Introduction to Econometrics
ECO4421

Department of Economics
Florida State University

Lecture Notes on
Review of Probability & Statistics
by

Farasat A.S. Bokhari ©

fbokhari@fsu.edu

Last updated on January 24, 2006

http://mailer.fsu.edu/~fbokhari/eco4421/

Sources for Lecture Notes

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.

A detailed version of the notes, with more complete discussions and derivations etc. is also available from the course website. I recommend that you download and use the detailed version as a study guide.

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
  1. $0 \leq P(A) \leq 1$ for every A.
  2. 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.
  3. 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.

Sample Space, Probabilities & Random Variables

Examples

- **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). 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.

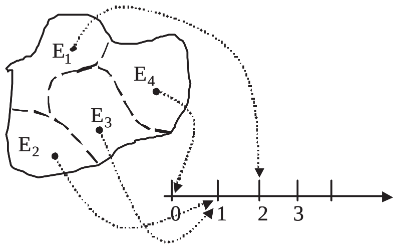


Figure 1: Tossing Two Coins

Note 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.

- **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.

Sample Space, Probabilities & Random Variables

- **Random Variables – Some Details**
  1. 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.
  2. 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.
  3. 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.
  4. Again, the converse is not true. A point in the sample space cannot map to more than one point on the real number line.
- Two Types of RVs
  1. Discrete random variables
     - Number of people at the symphony
     - Number of children in poverty
     - Number of Heads observed in the toss of 2 coins
  2. Continuous random variables
     - 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

Distribution Functions

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

- 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 $F(y) = P(Y \leq y)$. It is also known as the cumulative density function (CDF) or just the *distribution function*.

Theorem 1 (Properties of a Distribution Function). If $F(y)$ is a distribution function, then

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. $Prob(y_a < y \leq y_c) = F(y_c) - F(y_a)$

Probability Density Functions - Discrete

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

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

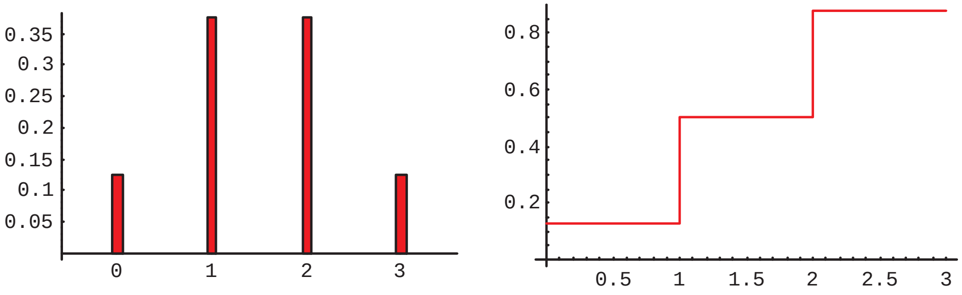


Figure 2: PDF and CDF of a Discrete Random Variable

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

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

- The probability of an event is given by the height of the PDF (eg. $P(Y = y_i) = f(y_i)$). The height of the CDF gives the probability upto and inclusive of that value (eg. $P(Y \leq y_i) = F(y_i)$).

Discrete Distributions - Uniform

- **Uniform:** 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,

$$f(y_n) = 1/n \tag{4}$$

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 is $f(y_n) = 1/6$ and the PDF and CDF are given also 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

Discrete Distributions - Bernoulli

- **Bernoulli:** A bernoulli process is one where there are only two outcomes, say $Y = 1$ or $Y = 0$ where the probabilities are p and $q = 1 - p$ respectively, i.e., the outcomes of Y and the probabilities are

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

  The PDF of a Bernoulli random variable may be written as

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

- Examples: (1)Tossing a fair coin once. probability of head = .5 and of tails = .5; (2) I take a driving test and the probability of passing is .8 and that of failing is .2.

Discrete Distributions - Binomial

- **Binomial:** 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. The PDF of a binomial random variable is

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

  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 (for instance $10!$ would be $10 \times 9 \ldots \times 2 \times 1$).

- Note the following:
  1. There are a fixed number of trials, n.
  2. The n trials are independent and repeated under identical conditions
  3. Each trial has only two outcomes: success (S) and failure (F) with probabilities p and $q = 1 - p$.
  4. For each individual trial, the probability of success is the same.
  5. 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.

- Examples: (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.

- What is the probability that we will observe exactly 2 heads if we flip a fair coin three times? Ans: $P(Y = 2) = C_{3,2}.5^2.5^{(3-2)} = 3 \times .25 \times .5 = .375$

Discrete Distributions - Binomial

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.

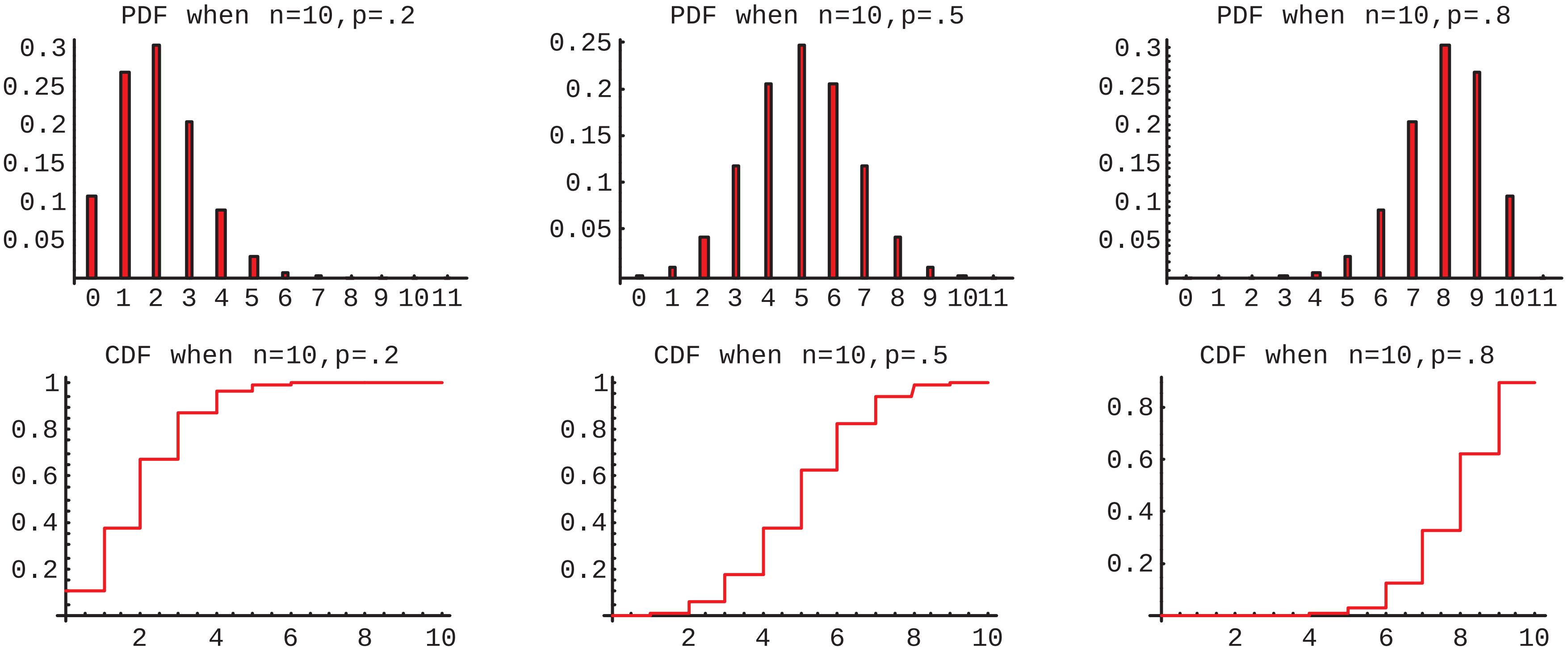


Figure 3: Binomial PDFs and CDFs

Some Discrete Distributions (PDFs)

1. **Uniform:** $f(y) = 1/n$
2. **Bernoulli:** $f(y) = p^y(1-p)^{(1-y)}$
3. **Binomial:**$f(y) = C_{n,y}p^y q^{(n-y)}$
   where $C_{n,y} = \frac{n!}{y!(n-y)!}$
4. **Geometric:**$f(y) = q^{(y-1)}p$
5. **Negative Binomial:** $f(y) = C_{(y-1),(k-1)}p^k q^{(y-k)}$
6. **Poisson Distribution:** $f(y)\frac{\lambda^y}{y!}e^{-\lambda}$

Check detailed notes (on the web) for more information on these distributions.

Probability Density Functions - Continuous

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:

$$f(y) \geq 0 \tag{8a}$$

$$\int_{-\infty}^{+\infty} f(y)dy = 1 \tag{8b}$$

$$\int_{a}^{c} f(y)dy = P(a \leq y \leq c) \tag{8c}$$

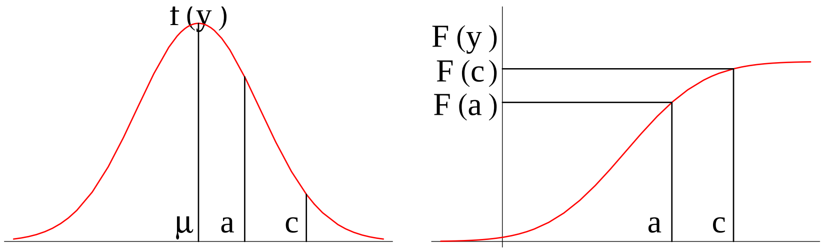


Figure 4: PDF and CDF of a Continuous Random Variable

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.

- Note the relationship between PDF and CDF of a continuous random variable. For a continuous random variable Y

$$F(y) = \int_{-\infty}^{y} f(t)dt \quad \text{and} \quad f(y) = \frac{dF(y)}{dy} \tag{9}$$

- The probability is given by the area under the PDF curve (eg. $P(a \leq y \leq c)$ is equal to the area under the curve $f(y)$ between a and c). The height of the CDF gives the probability upto and inclusive of that value (eg. $P(y \leq a) = F(a)$.)

Mean and Variance of a Random Variable

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

$$\mathsf{E}(Y) = \begin{cases} \sum_y y f(y) & \text{if Y is a discrete random variable} \\ \int_{-\infty}^{+\infty} y f(y) dy & \text{if Y is a continuous random variable} \end{cases} \tag{10}$$

- 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.

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. See detailed notes □

Mean and Variance of a Random Variable

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

$$\text{var}(Y) = \mathsf{E}[(Y - \mu_Y)^2]. \tag{11}$$

- The variance of a random variable Y, denoted var(Y) is a measure of the spread of the probability distribution.
- Note that, denoting var(Y) by σ_Y^2, $\mathsf{E}(Y)$ by μ_Y, and the PDF by $f(y)$, then definitions 7 and 8 imply that variance of Y is

$$\text{var}(Y) = \sigma^2 = \begin{cases} \sum_y (Y - \mu_Y)^2 f(y) & \text{if Y is a discrete random variable} \\ \int_{-\infty}^{+\infty} (Y - \mu_Y)^2 f(y) dy & \text{if Y is a continuous random variable} \end{cases} \tag{12}$$

- 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)).

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

$$\text{var}(Y) = E[Y^2] - (E[Y])^2 \tag{13}$$

Proof. Left as exercise □

Examples

- **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

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

- **Binomial Distribution:** Flip a fair coin three times and let Y be the number of heads observed in the three trials. Outcomes are

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

Then,

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

Moments of Discrete Distributions

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	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
Poisson	λ	λ

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

* Note also that if Y was a continuous variable than the variance would be just $(b-a)^2/12$

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.

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

$$Y = a + bX$$

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

$$E(Y) = \mu_Y = a + b\mu_X \tag{14a}$$

$$\mathsf{var}(Y) = \sigma_Y^2 = b^2\sigma_X^2 \tag{14b}$$

Proof. See detailed notes □

Linear Function of a Random Variable

- **Example:** 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

$$Y = 200 + .05X \qquad \text{then}$$

$$\mu_Y = \mathsf{E}(Y) = 200 + .05\mu_X = \$275$$

$$\sigma_Y^2 = \mathsf{var}(Y) = .05^2\sigma_X^2 = 2.5$$

$$\sigma_Y = \mathsf{std}(Y) = .05\sigma_X = \$1.58$$

Joint, Marginal & Conditional Distributions

Joint Distributions

- 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.
- 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.
- More generally, for two discrete random variables X and Y the joint probability distribution would be

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

- 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.,

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

Joint, Marginal & Conditional Distributions

Joint Distributions

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

$$\begin{aligned} f(x,y) &= \mathsf{P}(X = x \text{ and } Y = y) \qquad (15) \\ &= 0 \text{ when } X \neq x \text{ and } Y \neq y \\ \text{and } \sum_X \sum_Y f(x,y) &= 1 \end{aligned}$$

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

$$Prob(a \leq x \leq, c \leq y \leq d) = \int_a^b \int_c^d f(x,y)\, dydx \qquad (16)$$

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

Joint, Marginal & Conditional Distributions

Joint Distributions

- 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

$$E[g(x,y)] = \begin{cases} \sum_y \sum_x g(x,y)f(x,y) & \text{if X and Y are discrete random variables} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)f(x,y)dxdy & \text{if X and Y are continuous random variables} \end{cases} \tag{17}$$

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

$$F(x,y) = \begin{cases} \sum_{s \leq x} \sum_{t \leq y} f(s,t) & \text{if } X, Y \text{ discrete} \\ \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)\, dsdt & \text{if } X, Y \text{ continuous} \end{cases} \tag{18}$$

Joint, Marginal & Conditional Distributions

Examples of Joint Distributions

- **Joint PDF of 2 Dice:** 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 the table below.

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

Joint, Marginal & Conditional Distributions

Examples of Joint Distributions

- **Joint PDF of Difficulty on Exam and Passing the Exam:**Consider the case when a student 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.

Joint, Marginal & Conditional Distributions

Marginal Distributions

- Row and column sums are the marginal probabilities in the table 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

- Thus, the marginal probability distribution of a random variable is just another name for its probability distribution . . . something we have already seen before . . .
- More formally then . . .

Joint, Marginal & Conditional Distributions

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

$$f_X(x) = \begin{cases} \sum_y f(x,y) & \text{for the discrete case} \\ \int_y f(x,t)\,dt & \text{for the continuous case} \end{cases} \tag{19a}$$

$$f_Y(y) = \begin{cases} \sum_x f(x,y) & \text{for the discrete case} \\ \int_x f(s,y)\,ds & \text{for the continuous case} \end{cases} \tag{19b}$$

Table 4: Bivariate Probability Distribution

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

Joint, Marginal & Conditional Distributions

Marginal Distributions

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

$$E(X+Y) = E(X) + E(Y) \tag{20}$$

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:

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

□

Joint, Marginal & Conditional Distributions

Conditional Distributions

- 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 $f(y|x)$.
- 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., $\mathsf{P}(Y = y|X = x)$. 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 (1) for other values of X and (2) by reversing the process, i.e., compute probabilities of X for given values of Y.

Joint, Marginal & Conditional Distributions

Conditional Distributions

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

$$f(y|x) = \frac{f(x,y)}{f_X(x)} \text{ and } \quad f(x|y) = \frac{f(x,y)}{f_Y(y)}. \tag{21}$$

- These definitions provide a useful way of computing the joint distribution

$$f(x,y) = f(y|x) \cdot f_X(x) \ = f(x|y) \cdot f_Y(y) \tag{22}$$

- Example follows . . .

Joint, Marginal & 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					
			$f(y \mid x)$							
		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		$f(x \mid y)$	.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

Conditional Mean and Variance

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 $\mathsf{E}(Y|x)$, is

$$\mathsf{E}(Y|x) = \begin{cases} \int_y yf(y|x)\,dy & \text{if } y \text{ is continuous} \\ \sum_y yf(y|x) & \text{if } y \text{ is discrete} \end{cases} \tag{23}$$

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

$$\begin{aligned} \mathsf{Var}(Y|x) =& \mathsf{E}\big((Y - \mathsf{E}(Y|x))^2|x\big) = \mathsf{E}(Y^2|x) - (\mathsf{E}(Y|x))^2 \\ =& \begin{cases} \int_y (y - \mathsf{E}(Y|x))^2 f(y|x)\,dy & \text{if } y \text{ is continuous} \\ \sum_y (y - \mathsf{E}(Y|x))^2 f(y|x) & \text{if } y \text{ is discrete} \end{cases} \end{aligned} \tag{24}$$

- 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 $\mathsf{var}(Y|x)$ is the same as $\mathsf{var}(y)$. 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**.

Conditional Mean and Variance

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

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

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

- In the equation above, note that 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 and for the second (outer) expectation, we use the marginal distribution.

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

$$\text{var}(Y) = \text{var}_X[E(Y|x)] + \mathsf{E}_X[var(Y|x)] \tag{25}$$

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

Proof. Given in detailed notes. □

Conditional Mean and Variance

- **Example** For the data given in table 5, lets first compute (1) E(Y), then (2) E$(Y|x)$, then (3) var$(Y|x)$ and finally (4) E$_X$[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., $\sum_y y f_Y(y)$

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

So E$(Y) = \sum_y y f_Y(y) =$ 2.48

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 condtional 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	$(Y - 1.4)^2$	$f(y\|x)$	$f(y\|x)\times$ $(Y - 1.4)^2$	$(Y - 2.5)^2$	$f(y\|x)$	$f(y\|x)\times$ $(Y - 2.5)^2$	$(Y - 3.9)^2$	$f(y\|x)$	$f(y\|x)\times$ $(Y - 3.9)^2$
1	0.16	0.7	0.112	2.25	0.1	0.225	8.41	0	0
2	0.36	0.2	0.072	0.25	0.5	0.125	3.61	0.1	0.361
3	2.56	0.1	0.256	0.25	0.2	0.05	0.81	0.2	0.162
4	6.76	0	0	2.25	0.2	0.45	0.01	0.5	0.005
5	12.96	0	0	6.25	0	0	1.21	0.1	0.121
6	21.16	0	0	12.25	0	0	4.41	0.1	0.441
var$(Y\|x) = \sum (y - \mathsf{E}(Y\|x))^2 f(y\|x) =$			.44			0.85			1.09

4. To do the final computation, multiply $\mathsf{E}(Y|x)$ 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
$\mathsf{E}(Y\|x) =$	1.4	2.5	3.9
$f_X(x) \times \mathsf{E}(Y\|x) =$	.56	.75	1.17

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

Independence, Covariance & Correlation

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.

$$f(x,y) = f_X(x)f_Y(y) \Leftrightarrow x\, and\, y \text{ are statistically independent.} \tag{26}$$

- 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.,

$$f(y|x) = f_Y(y) \Leftrightarrow x\, and\, y \text{ are statistically independent.} \tag{27}$$

- Which ever of these you start with as the definition of independence, you can always derive the other condition from it (as the following theorem shows).

Independence, Covariance & Correlation

Theorem 8.

$$f(x,y) = f_X(x) f_Y(y) \Leftrightarrow f(y|x) = f_Y(y) \tag{28}$$

Proof. Since this is a double arrow ($\Leftrightarrow$) we must prove both sides, ie, starting with the statement $f(x,y) = f_X(x)f_Y(y)$ we must show that $f(y|x) = f_Y(y)$ is true. Next, starting with the statement that $f(y|x) = f_Y(y)$ we must show that $f(x,y) = f_X(x)f_Y(y)$ is true. I will do below only the first of these and the second is left as a homework problem.

($\Rightarrow$) Start with the given hypothesis:

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

But, from the definition of joint distribution (eqns.21 and 22) we know that $f(x,y) = f(y|x)f_X(x)$. Substituting this in the left hand side of equation above we get,

$$f(y|x)f_X(x) = f_X(x)f_Y(y)$$

$$\text{which gives } f(y|x) = f_Y(y)$$

($\Leftarrow$) Left as homework problem. □

Independence, Covariance & Correlation

- 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,

$$\begin{aligned} \text{cov}(X, Y) = \sigma_{XY} &= \mathsf{E}\big((X - \mu_X)(Y - \mu_Y)\big) \qquad (29) \\ &= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) & \text{if discrete} \\ \int_x \int_y (x - \mu_X)(y - \mu_Y) f(x, y) dx dy & \text{if continuous} \end{cases} \end{aligned}$$

- 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.

  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

$$\begin{aligned} \rho_{XY} =& \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \qquad (30) \\ =& \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \end{aligned}$$

  Note that the correlation is always between -1 and $+1$.

Independence, Covariance & Correlation

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. $f(x|y) = f_X(x)$ Proof of 1 and 2 follows from theorem 8
3. $cov(X, Y) = \sigma_{XY} = 0$ proof given below
4. $cor(X, Y) = \rho_{XY} = 0$ proof follows from 3
5. $E(Y|x) = E(Y) = \mu_Y$ proof given below
6. $var(Y|x) = var(Y) = \sigma_Y^2$ left as homework problem. See detailed notes for hint

Note: It is not necessarily true that if cov$(X, Y) = 0$ then X and Y are independent.

Proof. (For 3) Observe that since X and Y are independent then $f(x, y) = f_X(x)f_Y(y)$. Hence,

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

(For 5) Observe that $E(Y|x) = \sum_y yf(y|x) = \sum_y yf_Y(y) = E(Y)$. □

Independence, Covariance & Correlation

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

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

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

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

Functions of Random Variables

- Recall definition 11 and theorem 5

$$
\begin{array}{ll}
E[g(x,y)] = \sum_y \sum_x g(x,y) f(x,y) & \text{definition 11} \\
E(X+Y) = E(X) + E(Y) & \text{theorem 5}
\end{array}
$$

Based on these we can also derive the following result:

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$$

- More generally though . . .

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

Proof. See detailed online lecture notes □

Continuous Distributions - Normal

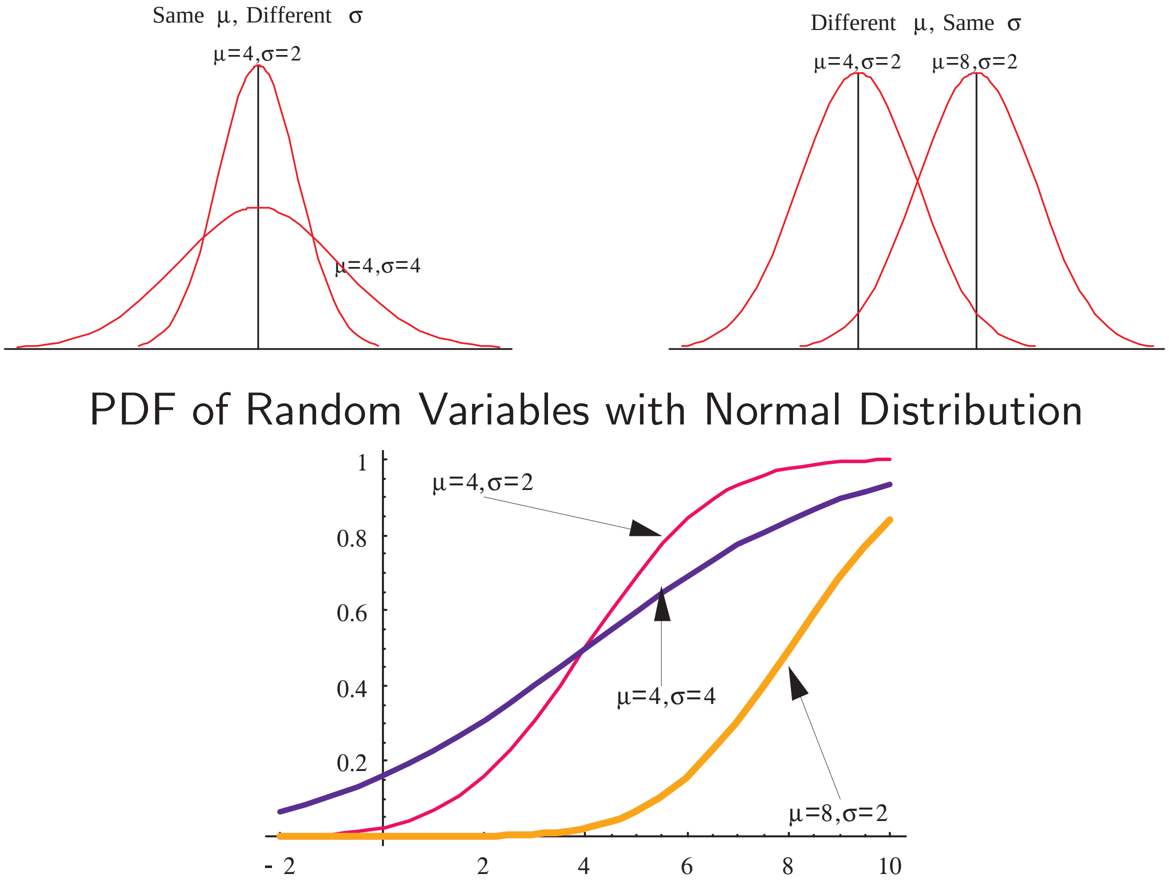
- **Normal Distribution:** A continuous random variable X with the **normal distribution** is the usual familiar bell shaped curve. Its parameters are the mean (μ_X) and variance (σ_X^2) of the random variable and the PDF is given by

$$N(\mu_X, \sigma_X^2) = f(x) = \frac{1}{\sigma_X\sqrt{2\pi}}\,\text{exp}\Big(-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\Big) \tag{31}$$

- The normal distribution with parameters μ and σ^2 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 $Y \sim N(1, 4)$.

Continuous Distributions - Normal

PDF and CDF for normal distributions $N(4,2)$, $N(4,4)$ and $N(8,2)$ are given in figure 5



PDF of Random Variables with Normal Distribution

CDF of Random Variables with Normal Distribution

Figure 5: Normal Distributions

Continuous Distributions - Normal

- 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 $\mu \pm 3\sigma$

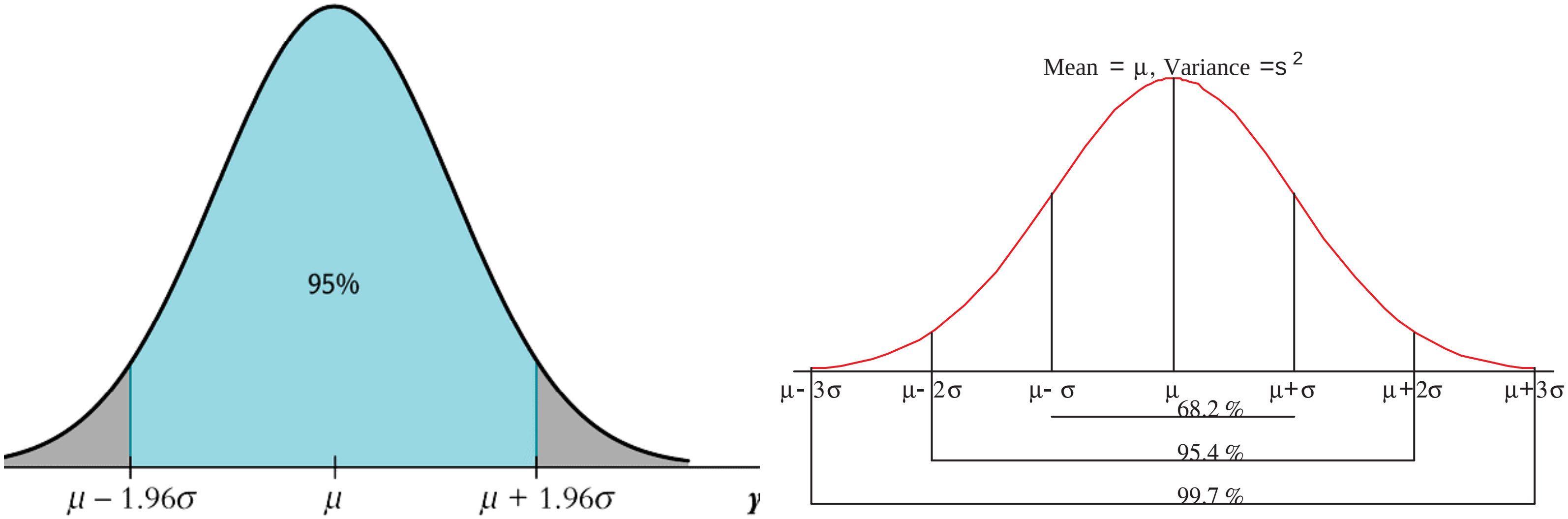


Figure 6: Area under a normal curve

Continuous Distributions - Normal

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.

Continuous Distributions - Standard Normal

- **Standard Normal Distribution** A standard normal is a normal distribution with mean 0 and variance 1. Thus,

$$N(0,1) = f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{32}$$

- The letter Z is usually reserved for the standard normal distribution and the greek letter Φ is often used for its CDF. 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 or is less than some specific value
- Example: 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 by first converting 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 p.960 in your text book). Here is how. Let

$$Z = \frac{Y - \mu}{\sigma} \tag{33}$$

  Then $Y = 2$ corresponds to $Z = \frac{2-1}{2} = .5$ To compute P$(Y \leq 2)$ we need to compute 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
- 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. P$(Y \leq c_2) = $ P$(Z \leq d_2) = \Phi(d_2)$
  2. P$(Y \geq c_1) = $ P$(Z \geq d_2) = 1 - \Phi(d_2)$
  3. P$(c_1 \leq Y \leq c_2) = $ P$(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1)$

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

Continuous Distributions - Bivariate Normal

- **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**.
- 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

- Earlier we noted that 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.

Continuous Distributions - χ^2

- 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

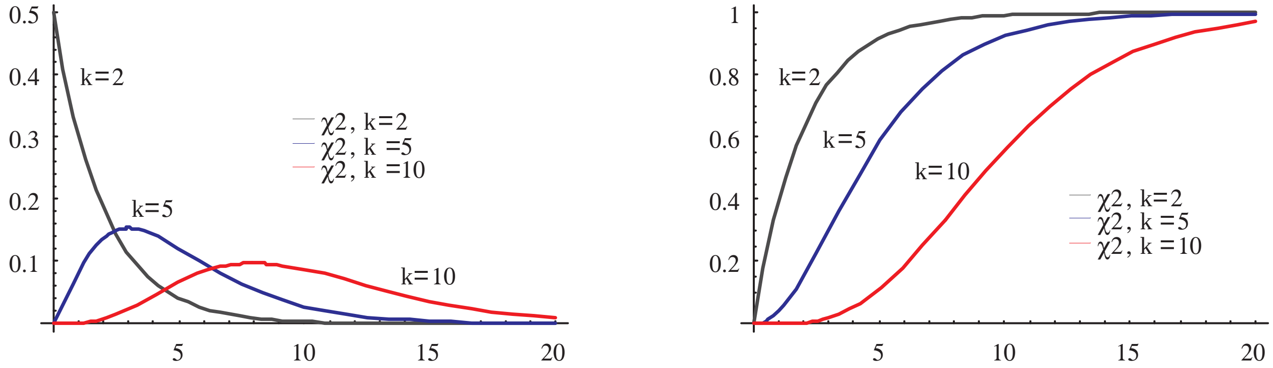
$$Y = \sum_{i=1}^{k} Z_i^2 \quad \text{then} \quad Y \sim \chi^2(k). \tag{34}$$

  i.e, Y has a χ^2 distribution with k degrees of freedom and is written as $Y \sim \chi^2(k)$.

- **Example** 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.

Continuous Distributions - χ^2

Figure 7 shows the PDF and CDF for the case when $k = 2, 5,$ and 10.



χ^2 PDF χ^2 CDF

Figure 7: χ^2 Distributions with k degrees of freedom

- 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.

Continuous Distributions - F

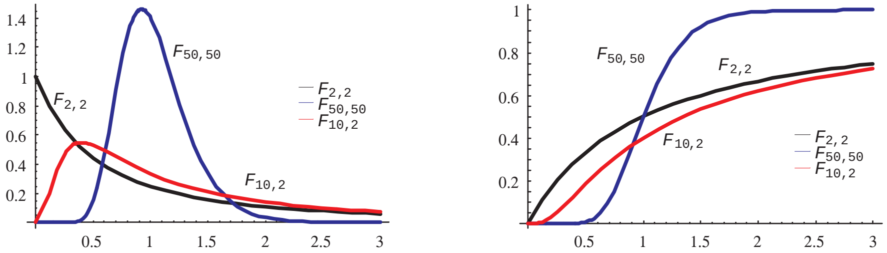
- **F Distribution** If Y_1 and Y_2 are two independent χ^2 variables with k_1 and k_2 degrees of freedom respectively, and you define a new variable X such that

$$X = \frac{Y_1/k_1}{Y_2/k_2} \text{ then } \quad X \sim F_{k_1 k_2} \tag{35}$$

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 8 shows the PDF and CDF for the case when
$(1)\ \ k_1 = 2, k_2 = 2; (2)\ \ k_1 = 50, k_2 = 50;$ and $(3)\ \ k_1 = 10, k_2 = 2.$



F Distribution PDF F Distribution CDF

Figure 8: F Distributions with k1, k2 degrees of freedom

Continuous Distributions - F

- Note the following properties:
  1. A random variable with a $F_{k_1 k_2}$ distribution only takes on positive values.
  2. Like the χ^2 distribution, $F_{k_1 k_2}$ is also right skewed but as k_1 and k_2 become large, the $F_{k_1 k_2}$ 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:
  $$k_1 F_{k_1 \infty} = \chi^2(k_1) \tag{36}$$

Continuous Distributions - F

- **Example** 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**
  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 χ^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;

  • 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!!

Continuous Distributions - t

- **t Distribution** If Z is standard normal random variable 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
$$Y = \frac{Z}{\sqrt{X/k}} \sim t_k \qquad (37)$$
has a students t-distribution with k degrees of freedom.
- Figure 9 shows the PDF for the case when $(1)\ \ k = 5 (2)\ \ k = 120$ which can be regarded as the normal distribution.

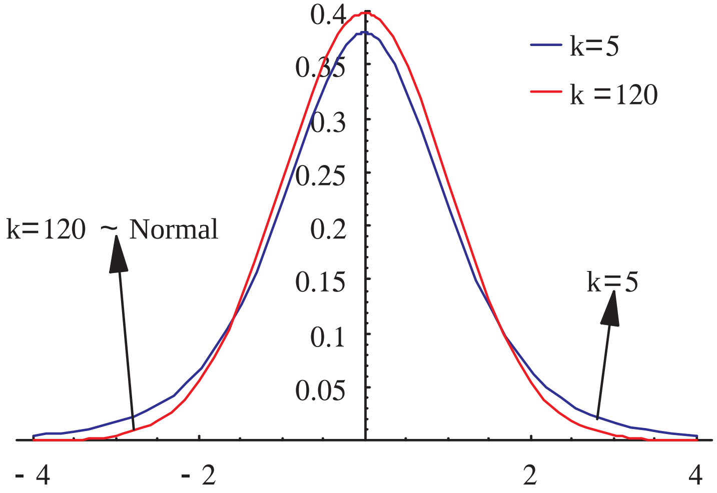


Figure 9: t Distributions (k=5 & k=120)

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.

Continuous Distributions - t

- **Note the following properties:**
  1. The t-distribution is symmetrical like the normal distribution but has more mass in the tails.
  2. As k become 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:** 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:** 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.

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.

Sampling and Sampling Distributions

- **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 $Y_1, Y_2, \ldots, Y_{20}$

Mean and Variance of a Sample

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

$$\overline{Y} = \frac{1}{n} \sum_{i}^{n} Y_i \tag{38}$$

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,

$$s^2 = \frac{1}{n-1} \sum_{i}^{n} (Y_i - \overline{Y})^2. \tag{39}$$

- Both, the sample average and the sample variance, are themselves random variables.

Mean and Variance of $\overline{Y}$

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. 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 $E(Y_i) = \mu_Y$

$$E(\overline{Y}) = E(\frac{Y_1 + Y_2 + \ldots + Y_n}{n}) = \frac{1}{n}E(Y_1 + Y_2 + \ldots + Y_n)$$

$$= \frac{1}{n}(E(Y_1) + E(Y_2) + \ldots + E(Y_n)) = \frac{1}{n}\sum_i^n E(Y_i) = \mu_Y$$

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,

$$var(\overline{Y}) = var(\frac{1}{n}\sum_i Y_i) = var(Y_1/n + Y_2/n + \ldots + Y_i/n)$$

$$= var(aY_1 + aY_2 + \ldots + aY_i) = a^2 var(Y_1 + Y_2 + \ldots + Y_n) = a^2(n\sigma_Y^2) = \sigma_Y^2/n$$

□

Convergence in Probability and 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 $n \rightarrow \infty$ for every $\delta > 0$. That is

$$S_n \xrightarrow{p} \mu \text{ if and only if } \underset{n\rightarrow\infty}{lim} Pr[|S_n - \mu| < \delta] = 1. \tag{40}$$

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

$$\overline{Y}_n \xrightarrow{p} \mu_Y \tag{41}$$

- **Law of Large Numbers:** 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.

Sampling Distribution of $\overline{Y}$ when $Y_i \sim N[\mu_Y, \sigma_Y^2]$

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,

$$\overline{Y} \sim N[\mu_Y, \sigma_Y^2/n] \tag{42}$$

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 $\underline{N[\mu_Y, \sigma_Y^2].}$ 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 & Central Limit Theorem

The basic idea of convergence in distribution is when you have a sequence of distribution functions i.e., a sequence of CDFs like $F_1, F_2, F_3, \ldots$ (or more compactly $\{F_n\}$) and it turns out that as $n \longrightarrow \infty$ then the sequence $\{F_n\}$ approaches a limiting distribution F.

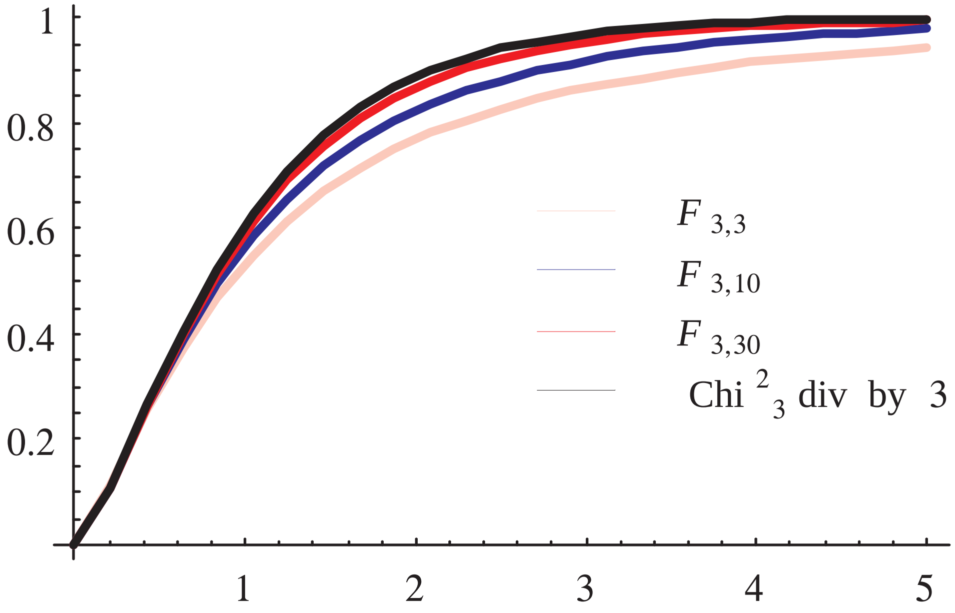


Figure 10: CDF of $\chi^2_3/3$ and of F_3, k where $k = 5, 10, 30$

Convergence in Distribution & Central Limit Theorem

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.

$$S_n \xrightarrow{d} S \text{ if and only if } \underset{n\to\infty}{lim} F_n(S_n) = F(S) \tag{43}$$

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

- 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, ie, $\underset{n\to\infty}{lim} |F_n(S_n) - F(S)| = 0$ 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.

Convergence in Distribution & Central Limit Theorem

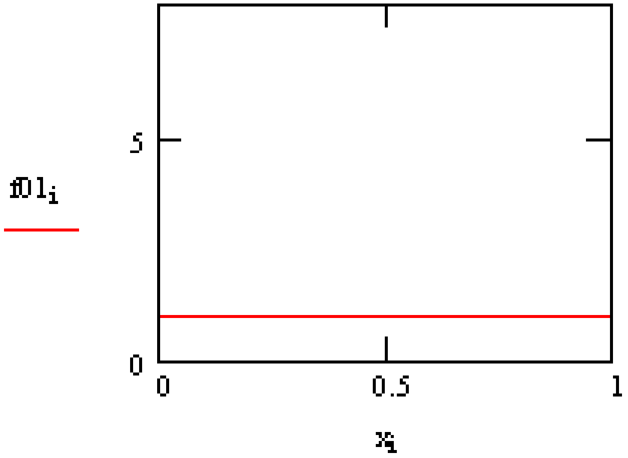
- Roughly, the central limit theorem states that the distribution of the sum of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

  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

$$\frac{\sqrt{n}(\overline{Y} - \mu_Y)}{\sigma_Y} \xrightarrow{d} N(0,1) \tag{44}$$

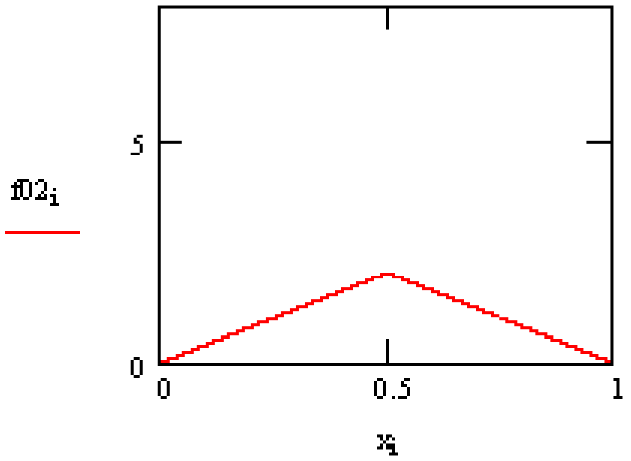
Central Limit Theorem - In Action

- Let the parent distribution be: Uniform[0,1] (Continuous)

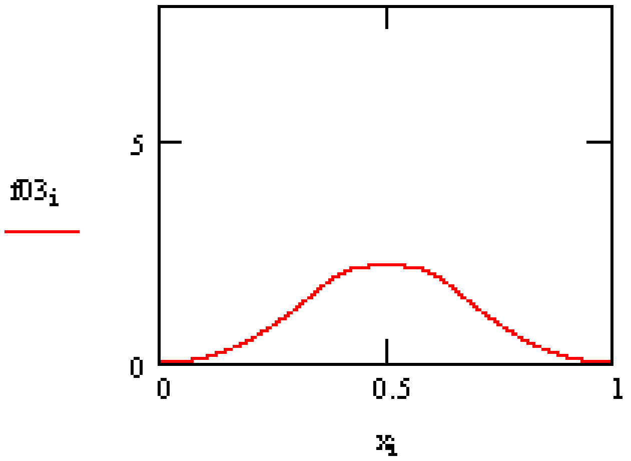


NonNormal Distribution of X

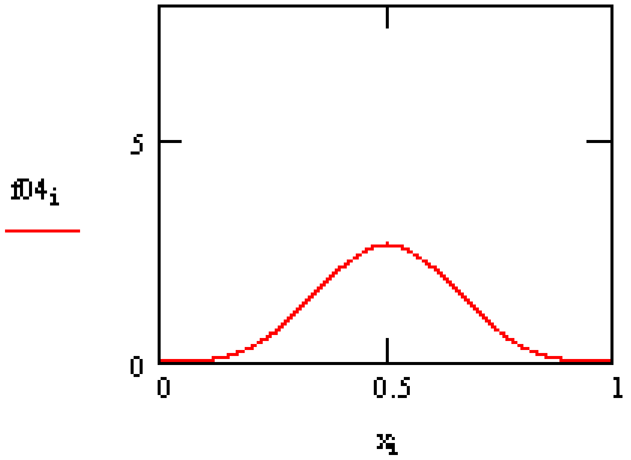
1. Draw a sample of size 2. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_2$.
2. Draw a sample of size 3. Compute the sample average. Repeat the process many times & draw the PDF of $\overline{X}_3$.
3. Draw a sample of size 4. Compute the sample average. Repeat 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 $\overline{X}_{32}$.



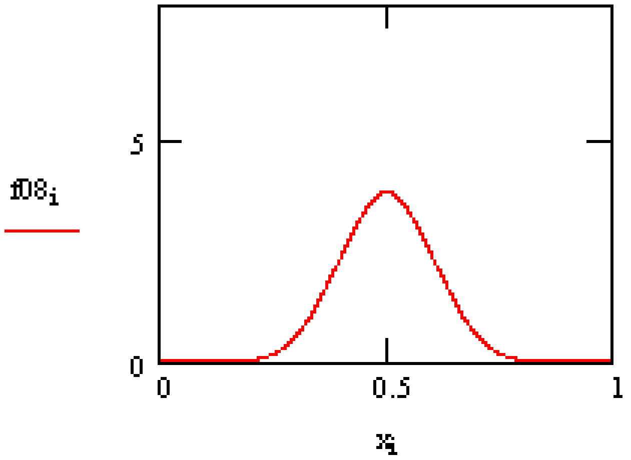
Distribution of Xbar when sample size is 2



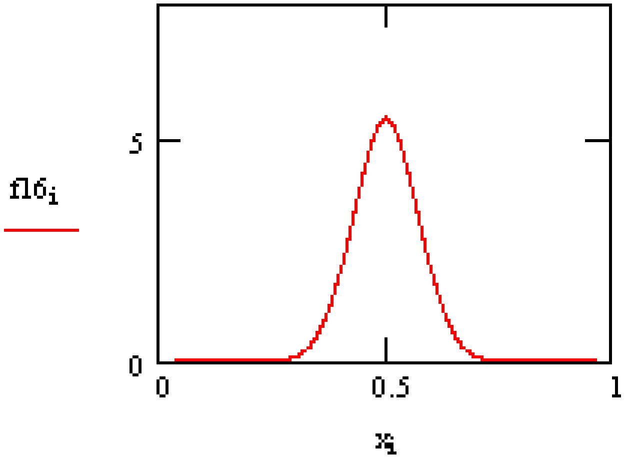
Distribution of Xbar when sample size is 3



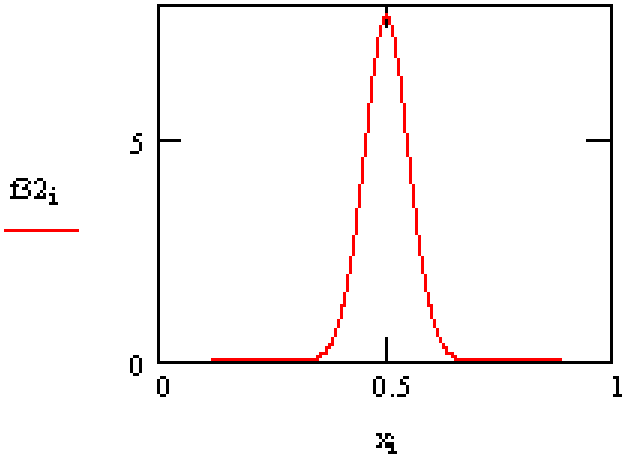
Distribution of Xbar when sample size is 4



Distribution of Xbar when sample size is 8



Distribution of Xbar when sample size is 16



Distribution of Xbar when sample size is 32

Estimators and Their Properties

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,

$$\widehat{\theta} = g(Y_1, Y_2, Y_3 \ldots, Y_n) \tag{45}$$

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

- Example: 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: 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.

Estimators and Their Properties

- There may be different estimators available for the same population parameter θ.
- 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. $\widehat{\theta}_3 = (1/n)[(1/2)Y_1 + (3/2)Y_2 + (1/2)Y_3 + (3/2)Y_4 + \ldots + (1/2)Y_{n-1} + (3/2)Y_n])$
- 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
  3. Efficiency

Estimators and Their Properties - Unbiasedness

- Criteria 1: 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

  $$E(\widehat{\theta}) = \theta \tag{46}$$

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

  $$B = Bias = E(\widehat{\theta}) - \theta \tag{47}$$

- Example: 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.

Estimators and Their Properties - Consistency

- Criteria 2: 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

  $$\widehat{\theta} \xrightarrow{p} \theta \tag{48}$$

- Example: 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.

Estimators and Their Properties - Efficiency

- Criteria 3: 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

$$var(\widehat{\theta_1}) < var(\widehat{\theta_2}) \tag{49}$$

- Example: 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 $var(\widehat{\theta_2})$.

Comparison Among Three Estimators of the Population Mean

Example 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. $\widehat{\theta}_3 = (1/n)[(1/2)Y_1 + (3/2)Y_2 + (1/2)Y_3 + (3/2)Y_4 + \ldots + (1/2)Y_{n-1} + (3/2)Y_n])$

Comparison Among Three Estimators of the Population Mean

1. **Biased?** Lets check the unbiasedness of the three estimators
   (a) $\widehat{\theta}_1 = \overline{Y}$: Since the draws are iid, then by theorem 14, $E(\overline{Y}) = \mu_Y$ and hence it is unbiased.
   (b) $\widehat{\theta}_2 = Y_1$: 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.
   (c) $\widehat{\theta}_3 =$ 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.

Comparison Among Three Estimators of the Population Mean

1. **Consistency?** Lets check the consistency of the three estimators
   (a) $\widehat{\theta}_1 = \overline{Y}$: 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.
   (b) $\widehat{\theta}_2 = Y_1$: 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, $var(Y_1) = var(Y_i) = \sigma^2_Y$. Since this is not a function of n, the variance does not approach zero as n increases. Thus, this estimator is not consistent.
   (c) $\widehat{\theta}_3 =$ 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 $1.25\sigma^2_Y/n$ (shown below). Because $var(\widehat{\theta}_3) \longrightarrow 0$ as $n \longrightarrow \infty$ hence $\widehat{\theta}_3$ is a consistent estimator.

$$\begin{aligned} var(\widehat{\theta}_3) &= var\Big[\frac{1}{n}(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \ldots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n)\Big] \\ &= \frac{1}{n^2}\Big[(\frac{1}{4}var(Y_1) + \frac{9}{4}var(Y_2) + \ldots + \frac{1}{4}var(Y_{n-1}) + \frac{9}{4}var(Y_n))\Big] \end{aligned}$$

but $var(Y_1) = var(Y_2) = var(Y_3) = \ldots = var(Y_n) = var(Y_i) = \sigma^2_Y$, and so

$$var(\widehat{\theta}_3) = \frac{1}{n^2}\Big[\frac{n}{4 \times 2}var(Y_i) + \frac{9n}{4 \times 2}var(Y_i)\Big] = \frac{5}{4n}var(Y_i) = 1.25\sigma^2_Y/n$$

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

Comparison Among Three Estimators of the Population Mean

1. **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.
   (a) $\widehat{\theta}_1 = \overline{Y}$: Since the draws are iid, then by theorem 14, $var(\widehat{\theta}_1) = \sigma_Y^2/n$
   (b) $\widehat{\theta}_2 = Y_1$: Similarly, $var(\widehat{\theta}_2) = \sigma_Y^2$
   (c) $\widehat{\theta}_3 =$ 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.
  - Also, 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 . . .

Most Efficient Estimator

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, if $\widehat{\theta}^*$ is unbiased, then $var(\widehat{\theta}) = var(\overline{Y}) < var(\widehat{\theta}^*)$. That is $\widehat{\theta} = \overline{Y}$ is the most efficient estimator among all unbiased estimators of the population mean that are weighted averages of the sample.

The End but . . .

The Big Picture

- In 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.