Introduction to Statistics

By:
Ewa Paszek

Introduction to Statistics

By:
Ewa Paszek

Online:
<http://cnx.org/content/col10343/1.3/ >

C O N N E X I O N S

Rice University, Houston, Texas

Table of Contents

Chapter 1

Discrete Distributions

1.1 DISCRETE DISTRIBUTION[1]

1.1.1 DISCRETE DISTRIBUTION

1.1.1.1 RANDOM VARIABLE OF DISCRETE TYPE

A **SAMPLE SPACE *S*** may be difficult to describe if the elements of S are not numbers. Let discuss how one can use a rule by which each simple outcome of a random experiment, an element s of S, may be associated with a real number x.

Definition 1: DEFINITION OF RANDOM VARIABLE

1. Given a random experiment with a sample space S, a function X that assigns to each element s in S one and only one real number $X(s) = x$ is called **a random variable**. The space of X is the set of real numbers $\{x : x = X(s), s \in S\}$, where s belongs to S means the element s belongs to the set S.

2. It may be that the set S has elements that are themselves real numbers. In such an instance we could write $X(s) = s$ so that X is **the identity function** and the space of X is also S. This is illustrated in the example below.

Example 1.1

Let the random experiment be the cast of a die, observing the number of spots on the side facing up. The sample space associated with this experiment is $S = (1, 2, 3, 4, 5, 6)$. For each s belongs to S, let $X(s) = s$. The space of the random variable X is then {1,2,3,4,5,6}.

If we associate a probability of 1/6 with each outcome, then, for example, $P(X = 5) = 1/6, P(2 \leq X \leq 5) = 4/6$, and s belongs to S seem to be reasonable assignments, where $(2 \leq X \leq 5)$ means (X = 2,3,4 or 5) and $(X \leq 2)$ means (X = 1 or 2), in this example.

We can recognize two major difficulties:

1. In many practical situations the probabilities assigned to the event are unknown.
2. Since there are many ways of defining a function X on S, which function do we want to use?

Let X denotes a random variable with one-dimensional space R, a subset of the real numbers. Suppose that the space R contains a countable number of points; that is, R contains either a finite number of points or the points of R can be put into a one-to- one correspondence with the positive integers. Such set R is called **a set of discrete points** or simply **a discrete sample space**.

Furthermore, the random variable X is called **a random variable of the discrete type**, and X is said to have **a distribution of the discrete type**. For a random variable X of the discrete type, the

[1] This content is available online at <http://cnx.org/content/m13114/1.5/>.

probability $P(X = x)$ is frequently denoted by *f(x)*, and is called **the probability density function** and it is abbreviated **p.d.f.**.

Let *f(x)* be the p.d.f. of the random variable X of the discrete type, and let R be the space of X. Since, $f(x) = P(X = x)$, x belongs to R, *f(x)* must be positive for x belongs to R and we want all these probabilities to add to 1 because each $P(X = x)$ represents the fraction of times x can be expected to occur. Moreover, to determine the probability associated with the event $A \subset R$, one would sum the probabilities of the x values in A.

That is, we want f(x) to satisfy the properties

- $P(X = x)$,
- $\sum_{x \in R} f(x) = 1;$
- $P(X \in A) = \sum_{x \in A} f(x)$, where $A \subset R$.

Usually let $f(x) = 0$ when $x \notin R$ and thus the domain of *f(x)* is the set of real numbers. When we define the p.d.f. of *f(x)* and do not say zero elsewhere, then we tacitly mean that *f(x)* has been defined at all x's in space R, and it is assumed that $f(x) = 0$ elsewhere, namely, $f(x) = 0$, $x \notin R$. Since the probability $P(X = x) = f(x) > 0$ when $x \in R$ and since R contains all the probabilities associated with X, R is sometimes referred to as **the support of X** as well as the space of X.

Example 1.2

Roll a four-sided die twice and let X equal the larger of the two outcomes if there are different and the common value if they are the same. The sample space for this experiment is $S = [(d_1, d_2) : d_1 = 1, 2, 3, 4; d_2 = 1, 2, 3, 4]$, where each of this 16 points has probability 1/16. Then $P(X = 1) = P[(1, 1)] = 1/16$, $P(X = 2) = P[(1, 2), (2, 1), (2, 2)] = 3/16$, and similarly $P(X = 3) = 5/16$ and $P(X = 4) = 7/16$. That is, the p. d.f. of X can be written simply as $f(x) = P(X = x) = \frac{2x-1}{16}, x = 1, 2, 3, 4.$

We could add that $f(x) = 0$ elsewhere; but if we do not, one should take *f(x)* to equal zero when $x \notin R$.

A better understanding of a particular probability distribution can often be obtained with a graph that depicts the p.d.f. of X.

NOTE THAT: the graph of the p.d.f. when $f(x) > 0$, would be simply the set of points $\{[x, f(x)] : x \in R\ \}$, where R is the space of X.

Two types of graphs can be used to give a better visual appreciation of the p.d.f., namely, **a bar graph** and **a probability histogram**. A bar graph of the p.d.f. *f(x)* of the random variable X is a graph having a vertical line segment drawn from $(x, 0)$ to $[x, f(x)]$ at each x in R, the space of X. If X can only assume integer values, **a probability histogram** of the p.d.f. f(x) is a graphical representation that has a rectangle of height *f(x)* and a base of length 1, centered at x, for each $x \in R$, the space of X.

Definition 2: CUMULATIVE DISTRIBUTION FUNCTION

1. Let X be a random variable of the discrete type with space R and p.d.f. $f(x) = P(X = x)$, $x \in R$. Now take x to be a real number and consider the set A of all points in R that are less than or equal to x. That is, $A = (t : t \leq x)$ and $t \in R$.
2. Let define the function *F(x)* by

$$F(x) = P(X \leq x) = \sum_{t \in A} f(t)\,. \tag{1.1}$$

The function *F(x)* is called **the distribution function** (sometimes **cumulative distribution function**) of the discrete-type random variable X.

Several properties of a distribution function *F(x)* can be listed as a consequence of the fact that probability must be a value between 0 and 1, inclusive:

- $0 \leq F(x) \leq 1$ because *F(x)* is a probability,
- *F(x)* is a nondecreasing function of x,
- $F(y) = 1$, where *y* is any value greater than or equal to the largest value in *R*; and $F(z) = 0$, where *z* is any value less than the smallest value in *R*;
- If *X* is a random variable of the discrete type, then *F(x)* is a step function, and the height at a step at x, $x \in R$, equals the probability $P(X = x)$.

NOTE: It is clear that the probability distribution associated with the random variable *X* can be described by either the distribution function *F(x)* or by the probability density function *f(x)*. The function used is a matter of convenience; in most instances, *f(x)* is easier to use than *F(x)*.

Graphical representation of the relationship between p.d.f. and c.d.f.

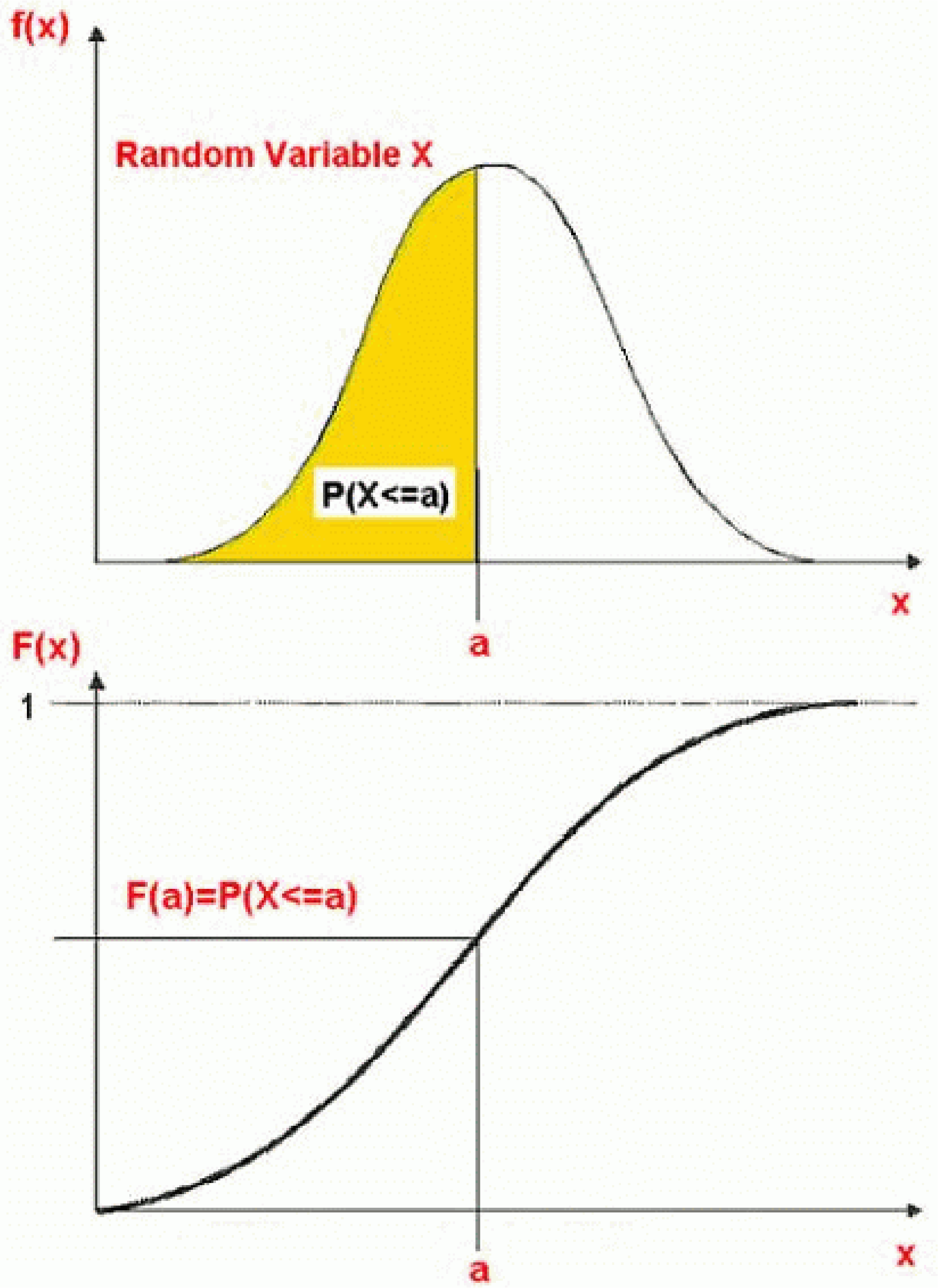


Figure 1.1: Area under p.d.f. curve to a equal to a value of c.d.f. curve at a point a.

Definition 3: MATHEMATICAL EXPECTATION
If *f(x)* is the p.d.f. of the random variable X of the discrete type with space R and if the summation

$$\sum_R u(x)\, f(x) = \sum_{x\in R} u(x)\, f(x) \tag{1.2}$$

exists, then the sum is called **the mathematical expectation** or **the expected value** of the function *u(X)*, and it is denoted by $E\left[u(X)\right]$. That is,

$$E\left[u(X)\right] = \sum_R u(x)\, f(x)\,. \tag{1.3}$$

We can think of the expected value $E\left[u(X)\right]$ as a weighted mean of *u(x)*, $x \in R$, where the weights are the probabilities $f(x) = P(X = x)$.

REMARK: The usual definition of the mathematical expectation of *u(X)* requires that the sum converges absolutely; that is, $\sum_{x\in R} |u(x)|f(x)$ exists.

There is another important observation that must be made about consistency of this definition. Certainly, this function *u(X)* of the random variable X is itself a random variable, say Y. Suppose that we find the p.d.f. of Y to be *g(y)* on the support R_1 . Then *E(Y)* is given by the summation $\sum_{y\in R_1} yg(y)$
In general it is true that

$$\sum_R u(x)\, f(x) = \sum_{y\in R_1} yg(y)\,;$$

that is, the same expectation is obtained by either method.

Example 1.3
Let X be the random variable defined by the outcome of the cast of the die. Thus the p.d.f. of X is
$f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6.$
In terms of the observed value x, the function is as follows

$$u(x) = \{\begin{array}{l} 1, x = 1, 2, 3, \\ 5, x = 4, 5, \\ 35, x = 6. \end{array}$$

The mathematical expectation is equal to

$$\sum_{x=1}^{6} u(x)\, f(x) = 1\left(\frac{1}{6}\right)+1\left(\frac{1}{6}\right)+1\left(\frac{1}{6}\right)+5\left(\frac{1}{6}\right)+5\left(\frac{1}{6}\right)+35\left(\frac{1}{6}\right) = 1\left(\frac{3}{6}\right)+5\left(\frac{2}{6}\right)+35\left(\frac{1}{6}\right) = 8. \tag{1.4}$$

Example 1.4
Let the random variable X have the p.d.f. $f(x) = \frac{1}{3}$, $x \in R$, where R ={-1,0,1}. Let $u(X) = X^2$. Then

$$\sum_{x\in R} x^2 f(x) = (-1)^2\left(\frac{1}{3}\right) + (0)^2\left(\frac{1}{3}\right) + (1)^2\left(\frac{1}{3}\right) = \frac{2}{3}. \tag{1.5}$$

However, the support of random variable $Y = X^2$ is $R_1 = (0, 1)$ and

$P(Y = 0) = P(X = 0) = \frac{1}{3}$
$P(Y = 1) = P(X = -1) + P(X = 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$

That is,

$$g\left(y\right)=\{\begin{array}{c}\frac{1}{3},y=0,\\ \frac{2}{3},y=1;\end{array}$$

and R_1. Hence
$\sum_{y\in R_1} yg\left(y\right)=0\left(\frac{1}{3}\right)+1\left(\frac{2}{3}\right)$, which illustrates the preceding observation.

Theorem 1.1:
When it exists, mathematical expectation E satisfies the following properties:

1. If c is a constant, $E(c)=c$,
2. If c is a constant and u is a function, $E\left[cu\left(X\right)\right]=cE\left[u\left(X\right)\right]$,
3. If c_1 and c_2 are constants and u_1 and u_2 are functions, then $E\left[c_1u_1\left(X\right)+c_2u_2\left(X\right)\right]=c_1E\left[u_1\left(X\right)\right]+c_2E\left[u_2\left(X\right)\right]$

Proof: First, we have for the proof of (1) that
$E\left(c\right)=\sum_R cf\left(x\right)=c\sum_R f\left(x\right)=c$
because $\sum_R f\left(x\right)=1.$

Proof: Next, to prove (2), we see that
$E\left[cu\left(X\right)\right]=\sum_R cu\left(x\right)f\left(x\right)=c\sum_R u\left(x\right)f\left(x\right)=cE\left[u\left(X\right)\right].$

Proof: Finally, the proof of (3) is given by
$E\left[c_1u_1\left(X\right)+c_2u_2\left(X\right)\right]=\sum_R\left[c_1u_1\left(x\right)+c_2u_2\left(x\right)\right]f\left(x\right)=\sum_R c_1u_1\left(x\right)f\left(x\right)+\sum_R c_2u_2\left(x\right)f\left(x\right).$
By applying (2), we obtain
$E\left[c_1u_1\left(X\right)+c_2u_2\left(X\right)\right]=c_1E\left[u_1\left(x\right)\right]+c_2E\left[u_2\left(x\right)\right].$
Property (3) can be extended to more than two terms by mathematical induction; That is, we have
3'. $E\left[\sum_{i=1}^k c_iu_i\left(X\right)\right]=\sum_{i=1}^k c_iE\left[u_i\left(X\right)\right].$

Because of property (3'), mathematical expectation E is called a **linear** or **distributive operator**.

Example 1.5
Let X have the p.d.f. $f\left(x\right)=\frac{x}{10}$, x=1,2,3,4.
then
$E\left(X\right)=\sum_{x=1}^4 x\left(\frac{x}{10}\right)=1\left(\frac{1}{10}\right)+2\left(\frac{2}{10}\right)+3\left(\frac{3}{10}\right)+4\left(\frac{4}{10}\right)=3$
$E\left(X^2\right)=\sum_{x=1}^4 x^2\left(\frac{x}{10}\right)=1^2\left(\frac{1}{10}\right)+2^2\left(\frac{2}{10}\right)+3^2\left(\frac{3}{10}\right)+4^2\left(\frac{4}{10}\right)=10,$
and
$E\left[X\left(5-X\right)\right]=5E\left(X\right)-E\left(X^2\right)=\left(5\right)\left(3\right)-10=5.$

SEE ALSO: the MEAN, VARIANCE, and STANDARD DEVIATION (Section 1.3.1: The MEAN, VARIANCE, and STANDARD DEVIATION)

1.2 MATHEMATICAL EXPECTATION[2]

1.2.1 MATHEMATICAL EXPECTIATION

Definition 4: MATHEMATICAL EXPECTIATION
If $f\left(x\right)$ is the p.d.f. of the random variable X of the discrete type with space R and if the summation

[2] This content is available online at <http://cnx.org/content/m13530/1.2/>.

$$\sum_{R} u(x) f(x) = \sum_{x \in R} u(x) f(x) . \tag{1.6}$$

exists, then the sum is called **the mathematical expectation** or **the expected value** of the function $u(X)$, and it is denoted by $E[u(x)]$. That is,

$$E[u(X)] = \sum_{R} u(x) f(x) . \tag{1.7}$$

We can think of the expected value $E[u(x)]$ as a weighted mean of $u(x)$, $x \in R$, where the weights are the probabilities $f(x) = P(X = x)$.

> REMARK: The usual definition of the mathematical expectation of $u(X)$ requires that the sum converges absolutely; that is, $\sum_{x \in R} |u(x)| f(x)$ exists.

There is another important observation that must be made about consistency of this definition. Certainly, this function $u(X)$ of the random variable X is itself a random variable, say Y. Suppose that we find the p.d.f. of Y to be $g(y)$ on the support R_1 . Then, $E(Y)$ is given by the summation $\sum_{y \in R_1} yg(y)$.

In general it is true that $\sum_{R} u(x) f(x) = \sum_{y \in R_1} yg(y)$.

This is, the same expectation is obtained by either method.

Example 1.6

Let X be the random variable defined by the outcome of the cast of the die. Thus the p.d.f. of X is

$f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6.$

In terms of the observed value x, the function is as follows

$$u(x) = \{ \begin{array}{l} 1, x = 1, 2, 3, \\ 5, x = 4, 5, \\ 35, x = 6. \end{array}$$

The mathematical expectation is equal to

$\sum_{x=1}^{6} u(x) f(x) = 1\left(\frac{1}{6}\right) + 1\left(\frac{1}{6}\right) + 1\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 35\left(\frac{1}{6}\right) = 1\left(\frac{3}{6}\right) + 5\left(\frac{2}{6}\right) + 35\left(\frac{1}{6}\right) = 8.$

Example 1.7

Let the random variable X have the p.d.f.

$f(x) = \frac{1}{3}$, $x \in R$,

where, $R = (-1, 0, 1)$. Let $u(X) = X^2$. Then

$\sum_{x \in R} x^2 f(x) = (-1)^2 \left(\frac{1}{3}\right) + (0)^2 \left(\frac{1}{3}\right) + (1)^2 \left(\frac{1}{3}\right) = \frac{2}{3}.$

However, the support of random variable $Y = X^2$ is $R_1 = (0, 1)$ and

$P(Y = 0) = P(X = 0) = \frac{1}{3}$

$P(Y = 1) = P(X = -1) + P(X = 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$

That is, $g(y) = \{ \begin{array}{l} \frac{1}{3}, y = 0, \\ \frac{2}{3}, y = 1; \end{array}$ and $R_1 = (0, 1)$. Hence

$$\sum_{y \in R_1} yg(y) = 0\left(\frac{1}{3}\right) + 1\left(\frac{2}{3}\right) = \frac{2}{3},$$

which illustrates the preceding observation.

Theorem 1.2:

When it exists, mathematical expectation E satisfies the following properties:

1. If c is a constant, $E(c) = c$,
2. If c is a constant and u is a function, $E[cu(X)] = cE[u(X)]$,

3. If c_1 and c_2 are constants and u_1 and u_2 are functions, then $E\left[c_1 u_1(X) + c_2 u_2(X)\right] = c_1 E\left[u_1(X)\right] + c_2 E\left[u_2(X)\right].$

Proof: First, we have for the proof of (1) that

$$E(c) = \sum_R c f(x) = c \sum_R f(x) = c,$$

because $\sum_R f(x) = 1$.

Proof: Next, to prove (2), we see that

$$E\left[cu(X)\right] = \sum_R cu(x) f(x) = c \sum_R u(x) f(x) = cE\left[u(X)\right].$$

Proof: Finally, the proof of (3) is given by

$$E\left[c_1 u_1(X) + c_2 u_2(X)\right] = \sum_R \left[c_1 u_1(x) + c_2 u_2(x)\right] f(x) = \sum_R c_1 u_1(x) f(x) + \sum_R c_2 u_2(x) f(x).$$

By applying (2), we obtain

$$E\left[c_1 u_1(X) + c_2 u_2(X)\right] = c_1 E\left[u_1(x)\right] + c_2 E\left[u_2(x)\right].$$

Property (3) can be extended to more than two terms by mathematical induction; that is, we have (3')

$$E\left[\sum_{i=1}^{k} c_i u_i(X)\right] = \sum_{i=1}^{k} c_i E\left[u_i(X)\right].$$

Because of property (3'), mathematical expectation E is called **a linear** or **distributive operator**.

Example 1.8

Let X have the p.d.f. $f(x) = \frac{x}{10}, x = 1, 2, 3, 4$, then

$$E(X) = \sum_{x=1}^{4} x\left(\frac{x}{10}\right) = 1\left(\frac{1}{10}\right) + 2\left(\frac{2}{10}\right) + 3\left(\frac{3}{10}\right) + 4\left(\frac{4}{10}\right) = 3,$$

$$E\left(X^2\right) = \sum_{x=1}^{4} x^2\left(\frac{x}{10}\right) = 1^2\left(\frac{1}{10}\right) + 2^2\left(\frac{2}{10}\right) + 3^2\left(\frac{3}{10}\right) + 4^2\left(\frac{4}{10}\right) = 10,$$

and

$$E\left[X(5-X)\right] = 5E(X) - E\left(X^2\right) = (5)(3) - 10 = 5.$$

1.3 THE MEAN, VARIANCE, AND STANDARD DEVIATION[3]

1.3.1 The MEAN, VARIANCE, and STANDARD DEVIATION

1.3.1.1 MEAN and VARIANCE

Certain mathematical expectations are so important that they have special names. In this section we consider two of them: the mean and the variance.

[3] This content is available online at <http://cnx.org/content/m13122/1.3/>.

Mean Value

If X is a random variable with p.d.f. $f(x)$ of the discrete type and space R=$(b_1, b_2, b_3, ...)$, then $E(X) = \sum_R xf(x) = b_1 f(b_1) + b_2 f(b_2) + b_3 f(b_3) + ...$ is the weighted average of the numbers belonging to R, where the weights are given by the p.d.f. $f(x)$.

We call $E(X)$ **the mean** of X (or **the mean of the distribution**) and denote it by μ. That is, $\mu = E(X)$.

REMARK: In mechanics, the weighted average of the points $b_1, b_2, b_3, ...$ in one-dimensional space is called the centroid of the system. Those without the mechanics background can think of the centroid as being the point of balance for the system in which the weights $f(b_1), f(b_2), f(b_3), ...$ are places upon the points $b_1, b_2, b_3, ...$

Example 1.9

Let X have the p.d.f.

$$f(x) = \{ \begin{array}{l} \frac{1}{8}, x = 0, 3, \\ \frac{3}{8}, x = 1, 2. \end{array}$$

The mean of X is

$$\mu = E\left[X = 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) = \frac{3}{2}.\right.$$

The example below shows that if the outcomes of X are equally likely (i.e., each of the outcomes has the same probability), then the mean of X is the arithmetic average of these outcomes.

Example 1.10

Roll a fair die and let X denote the outcome. Thus X has the p.d.f.

$$f(x) = \frac{1}{6}, x = 1, 2, 3, 4, 5, 6.$$

Then,

$$E(X) = \sum_{x=1}^{6} x\left(\frac{1}{6}\right) = \frac{1+2+3+4+5+6}{6} = \frac{7}{2},$$

which is the arithmetic average of the first six positive integers.

Variance

It was denoted that the mean $\mu = E(X)$ is the centroid of a system of weights of measure of the central location of the probability distribution of X. **A measure of the dispersion or spread of a distribution is defined as follows:**

If $u(x) = (x - \mu)^2$ and $E\left[(X - \mu)^2\right]$ exists, **the variance**, frequently denoted by σ^2 or $Var(X)$, of a random variable X of the discrete type (or variance of the distribution) is defined by

$$\sigma^2 = E\left[(X - \mu)^2\right] = \sum_R (x - \mu)^2 f(x). \tag{1.8}$$

The positive square root of the variance is called **the standard deviation of X** and is denoted by

$$\sigma = \sqrt{Var(X)} = \sqrt{E\left[(X - \mu)^2\right]}. \tag{1.9}$$

Example 1.11

Let the p.d.f. of X by defined by

$$f(x) = \frac{x}{6}, x = 1, 2, 3.$$

The mean of X is

$$\mu = E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{2}{6}\right) + 3\left(\frac{3}{6}\right) = \frac{7}{3}.$$

To find the variance and standard deviation of X we first find

$$E\left(X^2\right) = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{2}{6}\right) + 3^2\left(\frac{3}{6}\right) = \frac{36}{6} = 6.$$

Thus the variance of X is

$$\sigma^2 = E\left(X^2\right) - \mu^2 = 6 - \left(\frac{7}{3}\right)^2 = \frac{5}{9},$$

and the standard deviation of X is

Example 1.12

Let X be a random variable with mean μ_x and variance σ_x^2. Of course, $Y = aX + b$, where a and b are constants, is a random variable, too. The mean of Y is

$$\mu_Y = E(Y) = E(aX + b) = aE(X) + b = a\mu_X + b.$$

Moreover, the variance of Y is

$$\sigma_Y^2 = E\left[(Y - \mu_Y)^2\right] = E\left[(aX + b - a\mu_X - b)^2\right] = E\left[a^2(X - \mu_X)^2\right] = a^2\sigma_X^2.$$

Moments of the distribution

Let r be a positive integer. If

$$E\left(X^r\right) = \sum_R x^r f(x)$$

exists, it is called **the rth moment of the distribution** about the origin. The expression moment has its origin in the study of mechanics.

In addition, the expectation

$$E\left[(X - b)^r\right] = \sum_R x^r f(x)$$

is called **the rth moment of the distribution about b**. For a given positive integer r.

$$E\left[(X)_r\right] = E\left[X(X-1)(X-2)\cdots(X-r+1)\right]$$

is called **the rth factorial moment**.

NOTE THAT: The second factorial moment is equal to the difference of the second and first moments:

$$E\left[X(X-1)\right] = E\left(X^2\right) - E(X).$$

There is another formula that can be used for computing the variance that uses the second factorial moment and sometimes simplifies the calculations.

First find the values of $E(X)$ and $E\left[X(X-1)\right]$. Then

$$\sigma^2 = E\left[X(X-1)\right] + E(X) - \left[E(X)\right]^2,$$

since using the distributive property of E, this becomes

$$\sigma^2 = E\left(X^2\right) - E\left(X\right) + E\left(X\right) - \left[E\left(X\right)\right]^2 = E\left(X^2\right) - \mu^2.$$

Example 1.13
Let continue with example 4 (Example 1.12), it can be find that

$$E\left[X\left(X-1\right)\right] = 1\left(0\right)\left(\frac{1}{6}\right) + 2\left(1\right)\left(\frac{2}{6}\right) + 3\left(2\right)\left(\frac{3}{6}\right) = \frac{22}{6}.$$

Thus

$$\sigma^2 = E\left[X\left(X-1\right)\right] + E\left(X\right) - \left[E\left(X\right)\right]^2 = \frac{22}{6} + \frac{7}{3} - \left(\frac{7}{3}\right)^2 = \frac{5}{9}.$$

REMARK: Recall the empirical distribution is defined by placing the weight (probability) of $1/n$ on each of n observations $x_1, x_2, ..., x_n$. Then the mean of this empirical distribution is

$$\sum_{i=1}^{n} x_i \frac{1}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

The symbol $\overline{x}$ represents **the mean of the empirical distribution**. It is seen that $\overline{x}$ is usually close in value to $\mu = E\left(X\right)$; thus, when μ is unknown, $\overline{x}$ will be used to estimate μ.

Similarly, **the variance of the empirical distribution** can be computed. Let v denote this variance so that it is equal to

$$v = \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \frac{1}{n} = \sum_{i=1}^{n} x_i^2 \frac{1}{n} - \overline{x}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2.$$

This last statement is true because, in general,

$$\sigma^2 = E\left(X^2\right) - \mu^2.$$

NOTE THAT: There is a relationship between the sample variance s^2 and variance v of the empirical distribution, namely $s^2 = ns/\left(n-1\right)$. Of course, with large n, the difference between s^2 and v is very small. Usually, we use s^2 to estimate σ^2 when σ^2 is unknown.

SEE ALSO: BERNOULLI TRIALS and BINOMIAL DISTRIBUTION (Section 1.4.1: BERNOULLI TRIALS AND THE BINOMIAL DISTRIBUTION)

1.4 BERNOULLI TRIALS and the BINOMIAL DISTRIBUTION[4]

1.4.1 BERNOULLI TRIALS AND THE BINOMIAL DISTRIBUTION

A Bernoulli experiment is a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways, mainly, **success** or **failure** (e.g., female or male, life or death, nondefective or defective).

A sequence of **Bernoulli trials** occurs when a Bernoulli experiment is performed several independent times so that the probability of success, say, p, remains the same from trial to trial. That is, in such a sequence we let p denote the probability of success on each trial. In addition, frequently $q = 1 - p$ denote the probability of failure; that is, we shall use q and $1 - p$ interchangeably.

[4] This content is available online at <http://cnx.org/content/m13123/1.3/>.

1.4.1.1 Bernoulli distribution

Let X be a random variable associated with Bernoulli trial by defining it as follows:

X(success)=1 and X(failure)=0.

That is, the two outcomes, **success** and **failure**, are denoted by one and zero, respectively. The p.d.f. of X can be written as

$$f\left(x\right) = p^x{\left(1 - p\right)}^{1 - x}, \tag{1.10}$$

and we say that X has **a Bernoulli distribution**. The expected value of is

$$\mu = E\left(X\right) = \sum_{X=0}^{1} xp^x{\left(1 - p\right)}^{1 - x} = \left(0\right)\left(1 - p\right) + \left(1\right)\left(p\right) = p, \tag{1.11}$$

and the variance of X is

$$\sigma^2 = Var\left(X\right) = \sum_{x=0}^{1} {\left(x - p\right)}^2 p^x{\left(1 - p\right)}^{1 - x} = p^2\left(1 - p\right) + {\left(1 - p\right)}^2 p = p\left(1 - p\right) = pq. \tag{1.12}$$

It follows that the standard deviation of X is $\sigma = \sqrt{p\left(1 - p\right)} = \sqrt{pq}$.

In a sequence of n Bernoulli trials, we shall let X_i denote the Bernoulli random variable associated with the ith trial. An observed sequence of n Bernoulli trials will then be an n-tuple of zeros and ones.

Binomial Distribution

In a sequence of Bernoulli trials we are often interested in the total number of successes and not in the order of their occurrence. If we let the random variable X equal the number of observed successes in n Bernoulli trials, the possible values of X are 0,1,2,. . .,n. If x success occur, where $x = 0, 1, 2, ..., n$, then n-x failures occur. The number of ways of selecting x positions for the x successes in the x trials is

$$\left(\begin{array}{c} n \\ x \end{array}\right) = \frac{n!}{x!\left(n - x\right)!}.$$

Since the trials are independent and since the probabilities of success and failure on each trial are, respectively, p and $q = 1 - p$, the probability of each of these ways is $p^x{\left(1 - p\right)}^{n - x}$.. Thus the p.d.f. of X, say $f\left(x\right)$, is the sum of the probabilities of these $\left(\begin{array}{c} n \\ x \end{array}\right)$ mutually exclusive events; that is,

$$f\left(x\right) = \left(\begin{array}{c} n \\ x \end{array}\right) p^x{\left(1 - p\right)}^{n - x}, x = 0, 1, 2, ..., n.$$

These probabilities are called binomial probabilities, and the random variable X is said to have **a binomial distribution**.

Summarizing, a binomial experiment satisfies the following properties:

1. A Bernoulli (success-failure) experiment is performed n times.
2. The trials are independent.
3. The probability of success on each trial is a constant p; the probability of failure is $q = 1 - p$.
4. The random variable X counts the number of successes in the n trials.

A binomial distribution will be denoted by the symbol $b\left(n, p\right)$ and we say that the distribution of X is $b\left(n, p\right)$. The constants n and p are called **the parameters of the binomial distribution**, they correspond to the number n of independent trials and the probability p of success on each trial. Thus, if we say that the distribution of X is $b\left(12, 14\right)$, we mean that X is the number of successes in n =12 Bernoulli trials with probability $p = \frac{1}{4}$ of success on each trial.

Example 1.14

In the instant lottery with 20% winning tickets, if X is equal to the number of winning tickets among n =8 that are purchased, the probability of purchasing 2 winning tickets is

$$f(2) = P(X = 2) = \left(\begin{array}{c} 8 \\ 2 \end{array} \right) (0.2)^2 (0.8)^6 = 0.2936.$$

The distribution of the random variable X is $b(8, 0.2)$.

Example 1.15

Leghorn chickens are raised for lying eggs. If p =0.5 is the probability of female chick hatching, assuming independence, the probability that there are exactly 6 females out of 10 newly hatches chicks selected at random is

$$\left(\begin{array}{c} 10 \\ 6 \end{array} \right) \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^4 = P(X \leq 6) - P(X \leq 5) = 0.8281 - 0.6230 = 0.2051.$$

Since

$$P(X \leq 6) = 0.8281$$

and

$$P(X \leq 5) = 0.6230,$$

which are tabularized values, the probability of at least 6 females chicks is

$$\sum_{x=6}^{10} \left(\begin{array}{c} 10 \\ x \end{array} \right) \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = 1 - P(X \leq 5) = 1 - 0.6230 = 0.3770.$$

Example 1.16

Suppose that we are in those rare times when 65% of the American public approve of the way the President of The United states is handling his job. Take a random sample of n =8 Americans and let Y equal the number who give approval. Then the distribution of Y is $b(8, 0.65)$. To find

$$P(Y \geq 6)$$

note that

$$P(Y \geq 6) = P(8 - Y \leq 8 - 6) = P(X \leq 2),$$

where

$$X = 8 - Y$$

counts the number who disapprove. Since $q = 1 - p = 0.35$ equals the probability if disapproval by each person selected, the distribution of X is $b(8, 0.35)$. From the tables, since

$$P(X \leq 2) = 0.4278$$

it follows that

$$P(Y \geq 6)\, 0.4278.$$

Similarly,

$$P(Y \leq 5) = P(8 - Y \geq 8 - 5) = P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.4278 = 0.5722$$

and

$$P(Y=5)=P(8-Y=8-5)=P(X=3)=P(X\leq 3)-P(X\leq 2)=0.7064-0.4278=0.2786.$$

RECALL THAT: if n is a positive integer, then

$$(a+b)^n=\sum_{x=0}^{n}\left(\begin{array}{c}x\\n\end{array}\right)b^x a^{n-x}.$$

Thus the sum of the binomial probabilities, if we use the above binomial expansion with $b=p$ and $a=1-p$, is

$$\sum_{x=0}^{n}\left(\begin{array}{c}n\\x\end{array}\right)p^x(1-p)^{n-x}=[(1-p)+p]^n=1,$$

A result that had to follow from the fact that $f(x)$ is a p.d.f. We use the binomial expansion to find the mean and the variance of the binomial random variable X that is $b(n,p)$. The mean is given by

$$\mu=E(X)=\sum_{x=0}^{n}x\frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x}. \tag{1.13}$$

Since the first term of this sum is equal to zero, this can be written as

$$\mu=\sum_{x=0}^{n}\frac{n!}{(x-1)!\,(n-x)!}p^x(1-p)^{n-x}. \tag{1.14}$$

because $x/x!=1/(x-1)!$ when $x>0$.

To find the variance, we first determine the second factorial moment $E[X(X-1)]$:

$$E[X(X-1)]=\sum_{x=0}^{n}x(x-1)\frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x}. \tag{1.15}$$

The first two terms in this summation equal zero; thus we find that

$$E[X(X-1)]=\sum_{x=2}^{n}\frac{n!}{(x-2)!\,(n-x)!}p^x(1-p)^{n-x}.$$

After observing that $x(x-1)/x!=1/(x-2)!$ when $x>1$. Letting $k=x-2$, we obtain

$$E[X(X-1)]=\sum_{x=0}^{n-2}\frac{n!}{k!\,(n-k-2)!}p^{k+2}(1-p)^{n-k-2}.=n(n-1)p^2\sum_{x=0}^{n-2}\frac{(n-2)!}{k!\,(n-2-k)!}p^k(1-p)^{n-2-k}.$$

Since the last summand is that of the binomial p.d.f. $b(n-2,p)$, we obtain

$$E[X(X-1)]=n(n-1)p^2.$$

Thus,

$$\begin{aligned}\sigma^2&=Var(X)=E\left(X^2\right)-[E(X)]^2=E[X(X-1)]+E(X)-[E(X)]^2\\&=n(n-1)p^2+np-(np)^2=-np^2+np=np(1-p).\end{aligned}$$

Summarizing,
if X is $b(n,p)$, we obtain

$$\mu=np,\sigma^2=np(1-p)=npq,\sigma=\sqrt{np(1-p)}.$$

NOTE THAT: When p is the probability of success on each trial, the expected number of successes in n trials is np, a result that agrees with most of our intuitions.

1.5 GEOMETRIC DISTRIBUTION[5]

1.5.1 GEOMETRIC DISTRIBUTION

To obtain a binomial random variable, we observed a sequence of n Bernoulli trials and counted the number of successes. Suppose now that we do not fix the number of Bernoulli trials in advance but instead continue to observe the sequence of Bernoulli trials until a certain number r, of successes occurs. **The random variable of interest is the number of trials needed to observe the rth success**.

Let first discuss the problem when r =1. That is, consider a sequence of Bernoulli trials with probability p of success. This sequence is observed until the first success occurs. Let X denot the trial number on which the first success occurs.

For example, if F and S represent failure and success, respectively, and the sequence starts with F,F,F,S,..., then X =4. Moreover, because the trials are independent, the probability of such sequence is

$$P(X=4)=(q)(q)(q)(p)=q^3p=(1-p)^3p.$$

In general, the p.d.f. $f(x)=P(X=x)$, of X is given by $f(x)=(1-p)^{x-1}p$, $x=1,2,...$, because there must be x -1 failures before the first success that occurs on trail x. We say that X has **a geometric distribution**.

RECALL THAT: for a geometric series, the sum is given by

$$\sum_{k=0}^{\infty} ar^k = \sum_{k=1}^{\infty} ar^{k-1} = \frac{a}{1-r},$$

when $|r|<1$.

Thus,

$$\sum_{x=1}^{\infty} f(x) = \sum_{x=1}^{\infty} (1-p)^{k-1}p = \frac{p}{1-(1-p)} = 1,$$

so that $f(x)$ does satisfy the properties of a p.d.f..

From the sum of geometric series we also note that, when k is an integer,

$$P(X>k) = \sum_{x=k+1}^{\infty} (1-p)^{x-1}p = \frac{(1-p)^k p}{1-(1-p)} = (1-p)^k = q^k,$$

and thus the value of the distribution function at a positive integer k is

$$P(X\le k) = \sum_{x=k+1}^{\infty} (1-p)^{x-1}p = 1-P(X>k) = 1-(1-p)^k = 1-q^k.$$

Example 1.17

Some biology students were checking the eye color for a large number of fruit flies. For the individual fly, suppose that the probability of white eyes is 14 and the probability of red eyes is 34

[5] This content is available online at <http://cnx.org/content/m13124/1.3/>.

, and that we may treat these flies as independent Bernoulli trials. The probability that at least four flies have to be checked for eye color to observe a white-eyed fly is given by

$$P\left(X \geq 4\right) = P\left(X > 3\right) = q^3 = \left(\frac{3}{4}\right)^3 = 0.422.$$

The probability that at most four flies have to be checked for eye color to observe a white-eyed fly is given by

$$P\left(X \leq 4\right) = 1 - q^4 = 1 - \left(\frac{3}{4}\right)^4 = 0.684.$$

The probability that the first fly with white eyes is the fourth fly that is checked is

$$P\left(X = 4\right) = q^{4-1}p = \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right) = 0.105.$$

It is also true that

$$P\left(X = 4\right) = P\left(X \leq 4\right) - P\left(X \leq 3\right) = \left[1 - \left(\frac{3}{4}\right)^4\right] - \left[1 - \left(\frac{3}{4}\right)^3\right] = \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right).$$

In general,

$$f\left(x\right) = P\left(X = x\right) = \left(\frac{3}{4}\right)^{x-1} \left(\frac{1}{4}\right), x = 1, 2, 3, ...$$

To find a mean and variance for the geometric distribution, let use the following results about the sum and the first and second derivatives of a geometric series. For $-1 < r < 1$, let

$$g\left(r\right) = \sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}.$$

Then

$$g\text{'}\left(r\right) = \sum_{k=1}^{\infty} akr^{k-1} = \frac{a}{\left(1-r\right)^2},$$

and

$$g\text{''}\left(r\right) = \sum_{k=2}^{\infty} ak\left(k-1\right)r^{k-2} = \frac{2a}{\left(1-r\right)^3}.$$

If X has a geometric distribution and $0 < p < 1$, then the mean of X is given by

$$E\left(X\right) = \sum_{x=1}^{\infty} xq^{x-1}p = \frac{p}{\left(1-q\right)^2} = \frac{1}{p}, \tag{1.16}$$

using the formula for $g\text{'}\left(x\right)$ with $a = p$ and $r = q$.

NOTE: for example, that if p =1/4 is the probability of success, then

$$E\left(X\right) = 1/\left(1/4\right) = 4$$

trials are needed on the average to observe a success.

To find the variance of X, let first find the second factorial moment $E\left[X\left(X-1\right)\right]$. We have

$$E\left[X\left(X-1\right)\right]=\sum_{x=1}^{\infty}x\left(x-1\right)q^{x-1}p=\sum_{x=1}^{\infty}pqx\left(x-1\right)q^{x-2}=\frac{2pq}{\left(1-q\right)^{3}}=\frac{2q}{p^{2}}.$$

Using formula for $g"\left(x\right)$ with $a=pq$ and $r=q$. Thus the variance of X is

$$Var\left(X\right)=E\left(X^{2}\right)-\left[E\left(X\right)\right]^{2}=\left\{E\left[X\left(X-1\right)\right]+E\left(X\right)\right\}-\left[E\left(X\right)\right]^{2}=$$
$$=\tfrac{2q}{p^{2}}+\tfrac{1}{p}-\tfrac{1}{p^{2}}=\tfrac{2q+p-1}{p^{2}}=\tfrac{1-p}{p^{2}}.$$

The standard deviation of X is

$$\sigma=\sqrt{\left(1-p\right)/p^{2}}.$$

Example 1.18

Continuing with example 1 (Example 1.17), with p =1/4, we obtain

$$\mu=\tfrac{1}{1/4}=4,$$
$$\sigma^{2}=\tfrac{3/4}{\left(1/4\right)^{2}}=12,$$

and

$$\sigma=\sqrt{12}=3.464.$$

SEE ALSO: Binomial Distribution (Section 1.4.1.1.1)

SEE ALSO: Poisson Distribution (Section 1.6.1: POISSON DISTRIBUTION)

1.6 POISSON DISTRIBUTION[6]

1.6.1 POISSON DISTRIBUTION

Some experiments results in counting the number of times particular events occur in given times of on given physical objects. For example, we would count the number of phone calls arriving at a switch board between 9 and 10 am, the number of flaws in 100 feet of wire, the number of customers that arrive at a ticket window between 12 noon and 2 pm, or the number of defects in a 100-foot roll of aluminum screen that is 2 feet wide. Each count can be looked upon as a random variable associated with an approximate Poisson process provided the conditions in the definition below are satisfied.

Definition 5: POISSON PROCCESS

Let the number of changes that occur in a given continuous interval be counted. We have **an approximate Poisson process** with parameter $\lambda>0$ if the following are satisfied:

1. The number of changes occurring in nonoverlapping intervals are independent.
2. The probability of exactly one change in a sufficiently short interval of length h is approximately λh .
3. The probability of two or more changes in a sufficiently short interval is essentially zero.

[6] This content is available online at <http://cnx.org/content/m13125/1.3/>.

Suppose that an experiment satisfies the three points of an approximate Poisson process. Let X denote the number of changes in an interval of "length 1" (where "length 1" represents one unit of the quantity under consideration). We would like to find an approximation for $P(X = x)$, where x is a nonnegative integer. To achieve this, we partition the unit interval into n subintervals of equal length $1/n$. If N is sufficiently large (i.e., much larger than x), one shall approximate the probability that x changes occur in this unit interval by finding the probability that one change occurs exactly in each of exactly x of these n subintervals. The probability of one change occurring in any one subinterval of length $1/n$ is approximately $\lambda(1/n)$ by condition (2). The probability of two or more changes in any one subinterval is essentially zero by condition (3). So for each subinterval, exactly one change occurs with a probability of approximately $\lambda(1/n)$. Consider the occurrence or nonoccurrence of a change in each subinterval as a Bernoulli trial. By condition (1) we have a sequence of n Bernoulli trials with probability p approximately equal to $\lambda(1/n)$. Thus an approximation for $P(X = x)$ is given by the binomial probability

$$\frac{n!}{x!\,(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}.$$

In order to obtain a better approximation, choose a large value for n. If n increases without bound, we have that

$$\underset{n\to\infty}{lim}\frac{n!}{x!\,(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}=\underset{n\to\infty}{lim}\frac{n\,(n-1)\ldots(n-x+1)}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}.$$

Now, for fixed x, we have

$$\begin{aligned}&\underset{n\to\infty}{lim}\tfrac{n(n-1)\ldots(n-x+1)}{n^x}=\underset{n\to\infty}{lim}\left[1\left(1-\tfrac{1}{n}\right)\ldots\left(1-\tfrac{x-1}{n}\right)\right]=1,\\&\underset{n\to\infty}{lim}\left(1-\tfrac{\lambda}{n}\right)^n=e^{-\lambda},\end{aligned}$$

and

$$\underset{n\to\infty}{lim}\left(1-\frac{\lambda}{n}\right)^{-x}=1.$$

Thus,

$$\underset{n\to\infty}{lim}\frac{n!}{x!\,(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}=\frac{\lambda^x e^{-\lambda}}{x!}=P(X=x),$$

approximately. The distribution of probability associated with this process has a special name.

Definition 6: POISSON DISTRIBUTION

We say that the random variable X has **a Poisson distribution** if its p.d.f. is of the form

$$f(x)=\frac{\lambda^x e^{-\lambda}}{x!},x=0,1,2,...,$$

where $\lambda > 0$.

It is easy to see that $f(x)$ enjoys the properties pf a p.d.f. because clearly $f(x) \geq 0$ and, from the Maclaurin's series expansion of e^{λ} , we have

$$\sum_{x=0}^{\infty}\frac{\lambda^x e^{-\lambda}}{x!}=e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}=e^{-\lambda}e^{\lambda}=1.$$

To discover the exact role of the parameter $\lambda > 0$, let us find some of the characteristics of the Poisson distribution . The mean for the Poisson distribution is given by

$$E(X)=\sum_{x=0}^{\infty}x\frac{\lambda^x e^{-\lambda}}{x!}=e^{-\lambda}\sum_{x=1}^{\infty}x\frac{\lambda^x}{(x-1)!},$$

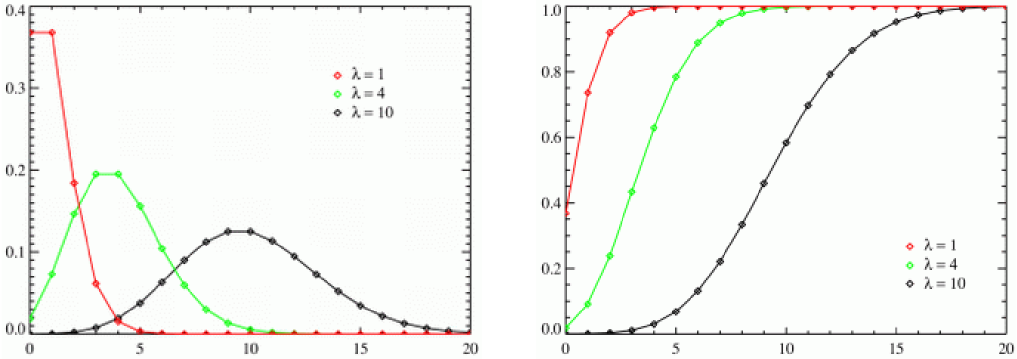
because $(0)\, f(0) = 0$ and $x/x! = 1/(x-1)!$, when $x > 0$.

If we let $k = x - 1$, then

$$E(X) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k}}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

That is, **the parameter λ is the mean of the Poisson distribution**. On the Figure 1 (Figure 1.2: Poisson Distribution) is shown the p.d.f. and c.d.f. of the Poisson Distribution for $\lambda = 1, \lambda = 4, \lambda = 10$.

Poisson Distribution



(a)

(b)

Figure 1.2: The p.d.f. and c.d.f. of the Poisson Distribution for $\lambda = 1, \lambda = 4, \lambda = 10$. (a) The p.d.f. function. (b) The c.d.f. function.

To find the variance, we first determine the second factorial moment $E[X(X-1)]$. We have,

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1) \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!},$$

because $(0)(0-1) f(0) = 0, (1)(1-1) f(1) = 0$, and $x(x-1)/x! = 1/(x-2)!$, when $x > 1$.

If we let $k = x - 2$, then

$$E[X(X-1)] = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+2}}{k!} = \lambda^2 e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{x}}{k!} = \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2.$$

Thus,

$$Var(X) = E\left(X^2\right) - [E(X)]^2 = E[X(X-1)] + E(X) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

That is, for the Poisson distribution, $\mu = \sigma^2 = \lambda$.

Example 1.19

Let X have a Poisson distribution with a mean of $\lambda = 5$, (it is possible to use the tabularized Poisson distribution).

$P(X \leq 6) = \sum_{x=0}^{6} \frac{5^x e^{-5}}{x!} = 0.762,$

$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.616 = 0.384,$

and

$$P\left(X=6\right)=P\left(X\leq 6\right)-P\left(X\leq 5\right)=0.762-0.616=0.146.$$

Example 1.20

Telephone calls enter a college switchboard on the average of two every 3 minutes. If one assumes an approximate Poisson process, what is the probability of five or more calls arriving in a 9-minute period? Let X denotes the number of calls in a 9-minute period. We see that $E\left(X\right)=6$; that is, on the average, sic calls will arrive during a 9-minute period. Thus using tabularized data,

$$P\left(X\geq 5\right)=1-P\left(X\leq 4\right)=1-\sum_{x=0}^{4}\frac{6^{x}e^{-6}}{x!}=1-0.285=0.715.$$

Note That: Not only is the Poisson distribution important in its own right, but it can also be used to approximate probabilities for a binomial distribution.

If X has a Poisson distribution with parameter λ , we saw that with n large,

$$P\left(X=x\right)\approx\left(\begin{array}{c}n\\x\end{array}\right)\left(\frac{\lambda}{n}\right)^{x}\left(1-\frac{\lambda}{n}\right)^{n-x},$$

where, $p=\lambda/n$ so that $\lambda=np$ in the above binomial probability. That is, if X has the binomial distribution $b\left(n,p\right)$ with large n, then

$$\frac{\left(np\right)^{x}e^{-np}}{x!}=\left(\begin{array}{c}n\\x\end{array}\right)p^{x}\left(1-p\right)^{n-x}.$$

This approximation is reasonably good if n is large. But since λ was fixed constant in that earlier argument, p should be small since $np=\lambda$. In particular, the approximation is quite accurate if $n\geq 20$ and $p\leq 0.05$, and it is very good if $n\geq 100$ and $np\leq 10$.

Example 1.21

A manufacturer of Christmas tree bulbs knows that 2% of its bulbs are defective. Approximate the probability that a box of 100 of these bulbs contains at most three defective bulbs. Assuming independence, we have binomial distribution with parameters p=0.02 and n=100. The Poisson distribution with $\lambda=100\left(0.02\right)=2$ gives

$$\sum_{x=0}^{3}\frac{2^{x}e^{-2}}{x!}=0.857,$$

using the binomial distribution, we obtain, after some tedious calculations,

$$\sum_{x=0}^{3}\left(\begin{array}{c}100\\x\end{array}\right)\left(0.02\right)^{x}\left(0.98\right)^{100-x}=0.859.$$

Hence, in this case, the Poisson approximation is extremely close to the true value, but much easier to find.

Chapter 2

Continuous Distributions

2.1 CONTINUOUS DISTRIBUTION[1]

2.1.1 CONTINUOUS DISTRIBUTION

2.1.1.1 RANDOM VARIABLES OF THE CONTINUOUS TYPE

Random variables whose spaces are not composed of a countable number of points but are intervals or a union of intervals are said to be of the **continuous type**. Recall that the relative frequency histogram $h(x)$ associated with n observations of a random variable of that type is a nonnegative function defined so that the total area between its graph and the x axis equals one. In addition, $h(x)$ is constructed so that the integral

$$\int_a^b h(x)\,dx \qquad (2.1)$$

is an estimate of the probability $P(a < X < b)$, where the interval (a, b) is a subset of the space R of the random variable X.

Let now consider what happens to the function $h(x)$ in the limit, as n increases without bound and as the lengths of the class intervals decrease to zero. It is to be hoped that $h(x)$ will become closer and closer to some function, say $f(x)$, that gives the true probabilities , such as $P(a < X < b)$, through the integral

$$P(a < X < b) = \int_a^b f(x)\,dx. \qquad (2.2)$$

Definition 7: PROBABILITY DENSITY FUNCTION

1. Function f(x) is a nonnegative function such that the total area between its graph and the x axis equals one.
2. The probability $P(a < X < b)$ is the area bounded by the graph of $f(x)$, the x axis, and the lines $x = a$ and $x = b$.
3. We say that **the probability density function (p.d.f.)** of the random variable X of the continuous type, with space R that is an interval or union of intervals, is an integrable function $f(x)$ satisfying the following conditions:

- $f(x) > 0$, x belongs to R,
- $\int_R f(x)\,dx = 1$,

[1] This content is available online at <http://cnx.org/content/m13127/1.4/>.

- The probability of the event A belongs to R is $P(X) \in A \int\limits_A f(x)\,dx$.

Example 2.1

Let the random variable X be the distance in feet between bad records on a used computer tape. Suppose that a reasonable probability model for X is given by the p.d.f.

$$f(x)\,\frac{1}{40}e^{-x/40}, 0 \leq x < \infty.$$

NOTE THAT: $R = (x : 0 \leq x < \infty)$ and $f(x)$ for x belonging to R,

$$\int_R f(x)\,dx = \int\limits_0^\infty \frac{1}{40}e^{-x/40}dx = \lim_{b\to\infty}\left[e^{-x/40}\right]_0^b = 1 - \lim_{b\to\infty} e^{-b/40} = 1.$$

The probability that the distance between bad records is greater than 40 feet is

$$P(X > 40) = \int\limits_{40}^\infty \frac{1}{40}e^{-x/40}dx = e^{-1} = 0.368.$$

The p.d.f. and the probability of interest are depicted in FIG.1 (Figure 2.1).

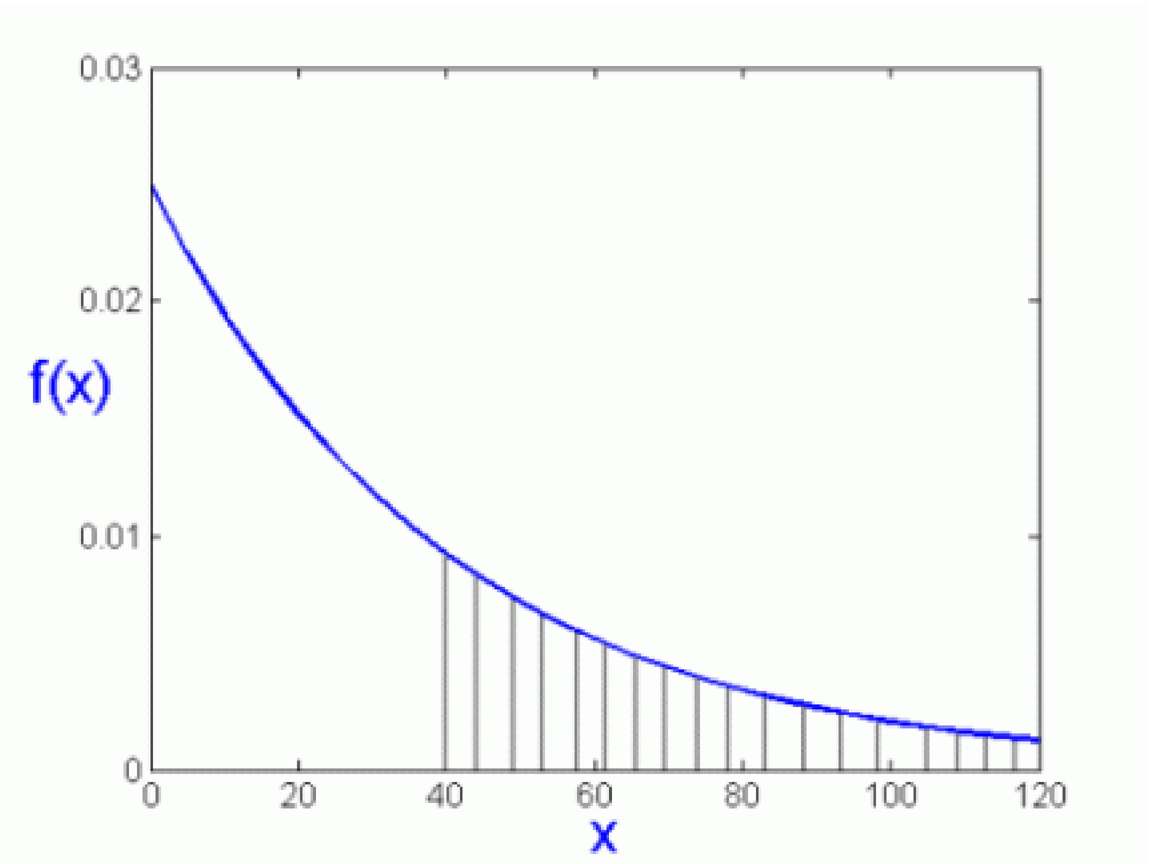


Figure 2.1: The p.d.f. and the probability of interest.

We can avoid repeated references to the space R of the random variable X, one shall adopt the same convention when describing probability density function of the continuous type as was in the discrete case.

Let extend the definition of the p.d.f. $f(x)$ to the entire set of real numbers by letting it equal zero when, x belongs to R. For example,

$$f(x) = \{ \begin{array}{c} \frac{1}{40}e^{-x/40} \\ 0, elsewhere, \end{array} ,0 \leq x < \infty,$$

has the properties of a p.d.f. of a continuous-type random variable x having support $(x : 0 \leq x < \infty)$. It will always be understood that $f(x) = 0$, when x belongs to R, even when this is not explicitly written out.

Definition 8: PROBABILITY DENSITY FUNCTION

1. The distribution function of a random variable X of the continuous type, is defined in terms of

the p.d.f. of X, and is given by

$$F(x) = P(X \leq x) = \int\limits_{-\infty}^{x} f(t)\,dt.$$

2. For the fundamental theorem of calculus we have, for x values for which the derivative $F'(x)$ exists, that $F'(x)=f(x)$.

Example 2.2

continuing with Example 1 (Example 2.1)

If the p.d.f. of X is

$$f(x) = \{ \begin{array}{l} 0, -\infty < x < 0, \\ \frac{1}{40}e^{-x/40}, 0 \leq x < \infty, \end{array}$$

The distribution function of X is $F(x) = 0$ for $x \leq 0$

$$F(x) = \int\limits_{-\infty}^{x} f(t)\,dt = \int\limits_{0}^{x} \frac{1}{40}e^{-t/40}dt = -e^{-t/40}|_0^x = 1 - e^{-x/40}.$$

Note That:

$$F(x) = \{ \begin{array}{l} 0, -\infty < x < 0, \\ \frac{1}{40}e^{-x/40}, 0 < x < \infty. \end{array}$$

Also $F'(0)$ does not exist. Since there are no steps or jumps in a distribution function $F(x)$, of the continuous type, it must be true that

$$P(X = b) = 0$$

for all real values of b. This agrees with the fact that the integral

$$\int\limits_{a}^{b} f(x)\,dx$$

is taken to be zero in calculus. Thus we see that

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = F(b) - F(a),$$

provided that X is a random variable of the continuous type. Moreover, we can change the definition of a p.d.f. of a random variable of the continuous type at a finite (actually countable) number of points without alerting the distribution of probability.

For illustration,

$$f(x) = \{ \begin{array}{l} 0, -\infty < x < 0, \\ \frac{1}{40}e^{-x/40}, 0 \leq x < \infty, \end{array}$$

and

$$f(x) = \{ \begin{array}{l} 0, -\infty < x \leq 0, \\ \frac{1}{40}e^{-x/40}, 0 < x < \infty, \end{array}$$

are equivalent in the computation of probabilities involving this random variable.

Example 2.3
Let Y be a continuous random variable with the p.d.f. $g(y) = 2y$, $0 < y < 1$. The distribution function of Y is defined by

$$G(y) = \left(\begin{array}{l} 0, y < 0, \\ 1, y \geq 1, \\ \int\limits_0^y 2tdt = y^2, 0 \leq y < 1. \end{array} \right.$$

Figure 2 (Figure 2.2) gives the graph of the p.d.f. $g(y)$ and the graph of the distribution function $G(y)$.

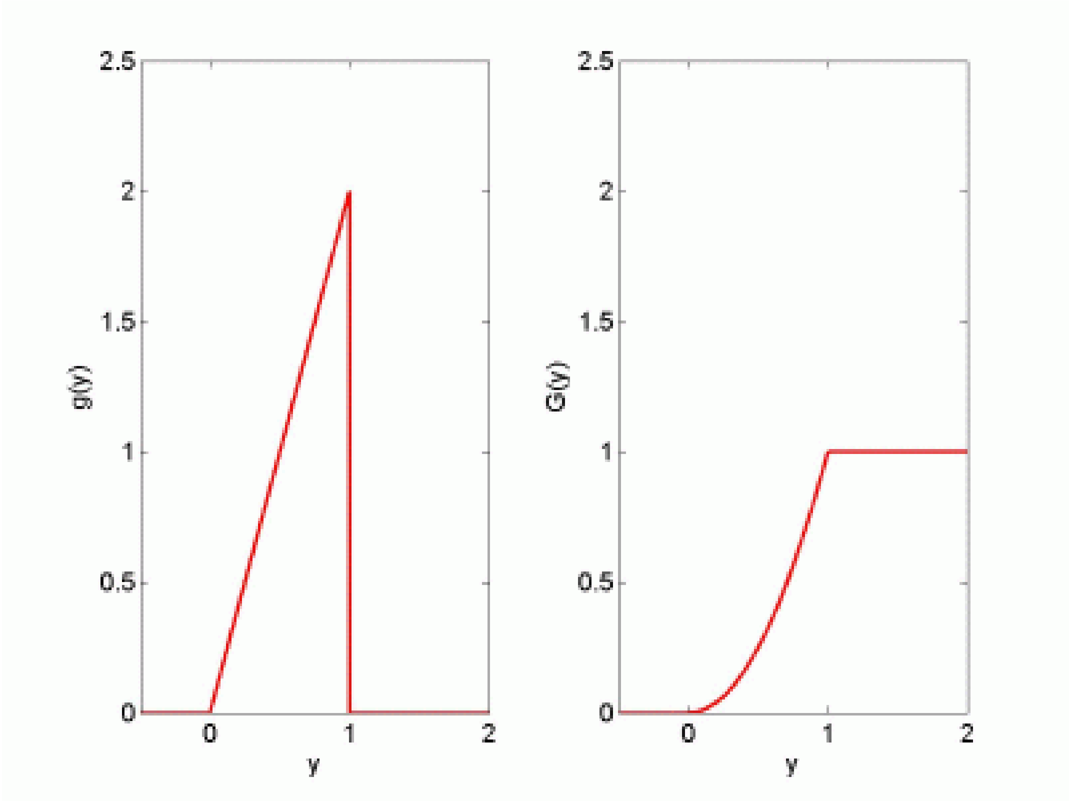


Figure 2.2: The p.d.f. and the probability of interest.

For illustration of computations of probabilities, consider

$$P\left(\frac{1}{2} < Y \leq \frac{3}{4}\right) = G\left(\frac{3}{4}\right) - G\left(\frac{1}{2}\right) = \left(\frac{3}{4}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{5}{16}$$

and

$$P\left(\frac{1}{4} \leq Y < 2\right) = G(2) - G\left(\frac{1}{4}\right) = 1 - \left(\frac{1}{4}\right)^2 = \frac{15}{16}.$$

Recall That: The p.d.f. $f(x)$ of a random variable of the discrete type is bounded by one because $f(x)$ gives a probability, namely $f(x) = P(X = x)$.

For random variables of the continuous type, the p.d.f. does not have to be bounded. The restriction is that the area between the p.d.f. and the x axis must equal one. Furthermore, it should be noted that the p.d.f. of a random variable X of the continuous type does not need to be a continuous function.

For example,

$$f(x) = \{ \begin{array}{l} \frac{1}{2}, 0 < x < 1 or 2 < x < 3, \\ 0, elsewhere, \end{array}$$

enjoys the properties of a p.d.f. of a distribution of the continuous type, and yet $f(x)$ had discontinuities at $x = 0, 1, 2,$ and 3. However, the distribution function associates with a distribution of the continuous type is always a continuous function. For continuous type random variables, the definitions associated with mathematical expectation are the same as those in the discrete case except that integrals replace summations.

FOR ILLUSTRATION, let X be a random variable with a p.d.f. $f(x)$. The **expected value** of X or **mean** of X is

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

The **variance** of X is

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx.$$

The **standard deviation** of X is

$$\sigma = \sqrt{Var(X)}.$$

Example 2.4

For the random variable Y in the Example 3 (Example 2.3).

$$\mu = E(Y) = \int_0^1 y(2y)\,dy = \left[\left(\frac{2}{3}y^3\right)\right]_0^1 = \frac{2}{3}$$

and

$$\begin{array}{l} \sigma^2 = Var(Y) = E\left(Y^2\right) - \mu^2 \\ = \int_0^1 y^2(2y)\,dy - \left(\frac{2}{3}\right)^2 = \left[\left(\frac{1}{2}y^4\right)\right]_0^1 - \frac{4}{9} = \frac{1}{18}. \end{array}$$

2.2 THE UNIFORM AND EXPONENTIAL DISTRIBUTIONS[2]

2.2.1 THE UNIFORM AND EXPONENTIAL DISTRIBUTIONS

2.2.1.1 The Uniform Distribution

Let the random variable X denote the outcome when a point is selected at random from the interval $[a, b]$, $-\infty < a < b < \infty$. If the experiment is performed in a fair manner, it is reasonable to assume that the

[2] This content is available online at <http://cnx.org/content/m13128/1.7/>.

probability that the point is selected from the interval $[a, x]$, $a \leq x < b$ is $(x-a)(b-a)$. That is, the probability is proportional to the length of the interval so that the distribution function of X is

$$F(x) = \left(\begin{array}{l} 0, x < a, \\ \frac{x-a}{b-a}, a \leq x < b, \\ 1, b \leq x. \end{array} \right.$$

Because X is a continuous-type random variable, $F\text{'}(x)$ is equal to the p.d.f. of X whenever $F\text{'}(x)$ exists; thus when $a < x < b$, we have

$$f(x) = F\text{'}(x) = 1/(b-a).$$

Definition 9: DEFINITION OF UNIFORM DISTRIBUTION

The random variable X has **a uniform distribution** if its p.d.f. is equal to a constant on its support. In particular, if the support is the interval $[a, b]$, then

$$f(x) = \frac{1}{b = a}, a \leq x \leq b. \tag{2.3}$$

Moreover, one shall say that X is $U(a, b)$. This distribution is referred to as **rectangular** because the graph of $f(x)$ suggest that name. See Figure1. (Figure 2.3) for the graph of $f(x)$ and the distribution function F(x).

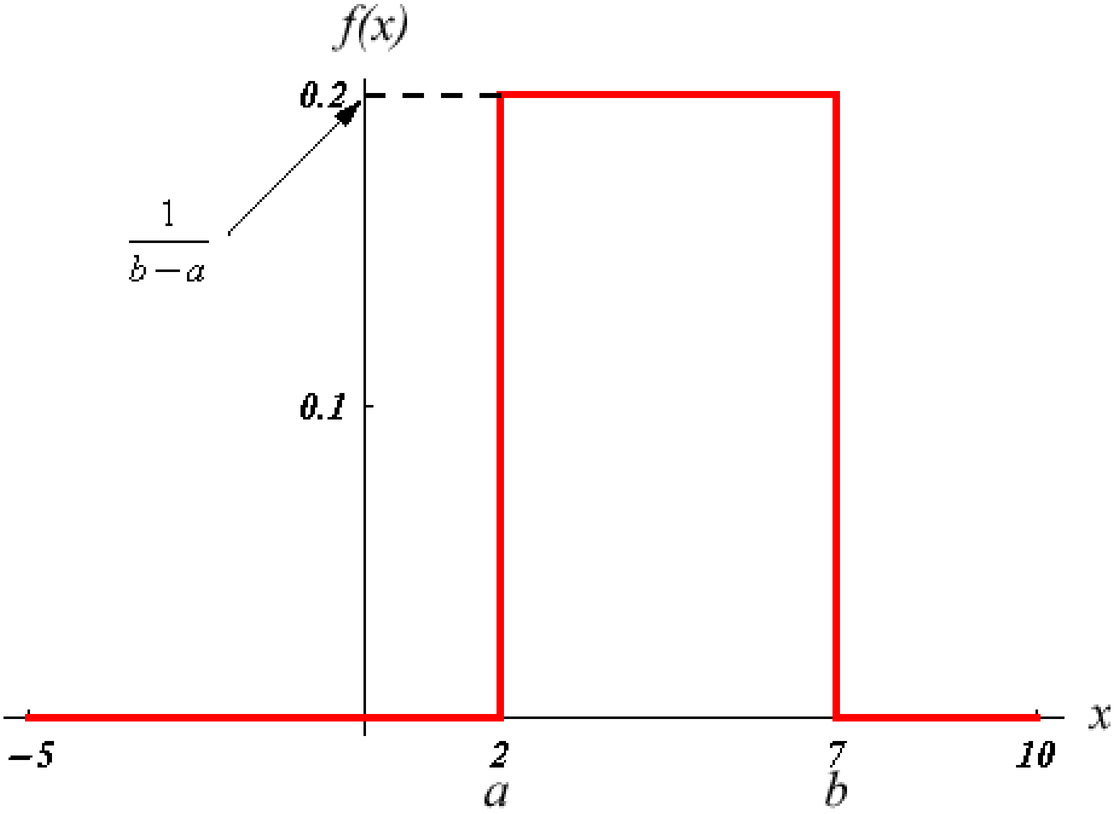


Figure 2.3: The graph of the p.d.f. of the uniform distriution.

Note that: We could have taken $f\left(a\right)=0$ or $f\left(b\right)=0$ without alerting the probabilities, since this is a continuous type distribution, and it can be done in some cases.

The **mean** and **variance** of X are as follows:

$$\mu=\frac{a+b}{2}$$

and

$$\sigma^2=\frac{\left(b-a\right)^2}{12}.$$

An important uniform distribution is that for which a=0 and b =1, namely $U\left(0,1\right)$. If X is $U\left(0,1\right)$, approximate values of X can be simulated on most computers using a random number generator. In fact, it should be called a pseudo-random number generator (see the pseudo-numbers generation (Section 5.3.1: THE IVERSE PROBABILITY METHOD FOR GENERATING RANDOM VARIABLES)) because the programs that produce the random numbers are usually such that if the starting number is known, all subsequent numbers in the sequence may be determined by simple arithmetical operations.

2.2.1.2 An Exponential Distribution

Let turn to the continuous distribution that is related to the Poisson distribution (Section 1.6.1: POISSON DISTRIBUTION). When previously observing a process of the approximate Poisson type, we counted the number of changes occurring in a given interval. This number was a discrete-type random variable with a Poisson distribution. But not only is the number of changes a random variable; **the waiting times** between successive changes are also random variables. However, the latter are of the continuous type, since each of then can assume any positive value.

Let W denote the waiting time until the first change occurs when observing the Poisson process (Definition: "POISSON PROCCESS", p. 17) in which the mean number of changes in the unit interval is λ. Then W is a continuous-type random variable, and let proceed to find its distribution function.

Because this waiting time is nonnegative, the distribution function $F(w) = 0$, $w < 0$. For $w \geq 0$,

$$F(w) = P(W \leq w) = 1 - P(W > w) = 1 - P(no\_changes\_in\_[0, w]) = 1 - e^{-\lambda w},$$

since that was previously discovered that $e^{-\lambda w}$ equals the probability of no changes in an interval of length w is proportional to w, namely, λw. Thus when w >0, the p.d.f. of W is given by

$$F\text{'}(w) = \lambda e^{-\lambda w} = f(w).$$

Definition 10: DEFINITION OF EXPONENTIAL DISTRIBUTION
Let $\lambda = 1/\theta$, then the random variable X has **an exponential distribution** and its p.d.f. id defined by

$$f(x) = \frac{1}{\theta}e^{-x/\theta}, 0 \leq x < \infty, \tag{2.4}$$

where the parameter $\theta > 0$.

Accordingly, the waiting time W until the first change in a Poisson process has an exponential distribution with $\theta = 1/\lambda$. The **mean** and **variance** for the exponential distribution are as follows: $\mu = \theta$ and $\sigma^2 = \theta^2$.

So if λ is the mean number of changes in the unit interval, then

$$\theta = 1/\lambda$$

is the mean waiting for the first change. Suppose that λ=7 is the mean number of changes per minute; then that mean waiting time for the first change is 1/7 of a minute.

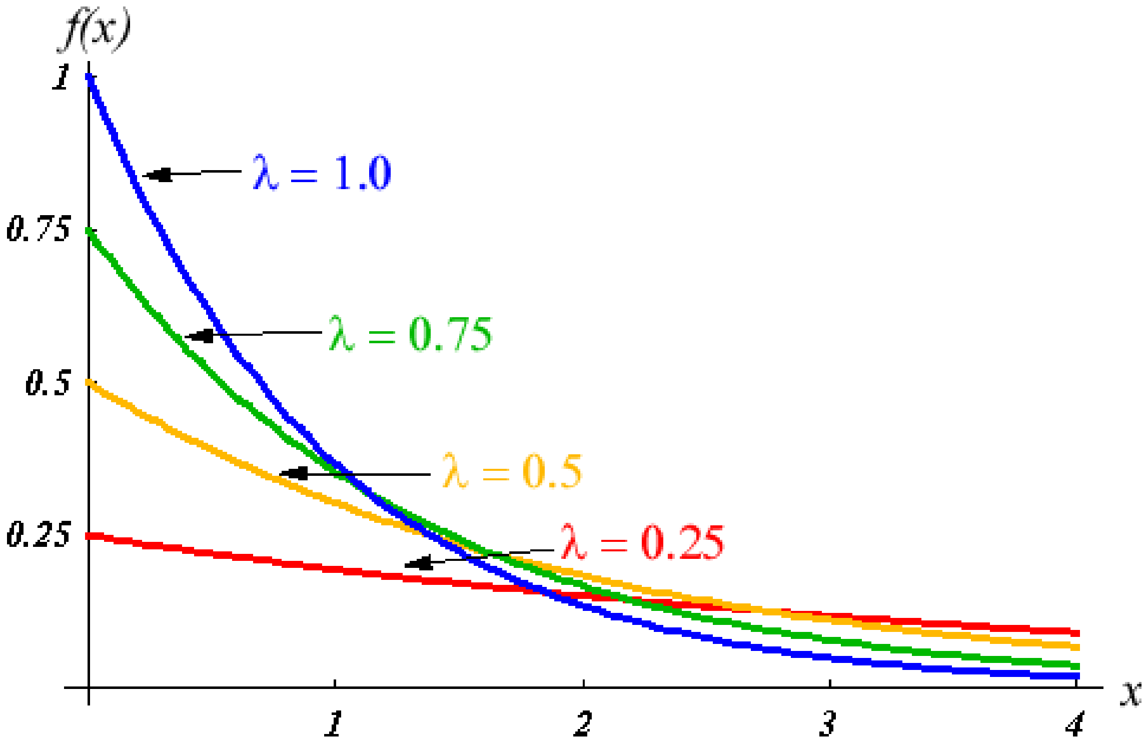


Figure 2.4: The graph of the p.d.f. of the exponential distriution.

Example 2.5
Let X have an exponential distribution with a mean of 40. The p.d.f. of X is

$$f(x) = \frac{1}{40}e^{-x/40}, 0 \leq x < \infty.$$

The probability that X is less than 36 is

$$P(X < 36) = \int_{0}^{36} \frac{1}{40}e^{-x/40}dx = 1 - e^{-36/40} = 0.593.$$

Example 2.6
Let X have an exponential distribution with mean $\mu = \theta$. Then the distribution function of X is

$$F(x) = \{ \begin{array}{l} 0, -\infty < x < 0, \\ 1 - e^{-x/\theta}, 0 \leq x < \infty. \end{array}$$

The p.d.f. and distribution function are graphed in the Figure 3 (Figure 2.5) for θ=5.

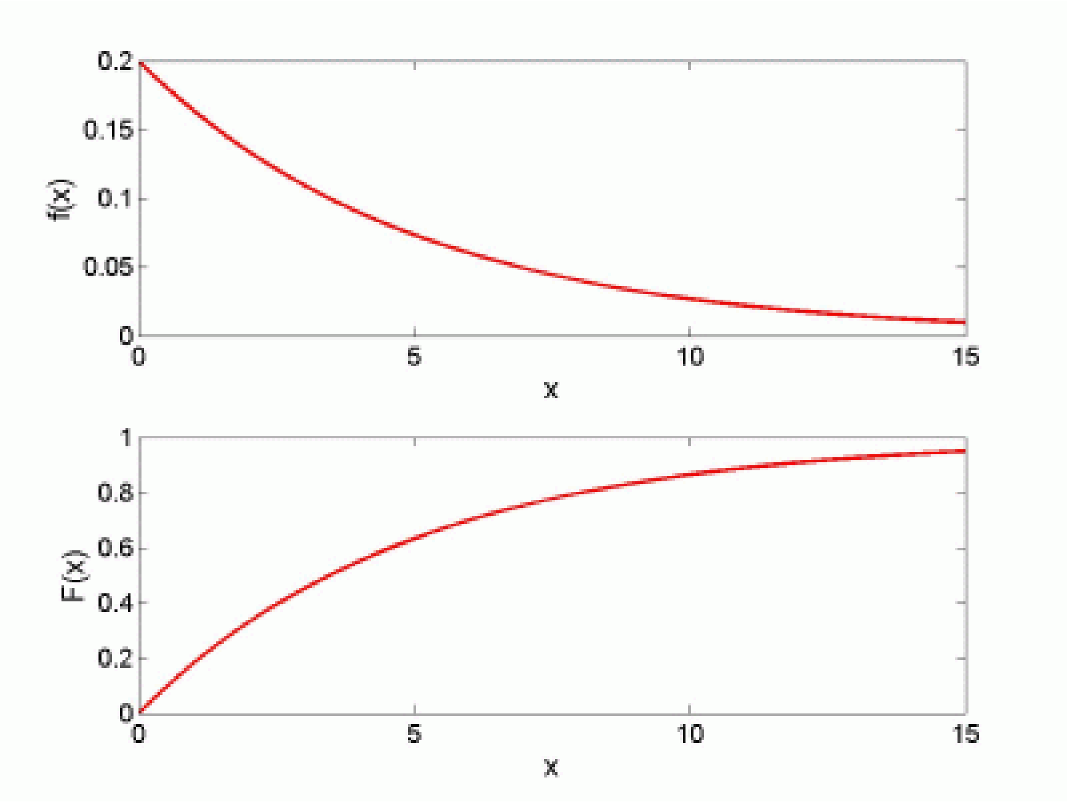


Figure 2.5: The p.d.f. and c.d.f. graphs of the exponential distriution with $\theta = 5$.

NOTE THAT: For an exponential random variable X, we have that

$$P\left(X > x\right) = 1 - F\left(x\right) = 1 - \left(1 - e^{-x/\theta}\right) = e^{-x/\theta}.$$

2.3 THE GAMMA AND CHI-SQUARE DISTRIBUTIONS[3]

2.3.1 GAMMA AND CHI-SQUARE DISTRIBUTIONS

In the (approximate) Poisson process (Definition: "POISSON PROCCESS", p. 17) with mean λ, we have seen that the waiting time until the first change has an exponential distribution (Section 2.2.1.2: An Exponential Distribution). Let now W denote the waiting time until the αth change occurs and let find the distribution of W. The distribution function of W ,when $w \geq 0$ is given by

$$F\left(w\right) = P\left(W \leq w\right) = 1 - P\left(W > w\right) = 1 - P\left(fewer\_than\_\alpha\_changes\_occur\_in\_\left[0, w\right]\right)$$
$$= 1 - \sum_{k=0}^{\alpha-1} \frac{\left(\lambda w\right)^k e^{-\lambda w}}{k!},$$

[3] This content is available online at <http://cnx.org/content/m13129/1.3/>.

since the number of changes in the interval $[0, w]$ has a Poisson distribution with mean λw. Because W is a continuous-type random variable, $F'\left(w\right)$ is equal to the p.d.f. of W whenever this derivative exists. We have, provided w>0, that

$$F'\left(w\right) = \lambda e^{-\lambda w} - e^{-\lambda w}\sum_{k=1}^{\alpha-1}\left[\frac{k(\lambda w)^{k-1}\lambda}{k!} - \frac{(\lambda w)^{k}\lambda}{k!}\right] = \lambda e^{-\lambda w} - e^{-\lambda w}\left[\lambda - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!}\right]$$
$$= \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!}e^{-\lambda w}.$$

2.3.1.1 Gamma Distribution

Definition 11:
1. If $w < 0$, then $F\left(w\right) = 0$ and $F'\left(w\right) = 0$, a p.d.f. of this form is said to be one of the **gamma type**, and the random variable W is said to have **the gamma distribution**.
2. The **gamma function** is defined by

$$\Gamma\left(t\right) = \int_{0}^{\infty} y^{t-1}e^{-y}dy, 0 < t.$$

This integral is positive for $0 < t$, because the integrand id positive. Values of it are often given in a table of integrals. If $t > 1$, integration of gamma fnction of t by parts yields

$$\Gamma\left(t\right) = \left[-y^{t-1}e^{-y}\right]_{0}^{\infty} + \int_{0}^{\infty}\left(t-1\right)y^{t-2}e^{-y}dy = \left(t-1\right)\int_{0}^{\infty} y^{t-2}e^{-y}dy = \left(t-1\right)\Gamma\left(t-1\right).$$

Example 2.7
Let $\Gamma\left(6\right) = 5\Gamma\left(5\right)$ and $\Gamma\left(3\right) = 2\Gamma\left(2\right) = \left(2\right)\left(1\right)\Gamma\left(1\right)$. Whenever $t = n$, a positive integer, we have, be repeated application of $\Gamma\left(t\right) = \left(t-1\right)\Gamma\left(t-1\right)$, that $\Gamma\left(n\right) = \left(n-1\right)\Gamma\left(n-1\right) = \left(n-1\right)\left(n-2\right)...\left(2\right)\left(1\right)\Gamma\left(1\right).$

However,

$$\Gamma\left(1\right) = \int_{0}^{\infty} e^{-y}dy = 1.$$

Thus when n is a positive integer, we have that $\Gamma\left(n\right) = \left(n-1\right)!$; and, for this reason, the gamma is called **the generalized factorial**.

Incidentally, $\Gamma\left(1\right)$ corresponds to 0!, and we have noted that $\Gamma\left(1\right) = 1$, which is consistent with earlier discussions.

2.3.1.1.1 SUMMARIZING

The random variable x has **a gamma distribution** if its p.d.f. is defined by

$$f\left(x\right) = \frac{1}{\Gamma\left(\alpha\right)\theta^{\alpha}}x^{\alpha-1}e^{-x/\theta}, 0 \le x < \infty. \qquad (2.5)$$

Hence, w, the waiting time until the α th change in a Poisson process, has a gamma distribution with parameters α and $\theta = 1/\lambda$.

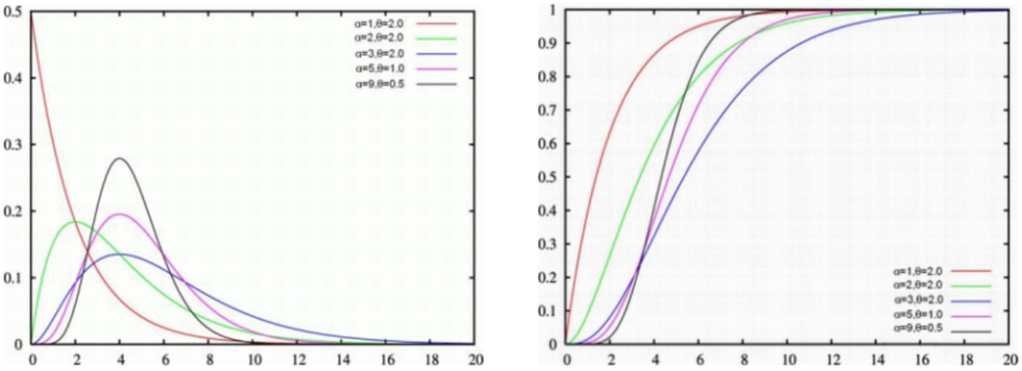
Function $f\left(x\right)$ actually has the properties of a p.d.f., because $f\left(x\right) \ge 0$ and

$$\int_{-\infty}^{\infty} f\left(x\right)dx = \int_{0}^{\infty}\frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma\left(\alpha\right)\theta^{\alpha}}dx,$$

which, by the change of variables $y = x/\theta$ equals

$$\int_0^\infty \frac{(\theta y)^{\alpha-1} e^{-y}}{\Gamma(\alpha)\,\theta^\alpha}\theta dy = \frac{1}{\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}e^{-y}dy = \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1.$$

The mean and variance are: $\mu = \alpha\theta$ and $\sigma^2 = \alpha\theta^2$.



(a) Gamma Distribution

(b)

Figure 2.6: The p.d.f. and c.d.f. graphs of the Gamma Distribution. (a) The c.d.f. graph. (b) The p.d.f. graph.

Example 2.8
Suppose that an average of 30 customers per hour arrive at a shop in accordance with Poisson process. That is, if a minute is our unit, then $\lambda = 1/2$. What is the probability that the shopkeeper will wait more than 5 minutes before both of the first two customers arrive? If X denotes the waiting time in minutes until the second customer arrives, then X has a gamma distribution with $\alpha = 2, \theta = 1/\lambda = 2$. Hence,

$$p(X > 5) = \int_5^\infty \frac{x^{2-1}e^{-x/2}}{\Gamma(2)\,2^2}dx = \int_5^\infty \frac{xe^{-x/2}}{4}dx = \frac{1}{4}\left[(-2)\,xe^{-x/2} - 4e^{-x/2}\right]_5^\infty = \frac{7}{2}e^{-5/2} = 0.287.$$

We could also have used equation with $\lambda = 1/\theta$, because α is an integer

$$P(X > x) = \sum_{k=0}^{\alpha-1} \frac{(x/\theta)^k e^{-x/\theta}}{k!}.$$

Thus, with x=5, α=2, and $\theta = 2$, this is equal to

$$P(X > x) = \sum_{k=0}^{2-1} \frac{(5/2)^k e^{-5/2}}{k!} = e^{-5/2}\left(1 + \frac{5}{2}\right) = \left(\frac{7}{2}\right)e^{-5/2}.$$

2.3.1.2 Chi-Square Distribution

Let now consider the special case of the gamma distribution that plays an important role in statistics.

Definition 12:

Let X have a gamma distribution with $\theta = 2$ and $\alpha = r/2$, where r is a positive integer. If the p.d.f. of X is

$$f(x) = \frac{1}{\Gamma(r/2)\, 2^{r/2}} x^{r/2-1} e^{-x/2}, 0 \leq x < \infty. \tag{2.6}$$

We say that X has **chi-square distribution** with r degrees of freedom, which we abbreviate by saying is $\chi^2(r)$.

The **mean** and the **variance** of this chi-square distributions are

$$\mu = \alpha\theta = \left(\frac{r}{2}\right) 2 = r$$

and

$$\sigma^2 = \alpha\theta^2 = \left(\frac{r}{2}\right) 2^2 = 2r.$$

That is, the mean equals the number of degrees of freedom and the variance equals twice the number of degrees of freedom.

In the fugure 2 (Figure 2.7) the graphs of chi-square p.d.f. for r=2,3,5, and 8 are given.

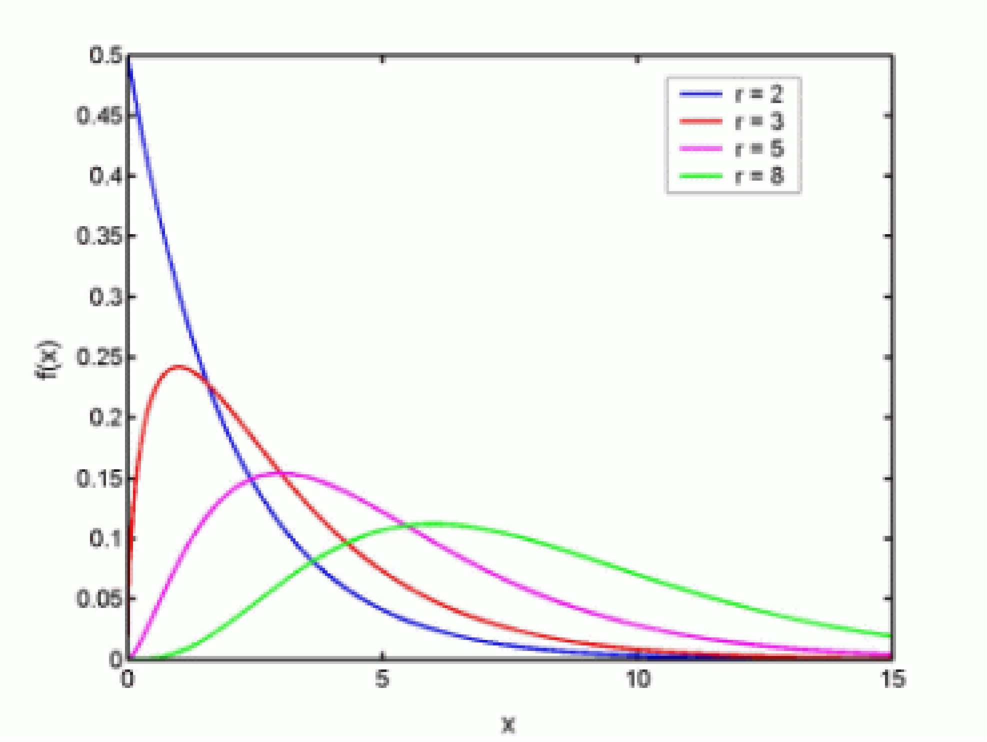


Figure 2.7: The p.d.f. of chi-square distribution for degrees of freedom r=2,3,5,8.

NOTE: the relationship between the mean $\mu = r$, and the point at which the p.d.f. obtains its maximum.

Because the chi-square distribution is so important in applications, tables have been prepared giving the values of the distribution function for selected value of r and x,

$$F(x) = \int_0^x \frac{1}{\Gamma(r/2)\,2^{r/2}} w^{r/2-1} e^{-w/2} dw. \tag{2.7}$$

Example 2.9
Let X have a chi-square distribution with r =5 degrees of freedom. Then, using tabularized values,

$$P(1.145 \leq X \leq 12.83) = F(12.83) - F(1.145) = 0.975 - 0.050 = 0.925$$

and

$$P(X > 15.09) = 1 - F(15.09) = 1 - 0.99 = 0.01.$$

Example 2.10
If X is $\chi^2(7)$, two constants, a and b, such that $P(a < X < b) = 0.95$, are a=1.690 and b=16.01.

Other constants a and b can be found, this above are only restricted in choices by the limited table.

Probabilities like that in Example 4 (Example 2.10) are so important in statistical applications that one uses special symbols for a and b. Let α be a positive probability (that is usually less than 0.5) and let X have a chi-square distribution with r degrees of freedom. Then $\chi^2_\alpha(r)$ is a number such that $P\left[X \geq \chi^2_\alpha(r)\right] = \alpha$

That is, $\chi^2_\alpha(r)$ is the 100(1-α) percentile (or upper 100a percent point) of the chi-square distribution with r degrees of freedom. Then the 100α percentile is the number $\chi^2_{1-\alpha}(r)$ such that $P\left[X \leq \chi^2_{1-\alpha}(r)\right] = \alpha$. This is, the probability to the right of $\chi^2_{1-\alpha}(r)$ is 1-α. SEE fugure 3 (Figure 2.8).

Example 2.11
Let X have a chi-square distribution with seven degrees of freedom. Then, using tabularized values, $\chi^2_{0.05}(7) = 14.07$ and $\chi^2_{0.95}(7) = 2.167$. These are the points that are indicated on Figure 3.

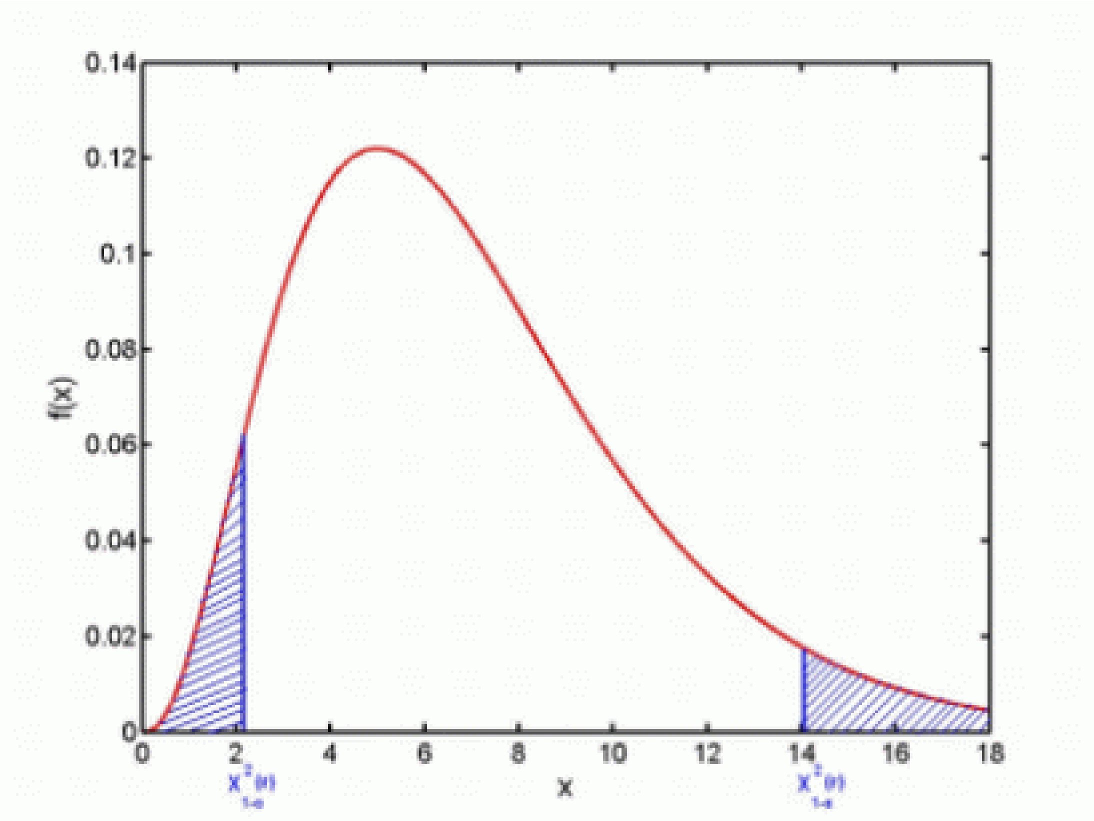


Figure 2.8: $\chi^2_{0.05}(7) = 14.07$ and $\chi^2_{0.95}(7) = 2.167$.

2.4 NORMAL DISTRIBUTION[4]

2.4.1 NORMAL DISTRIBUTION

The normal distribution is perhaps the most important distribution in statistical applications since many measurements have (approximate) normal distributions. One explanation of this fact is the role of the normal distribution in the Central Theorem.

Definition 13:

1. The random variable X has a normal distribution if its p.d.f. is defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty, \tag{2.8}$$

where μ and σ^2 are parameters satisfying $-\infty < \mu < \infty, 0 < \sigma < \infty$, and also where $exp[v]$ means e^v.

2. Briefly, we say that X is $N(\mu, \sigma^2)$

[4] This content is available online at <http://cnx.org/content/m13130/1.4/>.

2.4.1.1 Proof of the p.d.f. properties

Clearly, $f(x) > 0$. Let now evaluate the integral:

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

showing that it is equal to 1. In the integral, change the variables of integration by letting $z = (x-\mu)/\sigma$. Then,

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz,$$

since $I > 0$, if $I^2 = 1$, then $I = 1$.

Now

$$I^2 = \frac{1}{2\pi}\left[\int_{-\infty}^{\infty} e^{-x^2/2} dx\right]\left[\int_{-\infty}^{\infty} e^{-y^2/2} dy\right],$$

or equivalently,

$$I^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} exp\left(-\frac{x^2+y^2}{2}\right) dxdy.$$

Letting $x = rcos\theta, y = rsin\theta$ (i.e., using polar coordinates), we have

$$I^2 = \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} e^{-r^2/2} rdrd\theta = \frac{1}{2\pi}\int_{0}^{2\pi} d\theta = \frac{1}{2\pi}2\pi = 1.$$

The **mean** and the **variance** of the normal distribution is as follows:

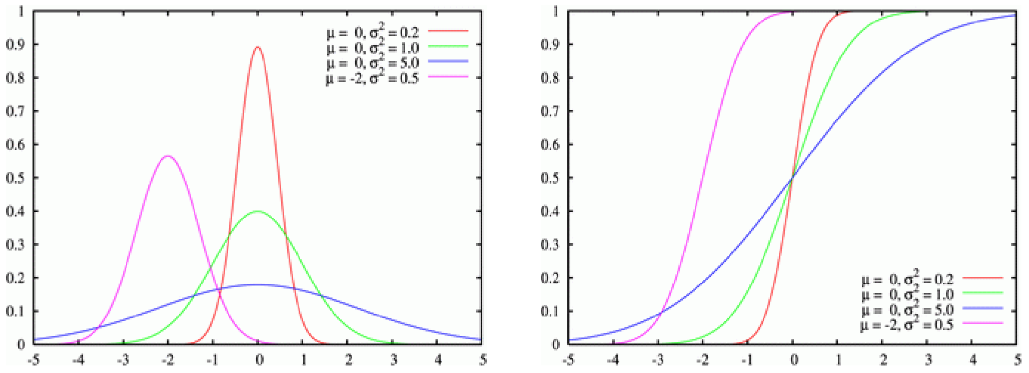
$$E(X) = \mu$$

and

$$Var(X) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2.$$

That is, the parameters μ and σ^2 in the p.d.f. are the mean and the variance of X.

Normal Distribution



(a) (b)

Figure 2.9: p.d.f. and c.d.f graphs of the Normal Distribution (a) Probability Density Function (b) Cumulative Distribution Function

Example 2.12
If the p.d.f. of X is

$$f\left(x\right)=\frac{1}{\sqrt{32\pi}}exp\left[-\frac{\left(x+7\right)^2}{32}\right],-\infty<x<\infty,$$

then X is $N\left(-7,16\right)$

That is, X has a normal distribution with a mean μ =-7, variance σ^2 =16, and the moment generating function

$$M\left(t\right)=exp\left(-7t+8t^2\right).$$

2.5 THE t DISTRIBUTION[5]

2.5.1 THE t DISTRIBUTION

In probability and statistics, the **t-distribution** or **Student's distribution** arises in the problem of estimating the mean of a normally distributed population when the sample size is small, as well as when (as in nearly all practical statistical work) the population standard deviation is unknown and has to be estimated from the data.

Textbook problems treating the standard deviation as if it were known are of two kinds:

1. those in which the sample size is so large that one may treat a data-based estimate of the variance as if it were certain,
2. those that illustrate mathematical reasoning, in which the problem of estimating the standard deviation is temporarily ignored because that is not the point that the author or instructor is then explaining.

[5] This content is available online at <http://cnx.org/content/m13495/1.3/>.

2.5.1.1 THE t DISTRIBUTION

Definition 14: t Distribution

If Z is a random variable that is $N\left(0,1\right)$, if U is a random variable that is $\chi^2\left(r\right)$, and if Z and U are independent, then

$$T = \frac{Z}{\sqrt{U/r}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{2.9}$$

has a t distribution with r degrees of freedom.

Where μ is the population mean, $\overline{x}$ is the sample mean and s is the estimator for population standard deviation (i.e., the sample variance) defined by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2. \tag{2.10}$$

If $\sigma = s$, $t = z$, the distribution becomes the normal distribution. As N increases, Student's t distribution approaches the normal distribution (Section 2.4.1: NORMAL DISTRIBUTION). It can be derived by transforming student's z-distribution using

$$z \equiv \frac{\overline{x} - \mu}{s}$$

and then defining

$$t = z\sqrt{n-1}.$$

The resulting probability and cumulative distribution functions are:

$$f\left(t\right) = \frac{\Gamma\left[\left(r+1\right)/2\right]}{\sqrt{\pi r}\Gamma\left(r/2\right)\left(1+t^2/r\right)^{\left(r+1\right)/2}}, \tag{2.11}$$

$$F\left(t\right) = \frac{1}{2}+\frac{1}{2}\left[I\left(1;\frac{1}{2}r,\frac{1}{2}\right) - I\left(\frac{r}{r+t^2},\frac{1}{2}r,\frac{1}{2}\right)\right] sgn\left(t\right) = \frac{1}{2} - \frac{itB\left(-\frac{t^2}{r};\frac{1}{2},\frac{1}{2}\left(1-r\right)\right)\Gamma\left(\frac{1}{2}\left(r+1\right)\right)}{2\sqrt{\pi}|t|\Gamma\left(\frac{1}{2}r\right)} \tag{2.12}$$

where,

- $r = n - 1$ is the number of degrees of freedom,
- $-\infty < t < \infty$,
- $\Gamma\left(z\right)$ is the gamma function,
- $B\left(a,b\right)$ is the bets function,
- $I\left(z;a,b\right)$ is the regularized beta function defined by

$$I\left(z;a,b\right) = \frac{B\left(z;a,b\right)}{B\left(a,b\right)}.$$

The effect of degree of freedom on the t distribution is illustrated in the four t distributions on the Figure 1 (Figure 2.10).

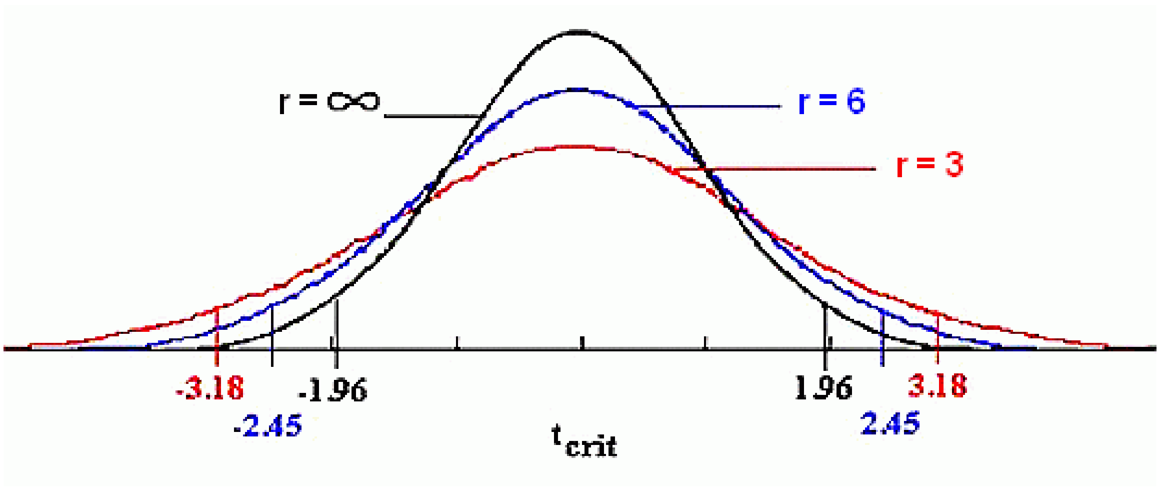


Figure 2.10: p.d.f. of the t distribution for degrees of freedom r=3, r=6, r=∞.

In general, it is difficult to evaluate the distribution function of T. Some values are usually given in the tables. Also observe that the graph of the p.d.f. of T is symmetrical with respect to the vertical axis t =0 and is very similar to the graph of the p.d.f. of the standard normal distribution $N\left(0,1\right)$. However the tails of the t distribution are heavier that those of a normal one; that is, there is more extreme probability in the t distribution than in the standardized normal one. Because of the symmetry of the t distribution about t =0, the mean (if it exists) must be equal to zero. That is, it can be shown that $E\left(T\right)=0$ when $r\geq 2$. When r=1 the t distribution is the **Cauchy distribution**, and thus both the variance and mean do not exist.

Chapter 3

Estimation

3.1 Estimation[1]

3.1.1 ESTIMATION

Once a model is specified with its parameters and data have been collected, one is in a position to evaluate the model's goodness of fit, that is, how well the model fits the observed pattern of data. Finding parameter values of a model that best fits the data — **a procedure called parameter estimation, which assesses goodness of fit**.

There are two generally accepted methods of parameter estimation. They are **least squares estimation (LSE)** and **maximum likelihood estimation (MLE)**. The former is well known as linear regression, the sum of squares error, and the root means squared deviation is tied to the method. On the other hand, MLE is not widely recognized among modelers in psychology, though it is, by far, the most commonly used method of parameter estimation in the statistics community. LSE might be useful for obtaining a descriptive measure for the purpose of summarizing observed data, but MLE is more suitable for statistical inference such as model comparison. LSE has no basis for constructing confidence intervals or testing hypotheses whereas both are naturally built into MLE.

3.1.1.1 Properties of Estimators

UNBIASED AND BIASED ESTIMATORS

Let consider random variables for which the functional form of the p.d.f. is know, but the distribution depends on an unknown parameter θ , that may have any value in a set θ , which is called the **parameter space**. In estimation the random sample from the distribution is taken to elicit some information about the unknown parameter θ. The experiment is repeated n independent times, the sample $X_1, X_2, ..., X_n$ is observed and one try to guess the value of θ using the observations $x_1, x_2, ...x_n$.

The function of $X_1, X_2, ..., X_n$ used to guess θ **is called an estimator of** θ . We want it to be such that the computed estimate $u\left(x_1, x_2, ...x_n\right)$ is usually close to θ. Let $Y = u\left(x_1, x_2, ...x_n\right)$ be an estimator of θ. If Y to be a good estimator of θ , a very desirable property is that **it means be equal to** θ , namely $E\left(Y\right) = \theta$.

Definition 15:
If $E\left[u\left(x_1, x_2, ..., x_n\right)\right] = \theta$ is called **an unbiased estimator of** θ. Otherwise, it is said to be **biased**.

It is required not only that an estimator has expectation equal to θ, but also the variance of the estimator should be as small as possible. If there are two unbiased estimators of θ, it could be probably possible to choose the one with the smaller variance. In general, with a random sample $X_1, X_2, ..., X_n$ of a fixed sample

[1] This content is available online at <http://cnx.org/content/m13524/1.2/>.

size n, a statistician might like to find the estimator $Y = u(X_1, X_2, ..., X_n)$ of an unknown parameter θ which minimizes the mean (expected) value of the square error (difference) $Y - \theta$ that is, minimizes

$$E\left[(Y-\theta)^2\right] = E\{[u(X_1, X_2, ..., X_n) - \theta]^2\}.$$

The statistic Y that minimizes $E\left[(Y-\theta)^2\right]$ is the one with minimum mean square error. If we restrict our attention to unbiased estimators only, then

$$Var(Y) = E\left[(Y-\theta)^2\right],$$

and the unbiased statistics Y that minimizes this expression is said to be **the unbiased minimum variance estimator of θ** .

3.1.1.2 Method of Moments

One of the oldest procedures for estimating parameters is **the method of moments**. Another method for finding an estimator of an unknown parameter is called **the method of maximum likelihood**. In general, in the method of moments, if there are k parameters that have to be estimated, the first k sample moments are set equal to the first k population moments that are given in terms of the unknown parameters.

Example 3.1

Let the distribution of X be $N\left(\mu, \sigma^2\right)$. Then $E(X) = \mu$ and $E\left(X^2\right) = \sigma^2 + \mu^2$. Given a random sample of size n, the first two moments are given by

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

We set $m_1 = E(X)$ and $m_2 = E\left(X^2\right)$ and solve for μ and σ^2,

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \mu$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \sigma^2 + \mu^2.$$

The first equation yields $\overline{x}$ as the estimate of μ . Replacing μ^2 with $\overline{x}^2$ in the second equation and solving for σ^2 ,

we obtain

$$\frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x}^2 = v$$

for the solution of σ^2 .

Thus the method of moment estimators for μ and σ^2 are $\tilde{\mu} = \overline{X}$ and $\tilde{\sigma}^2 = V$. Of course, $\tilde{\mu} = \overline{X}$ is unbiased whereas $\tilde{\sigma}^2 = V.$ is biased.

At this stage arises the question, which of two different estimators $\hat{\theta}$ and $\tilde{\theta}$, for a parameter θ one should use. Most statistician select he one that has the smallest mean square error, for example,

$$E\left[\left(\hat{\theta}-\theta\right)^2\right] < E\left[\left(\tilde{\theta}-\theta\right)^2\right],$$

then $\hat{\theta}$ seems to be preferred. This means that if $E\left(\hat{\theta}\right) = E\left(\tilde{\theta}\right) = \theta$, then one would select the one with the smallest variance.

Next, other questions should be considered. Namely, given an estimate for a parameter, how accurate is the estimate? How confident one is about the closeness of the estimate to the unknown parameter?

SEE: CONFIDENCE INTERVALS I (Section 3.2.1: CONFIDENCE INTERVALS I) and CONFIDENCE INTERVALS II (Section 3.3.1: CONFIDENCE INTERVALS II)

3.2 CONFIDENCE INTERVALS I[2]

3.2.1 CONFIDENCE INTERVALS I

Definition 16:

Given a random sample $X_1, X_2, ..., X_n$ from a normal distribution $N\left(\mu, \sigma^2\right)$, consider the closeness of $\overline{X}$, the unbiased estimator of μ, to the unknown μ. To do this, the error structure (distribution) of $\overline{X}$, namely that $\overline{X}$ is $N\left(\mu, \sigma^2/n\right)$, is used in order to construct what is called **a confidence interval** for the unknown parameter μ, when the variance σ^2 is known.

For the probability $1-\alpha$, it is possible to find a number $z_{\alpha/2}$, such that

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha.$$

For example, if $1-\alpha = 0.95$, then $z_{\alpha/2} = z_{0.025} = 1.96$ and if $1-\alpha = 0.90$, then $z_{\alpha/2} = z_{0.05} = 1.645$. Recalling that $\sigma > 0$, the following inequalities are equivalent :

$$-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}$$

and

$$-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \overline{X}-\mu \le z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

$$-\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le -\mu \le -\overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),$$

$$\overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \ge \mu \ge \overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Thus, since the probability of the first of these is 1-$1-\alpha$, the probability of the last must also be $1-\alpha$, because the latter is true if and only if the former is true. That is,

$$P\left[\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le -\overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1-\alpha.$$

So the probability that the random interval

$$\left[\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

includes the unknown mean μ is $1-\alpha$.

[2] This content is available online at <http://cnx.org/content/m13494/1.3/>.

Definition 17:
1. Once the sample is observed and the sample mean computed equal to $\overline{x}$, the interval

$$\overline{x} - z_{\alpha/2}\left(\sigma/\sqrt{n}\right), \overline{x} + z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$$

is a known interval. Since the probability that the random interval covers μ before the sample is drawn is equal to $1-\alpha$, call the computed interval, $\overline{x} \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$(for brevity), a $100\left(1-\alpha\right)\%$ **confidence interval** for the unknown mean μ.
2. The number $100\left(1-\alpha\right)\%$, or equivalently, $1-\alpha$, is called **the confidence coefficient**.

For illustration,

$$\overline{x} \pm 1.96\left(\sigma/\sqrt{n}\right)$$

is a 95% confidence interval for μ.

It can be seen that the confidence interval for μ is centered at the point estimate $\overline{x}$ and is completed by subtracting and adding the quantity $z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$.

NOTE THAT: as n increases, $z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$ decreases, resulting n a shorter confidence interval with the same confidence coefficient $1-\alpha$

A shorter confidence interval indicates that there is more reliance in $\overline{x}$ as an estimate of μ. For a fixed sample size n, the length of the confidence interval can also be shortened by decreasing the confidence coefficient $1-\alpha$. But if this is done, shorter confidence is achieved by losing some confidence.

Example 3.2
Let $\overline{x}$ be the observed sample mean of 16 items of a random sample from the normal distribution $N\left(\mu, \sigma^2\right)$. A 90% confidence interval for the unknown mean μ is

$$\left[\overline{x} - 1.645\sqrt{\frac{23.04}{16}}, \overline{x} + 1.645\sqrt{\frac{23.04}{16}}\right].$$

For a particular sample this interval either does or does not contain the mean μ. However, if many such intervals were calculated, it should be true that about 90% of them contain the mean μ.

If one cannot assume that the distribution from which the sample arose is normal, one can still obtain an approximate confidence interval for μ . By the Central Limit Theorem the ratio $\left(\overline{X}-\mu\right)/\left(\sigma/\sqrt{n}\right)$ has, provided that n is large enough, the approximate normal distribution $N\left(0,1\right)$ when the underlying distribution is not normal. In this case

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \approx 1-\alpha,$$

and

$$\left[\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

is an approximate $100\left(1-\alpha\right)\%$ confidence interval for μ. The closeness of the approximate probability $1-\alpha$ to the exact probability depends on both the underlying distribution and the sample size. When the underlying distribution is unimodal (has only one mode) and continuous, the approximation is usually quite good for even small n, such as $n=5$. As the underlying distribution becomes less normal (*i.e.*, badly skewed or discrete), a larger sample size might be required to keep reasonably accurate approximation. But, in all cases, an n of at least 30 is usually quite adequate.

SEE ALSO: Confidence Intervals II

3.3 CONFIDENCE INTERVALS II[3]

3.3.1 CONFIDENCE INTERVALS II

3.3.1.1 Confidence Intervals for Means

In the preceding considerations (Confidence Intervals I (Section 3.2.1: CONFIDENCE INTERVALS I)), the confidence interval for the mean μ of a normal distribution was found, assuming that the value of the standard deviation σ is known. However, in most applications, the value of the standard deviation σ is rather unknown, although in some cases one might have a very good idea about its value.

Suppose that the underlying distribution is normal and that σ^2 is unknown. It is shown that given random sample $X_1, X_2, ..., X_n$ from a normal distribution, the statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a **t distribution** with $r = n - 1$ degrees of freedom, where S^2 is the usual unbiased estimator of σ^2, (see, t distribution (Section 2.5.1: THE t DISTRIBUTION)).

Select $t_{\alpha/2}(n-1)$ so that

$$P\left[T \geq t_{\alpha/2}(n-1)\right] = \alpha/2.$$

Then

$$\begin{aligned} 1-\alpha = P\left[-t_{\alpha/2}(n-1) \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right] \\ = P\left[-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \leq \overline{X}-\mu \leq t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right] \\ = P\left[-\overline{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \leq -\mu \leq -\overline{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right] \\ = P\left[\overline{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \leq -\mu \leq \overline{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]. \end{aligned}$$

Thus the observations of a random sample provide a $\overline{x}$ and s^2 and $\overline{x}-t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \overline{x}+t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$ is a $100(1-\alpha)\%$ interval for μ.

Example 3.3

Let X equals the amount of butterfat in pound produced by a typical cow during a 305-day milk production period between her first and second claves. Assume the distribution of X is $N\left(\mu, \sigma^2\right)$. To estimate μ a farmer measures the butterfat production for n-20 cows yielding the following data:

481	537	513	583	453	510	570
500	487	555	618	327	350	643
499	421	505	637	599	392	-

For these data, $\overline{x} = 507.50$ and $s = 89.75$. Thus a point estimate of μ is $\overline{x} = 507.50$. Since $t_{0.05}(19) = 1.729$, a 90% confidence interval for μ is $507.50 \pm 1.729\left(\frac{89.75}{\sqrt{20}}\right)$, or equivalently, [472.80, 542.20].

Let T have a t distribution with n-1 degrees of freedom. Then, $t_{\alpha/2}(n-1) > z_{\alpha/2}$. Consequently, the interval $\overline{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$ is expected to be shorter than the interval $\overline{x} \pm t_{\alpha/2}(n-1)\,s/\sqrt{n}$. After all, there gives more information, namely the value of σ, in construction the first interval. However, the length of the second interval is very much dependent on the value of s. If the observed s is smaller than σ, a shorter confidence interval could result by the second scheme. But on the average, $\overline{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$is the shorter of the two confidence intervals.

[3] This content is available online at <http://cnx.org/content/m13496/1.4/>.

If it is not possible to assume that the underlying distribution is normal but μ and σ are both unknown, approximate confidence intervals for μ can still be constructed using

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

which now only has an approximate t distribution.

Generally, this approximation is quite good for many normal distributions, in particular, if the underlying distribution is symmetric, unimodal, and of the continuous type. However, if the distribution is **highly skewed**, there is a great danger using this approximation. In such a situation, it would be safer to use certain nonparametric method for finding a confidence interval for the median of the distribution.

3.3.1.3 Confidence Interval for Variances

The confidence interval for the variance σ^2 is based on the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

In order to find a confidence interval for σ^2, it is used that the distribution of $(n-1)\,S^2/\sigma^2$ is $\chi^2\,(n-1)$. The constants a and b should selected from tabularized Chi Squared Distribution (Section 2.3.1.2: Chi-Square Distribution) with n-1 degrees of freedom such that

$$P\left(a \leq \frac{(n-1)\,S^2}{\sigma^2} \leq b\right) = 1 - \alpha.$$

That is select a and b so that the probabilities in two tails are equal:

$$a = \chi^2_{1-\alpha/2}\,(n-1)$$

and

$$b = \chi^2_{\alpha/2}\,(n-1)\,.$$

Then, solving the inequalities, we have

$$1 - \alpha = P\left(\frac{a}{(n-1)\,S^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{(n-1)\,S^2}\right) = P\left(\frac{(n-1)\,S^2}{b} \leq \sigma^2 \leq \frac{(n-1)\,S^2}{a}\right).$$

Thus the probability that the random interval

$$[(\text{n-1})\text{S}^2/\text{b},\ (\text{n-1})\text{S}^2/\text{a}]$$

contains the unknown σ^2 is 1-α. Once the values of $X_1, X_2, ..., X_n$ are observed to be $x_1, x_2, ..., x_n$ and s^2 computed, then the interval

$$[(\text{n-1})\text{S}^2/\text{b},\ (\text{n-1})\text{S}^2/\text{a}]$$

is a $100\,(1-\alpha)\,\%$ confidence interval for σ^2.

It follows that

$$\left[\sqrt{(n-1)\,/bs}, \sqrt{(n-1)\,/as}\right]$$

is a $100\,(1-\alpha)\,\%$ confidence interval for σ, the standard deviation.

Example 3.4

Assume that the time in days required for maturation of seeds of a species of a flowering plant found in Mexico is $N\left(\mu, \sigma^2\right)$. A random sample of n=13 seeds, both parents having narrow leaves, yielded $\overline{x}$=18.97 days and $12s^2 = \sum_{i=1}^{13}\left(x - \overline{x}\right)^2 = 128.41$.

A confidence interval for σ^2 is $\left[\frac{128.41}{21.03}, \frac{128.41}{5.226}\right] = [6.11, 24.57]$, because $5.226 = \chi^2_{0.95}(12)$ and $21.03 = \chi^2_{0.055}(12)$, what can be read from the tabularized Chi Squared Distribution. The corresponding 90% confidence interval for σ is $\left[\sqrt{6.11}, \sqrt{24.57}\right] = [2.47, 4.96]$.

Although a and b are generally selected so that the probabilities in the two tails are equal, the resulting $100\,(1-\alpha)\,\%$ confidence interval is not the shortest that can be formed using the available data. The tables and appendixes gives solutions for a and b that yield confidence interval of minimum length for the standard deviation.

3.4 SAMPLE SIZE[4]

3.4.1 Size Sample

Very frequently asked question in statistical consulting is, **how large should the sample size be to estimate a mean?**

The answer will depend on the variation associated with the random variable under observation. The statistician could correctly respond, only one item is needed, provided that the standard deviation of the distribution is zero. That is, if σ is equal zero, then the value of that one item would necessarily equal the unknown mean of the distribution. This is the extreme case and one that is not met in practice. However, the smaller the variance, the smaller the sample size needed to achieve a given degree of accuracy.

Example 3.5

A mathematics department wishes to evaluate a new method of teaching calculus that does mathematics using a computer. At the end of the course, the evaluation will be made on the basis of scores of the participating students on a standard test. Because there is an interest in estimating the mean score μ, for students taking calculus using computer so there is a desire to determine the number of students, n, who are to be selected at random from a larger group. So, let find the sample size n such that we are fairly confident that $\overline{x} \pm 1$ contains the unknown test mean μ, from past experience it is believed that the standard deviation associated with this type of test is 15. Accordingly, using the fact that the sample mean of the test scores, $\overline{X}$, is approximately $N\left(\mu, \sigma^2/n\right)$, it is seen that the interval given by $\overline{x} \pm 1.96\,(15/\sqrt{n})$ will serve as an approximate 95% confidence interval for μ.

That is, $1.96\left(\frac{15}{\sqrt{n}}\right) = 1$ or equivalently $\sqrt{n} = 29.4$ and thus $n \approx 864.36$ or n=865 because n must be an integer. It is quite likely that it had not been anticipated that as many as 865 students would be needed in this study. If that is the case, the statistician must discuss with those involved in the experiment whether or not the accuracy and the confidence level could be relaxed some. For illustration, rather than requiring $\overline{x} \pm 1$ to be a 95% confidence interval for μ, possibly $\overline{x} \pm 2$ would be satisfactory for 80% one. If this modification is acceptable, we now have $1.282\left(\frac{15}{\sqrt{n}}\right) = 2$ or equivalently, $\sqrt{n} = 9.615$ and thus $n \approx 92.4$. Since n must be an integer = 93 is used in practice.

Most likely, the person involved in this project would find this a more reasonable sample size. Of course, any sample size greater than 93 could be used. Then either the length of the confidence interval could be decreased from that of $\overline{x} \pm 2$ or the confidence coefficient could be increased from 80% or a combination of both. Also, since there might be some question of whether the standard deviation σ actually equals 15, the sample standard deviations would no doubt be used in the construction of the interval.

For example, suppose that the sample characteristics observed are

$$n = 145, \overline{x} = 77.2, s = 13.2;$$

then, $\overline{x} \pm \frac{1.282s}{\sqrt{n}}$ or 77.2 ± 1.41 provides an approximate 80% confidence interval for μ.

[4] This content is available online at <http://cnx.org/content/m13531/1.2/>.

In general, if we want the $100\,(1-\alpha)\,\%$ confidence interval for μ, $\overline{x} \pm z_{\alpha/2}\,(\sigma/\sqrt{n})$, to be no longer than that given by $\overline{x} \pm \epsilon$, the sample size n is the solution of $\epsilon = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$, where $\Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$.

That is,

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{\epsilon^2},$$

where it is assumed that σ^2 is known.

Sometimes

$$\epsilon = z_{\alpha/2}\sigma/\sqrt{n}$$

is called **the maximum error of the estimate**. If the experimenter has no ideas about the value of σ^2, it may be necessary to first take a preliminary sample to estimate σ^2.

The type of statistic we see most often in newspaper and magazines is an estimate of a proportion p. We might, for example, want to know the percentage of the labor force that is unemployed or the percentage of voters favoring a certain candidate. Sometimes extremely important decisions are made on the basis of these estimates. If this is the case, we would most certainly desire short confidence intervals for p with large confidence coefficients. We recognize that these conditions will require a large sample size. On the other hand, if the fraction p being estimated is not too important, an estimate associated with a longer confidence interval with a smaller confidence coefficients is satisfactory; and thus a smaller sample size can be used.

In general, to find the required sample size to estimate p, recall that the point estimate of p is

$$\hat{p} = z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

Suppose we want an estimate of p that is within ϵ of the unknown p with $100\,(1-\alpha)\,\%$ confidence where $\epsilon = z_{\alpha/2}\sqrt{\hat{p}\,(1-\hat{p})\,/n}$ is **the maximum error of the point estimate** $\hat{p} = y/n$. Since $\hat{p}$ is unknown before the experiment is run, we cannot use the value of $\hat{p}$ in our determination of n. However, if it is known that p is about equal to p^*, the necessary sample size n is the solution of

$$\epsilon = \frac{z_{\alpha/2}\sqrt{p^*\,(1-p^*)}}{\sqrt{n}}.$$

That is,

$$n = \frac{z_{\alpha/2}^2 p^*\,(1-p^*)}{\epsilon^2}.$$

3.5 Maximum Likelihood Estimation (MLE)[5]

3.5.1 MAXIMUM LIKELIHOOD ESTIMATION (MLE)

3.5.1.1 Likelihood function

From a statistical standpoint, the data vector $x = (x_1, x_2, ..., x_n)$ as the outcome of an experiment is a random sample from an unknown population. **The goal of data analysis is to identify the population that is most likely to have generated the sample.** In statistics, each population is identified by a corresponding probability distribution. Associated with each probability distribution is a unique value of the model's parameter. As the parameter changes in value, different probability distributions are generated. Formally, a model is defined as the family of probability distributions indexed by the model's parameters.

Let denote the *probability distribution function* (PDF) by $f\,(x|\theta)$ that specifies the probability of observing data y given the parameter w. The parameter vector $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is a vector defined on a

[5] This content is available online at <http://cnx.org/content/m13501/1.3/>.

multi-dimensional parameter space. If individual observations, x_i's are statistically independent of one another, then according to the theory of probability, the PDF for the data $x = (x_1, x_2, ..., x_n)$ can be expressed as a multiplication of PDFs for individual observations,

$$f(x, \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta),$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

To illustrate the idea of a PDF, consider the simplest case with one observation and one parameter, that is, $n = k = 1$. Suppose that the data x represents the number of successes in a sequence of 10 independent binary trials (e.g., coin tossing experiment) and that the probability of a success on any one trial, represented by the parameter, θ is 0.2. The PDF in this case is then given by

$$f(x|\theta = 0.2) = \frac{10!}{x!\,(10-x)!}(0.2)^x (0.8)^{10-x}, (x = 0.1, ..., 10),$$

which is known as the binomial probability distribution. The shape of this PDF is shown in the top panel of Figure 1 (Figure 3.1). If the parameter value is changed to say $w = 0.7$, a new PDF is obtained as

$$f(x|\theta = 0.7) = \frac{10!}{x!\,(10-x)!}(0.7)^x (0.3)^{10-x}, (x = 0.1, ..., 10);$$

whose shape is shown in the bottom panel of Figure 1 (Figure 3.1). **The following is the general expression of the binomial PDF for arbitrary values of θ and n:**

$$f(x|\theta) = \frac{n!}{\theta!\,(n-x)!}\theta^x (1-\theta)^{n-x}, 0 \leq \theta \leq 1, x = 0.1, ..., n;$$

which as a function of y specifies the probability of data y for a given value of the parameter θ . The collection of all such PDFs generated by varying parameter across its range (0 - 1 in this case) defines a model.

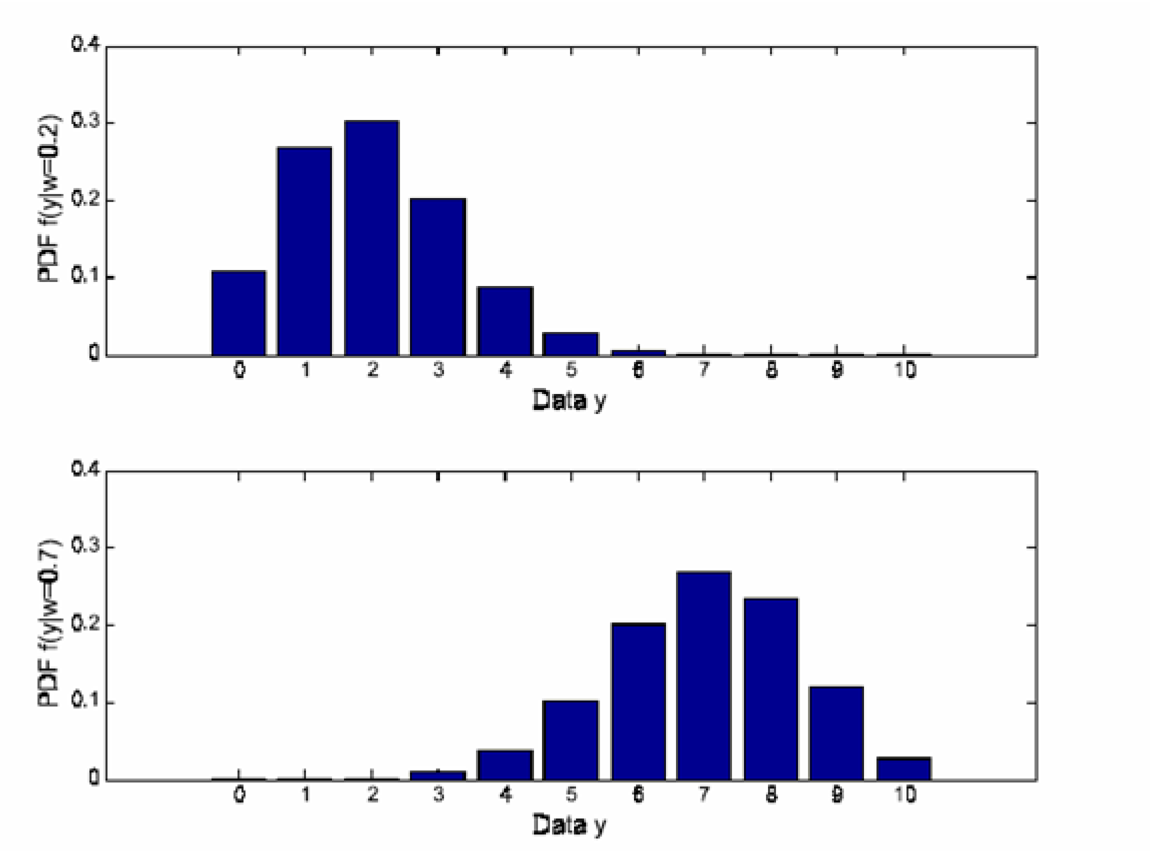


Figure 3.1: Binomial probability distributions of sample size $n = 10$ and probability parameter $\theta = 0.2$ (top) and $\theta = 0.7$ (bottom).

3.5.1.2 Maximum Likelihood Estimation

Once data have been collected and the likelihood function of a model given the data is determined, one is in a position to make statistical inferences about the population, that is, the probability distribution that underlies the data. Given that different parameter values index different probability distributions (Figure 1 (Figure 3.1)), we are interested in finding the parameter value that corresponds to the desired PDF.

The principle of **maximum likelihood estimation (MLE)**, originally developed by R. A. Fisher in the 1920s, states that the desired probability distribution be the one that makes the observed data most likely, which is obtained by seeking the value of the parameter vector that maximizes the likelihood function (Section 3.5.1: MAXIMUM LIKELIHOOD ESTIMATION (MLE)) $L(\theta)$. The resulting parameter, which is sought by searching the multidimensional parameter space, is called **the MLE estimate**, denoted by

$$\theta MLE = (\theta_1 MLE, ..., \theta_k MLE).$$

Let p equal the probability of success in a sequence of Bernoulli trials or the proportion of the large population with a certain characteristic. The method of moments estimate for p is relative frequency of success (having that characteristic). It will be shown below that the maximum likelihood estimate for p is also the relative frequency of success.

Suppose that X is $b\left(1,p\right)$ so that the p.d.f. of X is

$$f\left(x;p\right)=p^{x}\left(1-p\right)^{1-x},x=0,1,0\leq p\leq 1.$$

Sometimes is written

$$p\in\Omega=\left[p:0\leq p\leq 1\right],$$

where Ω is used to represent parameter space, that is, the space of all possible values of the parameter. A random sample $X_1,X_2,...,X_n$ is taken, and the problem is to find an estimator $\mathrm{u}\left(\mathrm{X}_1,\mathrm{X}_2,...,\mathrm{X}_\mathrm{n}\right)$ such that $\mathrm{u}\left(x_1,x_2,...,x_\mathrm{n}\right)$ is a good point estimate of p, where $x_1,x_2,...,x_n$ are the observed values of the random sample. Now the probability that $X_1,X_2,...,X_n$ takes the particular values is

$$P\left(X_1=x_1,...,X_n=x_n\right)=\prod_{i=1}^{n}p^{x_i}\left(1-p\right)^{1-x_i}=p\sum x_i\left(1-p\right)^{n-\sum x_i},$$

which is the joint p.d.f. of $X_1,X_2,...,X_n$ evaluated at the observed values. One reasonable way to proceed towards finding a good estimate of p is to regard this probability (or joint p.d.f.) as a function of p and find the value of p that maximizes it. That is, find the p value most likely to have produced these sample values. The joint p.d.f., when regarded as a function of p, is frequently called **the likelihood function**. Thus here the likelihood function is:

$$L\left(p\right)=L\left(p;x_1,x_2,...,x_n\right)=f\left(x_1;p\right)f\left(x_2;p\right)\cdots f\left(x_n;p\right)=p^{\sum x_i}\left(1-p\right)^{n-\sum x_i},0\leq p\leq 1.$$

To find the value of p that maximizes $L\left(p\right)$ first take its derivative for $0<p<1:$

$$\frac{dL\left(p\right)}{dp}=\left(\sum x_i\right)p^{n-\sum x_i}\left(1-p\right)^{n-\sum x_i}-\left(n-\sum x_i\right)p^{\sum x_i}\left(1-p\right)^{n-\sum x_i-1}.$$

Setting this first derivative equal to zero gives

$$p^{\sum x_i}\left(1-p\right)^{n-\sum x_i}\left[\frac{\sum x_i}{p}-\frac{n-\sum x_i}{1-p}\right]=0.$$

Since $0<p<1$, this equals zero when

$$\frac{\sum x_i}{p}-\frac{n-\sum x_i}{1-p}=0.$$

Or, equivalently,

$$p=\frac{\sum x_i}{n}=\overline{x}.$$

The corresponding statistics, namely $\sum X_i/n=\overline{X}$, is called **the maximum likelihood estimator** and is denoted by $\hat{p}$,that is,

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}.$$

When finding a maximum likelihood estimator, it is often easier to find the value of parameter that minimizes the natural logarithm of the likelihood function rather than the value of the parameter that minimizes the likelihood function itself. Because the natural logarithm function is an increasing function, the solution will be the same. To see this, the example which was considered above gives for $0<p<1$,

$$lnL\left(p\right)=\left(\sum_{i=1}^{n}x_i\right)lnp+\left(n-\sum_{i=1}^{n}x_i\right)ln\left(1-p\right).$$

To find the maximum, set the first derivative equal to zero to obtain

$$\frac{d\left[lnL\left(p\right)\right]}{dp}=\left(\sum_{i=1}^{n}x_i\right)\left(\frac{1}{p}\right)+\left(n-\sum_{i=1}^{n}x_i\right)\left(\frac{-1}{1-p}\right)=0,$$

which is the same as previous equation. Thus the solution is $p=\overline{x}$ and the maximum likelihood estimator for p is $\hat{p}=\overline{X}$.

Motivated by the preceding illustration, the formal definition of maximum likelihood estimators is presented. This definition is used in both the discrete and continuous cases. In many practical cases, these estimators (and estimates) are unique. For many applications there is just one unknown parameter. In this case the likelihood function is given by

$$L\left(\theta\right)=\prod_{i=1}^{n}f\left(x_i,\theta\right).$$

SEE ALSO: Maximum Likelihood Estimation - Examples (Section 3.6.1: MAXIMUM LIKELIHOOD ESTIMATION - EXAMPLES)

3.6 Maximum Likelihood Estimation - Examples[6]

3.6.1 MAXIMUM LIKELIHOOD ESTIMATION - EXAMPLES

3.6.1.1 EXPONENTIAL DISTRIBUTION

Let $X_1,X_2,...,X_n$ be a random sample from the exponential distribution with p.d.f.

$$f\left(x;\theta\right)=\frac{1}{\theta}e^{-x/\theta},0<x<\infty,\theta\in\Omega=\{\theta;0<\theta<\infty\}.$$

The likelihood function is given by

$$L\left(\theta\right)=L\left(\theta;x_1,x_2,...,x_n\right)=\left(\frac{1}{\theta}e^{-x_1/\theta}\right)\left(\frac{1}{\theta}e^{-x_2/\theta}\right)\cdots\left(\frac{1}{\theta}e^{-x_n/\theta}\right)=\frac{1}{\theta^n}exp\left(\frac{-\sum_{i=1}^{n}x_i}{\theta}\right),0<\theta<\infty.$$

The natural logarithm of $L\left(\theta\right)$ is

$$lnL\left(\theta\right)=-\left(n\right)ln\left(\theta\right)-\frac{1}{\theta}\sum_{i=1}^{n}x_i,0<\theta<\infty.$$

Thus,

$$\frac{d\left[lnL\left(\theta\right)\right]}{d\theta}=\frac{-n}{\theta}+\frac{\sum_{i=1}^{n}x_i}{\theta^2}=0.$$

The solution of this equation for θ is

$$\theta=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}.$$

Note that,

$$\frac{d\left[lnL\left(\theta\right)\right]}{d\theta}=\frac{1}{\theta}\left(-n+\frac{n\overline{x}}{\theta}\right)>0,\theta<\overline{x},$$

[6] This content is available online at <http://cnx.org/content/m13500/1.3/>.

$$\frac{d\left[lnL\left(\theta\right)\right]}{d\theta} = \frac{1}{\theta}\left(-n + \frac{n\overline{x}}{\theta}\right) = 0, \theta = \overline{x},$$

$$\frac{d\left[lnL\left(\theta\right)\right]}{d\theta} = \frac{1}{\theta}\left(-n + \frac{n\overline{x}}{\theta}\right) < 0, \theta > \overline{x},$$

Hence, $lnL\left(\theta\right)$ does have a maximum at $\overline{x}$, and thus the maximum likelihood estimator for θ is

$$\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This is both an unbiased estimator and the method of moments estimator for θ.

3.6.1.2 GEOMETRIC DISTRIBUTION

Let $X_1, X_2, ..., X_n$ be a random sample from the geometric distribution with p.d.f.

$$f\left(x;p\right) = \left(1-p\right)^{x-1}p, x = 1, 2, 3, ...$$

The likelihood function is given by

$$L\left(p\right) = \left(1-p\right)^{x_1-1}p(1-p)^{x_2-1}p\cdots\left(1-p\right)^{x_n-1}p = p^n(1-p)^{\sum x_i - n}, 0 \leq p \leq 1.$$

The natural logarithm of $L\left(\theta\right)$ is

$$lnL\left(p\right) = nlnp + \left(\sum_{i=1}^{n} x_i - n\right) ln\left(1-p\right), 0 < p < 1.$$

Thus restricting p to $0 < p < 1$ so as to be able to take the derivative, we have

$$\frac{dlnL\left(p\right)}{dp} = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i - n}{1-p} = 0.$$

Solving for p, we obtain

$$p = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}}.$$

So the maximum likelihood estimator of p is

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{X}$$

Again this estimator is the method of moments estimator, and it agrees with the intuition because, in n observations of a geometric random variable, there are n successes in the $\sum_{i=1}^{n} x_i$ trials. Thus the estimate of p is the number of successes divided by the total number of trials.

3.6.1.3 NORMAL DISTRIBUTION

Let $X_1, X_2, ..., X_n$ be a random sample from $N\left(\theta_1, \theta_2\right)$, where

$$\Omega = \left(\left(\theta_1, \theta_2\right) : -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\right).$$

That is, here let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Then

$$L\left(\theta_1, \theta_2\right) = \prod_{i-1}^{n}\left(\frac{1}{\sqrt{2\pi\theta_2}} exp\left[-\frac{\left(x_i - \theta_1\right)^2}{2\theta_2}\right]\right),$$

or equivalently,

$$L\left(\theta_1,\theta_2\right)=\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n exp\left[-\frac{-\sum_{i=1}^n\left(x_i-\theta_1\right)^2}{2\theta_2}\right],\left(\theta_1,\theta_2\right)\in\Omega.$$

The natural logarithm of the likelihood function is

$$lnL\left(\theta_1,\theta_2\right)=-\frac{n}{2}ln\left(2\pi\theta_2\right)-\frac{-\sum_{i=1}^n\left(x_i-\theta_1\right)^2}{2\theta_2}.$$

The partial derivatives with respect to θ_1 and θ_2 are

$$\frac{\partial\left(lnL\right)}{\partial\theta_1}=\frac{1}{\theta_2}\sum_{i=1}^n\left(x_i-\theta_1\right)$$

and

$$\frac{\partial\left(lnL\right)}{\partial\theta_2}=\frac{-n}{2\theta_2}+\frac{1}{2\theta_2^2}\sum_{i=1}^n\left(x_i-\theta_1\right)^2.$$

The equation $\frac{\partial(lnL)}{\partial\theta_1}=0$ has the solution $\theta_1=\overline{x}$. Setting $\frac{\partial(lnL)}{\partial\theta_2}=0$ and replacing θ_1 by $\overline{x}$ yields

$$\theta_2=\frac{1}{n}\sum_{i=1}^n\left(x_i-\overline{x}\right)^2.$$

By considering the usual condition on the second partial derivatives, these solutions do provide a maximum. Thus the maximum likelihood estimators

$$\mu=\theta_1$$

and

$$\sigma^2=\theta_2$$

are

$$\hat{\theta}_1=\overline{X}$$

and

$$\hat{\theta}_2=\frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2.$$

Where we compare the above example with the introductory one, we see that the method of moments estimators and the maximum likelihood estimators for μ and σ^2 are the same. But this is not always the case. If they are not the same, which is better? Due to the fact that the maximum likelihood estimator of θ has an approximate normal distribution with mean θ and a variance that is equal to a certain lower bound, thus at least approximately, it is unbiased minimum variance estimator. Accordingly, most statisticians prefer the maximum likelihood estimators than estimators found using the method of moments.

3.6.1.4 BINOMIAL DISTRIBUTION

Observations: k successes in n Bernoulli trials.

$$f\left(x\right)=\frac{n!}{x!\left(n-x\right)!}p^x\left(1-p\right)^{n-x}$$

$$L\left(p\right)=\prod_{i=1}^n f\left(x_i\right)=\prod_{i=1}^n\left(\frac{n!}{x_i!\left(n-x_i\right)!}p^{x_i}\left(1-p\right)^{n-x_i}\right)=\left(\prod_{i=1}^n\frac{n!}{x_i!\left(n-x_i\right)!}\right)p^{x_i}\left(1-p\right)^{n-\sum_{i=1}^n x_i}$$

$$lnL\left(p\right)=\sum_{i=1}^{n}x_{i}lnp+\left(n-\sum_{i=1}^{n}x_{i}\right)ln\left(1-p\right)$$

$$\frac{dlnL\left(p\right)}{dp}=\frac{1}{p}\sum_{i=1}^{n}x_{i}-\left(n-\sum_{i=1}^{n}x_{i}\right)\frac{1}{1-p}=0$$

$$\frac{\left(1-\hat{p}\right)\sum_{i=1}^{n}x_{i}-\left(n-\sum_{i=1}^{n}x_{i}\right)\hat{p}}{\hat{p}\left(1-\hat{p}\right)}=0$$

$$\sum_{i=1}^{n}x_{i}-\hat{p}\sum_{i=1}^{n}x_{i}-n\hat{p}+\sum_{i=1}^{n}x_{i}\hat{p}=0$$

$$\hat{p}=\frac{\sum_{i=1}^{n}x_{i}}{n}=\frac{k}{n}$$

3.6.1.5 POISSON DISTRIBUTION

Observations: $x_1, x_2, ..., x_n,$

$$f\left(x\right)=\frac{\lambda^{x}e^{-\lambda}}{x!},x=0,1,2,...$$

$$L\left(\lambda\right)=\prod_{i=1}^{n}\left(\frac{\lambda^{x_{i}}e^{-\lambda}}{x_{i}!}\right)=e^{-\lambda n}\frac{\lambda\sum_{i=1}^{n}x_{i}}{\prod_{i=1}^{n}x_{i}}$$

$$lnL\left(\lambda\right)=-\lambda n+\sum_{i=1}^{n}x_{i}ln\lambda-ln\left(\prod_{i=1}^{n}x_{i}\right)$$

$$\frac{dl}{d\lambda}=-n+\sum_{i=1}^{n}x_{i}\frac{1}{\lambda}$$

$$-n+\sum_{i=1}^{n}x_{i}\frac{1}{\lambda}=0$$

$$\hat{\lambda}=\frac{\sum_{i=1}^{n}x_{i}}{n}$$

3.7 ASYMPTOTIC DISTRIBUTION OF MAXIMUM LIKELIHOOD ESTIMATORS[7]

3.7.1 ASYMPTOTIC DISTRIBUTION OF MAXIMUM LIKELIHOOD ESTIMATORS

Let consider a distribution with p.d.f. $f\left(x;\theta\right)$ such that the parameter θ is not involved in the support of the distribution. We want to be able to find the maximum likelihood estimator $\hat{\theta}$ by solving

$$\frac{\partial\left[lnL\left(\theta\right)\right]}{\partial\theta}=0,$$

where here the partial derivative was used because $L\left(\theta\right)$ involves $x_1, x_2, ..., x_n$.

[7] This content is available online at <http://cnx.org/content/m13527/1.2/>.

That is,

$$\frac{\partial\left[lnL\left(\hat{\theta}\right)\right]}{\partial\theta}=0,$$

where now, with $\hat{\theta}$ in this expression,

$$L\left(\hat{\theta}\right)=f\left(X_1;\hat{\theta}\right)f\left(X_2;\hat{\theta}\right)\cdots f\left(X_n;\hat{\theta}\right).$$

We can approximate the left-hand member of this latter equation by a linear function found from the first two terms of a Taylor's series expanded about θ , namely

$$\frac{\partial\left[lnL\left(\theta\right)\right]}{\partial\theta}+\left(\hat{\theta}-\theta\right)\frac{\partial^2\left[lnL\left(\theta\right)\right]}{\partial\theta^2}\approx 0,$$

when $L\left(\theta\right)=f\left(X_1;\theta\right)f\left(X_2;\theta\right)\cdots f\left(X_n;\theta\right).$

Obviously, this approximation is good enough only if $\hat{\theta}$ is close to θ, and an adequate mathematical proof involves those conditions. But a heuristic argument can be made by solving for $\hat{\theta}-\theta$ to obtain

$$\hat{\theta}-\theta=\frac{\frac{\partial[lnL(\theta)]}{\partial\theta}}{-\frac{\partial^2[lnL(\theta)]}{\partial\theta^2}} \tag{3.1}$$

Recall that

$$lnL\left(\theta\right)=lnf\left(X_1;\theta\right)+lnf\left(X_2;\theta\right)+\cdots+lnf\left(X_n;\theta\right)$$

and

$$\frac{\partial lnL\left(\theta\right)}{\partial\theta}=\sum_{i=1}^{n}\frac{\partial\left[lnf\left(X_i;\theta\right)\right]}{\partial\theta}; \tag{3.2}$$

The expression (2) is the sum of the n independent and identically distributed random variables

$$Y_i=\frac{\partial\left[lnf\left(X_i;\theta\right)\right]}{\partial\theta},i=1,2,...,n.$$

and thus the Central Limit Theorem has an approximate normal distribution with mean (in the continuous case) equal to

$$\begin{gathered}\int\limits_{-\infty}^{\infty}\frac{\partial[lnf(x_i;\theta)]}{\partial\theta}f\left(x;\theta\right)dx=\int\limits_{-\infty}^{\infty}\frac{\partial[f(x_i;\theta)]}{\partial\theta}\frac{f(x_i;\theta)}{f(x_i;\theta)}dx=\int\limits_{-\infty}^{\infty}\frac{\partial[f(x_i;\theta)]}{\partial\theta}dx\\ =\frac{\partial}{d\partial}\left[\int\limits_{-\infty}^{\infty}f\left(x_i;\theta\right)dx\right]=\frac{\partial}{d\partial}\left[1\right]=0.\end{gathered} \tag{3.3}$$

Clearly, the mathematical condition is needed that it is permissible to interchange the operations of integration and differentiation in those last steps. Of course, the integral of $f\left(x_i;\theta\right)$ is equal to one because it is a p.d.f.

Since we know that the mean of each Y is

$$\int\limits_{-\infty}^{\infty}\frac{\partial\left[lnf\left(x_i;\theta\right)\right]}{\partial\theta}f\left(x;\theta\right)dx=0$$

let us take derivatives of each member of this equation with respect to θ obtaining

$$\int\limits_{-\infty}^{\infty}\{\frac{\partial^2\left[lnf\left(x_i;\theta\right)\right]}{\partial\theta^2}f\left(x;\theta\right)+\frac{\partial\left[lnf\left(x_i;\theta\right)\right]}{\partial\theta}\frac{\partial\left[f\left(x_i;\theta\right)\right]}{\partial\theta}\}dx=0.$$

However,

$$\frac{\partial\left[f\left(x_i;\theta\right)\right]}{\partial\theta}=\frac{\partial\left[lnf\left(x_i;\theta\right)\right]}{\partial\theta}f\left(x;\theta\right)$$

so

$$\int_{-\infty}^{\infty}\{\frac{\partial\left[lnf\left(x_i;\theta\right)\right]}{\partial\theta}\}^{2}\ f\left(x;\theta\right)dx=-\int_{-\infty}^{\infty}\frac{\partial^2\left[lnf\left(x_i;\theta\right)\right]}{\partial\theta^2}f\left(x_i;\theta\right)dx.$$

Since $E\left(Y\right)=0$, this last expression provides the variance of $Y=\partial\left[lnf\left(X;\theta\right)\right]/d\partial$. Then the variance of expression (2) is n times this value, namely

$$-nE\{\frac{\partial^2\left[lnf\left(x_i;\theta\right)\right]}{\partial\theta^2}\}.$$

Let us rewrite (1) (3.1) as

$$\frac{\sqrt{n}\left(\hat{\theta}-\theta\right)}{1-\sqrt{-E\{\partial^2\left[lnf\left(X;\theta\right)\right]/\partial\theta^2\}}}=\frac{\frac{\partial[lnL(\theta)]/\partial\theta}{\sqrt{-E\{\partial^2[lnf(X;\theta)]/\partial\theta^2\}}}}{\frac{-\frac{1}{n}\frac{\partial^2[lnL(\theta)]}{\partial\theta^2}}{E\{-\partial^2[lnf(X;\theta)]/\partial\theta^2\}}}\tag{3.4}$$

The numerator of (4) has an approximate $N\left(0,1\right)$ distribution; and those unstated mathematical condition require, in some sense for $-\frac{1}{n}\frac{\partial^2[lnL(\theta)]}{\partial\theta^2}$ to converge to $E\left[-\partial^2\left[lnf\left(X;\theta\right)\right]/\partial\theta^2\right]$. Accordingly, the ratios given in equation (4) must be approximately $N\left(0,1\right)$. That is, $\hat{\theta}$ has an approximate normal distribution with mean θ and standard deviation $\frac{1}{\sqrt{-nE\{\partial^2[lnf(X;\theta)]/\partial\theta^2\}}}$.

Example 3.6

With the underlying exponential p.d.f.

$$f\left(x;\theta\right)=\frac{1}{\theta}e^{-x/\theta},0<x<\infty,\theta\in\Omega=\{\theta;0<\theta<\infty\}.$$

$\overline{X}$ is the maximum likelihood estimator. Since $lnf\left(x;\theta\right)=-ln\theta-\frac{x}{\theta}$ and $\frac{\partial[lnf(x;\theta)]}{\partial\theta}=-\frac{1}{\theta}+\frac{x}{\theta^2}$ and $\frac{\partial^2[lnf(x;\theta)]}{\partial\theta}=\frac{1}{\theta^2}-\frac{2x}{\theta^3}$, we have

$$-E\left[\frac{1}{\theta^2}-\frac{2X}{\theta^3}\right]=-\frac{1}{\theta}+\frac{2\theta}{\theta^3}=\frac{1}{\theta^2}$$

because $E\left(X\right)=\theta$. That is, $\overline{X}$ has an approximate distribution with mean θ and standard deviation $\theta/\sqrt{n}$. Thus the random interval $\overline{X}\pm1.96\left(\theta/\sqrt{n}\right)$ has an approximate probability of 0.95 for covering θ. Substituting the observed $\overline{x}$ for θ , as well as for $\overline{X}$, we say that $\overline{x}\pm1.96\overline{x}/\sqrt{n}$ is an approximate 95% confidence interval for θ.

Example 3.7

The maximum likelihood estimator for λ in

$$f\left(x;\lambda\right)=\frac{\lambda^xe^{-\lambda}}{x!},x=0,1,2,...;\theta\in\Omega=\{\theta:0<\theta<\infty\}$$

is $\hat{\lambda}=\overline{X}$ Now $lnf\left(x;\lambda\right)=xln\lambda-\lambda-lnx!$ and $\frac{\partial[lnf(x;\lambda)]}{\partial\lambda}=\frac{x}{\lambda}-1$ and $\frac{\partial^2[lnf(x;\lambda)]}{\partial\lambda^2}=\frac{x}{\lambda^2}$. Thus $-E\left(-\frac{X}{\lambda^2}\right)=\frac{\lambda}{\lambda^2}=\frac{1}{\lambda}$ and $\hat{\lambda}=\overline{X}$ has an approximate normal distribution with mean λ and standard deviation $\sqrt{\lambda/n}$. Finally $\overline{x}\pm1.645\sqrt{\overline{x}/n}$ serves as an approximate 90% confidence interval for λ. With the data from example(...) $\overline{x}=2.225$ and hence this interval is from 1.887 to 2.563.

It is interesting that there is another theorem which is somewhat related to the preceding result in that the variance of $\hat{\theta}$ serves as a lower bound for the variance of every unbiased estimator of θ . Thus we know that if a certain unbiased estimator has a variance equal to that lower bound, we cannot find a better one and hence it is the best in the sense of being **the unbiased minimum variance estimator**. This is called **the Rao-Cramer Inequality**.

Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with p.d.f.

$$f(x;\theta), \theta \in \Omega = \{\theta : c < \theta < d\},$$

where the support X does not depend upon θso that we can differentiate, with respect to θ, under integral signs like that in the following integral:

$$\int\limits_{-\infty}^{\infty} f(x;\theta)\, dx = 1.$$

If $Y = u(X_1, X_2, ..., X_n)$ is an unbiased estimator of θ, then

$$Var(Y) \geq \frac{1}{n \int\limits_{-\infty}^{\infty} \{[\partial lnf(x;\theta)/\partial\theta]\}^2 f(x;\theta)\, dx} = \frac{-1}{n \int\limits_{-\infty}^{\infty} [\partial^2 lnf(x;\theta)/\partial\theta^2] f(x;\theta)\, dx}.$$

Note that the two integrals in the respective denominators are the expectations

$$E\{\left[\frac{\partial lnf(X;\theta)}{\partial\theta}\right]^2\}$$

and

$$E\left[\frac{\partial^2 lnf(X;\theta)}{\partial\theta^2}\right]$$

sometimes one is easier to compute that the other.

Note that above the lower bound of two distributions: exponential and Poisson was computed. Those respective lower bounds were $\theta^2 n$ and λn. Since in each case, the variance of $\overline{X}$ equals the lower bound, then $\overline{X}$ is the unbiased minimum variance estimator.

Example 3.8

The sample arises from a distribution with p.d.f.

$$f(x;\theta) = \theta x^{\theta-1}, 0 < x < 1, \theta \in \Omega = \{\theta : 0 < \theta < \infty\}.$$

We have

$$lnf(x;\theta) = ln\theta + (\theta - 1)\, lnx, \frac{\partial lnf(x;\theta)}{\partial\theta} = \frac{1}{\theta} + lnx,$$

and

$$\frac{\partial^2 lnf(x;\theta)}{\partial\theta^2} = -\frac{1}{\theta^2}.$$

Since $E\left(-1/\theta^2\right) = -1/\theta^2$, the lower bound of the variance of every unbiased estimator of θ is θ^2/n. Moreover, the maximum likelihood estimator

$$\hat{\theta} = -n/ln\prod_{i=1}^{n} X_i$$

has an approximate normal distribution with mean θ and variance θ^2/n. Thus, in a limiting sense, $\hat{\theta}$ is **the unbiased minimum variance estimator** of θ.

To measure the value of estimators; their variances are compared to the Rao-Cramer lower bound. The ratio of the Rao-Cramer lower bound to the actual variance of any unbiased estimator is called **the efficiency** of that estimator. As estimator with efficiency of 50% requires that 1/0.5=2 times as many sample observations are needed to do as well in estimation as can be done with the unbiased minimum variance estimator (then 100% efficient estimator).

Chapter 4

Tests of Statistical Hypotheses

4.1 TEST ABOUT PROPORTIONS[1]

4.1.1 TEST ABOUT PROPORTIONS

Tests of statistical hypotheses are a very important topic, let introduce it through an illustration.

Suppose a manufacturer of a certain printed circuit observes that about p=0.05 of the circuits fails. An engineer and statistician working together suggest some changes that might improve the design of the product. To test this new procedure, it was agreed that n=100 circuits would be produced using the proposed method and the checked. Let Y equal the number of these 200 circuits that fail. Clearly, if the number of failures, Y, is such that $Y/200$ is about to 0.05, then it seems that the new procedure has not resulted in an improvement. On the other hand, If Y is small so that $Y/200$ is about 0.01 or 0.02, we might believe that the new method is better than the old one. On the other hand, if $Y/200$ is 0.08 or 0.09, the proposed method has perhaps caused a greater proportion of failures. What is needed is to establish a formal rule that tells when to accept the new procedure as an improvement. For example, we could accept the new procedure as an improvement if $Y \leq 5$ of $Y/n \leq 0.025$. We do note, however, that the probability of the failure could still be about p=0.05 even with the new procedure, and yet we could observe 5 of fewer failures in n=200 trials.

That is, we would accept the new method as being an improvement when, in fact, it was not. This decision is a mistake which we call a **Type I error**. On the other hand, the new procedure might actually improve the product so that p is much smaller, say p=0.02, and yet we could observe y=7 failures so that $y/200$=0.035. Thus we would not accept the new method as resulting in an improvement when in fact it had. This decision would also be a mistake which we call a **Type II error**.

If it we believe these trials, using the new procedure, are independent and have about the same probability of failure on each trial, then Y is binomial $b\left(200, p\right)$. We wish to make a statistical inference about p using the unbiased $\hat{p} = Y/200$. We could also construct a confidence interval, say one that has 95% confidence, obtaining

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{200}}.$$

This inference is very appropriate and many statisticians simply do this. If the limits of this confidence interval contain 0.05, they would not say the new procedure is necessarily better, al least until more data are taken. If, on the other hand, the upper limit of this confidence interval is less than 0.05, then they fell 95% confident that the true p is now less than 0.05. Here, in this illustration, we are testing whether or not the probability of failure has or has not decreased from 0.05 when the new manufacturing procedure is used.

The *no change* hypothesis, $H_0 : p = 0.05$, is called **the null hypothesis**. Since $H_0 : p = 0.05$ completely specifies the distribution it is called **a simple hypothesis**; thus $H_0 : p = 0.05$ is **a simple null hypothesis**.

[1] This content is available online at <http://cnx.org/content/m13525/1.2/>.

The research worker's hypothesis $H_1 : p < 0.05$ is called **the alternative hypothesis**. Since $H_1 : p < 0.05$ does not completely specify the distribution, it is a composite hypothesis because it is composed of many simple hypotheses.

The rule of rejecting H_0 and accepting H_1 if $Y \leq 5$, and otherwise accepting H_0 is called **a test of a statistical hypothesis**.

It is clearly seen that two types of errors can be recorded

- **Type I error:** Rejecting H_0 and accepting H_1, when H_0 is true;
- **Type II error:** Accepting H_0 when H_1 is true, that is, when H_0 is false.

Since, in the example above, we make a Type I error if $Y \leq 5$ when in fact p=0.05. we can calculate the probability of this error, which we denote by α and call **the significance level of the test**. Under an assumption, it is

$$\alpha = P\,(Y \leq 5; p = 0.05) = \sum_{y=0}^{5} \left(\begin{array}{c} 200 \\ y \end{array} \right) (0.05)^y (0.95)^{200-y}.$$

.

Since n is rather large and p is small, these binomial probabilities can be approximated extremely well by Poisson probabilities with $\lambda = 200\,(0.05) = 10$. That is, from the Poisson table, the probability of the Type I error is

$$\alpha \approx \sum_{y=0}^{5} \frac{10^y e^{-10}}{y!} = 0.067.$$

Thus, the approximate significance level of this test is $\alpha = 0.067$. This value is reasonably small. However, what about the probability of Type II error in case p has been improved to 0.02, say? This error occurs if $Y > 5$ when, in fact, p=0.02; hence its probability, denoted by β, is

$$\beta = P\,(Y > 5; p = 0.02) = \sum_{y=6}^{200} \left(\begin{array}{c} 200 \\ y \end{array} \right) (0.02)^y (0.98)^{200-y}.$$

Again we use the Poisson approximation, here λ=200(0.02)=4, to obtain

$$\beta \approx 1 - \sum_{y=0}^{5} \frac{4^y e^{-4}}{y!} = 1 - 0.785 = 0.215.$$

The engineers and the statisticians who created this new procedure probably are not too pleased with this answer. That is, they note that if their new procedure of manufacturing circuits has actually decreased the probability of failure to 0.02 from 0.05 (a big improvement), there is still a good chance, 0.215, that H$_0$: p=0.05 is accepted and their improvement rejected. Thus, this test of H$_0$: p=0.05 against H$_1$: p=0.02 is unsatisfactory. Without worrying more about the probability of the Type II error, here, above was presented a frequently used procedure for testing H$_0$: p=p$_0$, where p$_0$ is some specified probability of success. This test is based upon the fact that the number of successes, Y, in n independent Bernoulli trials is such that Y/n has an approximate normal distribution, N[p$_0$, p$_0$(1- p$_0$)/n], provided H$_0$: p=p$_0$ is true and n is large. Suppose the alternative hypothesis is H$_0$: p>p$_0$; that is, it has been hypothesized by a research worker that something has been done to increase the probability of success. Consider the test of H$_0$: p=p$_0$ against H$_1$: p> p$_0$ that rejects H$_0$ and accepts H$_1$ if and only if

$$Z = \frac{Y/n - p_0}{\sqrt{p_0\,(1 - p_0)\,/n}} \geq z_\alpha.$$

That is, if Y/n exceeds p$_0$ by standard deviations of Y/n, we reject H$_0$ and accept the hypothesis H$_1$: p> p$_0$. Since, under H$_0$ Z is approximately N (0,1), the approximate probability of this occurring when

H_0: $p=p_0$ is true is α. That is the significance level of that test is approximately α. If the alternative is H_1: $p< p_0$ instead of H_1: $p> p_0$, then the appropriate α-level test is given by $Z \leq -z_\alpha$. That is, if Y/n is smaller than p_0 by standard deviations of Y/n, we accept H_1: $p< p_0$.

In general, without changing the sample size or the type of the test of the hypothesis, a decrease in α causes an increase in β, and a decrease in β causes an increase in α. Both probabilities α and β of the two types of errors can be decreased only by increasing the sample size or, in some way, constructing a better test of the hypothesis.

4.1.1.1.1 EXAMPLE

If n=100 and we desire a test with significance level α=0.05, then $\alpha = P\left(\overline{X} \geq c; \mu = 60\right) = 0.05$ means, since $\overline{X}$ is N(μ,100/100=1),

$$P\left(\frac{\overline{X}-60}{1} \geq \frac{c-60}{1}; \mu = 60\right) = 0.05$$

and $c-60 = 1.645$. Thus c=61.645. The power function is

$$K(\mu) = P\left(\overline{X} \geq 61.645; \mu\right) = P\left(\frac{\overline{X}-\mu}{1} \geq \frac{61.645-\mu}{1}; \mu\right) = 1 - \Phi(61.645-\mu).$$

In particular, this means that β at μ=65 is

$$= 1 - K(\mu) = \Phi(61.645 - 65) = \Phi(-3.355) \approx 0;$$

so, with n=100, both α and β have decreased from their respective original values of 0.1587 and 0.0668 when n=25. Rather than guess at the value of n, an ideal power function determines the sample size. Let us use a critical region of the form $\overline{x} \geq c$. Further, suppose that we want α=0.025 and, when μ=65, β=0.05. Thus, since $\overline{X}$ is N(μ,100/n),

$$0.025 = P\left(\overline{X} \geq c; \mu = 60\right) = 1 - \Phi\left(\frac{c-60}{10/\sqrt{n}}\right)$$

and

$$0.05 = 1 - P\left(\overline{X} \geq c; \mu = 65\right) = \Phi\left(\frac{c-65}{10/\sqrt{n}}\right).$$

That is, $\frac{c-60}{10/\sqrt{n}} = 1.96$ and $\frac{c-65}{10/\sqrt{n}} = -1.645$.

Solving these equations simultaneously for c and $10/\sqrt{n}$, we obtain

$$c = 60 + 1.96\frac{5}{3.605} = 62.718;$$

$$\frac{10}{\sqrt{n}} = \frac{5}{3.605}.$$

Thus, $\sqrt{n} = 7.21$ and $n = 51.98$. Since n must be an integer, we would use n=52 and obtain α=0.025 and β=0.05, approximately.

For a number of years there has been another value associated with a statistical test, and most statistical computer programs automatically print this out; it is called **the probability value** or, for brevity, ***p*-value**. The p-value associated with a test is the probability that we obtain the observed value of the test statistic or a value that is more extreme in the direction of the alternative hypothesis, calculated when H_0 is true. Rather than select the critical region ahead of time, the p-value of a test can be reported and the reader then makes a decision.

Say we are testing H_0: μ=60 against H_1: μ>60 with a sample mean $\overline{X}$ based on n=52 observations. Suppose that we obtain the observed sample mean of $\overline{x} = 62.75$. If we compute the probability of obtaining an $\overline{x}$ of that value of 62.75 or greater when μ=60, then we obtain the p-value associated with $\overline{x}$ = 62.75. That is,

$$p - value = P\left(\overline{X} \geq 62.75; \mu = 60\right) = P\left(\frac{\overline{X}-60}{10/\sqrt{52}} \geq \frac{62.75-60}{10/\sqrt{52}}; \mu = 60\right)$$
$$= 1 - \Phi\left(\frac{62.75-60}{10/\sqrt{52}}\right) = 1 - \Phi\left(1.983\right) = 0.0237.$$

If this p-value is small, we tend to reject the hypothesis H_0: μ=60 . For example, rejection of H_0: μ=60 if the p-value is less than or equal to 0.025 is exactly the same as rejection if $\overline{x}$ = 62.718.That is, $\overline{x}$ = 62.718 has a p-value of 0.025. To help keep the definition of p-value in mind, we note that it can be thought of as that **tail-end probability**, under H_0, of the distribution of the statistic, here $\overline{X}$, beyond the observed value of the statistic. See Figure 1 (Figure 4.1) for the p-value associated with $\overline{x}$ = 62.75.

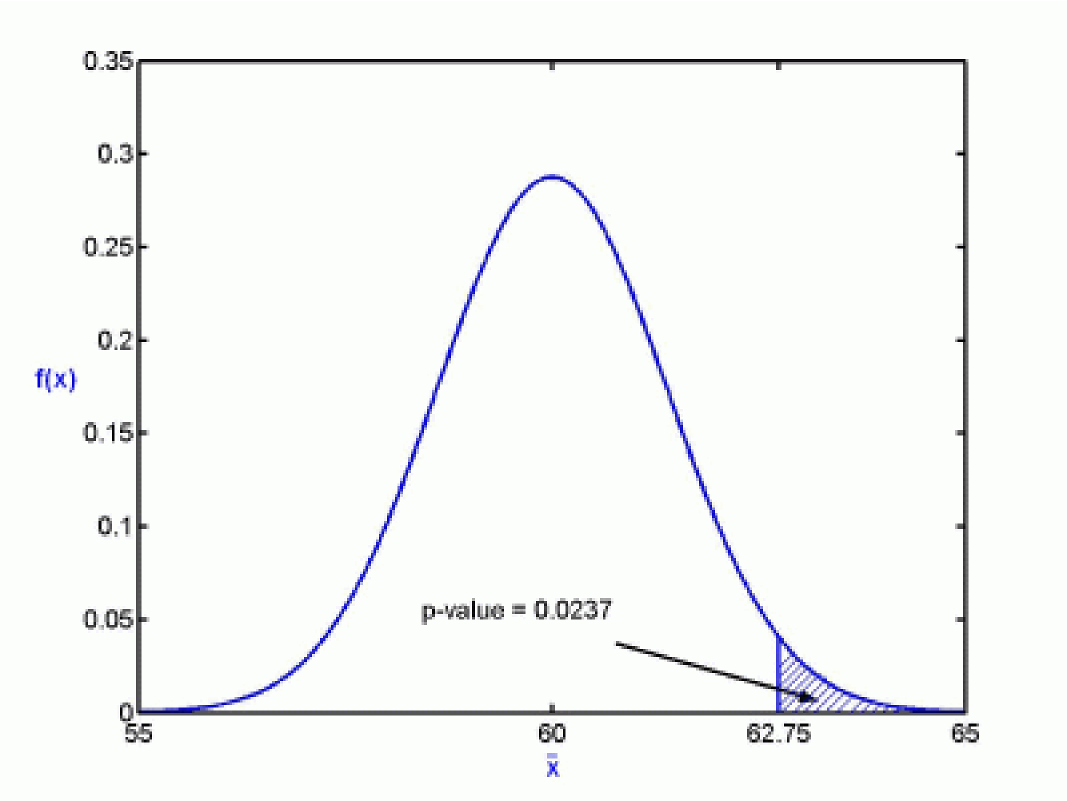


Figure 4.1: The p-value associated with $\overline{x}$ = 62.75.

Example 4.1
Suppose that in the past, a golfer's scores have been (approximately) normally distributed with mean μ=90 and σ^2=9. After taking some lessons, the golfer has reason to believe that the mean μ has decreased. (We assume that σ^2 is still about 9.) To test the null hypothesis H_0: μ=90 against the alternative hypothesis H_1: $\mu < 90$, the golfer plays 16 games, computing the sample mean $\overline{x}$.If

$\overline{x}$ is small, say $\overline{x} \leq c$, then H_0 is rejected and H_1 accepted; that is, it seems as if the mean μ has actually decreased after the lessons. If c=88.5, then the power function of the test is

$$K(\mu) = P\left(\overline{X} \leq 88.5; \mu\right) = P\left(\frac{\overline{X}-\mu}{3/4} \leq \frac{88.5-\mu}{3/4}; \mu\right) = \Phi\left(\frac{88.5-\mu}{3/4}\right).$$

Because 9/16 is the variance of $\overline{X}$. In particular,

$$\alpha = K(90) = \Phi(-2) = 1 - 0.9772 = 0.0228.$$

If, in fact, the true mean is equal to μ=88 after the lessons, the power is $K(88) = \Phi(2/3) = 0.7475$. If μ=87, then $K(87) = \Phi(2) = 0.9772$. An observed sample mean of $\overline{x} = 88.25$ has a

$$p - value = P\left(\overline{X} \leq 88.25; \mu = 90\right) = \Phi\left(\frac{88.25-90}{3/4}\right) = \Phi\left(-\frac{7}{3}\right) = 0.0098,$$

and this would lead to a rejection at α=0.0228 (or even α=0.01).

4.2 TESTS ABOUT ONE MEAN AND ONE VARIANCE[2]

4.2.1 TESTS ABOUT ONE MEAN AND ONE VARIANCE

In the previous paragraphs it was assumed that we were sampling from a normal distribution and the variance was known. The null hypothesis was generally of the form H$_0$: μ= μ_0.

There are essentially tree possibilities for the alternative hypothesis, namely that μ has increased,

1. H$_1$: $\mu >$ μ_0; μ has decreased,
2. H$_1$: $\mu <$ μ_0; μ has changed, but it is not known if it has increased or decreased, which leads to a two-sided alternative hypothesis
3. $H_1; \mu \neq \mu_0$.

To test $H_0; \mu = \mu_0$ against one of these tree alternative hypotheses, a random sample is taken from the distribution, and an observed sample mean, $\overline{x}$, that is close to μ_0 supports H$_0$. The closeness of $\overline{x}$ to μ_0 is measured in term of standard deviations of $\overline{X}$, $\sigma/\sqrt{n}$ which is sometimes called **the standard error of the mean**. Thus the statistic could be defined by

$$Z = \frac{\overline{X}-\mu_0}{\sqrt{\sigma 2/n}} = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}},$$

and the critical regions, at a significance level α, for the tree respective alternative hypotheses would be:

1. $z \geq z_\alpha$
2. $z \leq z_\alpha$
3. $|z| = z_{\alpha/2}$

In terms of $\overline{x}$ these tree critical regions become

1. $\overline{x} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$,
2. $\overline{x} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$,
3. $|\overline{x} - \mu_0| \geq z_\alpha \sigma/\sqrt{n}$

[2] This content is available online at <http://cnx.org/content/m13526/1.3/>.

These tests and critical regions are summarized in TABLE 1 (TABLE 1, p. 64). The underlying assumption is that the distribution is $N\left(\mu,\sigma^2\right)$ and σ^2 is known. Thus far we have assumed that the variance σ^2 was known. We now take a more realistic position and assume that the variance is unknown. Suppose our null hypothesis is $H_0; \mu = \mu_0$ and the two-sided alternative hypothesis is $H_1; \mu \neq \mu_0$. If a random sample $X_1, X_2, ..., X_n$ is taken from a normal distribution $N\left(\mu,\sigma^2\right)$,let recall that a confidence interval for μ was based on

$$T = \frac{\overline{X}-\mu}{\sqrt{S^2/n}} = \frac{\overline{X}-\mu}{S/\sqrt{n}}.$$

TABLE 1

H_0	H_1	**Critical Region**		
$\mu = \mu_0$	$\mu > \mu_0$	$z \geq z_\alpha$ or $\overline{x} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$		
$\mu = \mu_0$	$\mu < \mu_0$	$z \leq -z_\alpha$ or $\overline{x} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$		
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z\rvert \geq z_{\alpha/2}$ or $\lvert\overline{x} - \mu_0\rvert \geq z_{\alpha/2} \sigma/\sqrt{n}$		

This suggests that T might be a good statistic to use for the test $H_0; \mu = \mu_0$ with μ replaced by μ_0. In addition, it is the natural statistic to use if we replace σ^2/n by its unbiased estimator S^2/n in $\left(\overline{X}-\mu_0\right)/\sqrt{\sigma^2/n}$ in a proper equation. If $\mu = \mu_0$ we know that T has a t distribution with n-1 degrees of freedom. Thus, with $\mu = \mu_0$,

$$P\left[\lvert T\rvert \geq t_{\alpha/2}\left(n-1\right)\right] = P\left[\frac{\lvert\overline{X}-\mu_0\rvert}{S/\sqrt{n}} \geq t_{\alpha/2}\left(n-1\right)\right] = \alpha.$$

Accordingly, if $\overline{x}$ and s are the sample mean and the sample standard deviation, the rule that rejects $H_0; \mu = \mu_0$ if and only if

$$\lvert t\rvert = \frac{\lvert\overline{x}-\mu_0\rvert}{s/\sqrt{n}} \geq t_{\alpha/2}\left(n-1\right).$$

Provides the test of the hypothesis with significance level α. It should be noted that this rule is equivalent to rejecting $H_0; \mu = \mu_0$ if μ_0 is not in the open $100\left(1-\alpha\right)\%$ confidence interval

$$\left(\overline{x} - t_{\alpha/2}\left(n-1\right)s/\sqrt{n}, \overline{x} + t_{\alpha/2}\left(n-1\right)s/\sqrt{n}\right).$$

Table 2 (TABLE 2, p. 64) summarizes tests of hypotheses for a single mean, along with the three possible alternative hypotheses, when the underlying distribution is $N\left(\mu,\sigma^2\right)$, σ^2 is unknown, $t = \left(\overline{x}-\mu_0\right)/\left(s/\sqrt{n}\right)$ and $n \leq 31$. If n>31, use table 1 (TABLE 1, p. 64) for approximate tests with σ replaced by s.

TABLE 2

H_0	H_1	**Critical Region**		
$\mu = \mu_0$	$\mu > \mu_0$	$t \geq t_\alpha\left(n-1\right)$ or $\overline{x} \geq \mu_0 + t_\alpha\left(n-1\right)s/\sqrt{n}$		
$\mu = \mu_0$	$\mu < \mu_0$	$t \leq -t_\alpha\left(n-1\right)$ or $\overline{x} \leq \mu_0 - t_\alpha\left(n-1\right)s/\sqrt{n}$		
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t\rvert \geq t_{\alpha/2}\left(n-1\right)$ or $\lvert\overline{x} - \mu_0\rvert \geq t_{\alpha/2}\left(n-1\right)s/\sqrt{n}$		

Example 4.2

Let X (in millimeters) equal the growth in 15 days of a tumor induced in a mouse. Assume that the distribution of X is $N\left(\mu, \sigma^2\right)$. We shall test the null hypothesis $H_0 : \mu = \mu_0 = 4.0$ millimeters against the two-sided alternative hypothesis is $H_1 : \mu \neq 4.0$. If we use n=9 observations and a significance level of α =0.10, the critical region is

$$|t| = \frac{|\overline{x} - 4.0|}{s/\sqrt{9}} \geq t_{\alpha/2}(8) = t_{0.05}(8) = 1.860.$$

If we are given that n=9, $\overline{x}$=4.3, and s=1.2, we see that

$$t = \frac{4.3 - 4.0}{1.2/\sqrt{9}} = \frac{0.3}{0.4} = 0.75.$$

Thus $|t| = |0.75| < 1.860$ and we accept (do not reject) $H_0 : \mu = 4.0$ at the α=10% significance level. See Figure 1 (Figure 4.2).

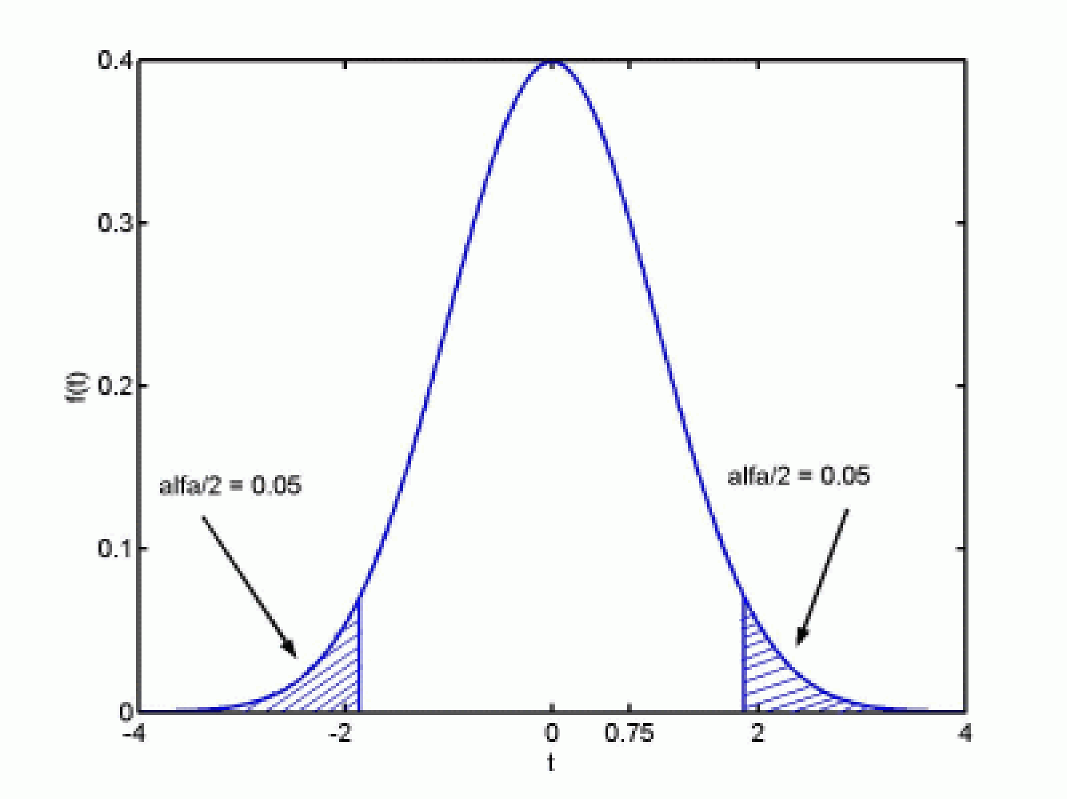


Figure 4.2: Rejection region at the $\alpha = 10\%$ significance level.

REMARK: In discussing the test of a statistical hypothesis, the word *accept* might better be replaced by *do not reject*. That is, in Example 1 (Example 4.2), $\overline{x}$ is close enough to 4.0 so that we accept μ=4.0, we do not want that acceptance to imply that μ is actually equal to 4.0. We want to say that the data do not deviate enough from μ=4.0 for us to reject that hypothesis; that is, we do not reject μ=4.0 with these observed data, With this understanding, one sometimes uses *accept* and sometimes *fail to reject* or *do not reject*, the null hypothesis.

In this example the use of the t-statistic with a one-sided alternative hypothesis will be illustrated.

Example 4.3

In attempting to control the strength of the wastes discharged into a nearby river, a paper firm has taken a number of measures. Members of the firm believe that they have reduced the oxygen-consuming power of their wastes from a previous mean μ of 500. They plan to test $H_0 : \mu = 500$ against $H_1 : \mu < 500$, using readings taken on n=25 consecutive days. If these 25 values can be treated as a random sample, then the critical region, for a significance level of α=0.01, is

$$t = \frac{\overline{x} - 500}{s/\sqrt{25}} \leq -t_{0.01}\left(24\right) = -2.492.$$

The observed values of the sample mean and sample standard deviation were $\overline{x}$=308.8 and s=115.15. Since

$$t = \frac{308.8 - 500}{115.15/\sqrt{25}} = -8.30 < \ -2.492,$$

we clearly reject the null hypothesis and accept $H_1 : \mu < 500$. It should be noted, however, that although an improvement has been made, there still might exist the question of whether the improvement is adequate. The 95% confidence interval $308.8 \pm 2.064\left(115.15/5\right)$ or [261.27, 356.33] for μ might the company answer that question.

4.3 TEST OF THE EQUALITY OF TWO INDEPENDENT NORMAL DISTRIBUTIONS[3]

4.3.1 TEST OF THE EQUALITY OF TWO INDEPENDENT NORMAL DISTRIBUTIONS

Let X and Y have independent normal distributions $N\left(\mu_x, \sigma_x^2\right)$ and $N\left(\mu_y, \sigma_y^2\right)$, respectively. There are times when we are interested in testing whether the distribution of X and Y are the same. So if the assumption of normality is valid, we would be interested in testing whether the two variances are equal and whether the two mean are equal.

Let first consider a test of the equality of the two means. When X and Y are independent and normally distributed, we can test hypotheses about their means using the same t-statistic that was used previously. Recall that the t-statistic used for constructing the confidence interval assumed that the variances of X and Y are equal. That is why we shall later consider a test for the equality of two variances.

Let start with an example and then let give a table that lists some hypotheses and critical regions.

Example 4.4

A botanist is interested in comparing the growth response of dwarf pea stems to two different levels of the hormone indoeacetic acid (IAA). Using 16-day-old pea plants, the botanist obtains 5-millimeter sections and floats these sections with different hormone concentrations to observe the effect of the hormone on the growth of the pea stem.

Let X and Y denote, respectively, the independent growths that can be attributed to the hormone during the first 26 hours after sectioning for $\left(0.5\right)\left(10\right)^{-4}$ and $\left(10\right)^{-4}$ levels of concentration of IAA. The botanist would like to test the null hypothesis $H_0 : \mu_x - \mu_y = 0$ against the alternative hypothesis $H_1 : \mu_x - \mu_y < 0$. If we can assume X and Y are independent and normally distributed with common variance, respective random samples of size n and m give a test based on the statistic

[3] This content is available online at <http://cnx.org/content/m13532/1.2/>.

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\left\{ \left[(n-1)\, S_x^2 + (m-1)\, S_{Y_y}^2 \right] / (n+m-2) \right\} (1/n + 1/m)}} = \frac{\overline{X} - \overline{Y}}{S_P \sqrt{1/n + 1/m}},$$

where

$$S_P = \sqrt{\frac{(n-1)\, S_X^2 + (m-1)\, S_Y^2}{n+m-2}}.$$

T has a t distribution with $r = n + m - 2$ degrees of freedom when H_0 is true and the variances are (approximately) equal. The hypothesis Ho will be rejected in favor of H_1 if the observed value of T is less than $-t_\alpha (n + m - 2)$.

Example 4.5

In the example 1 (Example 4.4), the botanist measured the growths of pea stem segments, in millimeters, for n=11 observations of X given in the Table 1:

Table 1

0.8	1.8	1.0	0.1	0.9 1.7	1.0	1.4	0.9	1.2	0.5	

and m=13 observations of Y given in the Table 2:

Table 2

1.0	0.8	1.6	2.6	1.3 1.1	2.4	1.8	2.5	1.4	1.9	2.0	1.2	

For these data, $\overline{x} = 1.03, s_x^2 = 0.24, \overline{y} = 1.66$, and $s_y^2 = 0.35$. The critical region for testing $H_0 : \mu_x - \mu_y = 0$ against $H_1 : \mu_x - \mu_y < 0$ is $t \leq -t_{0.05} (22) = -1.717$. Since H_0 is clearly rejected at α=0.05 significance level.

Notice that: an approximate p-value of this test is 0.005 because $-t_{0.05} (22) = -2.819$. Also, the sample variances do not differ too much; thus most statisticians would use this two sample t-test.

4.4 BEST CRITICAL REGIONS[4]

4.4.1 BEST CRITICAL REGIONS

In this paragraph, let consider the properties a satisfactory test should posses.

Definition 18:

1. Consider the test of the sample null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$.

2. Let C be a critical region of size α; that is, $\alpha = P(C; \theta_0)$. Then C is **a best critical region of size** α if, for every other critical region D of size $\alpha = P(D; \theta_0)$, we have that

$$P(C; \theta_1) \geq P(D; \theta_1).$$

[4] This content is available online at <http://cnx.org/content/m13528/1.2/>.

That is, when $H_1 : \theta = \theta_1$ is true, the probability of rejecting $H_0 : \theta = \theta_0$ using the critical region C is at least as great as the corresponding probability using any other critical region D of size α.

Thus a best critical region of size α is the critical region that has the greatest power among all critical regions for a best critical region of size α. **The Neyman-Pearson lemma** gives sufficient conditions for a best critical region of size α.

Theorem 4.1: Neyman-Pearson Lemma

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with p.d.f. $f(x;\theta)$, where θ_0 and θ_1 are two possible values of θ.

Denote the joint p.d.f. of $X_1, X_2, ..., X_n$ by the likelihood function

$$L(\theta) = L(\theta; x_1, x_2, ..., x_n) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_{n.};\theta).$$

If there exist a positive constant k and a subset C of the sample space such that

1. $P[(X_1, X_2, ..., X_n) \in C; \theta_0] = \alpha$,
2. $\frac{L(\theta_0)}{L(\theta_1)} \leq k$ for $(x_1, x_2, ..., x_n) \in C$,
3. $\frac{L(\theta_0)}{L(\theta_1)} \geq k$ for $(x_1, x_2, ..., x_n) \in C$'.

Then C is a best critical region of size α for testing the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$.

For a realistic application of the Neyman-Pearson lemma, consider the following, in which the test is based on a random sample from a normal distribution.

Example 4.6

Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution $N(\mu, 36)$. We shall find the best critical region for testing the simple hypothesis $H_0 : \mu = 50$ against the simple alternative hypothesis $H_1 : \mu = 55$. Using the ratio of the likelihood functions, namely $L(50)/L(55)$, we shall find those points in the sample space for which this ratio is less than or equal to some constant k.

That is, we shall solve the following inequality:

$$\frac{L(50)}{L(55)} = \frac{(72\pi)^{-n/2} exp\left[-\left(\frac{1}{72}\right)\sum_1^n (x_i-50)^2\right]}{(72\pi)^{-n/2} exp\left[-\left(\frac{1}{72}\right)\sum_1^n (x_i-55)^2\right]}$$
$$= exp\left[-\left(\frac{1}{72}\right)\left(10\sum_1^n x_i + n50^2 - n55^2\right)\right] \leq k.$$

If we take the natural logarithm of each member of the inequality, we find that

$$-10\sum_1^n x_i - n50^2 + n55^2 \leq (72)\, lnk.$$

Thus,

$$\frac{1}{n}\sum_1^n x_i \geq -\frac{1}{10n}\left[n50^2 - n55^2 + (72)\, lnk\right]$$

Or equivalently, $\overline{x} \geq c$, where $c = -\frac{1}{10n}\left[n50^2 - n55^2 + (72)\, lnk\right]$.

Thus $L(50)/L(55) \leq k$ is equivalent to $\overline{x} \geq c$.

A best critical region is, according to the Neyman-Pearson lemma,

$$C = \{(x_1, x_2, ..., x_n) : \overline{x} \geq c\},$$

where c is selected so that the size of the critical region is α. Say n=16 and c=53. Since $\overline{X}$ is $N(50, 36/16)$ under H$_0$ we have

$$\alpha = P\left(\overline{X} \geq 53; \mu = 50\right) = P\left[\frac{\overline{X} - 50}{6/4} \geq \frac{3}{6/4}; \mu = 50\right] = 1 - \Phi\left(2\right) = 0.0228.$$

The example 1 illustrates what is often true, namely, that the inequality

$$\frac{L\left(\theta_0\right)}{L\left(\theta_1\right)} \leq k$$

can be expressed in terms of a function $u\left(x_1, x_2, ..., x_n\right)$ say,

$$u\left(x_1, x_2, ..., x_n\right) \leq c_1$$

or

$$u\left(x_1, x_2, ..., x_n\right) \geq c_2,$$

where c_1 and c_2 is selected so that the size of the critical region is α . Thus the test can be based on the statistic $u\left(X_1, ..., X_n\right)$. Also, for illustration, if we want α to be a given value, say 0.05, we would then choose our c_1 and c_2. In example1, with α=0.05, we want

$$0.05 = P\left(\overline{X} \geq c; \mu = 50\right) = P\left(\frac{\overline{X} - 50}{6/4} \geq \frac{c - 50}{6/4}; \mu = 50\right) = 1 - \Phi\left(\frac{c - 50}{6/4}\right).$$

Hence it must be true that $(c - 50) / (3/2) = 1.645$, or equivalently, $c = 50 + \frac{3}{2}(1.645) \approx 52.47$.

Example 4.7

Let$X_1, X_2, ..., X_n$ denote a random sample of size n from a Poisson distribution with mean λ. A best critical region for testing $H_0 : \lambda = 2$ against $H_1 : \lambda = 5$ is given by

$$\frac{L\left(2\right)}{L\left(5\right)} = \frac{2^{\sum x_i} e^{-2n}}{x_1! x_2! \cdots x_n!} \frac{x_1! x_2! \cdots x_n!}{5^{\sum x_i} e^{-5n}} \leq k.$$

The inequality is equivalent to $\left(\frac{2}{5}\right)^{\sum x_i} e^{3n} \leq k$ and $\left(\sum x_i\right) ln\left(\frac{2}{5}\right) + 3n \leq lnk$. Since $ln\left(2/5\right) < 0$, this is the same as

$$\sum_{i=1}^{n} x_i \geq \frac{lnk - 3n}{ln\left(2/5\right)} = c.$$

If n=4 and c=13, then

$$\alpha = P\left(\sum_{i=1}^{4} X_i \geq 13; \lambda = 2\right) = 1 - 0.936 = 0.064,$$

from the tables, since $\sum_{i=1}^{4} X_i$ has a Poisson distribution with mean 8 when λ=2.

When $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ are both simple hypotheses, a critical region of size α is a best critical region if the probability of rejecting H$_0$ when H$_1$ is true is a maximum when compared with all other critical regions of size α. The test using the best critical region is called **a most powerful test** because it has the greatest value of the power function at $\theta = \theta_1$ when compared with that of other tests of significance level α. If H$_1$ is a composite hypothesis, the power of a test depends on each simple alternative in H$_1$.

Definition 19:

A test, defined by a critical region C of size α, is **a uniformly most powerful test** if it is a most powerful test against each simple alternative in H$_1$. The critical region C is called **a uniformly most powerful critical region of size** α.

Let now consider the example when the alternative is composite.

Example 4.8
Let $X_1, X_2, ..., X_n$ be a random sample from $N(\mu, 36)$. We have seen that when testing $H_0 : \mu = 50$ against $H_1 : \mu = 55$, a best critical region C is defined by

$$C = \{(x_1, x_2, ..., x_n) : \overline{x} \geq c\},$$

where c is selected so that the significance level is α. Now consider testing $H_0 : \mu = 50$ against the one-sided composite alternative hypothesis $H_1 : \mu > 50$. For each simple hypothesis in H_1, say $\mu = \mu_1$ the quotient of the likelihood functions is

$$\frac{L(50)}{L(\mu_1)} = \frac{(72\pi)^{-n/2} exp\left[-\left(\frac{1}{72}\right)\sum_1^n (x_i - 50)^2\right]}{(72\pi)^{-n/2} exp\left[-\left(\frac{1}{72}\right)\sum_1^n (x_i - \mu_1)^2\right]}$$
$$= exp\left[-\left(\frac{1}{72}\right)\left\{2(\mu_1 - 50)\sum_1^n x_i + n\left(50^2 - \mu_1^2\right)\right\}\right].$$

Now $L(50)/L(\mu_1) \leq k$ if and only if

$$\overline{x} \geq \frac{(-72)\, ln(k)}{2n(\mu_1 - 50)} + \frac{50 + \mu_1}{2} = c.$$

Thus the best critical region of size α for testing $H_0 : \mu = 50$ against $H_1 : \mu = \mu_1$, where $\mu_1 > 50$, is given by

$$C = \{(x_1, x_2, ..., x_n) : \overline{x} \geq c\},$$

where is selected such that

$$P\left(\overline{X} \geq c; H_0 : \mu = 50\right) = \alpha.$$

NOTE THAT: the same value of c can be used for each $\mu_1 > 50$, but of course k does not remain the same. Since the critical region C defines a test that is most powerful against each simple alternative $\mu_1 > 50$, this is a uniformly most powerful test, and C is a uniformly most powerful critical region if size α. Again if α=0.05, then $c \approx 52.47$.

4.5 HYPOTHESES TESTING[5]

4.5.1 Hypotheses Testing - Examples.

Example 4.9
We have tossed a coin 50 times and we got $\boldsymbol{k} = \mathbf{19}$ **heads**. Should we accept/reject the hypothesis that $p = 0.5$, provided taht the coin is fair?

Null versus Alternative Hypothesis:

- Null hypothesis $(H_0) : p = 0.5$.
- Alternative hypothesis $(H_1) : p \neq 0.5$.

[5] This content is available online at <http://cnx.org/content/m13533/1.2/>.

EXPERIMENT

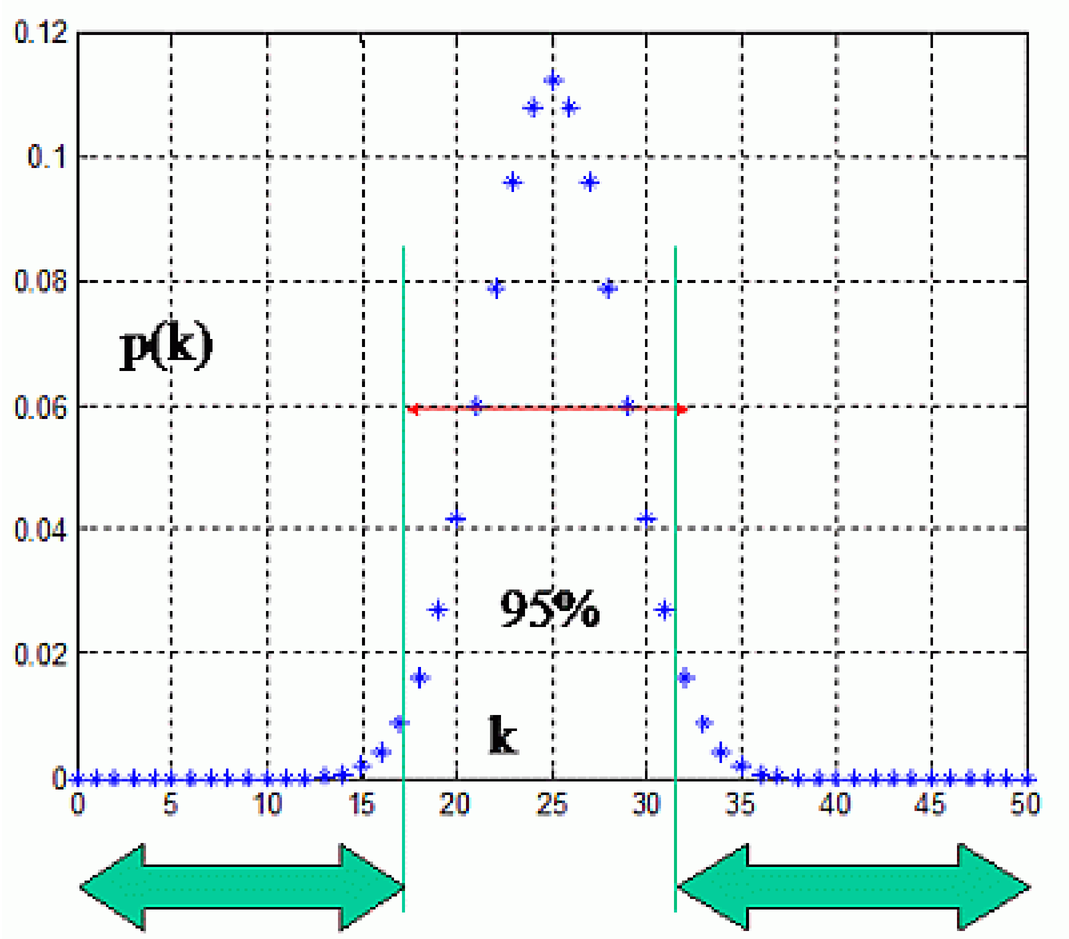


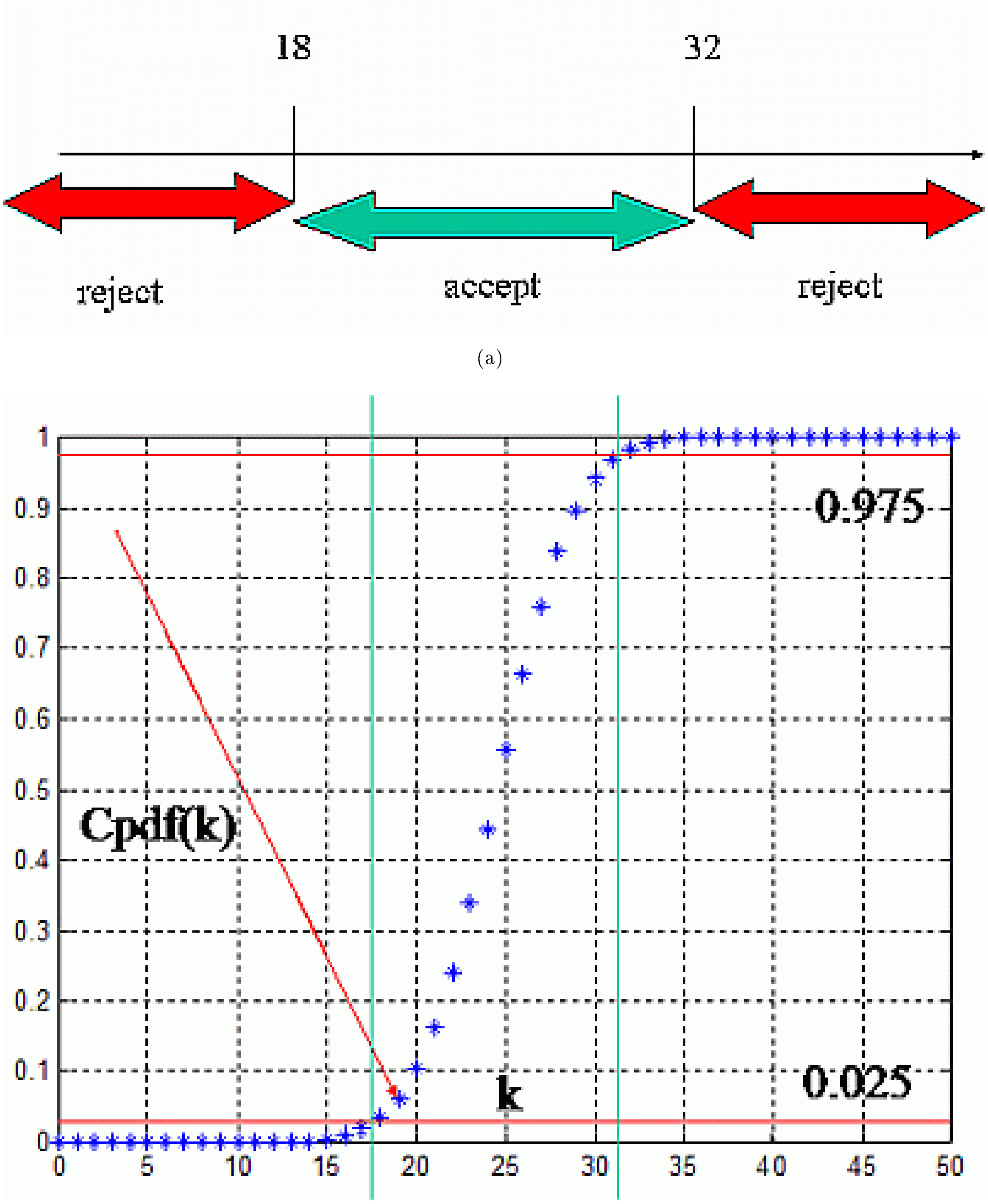
Figure 4.3

Significance level α = Probability of Type I error = Pr[rejecting H_0 | H_0 true]

$\boldsymbol{P}$**[**$k < 18$ **or** $k > 32$**]**< 0.05**.**

If $k < 18$ or $k > 32] < 0.05$, then under the null hypothesis the observed event falls into rejection region with the probability $\alpha < 0.05$.

Note that: We want α as small as possible.



(a)

(b)

Figure 4.4: (a) Test construction. (b) Cumulative distribution function.

Conclusion: No evidence to reject the null hypothesis.

Example 4.10
We have tossed a coin 50 times and we got $\boldsymbol{k} = \mathbf{10}$ **heads**. Should we accept/reject the hypothesis

that $p = 0.5$, provided taht the coin is fair?

EXPERIMENT

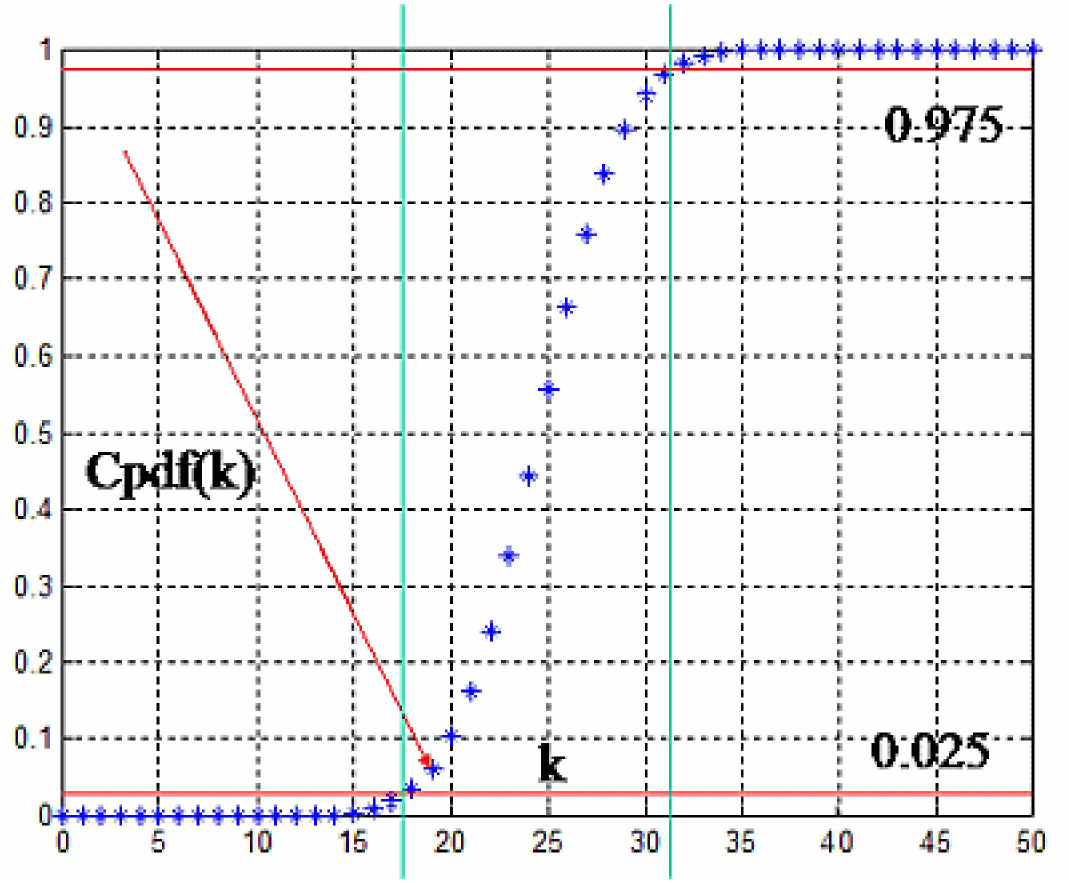


Figure 4.5: Cumulative distribution function.

***P*[$k \leq 10$ or $k \geq 40$]**≈ 0.000025. We could **reject** hypothesis H_0 at a significance level as low as $\alpha = 0.000025$.

NOTE THAT: p-value is the lowest attainable significance level.

REMARK: In STATISTICS, to prove something = **reject** the hypothesis that converse is true.

Example 4.11
We know that on average mouse tail is 5 cm long. We have a group of 10 mice, and give to each of them a dose of vitamin T everyday, from the birth, for the period of 6 months.

We want to prove that vitamin X makes mouse tail longer. We measure tail lengths of out group and we get the following sample:

Table 1

5.5	5.6	4.3	5.1	5.2 6.1	5.0	5.2	5.8	4.1	

- Hypothesis H_0 - sample = sample from normal distribution with $\mu = 5$ cm.
- Alternative H_1 - sample = sample from normal distribution with $\mu > 5$ cm.

CONSTRUCTION OF THE TEST

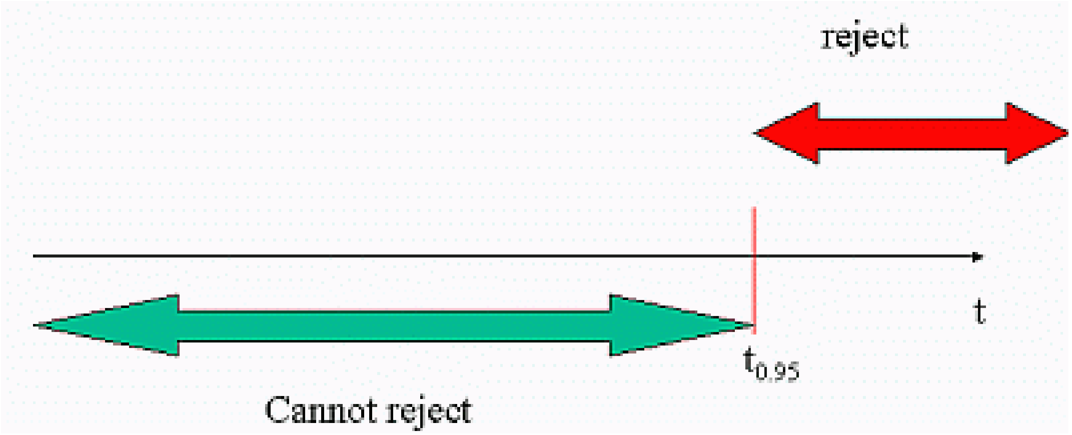


Figure 4.6

We do not know population variance, and/or we suspect that vitamin treatment may change the variance - so we use t distribution (Section 2.5.1: THE t DISTRIBUTION).

- $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$,
- $S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$,
- $t = \frac{\overline{X}-\mu}{S}\sqrt{N-1}$.

Example 4.12
χ^2 **test (K. Pearson, 1900)**

To test the hypothesis that a given data actually come from a population with the proposed distribution. Data is given in the Table 2 (DATA, p. 74).

DATA

0.4319	0.6874	0.5301	0.8774	0.6698	1.1900	0.4360	0.2192	0.5082	
0.3564	1.2521	0.7744	0.1954	0.3075	0.6193	0.4527	0.1843	2.2617	
0.4048	2.3923	0.7029	0.9500	0.1074	3.3593	0.2112	0.0237	0.0080	
0.1897	0.6592	0.5572	1.2336	0.3527	0.9115	0.0326	0.2555	0.7095	
0.2360	1.0536	0.6569	0.0552	0.3046	1.2388	0.1402	0.3712	1.6093	
1.2595	0.3991	0.3698	0.7944	0.4425	0.6363	2.5008	2.8841	0.9300	
3.4827	0.7658	0.3049	1.9015	2.6742	0.3923	0.3974	3.3202	3.2906	
1.3283	0.4263	2.2836	0.8007	0.3678	0.2654	0.2938	1.9808	0.6311	
0.6535	0.8325	1.4987	0.3137	0.2862	0.2545	0.5899	0.4713	1.6893	
0.6375	0.2674	0.0907	1.0383	1.0939	0.1155	1.1676	0.1737	0.0769	
1.1692	1.1440	2.4005	2.0369	0.3560	1.3249	0.1358	1.3994	1.4138	
0.0046	-	-	-	-	-	-	-	-	

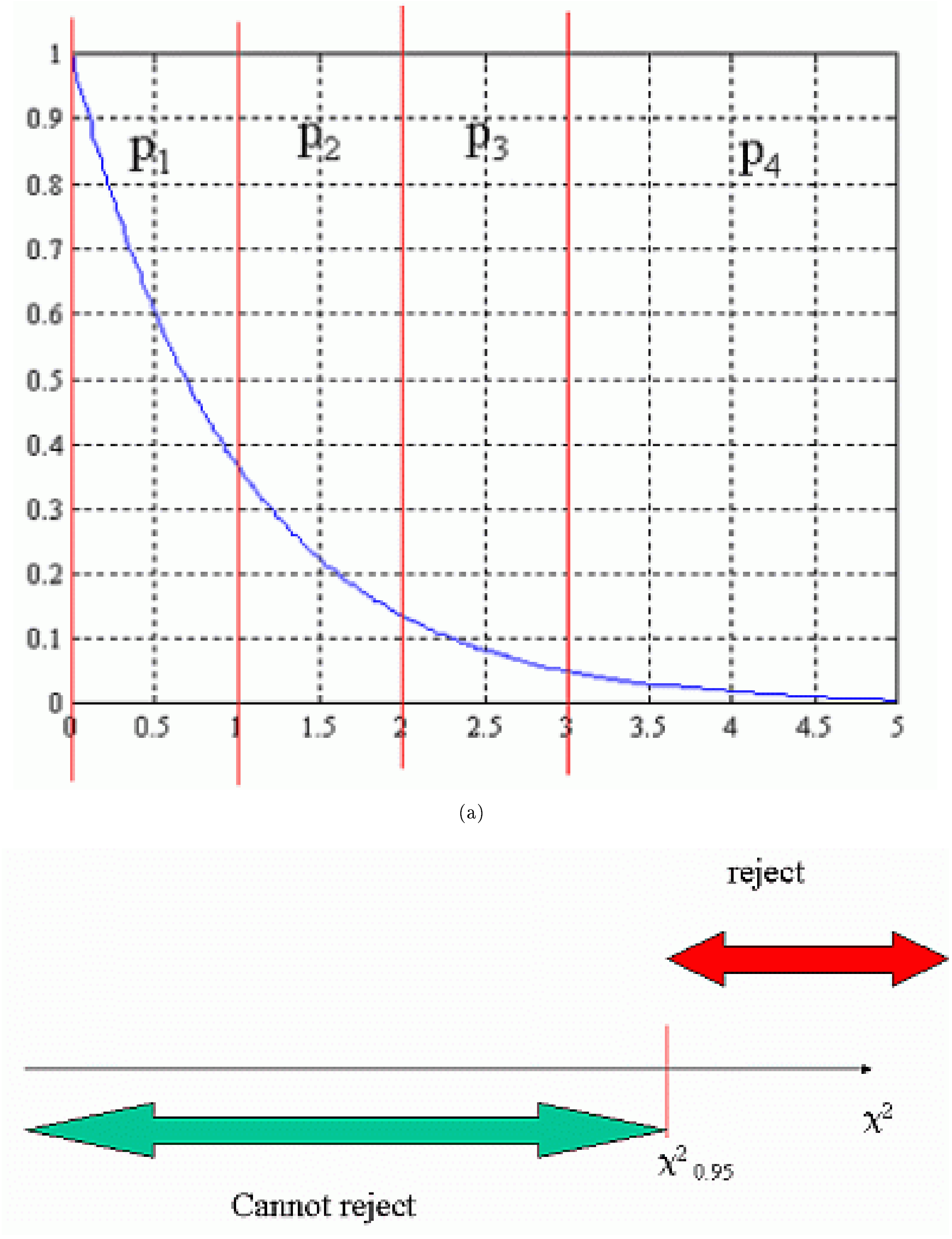
Problem

Are these data sampled from population with exponential p.d.f.?

Solution

$f(x) = e^{-x}$.

CONSTRUCTION OF THE TEST



(a)

(b)

Figure 4.7

Exercise 4.1 ***(Solution on p. 78.)***

Are these data sampled from population with exponential p.d.f.?

TABLE 1

Actual Situation	H_o **true**		H_o **false**	
decision	accept	Reject = error t. I	reject	Accept = error t. II
probability	$1 - \alpha$	α = significance level	$1 - \beta$ = power of the test	β

Solutions to Exercises in Chapter 4

Solution to Exercise 4.1 (p. 77)
$f(x) = ae^{-ax}$.

1. Estimate a.
2. Use χ^2 test.
3. Remember d.f. $=$ K-2.

Chapter 5

Pseudo - Numbers

5.1 PSEUDO-NUMBERS[1]

5.1.1 UNIFORM PSEUDO-RANDOM VARIABLE GENERATION

In this paragraph, our goals will be to look at, in more detail, how and whether particular types of pseudo-random variable generators work, and how, if necessary, we can implement a generator of our own choosing. Below a list of requirements is listed for our uniform random variable generator:

1. A uniform marginal distribution,
2. Independence of the uniform variables,
3. Repeatability and portability,
4. Computational speed.

5.1.1.1 CURRENT ALGORITHMS

The generation of pseudo-random variates through algorithmic methods is a mature field in the sense that a great deal is known theoretically about different classes of algorithms, and in the sense that particular algorithms in each of those classes have been shown, upon testing, to have good statistical properties. In this section, let describe the main classes of generators, and then let make specific recommendation about which generators should be implemented.

Congruential Generators

The most widely used and best understood class of pseudo-random number generators are those based on the linear congruential method introduced by *Lehmer (1951)*. Such generators are based on the following formula:

$$U_i = (aU_{i-1} + c)\, modm, \tag{5.1}$$

where $U_i, i = 1, 2, ...$ are the output random integers; U_0 is the chosen starting value for the recursion, called **the seed** and a,c, and m are prechosen constants.

> NOTICE THAT: to convert to uniform $(0, 1)$ variates, we need only divide by **modulus m**, that is, we use the sequence $\{U_i/m\}$.

The following properties of the algorithm are worth stating explicitly:

[1] This content is available online at <http://cnx.org/content/m13103/1.6/>.

1. Because of the "mod m" operation (for background on modular operations, see *Knuth, (1981)*), the only possible values the algorithm can produce are the integers $0, 1, 2, ..., m - 1$. This follows because, by definition, x mod m is the remainder after x is divided by m.
2. Because the current random integer U_i depends only on the previous random integer U_{i-1} once a previous value has been repeated, the entire sequence after it must be repeated. Such a repeating sequence is called **a cycle**, and its **period** is **the cycle length**. Clearly, **the maximum period** of the congruential generator is m. For given choices of a, c, and m, a generator may contain many short cycles, (see the Example 1 below), and the cycle you enter will depend on the seed you start with. Notice that the generator with many short cycles is not a good one, since the output sequence will be one of a number of short series, each of which may not be uniformly distributed or randomly dispersed on the line or the plane. Moreover, if the simulation is long enough to cause the random numbers to repeat because of the short cycle length, the outputs will not be independent.
3. If we are concern with a uniform $(0, 1)$ variates, the finest partition of the interval $(0, 1)$ that this generator can provide is $[0, 1/m, 2/m, ..., (m - 1/m)]$. This is, of course, not truly a uniform $(0, 1)$ distribution since, for any k in $(0, m - 1)$, we have $P\left[k/m < U < (k+1)/m\right] = 0$, not $1/m$ are required by theory for continuous random variables.
4. Choices of a,c, and m, will determine not only the fineness of the partition of $(0, 1)$ and the cycle length, and therefore, the uniformity of the marginal distribution, but also the independence properties of the output sequence. Properly choosing a,c, and m is a science that incorporates both theoretical results and empirical tests. The first rule is to select the modulus m to be "as large as possible", so that there is some hope to address point 3 above and to generate uniform variates with an approximately uniform marginal distribution. However, simply having m large is not enough; one may still find that the generator has many short cycles, or that the sequence is not approximately independent. See example 1 (Example 5.1) below.

Example 5.1
Consider

$$U_i = 2U_{i-1} mod 2^{32} \tag{5.2}$$

Where a seed of the form 2^k creates a loop containing only integers that are powers of 2, or

$$U_i = (U_{i-1} + 1)\, mod 2^{32} \tag{5.3}$$

which generates the nonrandom sequence of increasing integers. Therefore, the second equation gives a generator that has the maximum possible cycle length but is useless for simulating a random sequence.

Fortunately, one a value of the m has been selected; theoretical results exist that give conditions for choosing values of the multiplier a and the additive constant c such that all the possible integers, 0 through $m - 1$, are generated before any are repeated.

Notice, that: this does not eliminate the second counterexample above, which already has the maximal cycle length, but is a useless random number generator.

THEOREM I
A linear congruential generator will have maximal cycle length m, if and only if:

- c is nonzero and is relatively prime to m (i.e., c and m have no common prime factors).
- $(a mod q) = 1$ for each prime factor q of m.
- $(a mod 4) = 1$ if 4 is a factor of m.

PROOF

SEE: *Knuth (1981, p.16).*

As a mathematical note, c is called relatively prime to m if and only if c and m have no common divisor other than 1, which is equivalent to c and m having no common prime factor.

A related result concerns the case of c chosen to be 0. This case does not conform to condition in a Theorem I (p. 80), a value U_i of zero must be avoided because the generator will continue to produce zero after the first occurrence of a zero. In particular, a seed of zero is not allowable. By Theorem I (p. 80), a generator with $c = 0$, which is called **a multiplicative congruential generator**, cannot have maximal cycle length m. However, By Theorem II (p. 81). It can have cycle length $m - 1$.

THEOREM II

If $c = 0$ in a linear congruential generator, then $U_i = 0$ can never be included in a cycle, since the 0 will always repeat. However, the generator will cycle through all $m - 1$ integers in the set $(amodq)$ if and only if:

- m is a prime integer and
- m is a primitive element modulo m .

PROOF

SEE: *Knuth (1981, p.19).*

A formal definition or primitive elements modulo m, as well as theoretical results for finding them, are given in *Knuth (1981).* In effect, when m is a prime, a is a primitive element if the cycle is of length $m - 1$. The results of Theorem II (p. 81) are not intuitively useful, but for our purposes, it is enough to note that such primitive elements exist and have veen computed by researchers,

SEE: e.g., Table24.8 in Abramowitz and Stegun, 1965.

Hence, we now must select one of two possibilities:

- Choose a, c, and m according to Theorem I (p. 80) and work with a generator whose cycle length is known to be m.
- Choose $c = 0$, take a and m according to Theorem II (p. 81), use a number other than zero as the seed, and work with a generator whose cycle length is known to be $m-1$. A generator satisfying these conditions is known as **a prime-modulus multiplicative congruential generator** and, because of the simpler computation, it usually has an advantage in terms of speed over the mixed congruential generator.

Another method frequency speeding up a random number generator that has $c = 0$ is to choose the modulus m to be computationally convenient. For instance, consider $m = 2^k$. This is clearly not a prime number, but on a computer the modulus operation becomes a bit-shift operation in machine code. In such cases, Theorem III gives a guise to the maximal cycle length.

THEOREM III

If $c = 0$ and $m = 2^k$ with $k > 2$, then the maximal possible cycle length is 2^{k-2}. This is achieved if and only if two conditions hold:

- a is a primitive element modulo m.
- the seed is odd.

PROOF

SEE: *Knuth (1981, p.19).*

Notice that we sacrifice some of the cycle length and, as we will se in Theorem IV, we also lose some randomness in the low-order bits of the random variates. Having use any of Theorems I (p. 80), II (p. 81), or III (p. 81) to select triples (a, c, m) that lead to generators with sufficiently long cycles of known length, we can ask which triple gives the most random (i.e., approximately independent) sequence. Although some theoretical results exist for generators as a whole, these are generally too weak to eliminate any but the worst generators. *Marsaglia (1985)* and *Knuth(1981, Chap. 3.3.3)* are good sources for material on that results.

THEOREM IV

If $U_i = aU_{i-1} mod 2^k$, and we define

$$Y_i = U_i mod 2^j, 0 < j < k \tag{5.4}$$

then

$$Y_i = aY_{i-1} mod 2^j. \tag{5.5}$$

In practical terms, this means that the sequence of j-lo-order binary bits of the U_i sequence, namely Y_i cycle with cycle length at most 2^j. In particular, sequence of the least significant bit (i.e., j=1) in $(U_1, U_2, U_3, ...)$ must behave as $(0,0,0,0,...),(1,1,1,1,...),(0,1,0,1,...)$ or $(1,0,1,0,...)$.

PROOF

SEE: *Knuth (1981, pp. 12-14).*

Such normal behavior in the low-order bits of a congruential generator with non-prime-modulus *m* is an undesirably property, which may be aggravated by techniques such as the recycling of uniform variates. It has been observed *(Hutchinson, 1966)* that prime-modulus multiplicative congruential generators with full cycle (i.e., when *m* is a positive primitive element) tend to have fairly randomly distributed low-order bits, although no theory exists to explain this.

THEOREM V

If our congruential generator produces the sequence $(U_1, U_2, ...)$, and we look at the following sequence of points in n dimensions:

$$(U_1, U_2, U_3, ..., U_n), (U_2, U_3, U_4, ..., U_{n+1}), (U_3, U_4, U_5, ..., U_{n+2}), ... \tag{5.6}$$

then the points will all lie in fewer than $(n|m)^{1/n}$ parallel hyper planes.

PROOF

SEE: *Marsaglia (1976).*

Given these known limitations of congruential generator, we are still left with the question of how to choose the "best" values for *a*, *c*, and *m*. To do this, researchers have followed a straightforward but time-consuming procedure:

1. Take values *a*, *c*, and *m* that give a sufficiently long, known cycle length and usa the generator to produce sequences of uniform variates.
2. Subject the output sequences to batteries of statistical tests for independence and a uniform marginal distribution. Document the results.
3. Subject the generator to theoretical tests. In particular, the spectral test of *Coveyou and MacPherson (1967)* is currently widely used and recognized as a very sensitive structural test for distinguishing between good and bad generators. Document the results.
4. As new, more sensitive tests appear, subject to generator to those tests. Several such tests are discussed in *Marsaglia(1985).*

SEE ALSO: Other Types of Generators

5.2 PSEUDO-RANDOM VARIABLE GENERATORS, cont.[2]

5.2.1 PSEUDO-RANDOM VARIABLE GENERATORS, cont.

5.2.1.1 A Shift-Register Generator

An alternative class of pseudo-numbers generators are **shift-register** or **Tausworthe generators**, which have their origins in the work of *Golomb (1967)*. These algorithms operate on n-bit, pseudo-random binary vectors, just as congruential generators (p. 79) operate on pseudo-random integers. To return a uniform $(0, 1)$ variate, the binary vector must be converted to an integer and divided by one plus the largest possible number, 2^n.

5.2.1.2 Fibonacci Generators

The final major class of generators to be considered are **the lagged Fibonacci generators**, which take their name from the famous Fibonacci sequence $U_i = U_{i-1} + U_{i-2}$. This recursion is reminiscent of the congruential generators, which the added feature that the current value depends on the two previous values.

The integer generator based directly on the Fibonacci formula

$$2^n \tag{5.7}$$

has been investigated, but not found to be satisfactory random. A more general formulation can be given by the equation:

$$U_i = U_{i-r} \cdot U_{i-s}, r \geq 1, s \geq 1, r \neq s, \tag{5.8}$$

where the symbol 'square' represents an arbitrary mathematical operation. We can think of the $U_i = 0$ as either binary vectors, integers, or real numbers between 0 and 1, depending on the operation involved.

As examples:

1. The $U_i = 0$ are real and dot represents either mod 1 addition or subtraction.
2. The $U_i = 0$ are $(n-1)$ –bit integers and dot represents either mod 2^n addition, subtraction or multiplication.
3. The $U_i = 0$ are binary vectors and dot represents any of binary addition, binary subtraction, exclusive-or addition, or multiplication.

Other generators that generalize even further on the Fibonacci idea by using a linear combination of previous random integers to generate the current random integer are discussed in *Knuth (1981, Chap 3.2.2)*.

5.2.1.3 Combinations of Generators (Shuffling)

Intuitively, it is tempting to believe that "combining" two sequences of pseudo-random variables will produce one sequence with better uniformity and randomness properties than either of the two originals. In fact, even though good congruential (p. 79), Tausworthe (Section 5.2.1.1: A Shift-Register Generator), and Fibonacci (Section 5.2.1.2: Fibonacci Generators) generators exist, combination generators may be better for a number of reasons. The individual generators with short cycle length can be combined intone with a very long cycle. This can be a great advantage, especially on computers with limited mathematical precision. These potential advantages have led to the development of a number of successful combination generators and research into many others.

One of such generator, is a combination of three congruential generators, developed and tested by *Wichmann and Hill (1982)*.

Another generator, **Super-Duper**, developed by G.Marsaglia, combines the binary form of the output form the multiplicative congruenatial generator with a multiplier a=69.069 and modulus $m = 2^{32}$ with the

[2] This content is available online at <http://cnx.org/content/m13104/1.4/>.

output of the 32-bit Tausworthe generator using a left-shift of 17 and a right shift of 15. This generator performs well, though not perfectly, and suffers from some practical drawbacks.

A third general variation, **a shuffled generator**, randomizes the order in which a generator's variates are output. Specifically, we consider one pseudo-random variate generator that produces the sequence $(U_1, U_2, ...)$ of uniform (0,1) variates, and a second generator that outputs random integers , say between 1 and 16.

The algorithm for the combined, shuffled generator is as follows:

1. Set up a "table" in memory of locations 1 through 16 and store the values $U_1, U_2, ..., U_{16}$ sequentially in the table.
2. Generate one value, V, between 1 and 16 from the second generator.
3. Return the U variate from location V in the table as the desired output pseudo-random variate.
4. Generate a new U variate and store it in the location V that was just accessed.
5. If more random variates are desired, return to Step 2.

> NOTICE: the size of the table can be any value, with larger tables creating more randomness but requiring more memory allocation

This method of shuffling by randomly accessing and filling a table is due to *MacLaren and Marsaglia (1965)*. Another scheme, attributed to *M.Gentlemanin Andrews et al. (1972)*, is to permute the table of 128 random numbers before returning them for use. The use of this type of combination of generators has also been described in the contexts of simulation problems in physics by *Binder and Stauffer (1984)*.

5.3 THE IVERSE PROBABILITY METHOD FOR GENERATING RANDOM VARIABLES[3]

5.3.1 THE IVERSE PROBABILITY METHOD FOR GENERATING RANDOM VARIABLES

Once the generation of the uniform random variable (Section 5.1.1: UNIFORM PSEUDO-RANDOM VARIABLE GENERATION) is established, it can be used to generate other types of random variables.

5.3.1.1 The Continuous Case

THEOREM I

Let X have a continuous distribution $F_X(x)$, so that $F_X^{-1}(\alpha)$ exists for $0 < \alpha < 1$ (and is hopefully countable). Then the random variable $F_X^{-1}(U)$ has distribution $F_X(x)$, U is uniformly distributed on (0,1).

PROOF

$$P\left(F_X^{-1}(U) \leq x\right) = P\left(F_X\left(F_X^{-1}(U)\right) \leq F_X(x)\right). \tag{5.9}$$

Because $F_X(x)$ is monotone. Thus,

$$P\left(F_X^{-1}(U) \leq x\right) = P\left(U \leq F_X(x)\right) = F_X(x). \tag{5.10}$$

The last step follows because U is uniformly distributed on (0,1). Diagrammatically, we have that $(X \leq x)$ if and only if $[U \leq F_X(x)]$, an event of probability $F_X(x)$.

As long as we can invert the distribution function $F_X(x)$ to get the inverse distribution function $F_X^{-1}(\alpha)$, the theorem assures us we can start with a pseudo-random uniform variable U and turn into a random variable $F_X^{-1}(U)$, which has the required distribution $F_X(x)$.

Example 5.2
The Exponential Distribution

[3] This content is available online at <http://cnx.org/content/m13113/1.3/>.

Consider the exponential distribution defined as

$$\alpha = F_X\left(x\right) = \{ \begin{array}{l} 1 - e^{-\lambda x}, \lambda > 0, x \geq 0, \\ 0, x < 0. \end{array} \tag{5.11}$$

Then f or the inverse distribution function we have

$$x = -\frac{1}{\lambda}ln\left(1 - \alpha\right) = F^{-1}\left(\alpha\right). \tag{5.12}$$

Thus if U is uniformly distributed on 0 to 1, then $X = -\frac{1}{\lambda}ln\left(1 - U\right)$ has the distribution of an exponential random variable with parameter λ. We say, for convenience, that X is exponential (λ).

Note that: If U is uniform (0,1), then so is (1-U), and the pair U and (1-U) are interchangeable in terms of distribution. Hence, X' $= -\frac{1}{\lambda}ln\left(U\right)$ is exponential. However, the two variables X and X' are correlated and are known as **an antithetic pair**.

Example 5.3
Normal and Gamma Distributions

For both these cases there is no simple functional form for the inverse distribution $F_X^{-1}\left(\alpha\right)$, but because of the importance of the Normal and Gamma distribution models, a great deal of effort has been expended in deriving good approximations.

The Normal distribution is defined through its density,

$$f_X\left(x\right) = \frac{1}{\sqrt{2\pi\sigma}}exp\left[\frac{-\left(x-\mu\right)^2}{2\sigma^2}\right]. \tag{5.13}$$

So that,

$$F_X\left(x\right) = \int\limits_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma}}exp\left[\frac{-\left(x-u\right)^2}{2\sigma^2}\right]dv. \tag{5.14}$$

The normal distribution function $F_X\left(x\right)$ is also often denoted $\Phi\left(x\right)$, when the parameter u and σ are set to 0 to 1, respectively. The distribution has no closed-form inverse, $F_X^{-1}\left(\alpha\right)$, but the inverse is needed do often that $\Phi^{-1}\left(\alpha\right)$, like logarithms or exponentials, is a system function.

The inverse of the Gamma distribution function, which is given by

$$F_X\left(x\right) = \frac{1}{\Gamma\left(k\right)}\int\limits_{0}^{kx/u}v^{k-1}e^{-v}dv, x \geq 0, k > 0, u > 0. \tag{5.15}$$

Is more difficult to compute because its shape changes radically with the value of k. It is however available on most computers as a numerically reliable function.

Example 5.4
The Normal and Gamma Distributions

A commonly used symmetric distribution, which has a shape very much like that of the Normal distribution, is the standardized logistic distribution.

$$F_X\left(x\right) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}, -\infty < x < \infty, \tag{5.16}$$

with probability density function

$$F_X\left(x\right) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}, -\infty < x < \infty. \tag{5.17}$$

NOTE THAT: $F_X(-\infty) = e^{-\infty}/(1+e^{-\infty}) = 0$ and $F_X(\infty) = 1$ by using the second form for $F_X(x)$.

The inverse is obtained by setting $\alpha = \frac{e^x}{1+e^x}$. Then, $\alpha + \alpha e^x = e^x$ or $\alpha = e^x(1-\alpha)$.

Therefore,

$$x = F_X^{-1}(\alpha) = ln\alpha - ln(1-\alpha).$$

And the random variable is generated, using the inverse probability integral method. As follows $X = lnU - ln(1-U)$.

5.3.1.2 The Discrete Case

Let X have a discrete distribution $F_X(x)$ that is, $F_X(x)$ jumps at points $x_k = 0, 1, 2, \ldots$. Usually we have the case that $x_k = k$, so that X is an integer value.

Let the probability function be denoted by

$$p_k = P(X = x_k), k = 0, 1, \ldots. \tag{5.18}$$

The probability distribution function is then,

$$F_X(x_k) = P(X \le x_k) = \sum_{j \le k} p_j, k = 0, 1, \ldots, \tag{5.19}$$

and the reliability or survivor function is

$$R_X(x_k) = 1 - F_X(x_k) = P(X > x_k), k = 0, 1, \ldots. \tag{5.20}$$

The survivor function is sometimes easier to work with than the distribution function, and in fields such as reliability, it is habitually used. The inverse probability integral transform method of generating discrete random variables is based on the following theorem.

THEOREM

Let U be uniformly distributed in the interval (0,1). Set $X = x_k$ whenever $F_X(x_{k-1}) < U \le F_X(x_k)$, for $k = 0, 1, 2, \ldots$ with $F_X(x_{-1}) = 0$. Then X has probability function p_k.

PROOF

By definition of the procedure,

$X = x_k$ if and only if $F_X(x_{k-1}) < U \le F_X(x_k)$.

Therefore,

$$P(X = x_k) = PF_X((x_{k-1}) < U \le F_X(x_k)) = F_X(x_k) - F(x_{k-1}) = p_k. \tag{5.21}$$

By the definition of the distribution function of a uniform (0,1) random variable.

Thus the inverse probability integral transform algorithm for generating X is to find x_k such that $U \le F_X(x_k)$ and $U > F_X(x_{k-1})$ and then set $X = x_k$.

In the discrete case, there is never any problem of numerically computing the inverse distribution function, but the search to find the values $F_X = (x_k)$ and $F_X(x_{k-1})$ between which U lies can be time-consuming, generally, sophisticated search procedures are required. In implementing this procedure, we try to minimize the number of times one compares U to $F_X = (x_k)$. If we want to generate many of X, and $F_X = (x_k)$ is not easily computable, we may also want to store $F_X = (x_k)$ for all k rather than recomputed it. Then we have to worry about minimizing the total memory to store values of $F_X = (x_k)$.

Example 5.5

The Binary Random Variable

To generate a binary-valued random variable X that is 1 with probability p and 0 with probability 1-p, the algorithm is:

- If $U \leq p$, set X=1.
- Else set X=0.

Example 5.6

The Discrete Uniform Random Variable

Let X take on integer values between and including the integers a and b, where $a \leq b$, with equal probabilities. Since there are $(b - a + 1)$ distinct values for X, the probability of getting any one of these values is, by definition, $1/(b - a + 1)$. If we start with a continuous uniform (0,1) random number U, then the discrete inverse probability integral transform shows that

X= integer part of $[(b - a + 1)U + a]$.

NOTE THAT: The continuous random variable $[(b - a + 1)U + a]$ is uniformly distributed in the open interval $(a, b + 1)$.

Example 5.7

The Geometric Distribution

Let X take values on zero and the positive integers with a geometric distribution. Thus,

$$P(X = k) = p_k = (1 - \rho)\rho^k, k = 0, 1, 2, ..., 0 < \rho < 1, \tag{5.22}$$

and

$$P(X \leq k) = F_X(k) = 1 - \rho^{k+1}, k = 0, 1, 2, ..., 0 < \rho < 1. \tag{5.23}$$

To generate geometrically distributed random variables then, you can proceed successively according to the following algorithm:

- Compute $F_X(0) = 1 - \rho$. Generate U.
- If $U \leq F_X(0)$ set X=0 and exit.
- Otherwise compute $F_X(1) = 1 - \rho^2$.
- If $U \leq F_X(1)$ set X=1, and exit.
- Otherwise compute $F_X(2)$, and so on.

Glossary

A

A test, defined by a critical region C of size α, is **a uniformly most powerful test** if it is a most powerful test against each simple alternative in H$_1$. The critical region C is called **a uniformly most powerful critical region of size** α.

C

1. Consider the test of the sample null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$.
2. Let C be a critical region of size α; that is, $\alpha = P\left(C; \theta_0\right)$. Then C is **a best critical region of size** α if, for every other critical region D of size $\alpha = P\left(D; \theta_0\right)$, we have that

$$P\left(C; \theta_1\right) \geq P\left(D; \theta_1\right).$$

CUMULATIVE DISTRIBUTION FUNCTION

1. Let X be a random variable of the discrete type with space R and p.d.f. $f\left(x\right) = P\left(X = x\right)$, $x \in R$. Now take x to be a real number and consider the set A of all points in R that are less than or equal to x. That is, $A = \left(t : t \leq x\right)$ and $t \in R$.
2. Let define the function *F(x)* by

$$F\left(x\right) = P\left(X \leq x\right) = \sum_{t \in A} f\left(t\right). \tag{1.1}$$

The function *F(x)* is called **the distribution function** (sometimes **cumulative distribution function**) of the discrete-type random variable X.

D

DEFINITION OF EXPONENTIAL DISTRIBUTION

Let $\lambda = 1/\theta$, then the random variable X has **an exponential distribution** and its p.d.f. id defined by

$$f\left(x\right) = \frac{1}{\theta} e^{-x/\theta}, 0 \leq x < \infty, \tag{2.4}$$

where the parameter $\theta > 0$.

DEFINITION OF RANDOM VARIABLE

1. Given a random experiment with a sample space S, a function X that assigns to each element s in S one and only one real number $X\left(s\right) = x$ is called **a random variable**. The space of X is the set of real numbers $\{x : x = X\left(s\right), s \in S\}$, where s belongs to S means the element s belongs to the set S.
2. It may be that the set S has elements that are themselves real numbers. In such an instance we could write $X\left(s\right) = s$ so that X is **the identity function** and the space of X is also S. This is illustrated in the example below.

DEFINITION OF UNIFORM DISTRIBUTION

The random variable X has **a uniform distribution** if its p.d.f. is equal to a constant on its support. In particular, if the support is the interval $[a, b]$, then

$$f(x) = \frac{1}{b = a}, a \leq x \leq b. \tag{2.3}$$

G

Given a random sample $X_1, X_2, ..., X_n$ from a normal distribution $N(\mu, \sigma^2)$, consider the closeness of $\overline{X}$, the unbiased estimator of μ, to the unknown μ. To do this, the error structure (distribution) of $\overline{X}$, namely that $\overline{X}$ is $N(\mu, \sigma^2/n)$, is used in order to construct what is called **a confidence interval** for the unknown parameter μ, when the variance σ^2 is known.

I

If $E[u(x_1, x_2, ..., x_n)] = \theta$ is called **an unbiased estimator of** θ. Otherwise, it is said to be **biased**.

1. If $w < 0$, then $F(w) = 0$ and $F'(w) = 0$, a p.d.f. of this form is said to be one of the **gamma type**, and the random variable W is said to have **the gamma distribution**.
2. The **gamma function** is defined by

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy, 0 < t.$$

L

Let X have a gamma distribution with $\theta = 2$ and $\alpha = r/2$, where r is a positive integer. If the p.d.f. of X is

$$f(x) = \frac{1}{\Gamma(r/2) 2^{r/2}} x^{r/2-1} e^{-x/2}, 0 \leq x < \infty. \tag{2.6}$$

We say that X has **chi-square distribution** with r degrees of freedom, which we abbreviate by saying is $\chi^2(r)$.

M MATHEMATICAL EXPECTATION

If *f(x)* is the p.d.f. of the random variable X of the discrete type with space R and if the summation

$$\sum_R u(x) f(x) = \sum_{x \in R} u(x) f(x) \tag{1.2}$$

exists, then the sum is called **the mathematical expectation** or **the expected value** of the function *u(X)*, and it is denoted by $E[u(X)]$. That is,

$$E[u(X)] = \sum_R u(x) f(x). \tag{1.3}$$

We can think of the expected value $E[u(X)]$ as a weighted mean of *u(x)*, $x \in R$, where the weights are the probabilities $f(x) = P(X = x)$.

MATHEMATICAL EXPECTIATION

If $f(x)$ is the p.d.f. of the random variable X of the discrete type with space R and if the summation

O

1. Once the sample is observed and the sample mean computed equal to $\overline{x}$, the interval

$$\overline{x} - z_{\alpha/2}\left(\sigma/\sqrt{n}\right), \overline{x} + z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$$

is a known interval. Since the probability that the random interval covers μ before the sample is drawn is equal to $1-\alpha$, call the computed interval, $\overline{x} \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$(for brevity), a $100\left(1-\alpha\right)\%$ **confidence interval** for the unknown mean μ.

2. The number $100\left(1-\alpha\right)\%$, or equivalently, $1-\alpha$, is called **the confidence coefficient**.

P

POISSON DISTRIBUTION

We say that the random variable X has **a Poisson distribution** if its p.d.f. is of the form

$$f\left(x\right) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, ...,$$

where $\lambda > 0$.

POISSON PROCCESS

Let the number of changes that occur in a given continuous interval be counted. We have **an approximate Poisson process** with parameter $\lambda > 0$ if the following are satisfied:

PROBABILITY DENSITY FUNCTION

1. Function f(x) is a nonnegative function such that the total area between its graph and the x axis equals one.
2. The probability $P\left(a < X < b\right)$ is the area bounded by the graph of $f\left(x\right)$, the x axis, and the lines $x = a$ and $x = b$.
3. We say that **the probability density function (p.d.f.)** of the random variable X of the continuous type, with space R that is an interval or union of intervals, is an integrable function $f\left(x\right)$ satisfying the following conditions:
   - $f\left(x\right) > 0$, x belongs to R,
   - $\int\limits_R f\left(x\right)dx = 1$,
   - The probability of the event A belongs to R is $P\left(X\right) \in A \int\limits_A f\left(x\right)dx$.

PROBABILITY DENSITY FUNCTION

1. The distribution function of a random variable X of the continuous type, is defined in terms of the p.d.f. of X, and is given by

$$F\left(x\right) = P\left(X \leq x\right) = \int\limits_{-\infty}^{x} f\left(t\right)dt.$$

2. For the fundamental theorem of calculus we have, for x values for which the derivative F' $\left(x\right)$ exists, that F'(x)=$f(x)$.

T

t Distribution

If Z is a random variable that is $N\left(0,1\right)$, if U is a random variable that is $\chi^2\left(r\right)$, and if Z and U are independent, then

$$T = \frac{Z}{\sqrt{U/r}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{2.9}$$

has a t distribution with r degrees of freedom.

1. The random variable X has a normal distribution if its p.d.f. is defined by

$$f\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}exp\left[-\frac{\left(x-\mu\right)^{2}}{2\sigma^{2}}\right],-\infty<x<\infty, \qquad (2.8)$$

   where μ and σ^2 are parameters satisfying $-\infty<\mu<\infty, 0<\sigma<\infty$, and also where $exp\left[v\right]$ means e^v.

2. Briefly, we say that X is $N\left(\mu,\sigma^2\right)$

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

Attributions

Collection: *Introduction to Statistics*
Edited by: Ewa Paszek
URL: http://cnx.org/content/col10343/1.3/
License: http://creativecommons.org/licenses/by/2.0/

Module: "DISCRETE DISTRIBUTION"
By: Ewa Paszek
URL: http://cnx.org/content/m13114/1.5/
Pages: 1-6
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "MATHEMATICAL EXPECTATION"
By: Ewa Paszek
URL: http://cnx.org/content/m13530/1.2/
Pages: 6-8
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "THE MEAN, VARIANCE, AND STANDARD DEVIATION"
By: Ewa Paszek
URL: http://cnx.org/content/m13122/1.3/
Pages: 8-11
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "BERNOULLI TRIALS and the BINOMIAL DISTRIBUTION"
By: Ewa Paszek
URL: http://cnx.org/content/m13123/1.3/
Pages: 11-15
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "GEOMETRIC DISTRIBUTION"
By: Ewa Paszek
URL: http://cnx.org/content/m13124/1.3/
Pages: 15-17
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "POISSON DISTRIBUTION"
By: Ewa Paszek
URL: http://cnx.org/content/m13125/1.3/
Pages: 17-20
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "CONTINUOUS DISTRIBUTION"
By: Ewa Paszek
URL: http://cnx.org/content/m13127/1.4/
Pages: 21-26
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "THE UNIFORM AND EXPONENTIAL DISTRIBUTIONS"
By: Ewa Paszek
URL: http://cnx.org/content/m13128/1.7/
Pages: 26-31
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "THE GAMMA AND CHI-SQUARE DISTRIBUTIONS"
By: Ewa Paszek
URL: http://cnx.org/content/m13129/1.3/
Pages: 31-36
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "NORMAL DISTRIBUTION"
By: Ewa Paszek
URL: http://cnx.org/content/m13130/1.4/
Pages: 36-38
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "THE t DISTRIBUTION"
By: Ewa Paszek
URL: http://cnx.org/content/m13495/1.3/
Pages: 38-40
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "Estimation"
By: Ewa Paszek
URL: http://cnx.org/content/m13524/1.2/
Pages: 41-43
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "CONFIDENCE INTERVALS I"
By: Ewa Paszek
URL: http://cnx.org/content/m13494/1.3/
Pages: 43-44
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "CONFIDENCE INTERVALS II"
By: Ewa Paszek
URL: http://cnx.org/content/m13496/1.4/
Pages: 45-47
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "SAMPLE SIZE"
By: Ewa Paszek
URL: http://cnx.org/content/m13531/1.2/
Pages: 47-48
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "Maximum Likelihood Estimation (MLE)"
By: Ewa Paszek
URL: http://cnx.org/content/m13501/1.3/
Pages: 48-52
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "Maximum Likelihood Estimation - Examples"
By: Ewa Paszek
URL: http://cnx.org/content/m13500/1.3/
Pages: 52-55
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "ASYMPTOTIC DISTRIBUTION OF MAXIMUM LIKELIHOOD ESTIMATORS"
By: Ewa Paszek
URL: http://cnx.org/content/m13527/1.2/
Pages: 55-58
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "TEST ABOUT PROPORTIONS"
By: Ewa Paszek
URL: http://cnx.org/content/m13525/1.2/
Pages: 59-63
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "TESTS ABOUT ONE MEAN AND ONE VARIANCE"
By: Ewa Paszek
URL: http://cnx.org/content/m13526/1.3/
Pages: 63-66
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "TEST OF THE EQUALITY OF TWO INDEPENDENT NORMAL DISTRIBUTIONS"
By: Ewa Paszek
URL: http://cnx.org/content/m13532/1.2/
Pages: 66-67
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "BEST CRITICAL REGIONS"
By: Ewa Paszek
URL: http://cnx.org/content/m13528/1.2/
Pages: 67-70
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "HYPOTHESES TESTING"
By: Ewa Paszek
URL: http://cnx.org/content/m13533/1.2/
Pages: 70-77
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "PSEUDO-NUMBERS"
By: Ewa Paszek
URL: http://cnx.org/content/m13103/1.6/
Pages: 79-82
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "PSEUDO-RANDOM VARIABLE GENERATORS, cont."
By: Ewa Paszek
URL: http://cnx.org/content/m13104/1.4/
Pages: 83-84
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Module: "THE IVERSE PROBABILITY METHOD FOR GENERATING RANDOM VARIABLES"
By: Ewa Paszek
URL: http://cnx.org/content/m13113/1.3/
Pages: 84-87
Copyright: Ewa Paszek
License: http://creativecommons.org/licenses/by/2.0/

Introduction to Statistics
This course is a short series of lectures on Introductory Statistics. Topics covered are listed in the Table of Contents. The notes were prepared by Ewa Paszek and Marek Kimmel. The development of this course has been supported by NSF 0203396 grant.