerator degrees of freedom were some fixed number,say 3, then the random numbers drawn from this distribution would be the same as if we drew random numbers from a χ^2 distribution and divided them by 3. More precisely, the distribution functions,

or the CDFs would be the same. To understand the idea of convergence in distribution, consider the of CDF random numbers drawn from a F distribution with 3 degrees of freedom in numerator and k degrees of freedom in the denominator. Now start increasing the denominator degrees of freedom, from say $k = 3, k = 10, k = 30$, and so on. As k gets larger, the CDF would start converging in distribution to the CDF of random numbers drawn from a χ^2 distribution with 3 degrees of freedom where each random variable is divided by 3. This is shown graphically below.

Definition 24 (**Convergence in Distribution**). Let $S_1, S_2, \ldots, S_n, \ldots$ be a sequence of random numbers indexed by the sample size and each with a cumulative distribution function $F_1(S_1), F_2(S_2), \ldots, F_n(S_n), \ldots$. Then the sequence of random variables $\{S_n\}$ is said to **converge in distribution** to S (denoted $S_n \xrightarrow{d} S$) if the distribution functions $\{F_n(S_n)\}$ converge to F(S), the distribution of S. Thus.

$$S_n \xrightarrow{d} S \text{ if and only if } \lim_{n\to\infty} F_n(S_n) = F(S) \tag{46}$$

where the limit holds at all points S at which the limiting distribution F is continuous.

To make the definition above a little easier to understand, think of $S_1, S_2, \ldots, S_n, \ldots$ as standardized sample averages. Thus, S_1 is $\sqrt{n}(\overline{Y}_1 - \mu_Y)/\sigma_Y$ and is a random variable which has a cumulative distribution function. Call it $F_1(S_1)$. Similarly, S_2 is also a standardized sample average from a different sample size. This too is a random number and has a cumulative distribution. Call it $F_2(S_2)$. And so on. Then the definition above says that we say that the sequence S_n converges in distribution to S, if the cumulative distribution $F_n(S_n)$ becomes closer and closer to a limiting distribution $F(S)$ as n increases, i.e., $\lim_{n\to\infty} |F_n(S_n) - F(S)| = 0$ at all (continuity) points of F(S). Also, compare convergence in probability with convergence in distribution. If $S_n \xrightarrow{p} \mu$ then as n increases, S_n becomes close to μ with high probability while if $S_n \xrightarrow{d} S$ then the distribution of S_n becomes close to the distribution of S.

Theorem 17 (**Central Limit Theorem**). Let $Y_1, Y_2, \ldots, Y_n$ be iid with $E(Y_i) = \mu$ and $var(Y_i) = \sigma_Y^2$ such that $0 < \sigma_Y^2 < \infty$. Then, as $n \to \infty$ the distribution

$$\frac{\sqrt{n}(\overline{Y} - \mu_Y)}{\sigma_Y} \xrightarrow{d} N(0, 1) \tag{47}$$

Roughly, the central limit theorem states that the distribution of the sum of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

Lets see the Central Limit Theorem in action

Example 24. Let X be a random variable such that it is drawn from a continuous uniform distribution between 0 and 1, i.e., let $X \sim$ UniformContinious[0,1] (the PDF is shown in figure below). Then do the following:

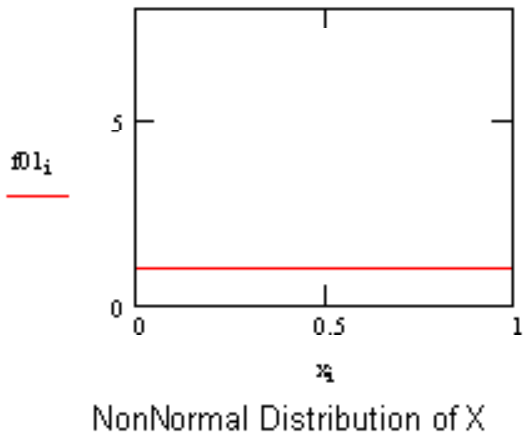
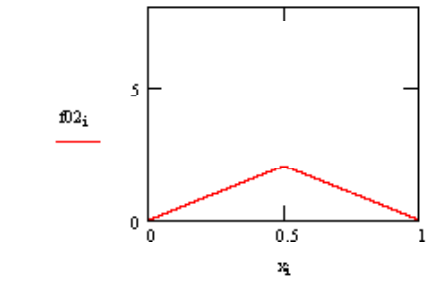


Figure 15. Parent Distribution, Uniform[0,1]

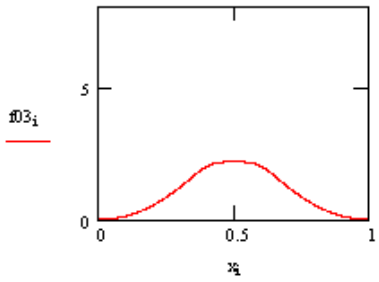
(1) Draw a sample of size 2. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_2$.
(2) Draw a sample of size 3. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_3$.
(3) Draw a sample of size 4. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_4$.
(4) Draw a sample of size 8. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_8$.
(5) Draw a sample of size 16. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_{16}$.
(6) Draw a sample of size 32. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_{32}$.

The PDFS for the cases above are given below.



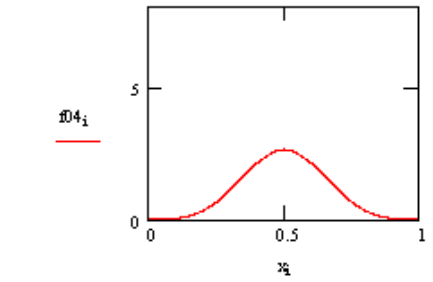
Distribution of Xbar when sample size is 2

PDF of $\overline{X}_2$



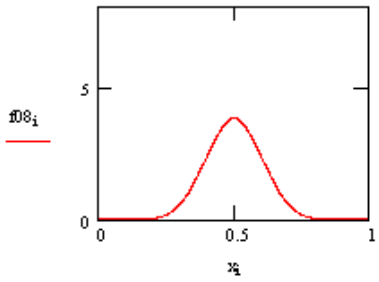
Distribution of Xbar when sample size is 3

PDF of $\overline{X}_3$



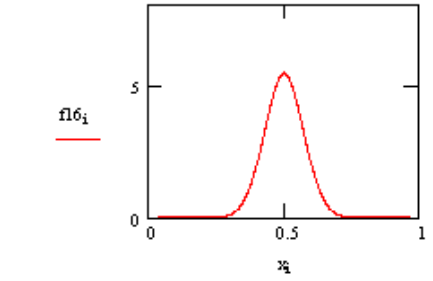
Distribution of Xbar when sample size is 4

PDF of $\overline{X}_4$



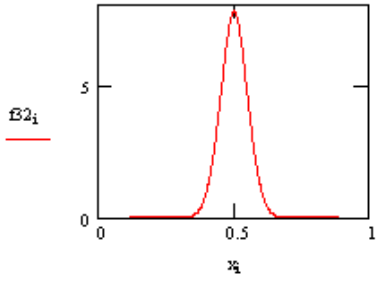
Distribution of Xbar when sample size is 8

PDF of $\overline{X}_8$



Distribution of Xbar when sample size is 16

PDF of $\overline{X}_{16}$



Distribution of Xbar when sample size is 32

PDF of $\overline{X}_{32}$

FIGURE 16. PDF of Sample Averages - CLT

7. Estimators

Definition 25 (Estimator)**.** Let θ be a population parameter. Then, $\widehat{\theta}$ is an estimator of θ if it is a function of the sample data and does not depend on the population parameter θ. Thus,

$$\widehat{\theta} = g(Y_1, Y_2, Y_3 \ldots, Y_n) \tag{48}$$

The numerical value of the estimator is an estimate of the population parameter.

Example 25. Let the mean value of annual income of all employed women in a given population be μ_f. Then, μ_f is a population parameter. If you collected data on a sample of size n and used this data to compute the sample average,

$$\overline{Y} = \frac{1}{n}\sum_{i}^{n}(Y_1 + Y_2 + Y_3 + \ldots + Y_n)$$

then the sample average would be an estimator of the population parameter. In this example, the population parameter θ is μ_f, the estimator $\widehat{\theta}$ is the sample average $\overline{Y}$ and the numerical value of computed sample average is an estimate of the population parameter.

Other examples include cases where population parameter of interest θ is (i) the population variance, (ii) the maximum value in the population, (iii) the minimum value etc. For each of these population parameters you can construct estimators $\widehat{\theta}$ such as the sample variance, the maximum value observed in the sample or the minimum value observed in the sample.

7.1. Properties of Estimators

In particular note that for the same population parameter, we may have more than 1 estimator available for it. For instance, let θ be the true population average μ, i.e, $\theta = \mu$. Then you can imagine constructing three different estimators $\widehat{\theta}$ for the population parameter θ. Let these be $\widehat{\theta_1}, \widehat{\theta_2}, \widehat{\theta_3}$ where

(1) $\widehat{\theta_1}$ is just the sample average $\overline{Y}$.
(2) $\widehat{\theta_2}$ is the *first* observation in a sample, and
(3) $\widehat{\theta_3}$ is the weighted average of n even number of observations where each odd observation is weighted 1/2 and each even observation is weighted 3/2, (ie. $\widehat{\theta_3} = (1/n)[(1/2)Y_1 + (3/2)Y_2 + (1/2)Y_3 + (3/2)Y_4 + \ldots + (1/2)Y_{n-1} + (3/2)Y_n]$)

The obvious question is which of these estimators is a good one? Which one should we use?

Since we are interested in knowing the value of the true population parameter (i.e., the population mean) based on a a sample, intuitively, it seems that the first estimator $\widehat{\theta}_1$ is in some sense better than the other two. For instance, in the second estimator we are not using all the available information (i.e. different values in the sample) but instead only one observation. In the third estimator, we are applying weights to the observations in some arbitrary fashion. We compare estimators based on three criteria: (1) Unbiasedness, (2) Consistency and (3) Efficiency.

Unbiasedness. If we come up with an estimator for some population parameter, then we would like the estimator to be such that if we repeated the exercise of sampling many times over and used the same estimator, then the average value of the estimator, i.e., its expected value, should be the same as the value of the underling population parameter. In other words, we would like the estimator to be not biased.

Definition 26. (**Unbiasedness**) An estimator $\widehat{\theta}$ is an unbiased estimator of θ if

$$E(\widehat{\theta}) = \theta \tag{49}$$

and if the estimator is biased then the amount of bias can be measured as

$$B = Bias = E(\widehat{\theta}) - \theta \tag{50}$$

Example 26. Suppose that the population parameter of interest is the population mean, i.e., $\theta = \mu$. Then if we construct as estimator $\widehat{\theta}$ so that it is just the sample average, i.e., $\widehat{\theta} = \overline{Y}$, then is this estimator unbiased? If the sample was drawn randomly, then the values of Y, $Y_1, Y_2, \ldots, Y_n$, will be be iid draws from a population and hence by theorem 14, $E(\overline{Y}) = \mu_Y$. Thus, the estimator $\widehat{\theta} = \overline{Y}$ is unbiased.

Consistency. Another desirable property that we would like for an estimator is that if we increased the sample size, then the value of the estimator should start getting close to the value of the population parameter. In other words, if we increased the sample size, then we would like the our estimate to converge to the value of the population parameter. Simply put, we want our estimator to be consistent. We have already seen the definition of consistency (recall convergence in probability). Thus,

Definition 27. (**Consistent**) An estimator $\widehat{\theta}$ is a consistent estimator of θ if

$$\widehat{\theta} \xrightarrow{p} \theta \tag{51}$$

Example 27. Suppose that the population parameter of interest is the population mean, i.e., $\theta = \mu$. Then if we construct as estimator $\widehat{\theta}$ so that it is just the sample average, i.e., $\widehat{\theta} = \overline{Y}$, then is this estimator consistent? If the sample was drawn randomly, then the values of Y, $Y_1, Y_2, \ldots, Y_n$, will be be iid draws and $E(Y_i) = \mu$ and as long as $var(Y_i) < \infty$, then from the law of large numbers (theorem 15) the estimator $\widehat{\theta} = \overline{Y}$ is consistent.

Efficiency. Suppose that you have two candidates for an estimator of θ (say $\widehat{\theta_1}$ and $\widehat{\theta_2}$) and suppose that both are unbiased and consistent. How do we choose between them? One way would be to look at the distributions of $\widehat{\theta_1}$ and $\widehat{\theta_2}$ and pick the one which has a smaller spread around its mean, ie., one with a smaller variance. The estimator with the smaller variance would be called more efficient compared to the other one. This so, because it uses the data in the sample more efficiently. Thus,

Definition 28. (**Efficiency**) Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be two estimators of a population parameter θ and suppose that both are unbiased (ie, $E(\widehat{\theta_1}) = E(\widehat{\theta_2}) = \theta$). Then $\widehat{\theta_1}$ is more efficient relative to $\widehat{\theta_2}$ if

$$var(\widehat{\theta_1}) < var(\widehat{\theta_2}) \tag{52}$$

Example 28. Suppose that the population parameter of interest is the population mean, i.e., $\theta = \mu$ and suppose that we construct two estimators $\widehat{\theta_1} = \overline{Y}$ and $\widehat{\theta_2}$ in which the observations are weighted alternatively as (1/3) and (2/3), i.e. $\widehat{\theta_2} = (1/3)Y_1 + (2/3)Y_2 + (1/3)Y_3 + (2/3)Y_4 + \ldots.)$. Then in order to compare the efficiency of the two estimators, we need to compare the variance of the two, ie., $var(\widehat{\theta_1})$ with $var(\widehat{\theta_2})$.

Example 29. Suppose that the population parameter of interest is the population mean μ_Y, ie, $\theta = \mu_Y$. If we collect a random sample from the population so that the draws are iid and construct three estimators for θ given below, then based on unbiasedness, consistency and efficiency, which of these three should we use?

(1) $\widehat{\theta_1} = \overline{Y}$.
(2) $\widehat{\theta_2} = Y_1$ (ie, take the *first* observation in the sample as an estimate of population mean), and
(3) $\widehat{\theta_3}$ is the weighted average of n even number of observations where each odd observation is weighted 1/2 and each even observation is weighted 3/2, (ie. $\widehat{\theta_3} = (1/n)[(1/2)Y_1 + (3/2)Y_2 + (1/2)Y_3 + (3/2)Y_4 + \ldots + (1/2)Y_{n-1} + (3/2)Y_n]$)

Biased? Lets check the unbiasedness of the three estimators

(1) $\widehat{\theta_1} = \overline{Y}$: Since the draws are iid, then by theorem 14, $E(\overline{Y}) = \mu_Y$ and hence it is unbiased.

(2) $\widehat{\theta}_2 = Y_1$: Again, since the draws are iid, then $E(Y_i) = \mu_Y$ for all i and hence also for i=1, $E(Y_1) = \mu_Y$. So, this estimator is also unbiased.

(3) $\widehat{\theta}_3 =$ as given earlier : For this one lets explicitly compute the expected value. Then,

$$\begin{aligned}
E(\widehat{\theta}_3) &= E\Big[\frac{1}{n}\big((1/2)Y_1 + (3/2)Y_2 + (1/2)Y_3 + (3/2)Y_4 + \ldots + (1/2)Y_{n-1} + (3/2)Y_n\big)\Big] \\
&= \frac{1}{n}E\Big[\big((1/2)Y_1 + (3/2)Y_2 + (1/2)Y_3 + (3/2)Y_4 + \ldots + (1/2)Y_{n-1} + (3/2)Y_n\big)\Big] \\
&= \frac{1}{n}\Big[\frac{1}{2}E(Y_1) + \frac{3}{2}E(Y_2) + \frac{1}{2}E(Y_3) + \frac{3}{2}E(Y_4) + \ldots + \frac{1}{2}E(Y_{n-1}) + \frac{3}{2}E(Y_n)\Big] \\
&= \frac{1}{n}\Big[(\frac{1}{2}\mu_Y + \frac{3}{2}\mu_Y) + (\frac{1}{2}\mu_Y + \frac{3}{2}\mu_Y) + \ldots + (\frac{1}{2}\mu_Y + \frac{3}{2}\mu_Y)\Big] \\
&= \frac{1}{n}\Big[\frac{n(2\mu_Y)}{2}\Big] \\
&= \mu_Y.
\end{aligned}$$

So, this estimator is also unbiased.

Consistency? Lets check the consistency of the three estimators

(1) $\widehat{\theta}_1 = \overline{Y}$: Since the draws are iid, then by law of large numbers (theorem 15) $\overline{Y_n} \xrightarrow{p} \mu_Y$ and hence it is a consistent estimator.

(2) $\widehat{\theta}_2 = Y_1$: For this estimator to be consistent, it would mean that as we drew larger and larger samples, then the probability that first observation takes the same value as the mean of the population approaches 1. This is clearly not true. To show this, you could argue that the variance of $\widehat{\theta}_2$ stays constant as n gets large: $var(\widehat{\theta}_2) = var(Y_1)$ and since these are iid draws, $var(Y_1) = var(Y_i) = \sigma_Y^2$. Since this is not a function of n, the variance does not approach zero as n increases. Thus, this estimator is not consistent.

(3) $\widehat{\theta}_3 =$ as given earlier : To check for the consistency of this estimator, we can first compute its variance. The variance of this estimator turns out to be $1.25\sigma_Y^2/n$ (shown below). Because $var(\widehat{\theta}_3) \longrightarrow 0$ as $n \longrightarrow \infty$ hence $\widehat{\theta}_3$ is a consistent estimator.

$$\begin{aligned}
var(\widehat{\theta}_3) &= var\Big[\frac{1}{n}\big(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \ldots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\big)\Big] \\
&= \frac{1}{n^2}\Big[\big(\frac{1}{4}var(Y_1) + \frac{9}{4}var(Y_2) + \ldots + \frac{1}{4}var(Y_{n-1}) + \frac{9}{4}var(Y_n)\big)\Big] \\
&\quad \text{but } var(Y_1) = var(Y_2) = var(Y_3) = \ldots = var(Y_n) = var(Y_i) = \sigma_Y^2, \text{ and so} \\
var(\widehat{\theta}_3) &= \frac{1}{n^2}\Big[\frac{n}{4\times 2}var(Y_i) + \frac{9n}{4\times 2}var(Y_i)\Big] = \frac{5}{4n}var(Y_i) = 1.25\sigma_Y^2/n.
\end{aligned}$$

So $\widehat{\theta}_1$ and $\widehat{\theta}_3$ are consistent estimators but $\widehat{\theta}_2$ is not.

Efficiency? Lets check the relative efficiency of the three estimators (relative to each other) even though we know that the second estimator is not even consistent. To do this we need the variance of the three estimators.

(1) $\widehat{\theta}_1 = \overline{Y}$: Since the draws are iid, then by theorem 14, $var(\widehat{\theta}_1) = \sigma_Y^2/n$

(2) $\widehat{\theta}_2 = Y_1$: Similarly, $var(\widehat{\theta}_2) = \sigma_Y^2$

(3) $\widehat{\theta}_3 =$ as given earlier : We already calculated the variance of this estimator as $var(\widehat{\theta}_3) = 1.25\sigma_Y^2/n$.

- For $n \geq 2$ the variance of $\widehat{\theta}_1$ and $\widehat{\theta}_3$ is less than the variance of $\widehat{\theta}_2$. Hence, for $n \geq 2$ $\widehat{\theta}_1$ and $\widehat{\theta}_3$ are relatively more efficient than $\widehat{\theta}_2$.
- For all values of $n, 1.25\sigma_Y^2/n > \sigma_Y^2/n$, i.e., $var(\widehat{\theta}_1) < var(\widehat{\theta}_3)$. Hence $\widehat{\theta}_1$ is efficient relative to $\widehat{\theta}_3$

Conclusion: For the three estimators above,

- all three are unbiased,
- the second one is not consistent, and
- the sample average estimator (the first one) is the most efficient estimator of the three.

Notice that all three estimators are weighted averages of the sample data (even the second estimator which can be thought of as a weighted average where we multiply the first observation with n and all other observations with 0, i.e. $\widehat{\theta}_2 = Y_1 = \frac{1}{n}(nY_1 + 0.Y_1 + \ldots + 0.Y_n)$.) The conclusions that we reached here can be generalized to the class of all unbiased and weighted averages of the data ...

Theorem 18 (**Efficiency of** $\overline{Y}$)**.** Let $\widehat{\theta}$ be an estimator of the population mean ($\theta = \mu$) where $\widehat{\theta} = \overline{Y}$ (ie, the estimator is the sample average). Now let $\widehat{\theta^*}$ be any other estimator of the population mean such that it is some weighted average of the data, ie., $\widehat{\theta^*} = \frac{1}{n}\sum_i^n a_i Y_i$ where $a_1, a_2, \ldots, a_n$ are non random constants not equal to 1. Then, <u>if</u> $\widehat{\theta^*}$ is unbiased, then $var(\widehat{\theta}) = var(\overline{Y}) < var(\widehat{\theta^*})$. That is $\widehat{\theta} = \overline{Y}$ is the most efficient estimator among <u>all</u> unbiased estimators of the population mean that are weighted averages of the sample.

The End

PS. During the rest of the semester, we will be estimating equations of the type

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

where $X_1, X_2, \ldots, X_k$ are the variables that influence Y (in some causal way) and $\beta_1, \beta_2, \ldots \beta_k$ are the true underlying population parameters.

By estimating, we mean constructing estimators for β s. Thus our estimated equations will look like

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \ldots + \widehat{\beta}_k X_{ki} + \widehat{u}_i$$

We will be concerned about the properties of these estimators ($\widehat{\beta}$s), ie, are they unbiased, consistent, efficient etc. Also, we will often be interested in various functions of population parameters (ex: $\beta_2 + \beta_3$). What are the properties of similar function of estimators, ie is $\widehat{\beta}_2 + \widehat{\beta}_3$ an unbiased, consistent and efficient estimator of $\beta_2 + \beta_3$?. Material covered in these last few lectures will be useful in answering these questions.