Contents

1 STATISTICAL REVIEW

1.1 RANDOM VARIABLES AND DISTRIBUTIONS

Unlike the deterministic functions that mathematicians usually deal with, statisticians and probabilists need to be able to handle quantities that vary randomly. These are referred to as *random variables*. Random variables are described through their probability distributions. Probability distributions tell us the probabilities with which the variable takes various values.

This chapter introduces a few probability distributions that play an important role in understanding and analyzing healthcare data. While theoretical in nature, these distributions model real life if appropriately chosen. Probability distributions are mathematical representations of how statistical data behave.

Random variables and their distributions can be broadly classified as *discrete or continuous*. If the variable under study takes a finite or countably infinite number of values (countably infinite means that the values the variable takes can be put in one-to-one correspondence with positive integers), then the distribution is discrete. Any variable that involves counting the number of occurrences of an event will automatically be a discrete variable. The number of insurance policies where at least one claim is made, the number of days elapsed before the first claim is made for a particular policyholder or the total numbers of claims made by all policyholders in a year are examples of discrete random variables.

On the other hand, if the variable takes all values in an interval or union of intervals, then the distribution is continuous. Time between occurrences of accidents on a highway, the amount of rainfall in Santa Barbara County, the height and weight of people or the amount of carbon dioxide emissions are all examples of continuous random variables. One of the important distinctions of continuous distributions from the discrete is that the probability of a continuous random variable taking a specific value is infinitesimally small. In other words while $P(X = x)$ can be defined for a discrete random variable X, this probability will always be zero for a continuous random variable. So for a continuous random variable X, we talk about $P(X \le x)$, $P(X > x)$ or $P(x \le X \le y)$.

1.1.1 Discrete Random Variables

Definition 1.1.1 The *probability distribution* of a discrete random variable is a listing of all values the r.v. takes, along with the corresponding probabilities.

Definition 1.1.2 The function $p_x(x) = P(X = x)$, or simply $p(x)$, is called the *probability mass function* of X.

Note that

- X denotes the random variable whereas x denotes a possible value for X.

- $p(x)$ is always non-negative.
- $\sum_x \mathrm{P}(X = x) = 1$, where the summation is over all possible values of x.

Definition 1.1.3 The *distribution function* or the *cumulative distribution function* (CDF) F_X or just F of a random variable X is a function given by $F_X(x) = \mathrm{P}(X \le x)$. If X is discrete, then $F_X(x) = \sum_{x \le y} p_x(y)$.

Example 1.1.4 Toss a pair of coins. The sample space is given by $S = \{HH, HT, TH, TT\}$. Let X denote the number of heads obtained. Then $X(HH) = 2, X(HT) = 1, X(TH) = 1,$ and $X(TT) = 0.$ If the coins are fair, you associate a probability of 0.25 with each sample point. Then $P(X = 0) = 0.25$, $P(X = 1) = 0.5$ and $P(X = 2) = 0.25.$

So the distribution of X defined in this example can be written as

X	0	1	2
p	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example 1.1.5 For the earlier example the distribution function of X is given by

$$\begin{aligned} F(x) &= 0 && x \text{ if} < \\ &= \frac{1}{4} && \le x \text{ if} < 0 \\ &= \frac{3}{4} && \le x \text{ if} < 1 \\ &= 1 && x \ge \end{aligned}$$

1.1.2 Expectation of Discrete Random Variables

Definition 1.2.1 Let X be a discrete random variable with probability mass function p. Then the *expectation* (or the expected value or mean) of X is given by

$$\mu_x = E(X) = \sum_x xp(x)$$

where the summation is taken over all possible values x of X. If g is a real-valued function, then the expectation of $g(X)$ is given by $E\left(g\left(X\right)\right) = \sum_{x} g\left(x\right) p(x)$.

Example 1.2.2 Suppose Y is a random variable with the following distribution:

y	0	1	2	3
p	0.1	0.2	0.4	0.3

Then $E(Y) = 0(0.1) + 1(0.2) + 2(0.4) + 3(0.3) = 1.9$

Example 1.2.3 Let Y be as in Example 1.2.2. Then $E\left(Y^2\right) = \sum_{y=0}^{3}\left(y^2\right) p\left(y\right) = \left(0^2\right)(0.1) + \left(1^2\right)(0.2) +$ $\left(2^2\right)(0.4) + \left(3^2\right)(0.3) = 4.5.$

Definition 1.2.4 If X is a random variable then the *variance* of X is defined by

$$\pm_X^2 = V\left(X\right) = E(X^2) - (E(X))^2$$

and the *standard deviation* of X is defined by

$$\pm_X = SD\left(X\right) = \sqrt{V\left(X\right)}$$

Example 1.2.5 Let Y be as in Example 1.2.2. $E(Y^2) = 4.5$. Thus $V(Y) = 4.5 - 1.9^2 = 0.89$ and hence $SD(Y) = \sqrt{0.89} = 0.94$

Properties of Expectation and Variance: Let X be any random variable, continuous or discrete. For any two real numbers a and b,

1. $E(aX + b) = aE(X) + b$

2. $V\left(aX + b\right) = a^2 V(X)$

3. $SD(aX + b) = \left|a\right| SD(X)$

1.1.3 Important Discrete Distributions

Some commonly occurring distributions have special names. Some famous names are given below. They are all discrete distributions.

1. Discrete Uniform

2. Bernoulli

3. Binomial

4. Hypergeometric

5. Geometric

6. Negative Binomial

7. Poisson

1.1.3.1 Discrete Uniform

If a random variable X takes n possible values, $a_1, a_2, ..., a_n$, all with probability $\frac{1}{n}$ then X is said to have a *discrete uniform* distribution on $\{a_1, a_2, ..., a_n\}$.

Example 1.3.1 Roll a die and let X be the number that comes up. If the die is well balanced, then X has discrete uniform distribution on $\{1, 2, 3, 4, 5,$ Then:

$$(a) E(X) = \frac{a+b}{2}$$

$$(b) V(X) = \frac{(b-a+1)^2 - 1}{12}$$

1. Let X Discrete Uniform$\{1,2,...n\}$ so that $p(x) = \frac{1}{n}$ for $x = 1,2,...n$. Then:

$$(a) E(X) = \frac{n+1}{2}$$

$$(b) V(X) = \frac{n^2 - 1}{12}$$

2. Let X ~ Discrete Uniform $\{0,1,2,...n\}$ so that $p(x) = \frac{1}{(n+1)}$ for $x = 1,2,...n$. Then:

$$(a) E(X) = \frac{n}{2}$$

$$(b) V(X) = \frac{n(n+2)}{12}$$

Example 1.3.2 Let X be the number that comes up when a die is rolled once. Here X has a uniform distribution over the set $\{1,2,3,4,5,6\}$. Hence $E(X) = \frac{6+1}{2} = 3.5$ and

$V(X) = \frac{6^2 - 1}{12} = 35/12 = 2.92$.

1.1.3.2 Bernoulli

A *Bernoulli* random variable is one that takes values 1 and 0 with probabilities p and $1-p$.

Example 1.3.3 Toss a coin whose probability of heads is p. Let X be the number of heads obtained. Then X is distributed as $\text{Bernoulli}(p)$.

If $X \sim \text{Bernoulli}(p)$, then

1. $E(X) = p$
2. $V(X) = p(1-p)$

Example 1.3.4 Let X be the number of sixes that come up when a die is rolled once. Here X has Bernoulli distribution with parameter $\frac{1}{6}$. Hence $E(X)=\frac{1}{6}$ and $V(X)=\left(\frac{1}{6}\times\frac{5}{6}\right)=\frac{5}{36}$.

1.1.3.3 Binomial

The *Binomial* distribution occurs in the following way. Suppose an experiment consists of n independent identical trials, each having two possible outcomes: success or failure. For a given trial the probability of success is p. (Such trials are called Bernoulli trials). Then X the total number of successes in n trials, has binomial distribution with parameters n and p. Note that when $n=1$, $\text{Binomial}(n,p)$ reduces to $\text{Bernoulli}(p)$.

If $X \sim \text{Bin}(n,p)$ that is, X is distributed as $\text{Binomial}(n,p)$**,** then for $k = 0,1,2,\ldots,n,$

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=\binom{n}{k}p^k(q)^{n-k}$$

where $q=1-p.$

Example 1.3.5 Take a coin whose probability of heads is $\frac{1}{3}$ and toss it 10 times. If X is the number of heads in 10 tosses then X is distributed as $\text{Bin}\left(10,\frac{1}{3}\right)$.

Example 1.3.6 A coin has $\frac{1}{4}$ probability of heads. If this coin is tossed 20 times, find the probability that you get

1. No heads
2. Exactly two heads
3. At least one head
4. Less than or equal to 18 heads
5. Between 3 and 5 heads, both inclusive.

Solution:

Let X be the number of heads obtained. Then X is a random variable distributed as $\text{Bin}\left(20, \frac{1}{4}\right)$.

1. $P\left(X = 0\right) = \binom{20}{0}\left(\frac{1}{4}\right)^{0}\left(\frac{3}{4}\right)^{20} = 0.003171$

2. $P\left(X = 2\right) = \binom{20}{2}\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right)^{18} = 0.06695$

3. $P(X \geq 1) = 1 - P\left(X = 0\right) = 0.996829$

4. $P(X \leq 18) = 1 - \left[P\left(X = 19\right) + P\left(X = 20\right)\right] = 1 - \left[\binom{20}{19}\left(\frac{1}{4}\right)^{19}\left(\frac{3}{4}\right)^{1} + \binom{20}{20}\left(\frac{1}{4}\right)^{20}\left(\frac{3}{4}\right)^{0}\right] = 1 - 5.548 \times 10^{-11}$

5. $P(3 \leq X \leq 5) = \left[\binom{20}{3}\left(\frac{1}{4}\right)^{3}\left(\frac{3}{4}\right)^{17} + \binom{20}{4}\left(\frac{1}{4}\right)^{4}\left(\frac{3}{4}\right)^{16} + \binom{20}{4}\left(\frac{1}{4}\right)^{5}\left(\frac{3}{4}\right)^{15}\right] = 0.5259.$

Example 1.3.7 Suppose a box has 6 balls where two are white and four are black. We choose 10 balls from it with replacement. Let X be the number of white balls. The fact that the sampling is done with replacement makes this experiment conformable to the binomial model. Thus $X \sim \text{Bin}\left(10, \frac{1}{3}\right)$.

If $X \sim \text{Bin}(n, p)$, then

1. $E(X) = np$

2. $V(x) = np(1 - p)$

If you toss a coin 100 times, and if the probability of heads is 0.4, on average you can expect 40 heads. Similarly, if the number of tosses is n and the probability of heads is p, you expect np heads, so the expectation formula makes sense. You may also think of Binomial(n,p) as a sum of n independent Bernoulli(p) random variables. Therefore the expectation and variance of binomial are n times the expectation and variance of Bernoulli.

Example 1.3.8 Let X be the number of sixes that came up when a die is rolled 30 times. Here X has binomial distribution with parameters $n = 30$ and $p = \frac{1}{6}$. Hence $E(X) = 30\left(\frac{1}{6}\right) = 5$ and $V(X) = 30\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) = \frac{25}{6}$.

1.3.4 Hypergeometric

If, in an example similar to the above, the sampling were done without replacement, one can easily see that the resulting random variable will not be binomial. Let us consider a case where 2 balls are chosen from the six balls. As there are only two white balls, the number of white balls can only be 0, 1, or 2. The probability that it is 1, for instance, is given by

$$\frac{\binom{2}{1}\binom{4}{1}}{\binom{6}{2}}.$$

Now let us consider the general situation. Assume there are N balls, of which d are white and the rest black. We choose n balls from these without replacement and let X be the number of white balls in the sample. Note that both d and n are integers between 0 and N. This X is then distributed as $hypergeometric(N,d,n)$.

In this case the probability that there are exactly k white balls is calculated as follows:

$$P(X=k)=\frac{\binom{d}{k}\binom{N-d}{n-k}}{\binom{N}{n}}.$$

For what values of k is this value non-zero? The necessary conditions are $0\le k\le d$ and $0\le n-k\le N-d$. This implies that k must be between $\max(0,n+d-N)$ and $\min(n,d)$.

Example 1.3.9 A box has 25 bolts, of which 10 are defective. A sample of size 5 is chosen without replacement and X is the number of defective bolts in the sample. Find the distribution of X.

Here X is hypergeometric with $N=25$, $d=10$, and $n=5$. Thus

$$P(X=k)=\frac{\binom{10}{k}\binom{15}{5-k}}{\binom{25}{5}}$$

for $k=0$, 1, 2, 3, 4 and 5. In particular,

$$P(X=2)=\frac{\binom{10}{2}\binom{15}{3}}{\binom{25}{5}}=.3854$$

For problems where a sample of size n is chosen from a population of N objects of which d are of a special type and X is the number of objects in the sample that are of that special type, the rule is the following: If the sampling is without replacement what we have is $\text{hypergeometric}(N,d,n)$; if the sampling is with replacement, then it is $\text{Bin}\left(n,\frac{d}{n}\right)$.

If $X\sim\text{Hypergeometric}(N,d,n)$, then

1. $E(X)=n\left(\frac{d}{N}\right)$
2. $V(X)=n\left(\frac{d}{N}\right)\left(1-\frac{d}{N}\right)\left(\frac{N-n}{N-1}\right)$

Similar to binomial, the expectation is np where $p = \frac{d}{N}$ is the proportion of special items. The formula for the variance is $npqc$ where $c = \frac{N-n}{N-1}$ is called the *finite population correction*. If *N* is very large compared to n, c is very close to 1, and can be ignored.

Example 1.3.10 A box contains 7 white balls, 3 black balls, and 2 red balls. A sample of 4 balls is drawn from the box without replacement. Let X represent the number of white balls in the sample. Here X has hypergeometric distribution with parameters $N = 12$, $d = 7$, and $n = 4$. Hence $E(X) = 4\left(\frac{7}{12}\right) = \frac{7}{3}$ and $V(X) = 4\left(\frac{7}{12}\right)\left(\frac{5}{12}\right)\left(\frac{8}{11}\right) = \frac{77}{99}$.

1.3.5 Geometric

Suppose we perform Bernoulli trials until the first success is obtained. Let X be the number of failures. X then is said to have *geometric* distribution.

Note that the sample space is given by $\{S, FS, FFS, FFFS, FFFFS, \ldots\}$, where S indicates success and F indicates failure. Note that the sample space is infinite in this case. Thus X takes values 0, 1, 2, 3, . . . with probabilities given by

$$\begin{aligned}
P\left(X=0\right) &= P\left(S\right) = p \\
P\left(X=1\right) &= P\left(FS\right) = p(1-p) = pq \\
P\left(X=2\right) &= P\left(FFS\right) = p(1-p)^2 = pq^2 \\
P\left(X=k\right) &= P\left(FF\ldots FS\right) = pq^k \;, k = 0,1,2,3,\ldots
\end{aligned}$$

Thus we have a formula to compute the probabilities for geometric distribution.

Example 1.3.11 Roll a die until a six is obtained. Find the probability that the first six occurs

1. at the fourth trial.

2. at or before the third trial.

3. after the fourth trial.

Solution: Let X be the number of failures.

1. $P(X=3) = pq^3 = \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^3 = \frac{125}{1296}$

2. $P(X \le 2) = P(X=0) + P(X=1) + P(X=2) = \frac{1}{6} + \frac{1}{6}\cdot\frac{5}{6} + \frac{1}{6}\left(\frac{5}{6}\right)^2 = \frac{91}{216}$

3. $P(X \ge 4) = 1 - P(X \le 3) = 1 - \left[\frac{125}{1296} + \frac{91}{216}\right] = \frac{625}{1296}$

Note: There is a faster way of computing (3) and hence (2). You will see how in the next result.

Properties of Geometric Distribution: Let X have Geometric(p) distribution. Then for nonnegative integer k,

1. $P(X \geq k) = q^k$

2. $P\left(X < k\right) = 1 - q^k$

3. $P\left(X > k\right) = q^{k+1}$

4. $P(X \leq k) = 1 - q^{k+1}$

If X ~ Geometric(p), then

1. $E(X) = \dfrac{p}{q}$

2. $V(X) = \dfrac{q}{p^2}$

Unsurprisingly, the average waiting time, as seen in the expression for expectation, increases as the probability p of occurrence decreases.

Example 1.3.12 An experiment whose success probability is 0.25 is repeated until a success is obtained. Let X represent the number of failures and Y represent the number of trials. Find the mean and variance for X and Y.

Here X has geometric distribution with $p = .25$ and $Y = X + 1$. So $E(X) = \dfrac{q}{p} = \dfrac{.75}{.25} = 3$ and $E(Y) = 3 + 1 = 4$. $V(X) = \dfrac{q}{p^2} = \dfrac{.75}{.0625} = 12$. It is one of the properties of variance that adding a constant does not change the variance. Thus $V(Y) = V(X+1) = V(X) = 12$.

1.3.6 Negative Binomial

Suppose we perform Bernoulli trials until r successes are obtained. Let X be the number of failures. Then X is said to have *negative binomial* distribution with parameters r and p, where p is the probability of success for the Bernoulli trials.

Geometric distribution is a special case of negative binomial distribution; negative binomial with $r=1$ is geometric.

The probability distribution of X is computed as follows. For X to be equal to k, S should be the last link in a chain of S and F of length $k+r$ such that exactly $r-1$ of the first $k+r-1$ is S. This way we make sure that before the r^{th} success there are k failures. As there are $\binom{k+r-1}{r-1}$ ways of selecting such chains and each has probability $p^r\left(1-p\right)^k$, we get

$$P(X=k)=\binom{k+r-1}{r-1}p^r\left(1-p\right)^k$$

for $k=0,1,2....$

If X has *negative binomial* distribution with parameters r and p and we need to compute the probability that the total number of trials is n, then the formula is given by

$$\binom{n-1}{r-1}p^r\left(1-p\right)^{n-r}$$

Example 1.3.13 A coin has probability $\frac{2}{3}$ of heads. It is tossed until 5 heads are obtained. Find the probability that exactly 8 tosses are required.

Solution: Let X is the number of tails. Then X has *negative binomial* distribution with parameters $r=5$ and $p=\frac{2}{3}$. Exactly 8 tosses are required if and only if $X=3$.

$$P(X=3)=\binom{3+5-1}{5-1}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^3=\frac{1120}{6561}$$

In general, for $n = 5, 6, 7 \ldots$, the probability that the number of tosses is n is given by

$$\binom{n-1}{4}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^{n-5}$$

If $X \sim \text{Negative Binomial}(r, p)$, then

1. $E(X) = \dfrac{rq}{p}$
2. $V(X) = \dfrac{rq}{p^2}$

Example 1.3.14 A die is rolled until 3 sixes are obtained. Let X represent the number of non-sixes and Y represent the number of rolls. Find the mean and variance for X and Y.

Here X has negative binomial distribution with $r = 3$ and $p = \dfrac{1}{6}$, and Y is $X + 3$. So

$E(X) = \dfrac{rq}{p} = \dfrac{3\left(\frac{5}{6}\right)}{\frac{1}{6}} = 15$ and $E(Y) = 15 + 3 = 18$. $V(X) = \dfrac{q}{p^2} = \dfrac{3\left(\frac{5}{6}\right)}{\left(\frac{1}{6}\right)^2} = 90$.

$V(Y) = V(X + 3) = V(X) = 90$.

1.3.7 Poisson

Poisson distribution is used when we count the number of occurrences of an event in a specified time interval. Some examples are the number of traffic accidents on a highway per month, the number of phone calls that go through a telephone switchboard in an hour and the number of customers arriving at a bank in a day. The probability mass function of a Poisson random variable X with parameter ª is given by

$$P(X = k) = \frac{e^{-ª} ª^{k}}{k!},$$

for $k = 0, 1, 2, 3, \ldots$. Here the parameter ª indicates the average number of occurrences of the event.

Example 1.3.15 Assume that X, the number of calls that go through a telephone switchboard in an hour, follows Poisson distribution with parameter $ª = 5$. Find

1. $P(X \leq 1)$
2. the probability that at least one call comes through between 10:00 and 10:30.

Solution:

1. $P(X \leq 1) = P(X = 0) + P(X = 1) = \dfrac{e^{-5}5^0}{0!} + \dfrac{e^{-5}5^1}{1!} = 6e^{-5}$.
2. If Y is the number of calls in the half-hour then Y is distributed as $\text{Poisson}\left(2.5\right)$. $P(Y \geq 1) =$

$1-P(Y=0)=1-e^{-2.5}$

If X ~ Poisson(ª), then

1. $E(X) =$ ª
2. $V(X) =$ ª

Example 1.3.16 Assume that the number of typing errors in a page of a book is distributed as Poisson(9).

1. Find the mean and standard deviation of the number of errors in a randomly chosen page of the book.
2. Find the mean and standard deviation of the total number of errors in four randomly chosen pages of the book.

If X denotes the number of errors in one randomly chosen page, $E(X)=V(X)=9$ and the standard deviation of X is 3. If Y denotes the number of errors in five randomly chosen pages, $E(X)=V(X)=9(4)=36$ and the standard deviation of Y is 6.

1.4 Continuous Random Variables

As explained previously, a random variable is said to be continuous if $F_X(x)=P(X\le x)$ is a continuous function of x.

Note: All continuous random variables take infinitely many values. Also, for a continuous random variable X, $P(X=x)=0$ for all x, and consequently, $P(a\le X\le b)=P(a\le X<b)=P(a<X\le b)=P(a<X<b)$.

Definition 1.4.1 A function f is said to be the *probability density function (pdf)* (or just the *density function*) of a continuous random variable X if $P(X\le x)=\int_{-\infty}^{x} f(t)\,dt$

Proposition 1.4.2 If f is the probability density function of a continuous random variable X, then

1. $f(x)\ge 0$ for all x
2. $\int\limits_{-\infty}^{\infty} f(x)\,dx=1$
3. $\int\limits_{a}^{b} f(x)\,dx=P(a\le X\le b)$

1.5 Expectation and Variance of Continuous Random Variables

Definition 1.5.1 let X be a continuous random variable with density function f. Then the expectation (or the expected value) of X is given by

$$E(X)=\int\limits_{-\infty}^{\infty} xf(x)\,dx$$

$$V(X)=E(X^2)-(E(X))^2=\int\limits_{-\infty}^{\infty} x^2 f(x)\,dx-(E(X))^2$$

$$SD(X)=\sqrt{V(X)}$$

Example 1.5.2 Let X be a random variable with density function f given by $f(x) = 3x2$ if $0\leq x \leq$ 1and zero otherwise. Find $E(X)$ and $E\left(X^2\right)$.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_0^1 3x^3 dx \\ &= \left[\frac{3x^4}{4}\right]_0^1 \\ &= \frac{3}{4} \\ E\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 f(x)dx \\ &= \int_0^1 3x^4 dx \\ &= \left[\frac{3x^5}{5}\right]_0^1 \\ &= \frac{3}{5} \end{aligned}$$

1.6 Important Continuous Distributions

The following are some of the commonly occurring continuous distributions.

1. Continuous Uniform
2. Exponential
3. Gamma
4. Beta
5. Normal
6. Weibull
7. *t*
8. Chi-square
9. *F*

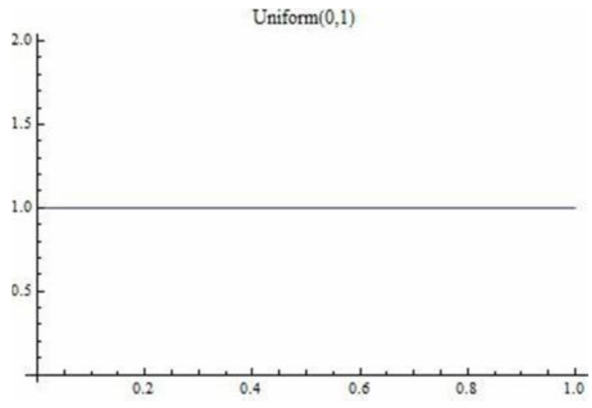
1.6.1 Continuous Uniform

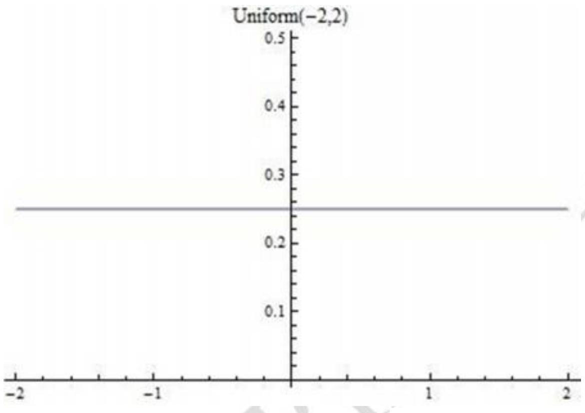
A random variable X has $uniform(a,b)$ distribution if its pdf is given by

$$f(x) = \frac{1}{b-a}$$

for $a \le x \le b$. That the above function integrates to 1 is easy to verify by a simple integration, or by the fact that the area of a rectangle with length $b-a$ and height $\frac{1}{b-a}$ is 1.

The following are examples of uniform distribution:





Example 1.6.1 Let X be a random real number chosen from the interval (3, 5). Find the probability that X is less than or equal 4.5.

Solution: The X has uniform$(3,5)$ distribution. $P(X \le 4.5) = \int_3^{4.5} \frac{1}{2}dx = \frac{1.5}{2} = 0.75$

Proposition 1.6.2 Let X be a random variable with $uniform(a,b)$ distribution. Then

1. $E(X) = \frac{a+b}{2}$
2. $V(X) = \frac{(b-a)^2}{12}$

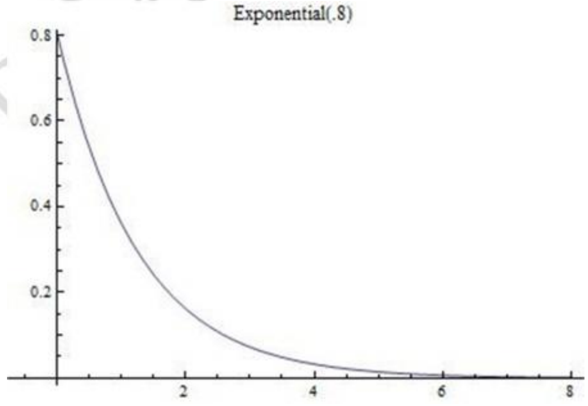
Example 1.6.3 Let X be the random variable in Example 1.6.1. Then $E(X) = \frac{3+5}{2}$ and $V(X) = \frac{2^2}{12} = \frac{1}{3}$

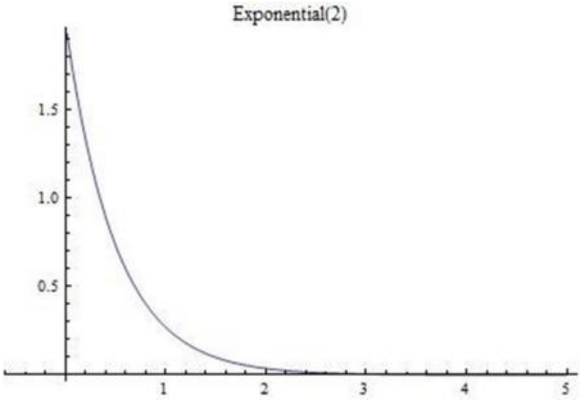
1.6.2 Exponential

A random variable X has *exponential*(ª) distribution if its pdf is given by

$$f(x) = \text{ª}e^{-\text{ª}x}$$

for $x \ge 0$. Exponential distribution is the simplest distribution that can be used to model survival time waiting time for occurrences of events. Here, ª is called the rate parameter, measuring the rate at which events occur.





Example 1.6.4 The life time in years of a light bulb is represented by a random variable X with exponential(.5) distribution. Find the probability that the bulb will last for less than 18 months.

$$\begin{aligned} P(X<1.5) &= \int_0^{1.5} .5e^{-.5x}dx \\ &= \left[-e^{-.5x}\right]_0^{1.5} \\ &= 1-e^{-.75} \\ &= 0.5276 \end{aligned}$$

Proposition 1.6.5 Let X be a random variable with $\text{exponential}(^{\text{a}})$ distribution. Then

1. $E(X)=\frac{1}{\text{a}}$

2. $V(X)=\frac{1}{\text{a}^2}$

Note that the expression for the expectation tells us that the average waiting time is inversely related to the rate of occurrence of the event, which makes intuitive sense. If on the average, 10 events occur per unit time, the average waiting time or the next occurrence is $\frac{1}{10}$ units of time.

The exponential distribution has the *memoryless* property: if at a particular time point, we know that no event has occurred, the waiting time for the next event is unchanged. Mathematically $P(X>s+t\mid X>s)=P(X>t)$. In other words, if we are told that until time s no event has occurred, the probability that we need to wait at least t more units of time is exactly equal to the probability that we need to wait at least t units in the first place.

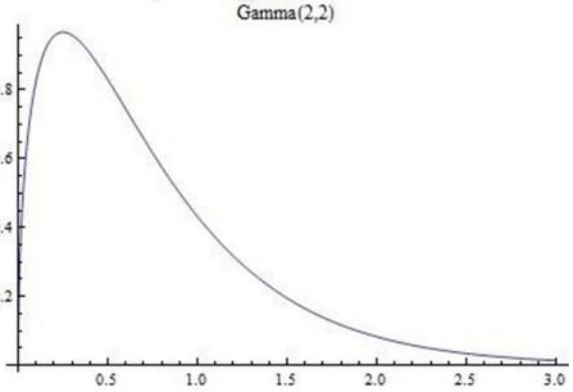
Example 1.6.6 Let X be the random variable in Example 1.6.4. Then $E(X)=\frac{1}{.5}=2$ and $V(X)=\frac{1}{.5^2}=4$.

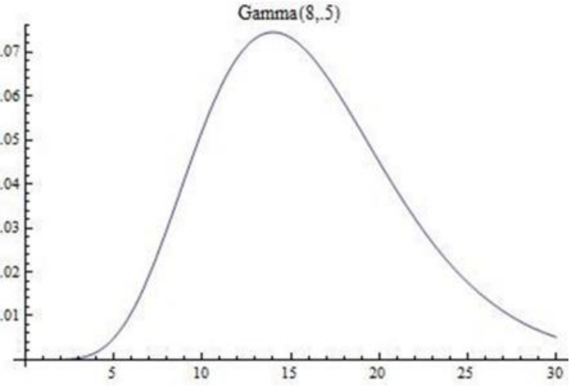
1.6.3 Gamma

The Gamma Function, represented by $\Gamma(X)$, is an extension of the factorial function. If $a\in\{1,2,...n\}$ then $\Gamma(a)=(a-1)!$. A random variable X has $gamma(\sim,^{\text{a}})$ distribution if its pdf is given by

$$f(x)=\frac{\text{a}^{\sim}}{\Gamma(\sim)}x^{\sim-1}e^{-\text{a}x}$$

for $x\geq 0.^{\text{a}}$ is the rate parameter, and $\sim$ is called the shape parameter.





Example 1.6.7 The time in years a transplanted heart lasts is given by a $\text{gamma}(\sim=2,^{\text{a}}=.5)$ random variable X. Find the probability that the heart will last for more than 3 years.

The density function of X is given by

$$f(x) = \frac{.5^2}{\Gamma(2)} xe^{-.5x} = \frac{1}{4} xe^{-.5x}$$

for $x \geq 0$.

$$\begin{aligned} P(X > 3) &= \int_3^\infty \frac{1}{4} xe^{-.5x} dx \\ &= \frac{1}{4}\left[-2xe^{-.5x}\right]_3^\infty + \frac{1}{4}\int_3^\infty 2e^{-.5x} dx \\ &= 1.5e^{-1.5} - \left[e^{-.5x}\right]_3^\infty \\ &= 1.5e^{-1.5} + e^{-1.5} \\ &= 2.5e^{-1.5} \\ &= 0.5578 \end{aligned}$$

Proposition 1.6.8 Let X be a random variable with $\text{gamma}(\sim, ª)$ distribution. Then

1. $E(X) = \frac{\sim}{ª}$
2. $V(X) = \frac{\sim}{ª^2}$

Example 1.6.9 Let X be the random variable in Example 1.6.7. Then $E(X) = \frac{2}{.5} = 4$ and $V(X) = \frac{2}{.5^2} = 8$

Note that some books use the notations $\text{exponential}(¯)$ and $\text{gamma}(\sim, ¯)$ in place of $\text{exponential}(ª)$ and $\text{gamma}(\sim, ª)$ where $¯ = \frac{1}{ª}$. Then the expectation and variance of exponential will change to $¯$ and $¯^2$ respectively, and the expectation and variance of gamma will change to $\sim¯$ and $\sim¯^2$ respectively. The parameter $ª$ is referred to as the *rate* parameter whereas the parameter $¯$, in the case of exponential, is referred to as the *mean* parameter.

Example 1.6.10 Let X be exponentially distributed with mean 0.2. Find the probability that X is less than 0.3.

Here the *mean* is 0.2, so $¯ = 0.2$. $ª$, the rate, is $\frac{1}{¯} = 5$. Now $P(X < .3)$ is easily seen to be $1 - e^{-1.5}$.

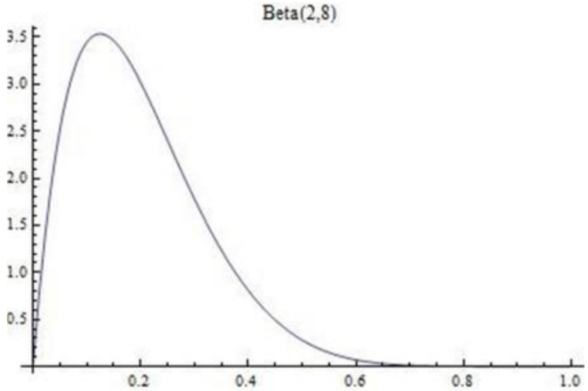
1.6.4 Beta

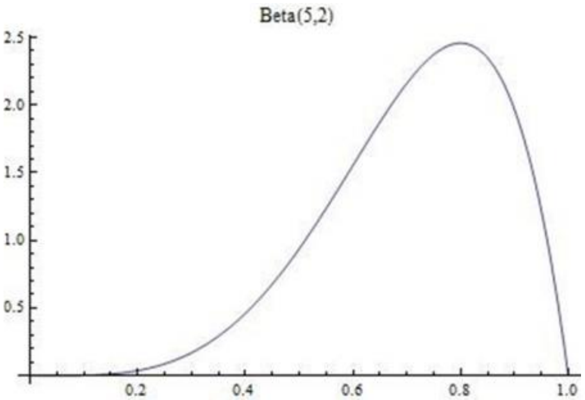
A random variable X has $\textit{beta}(\sim, \ )$ distribution if its pdf is given by

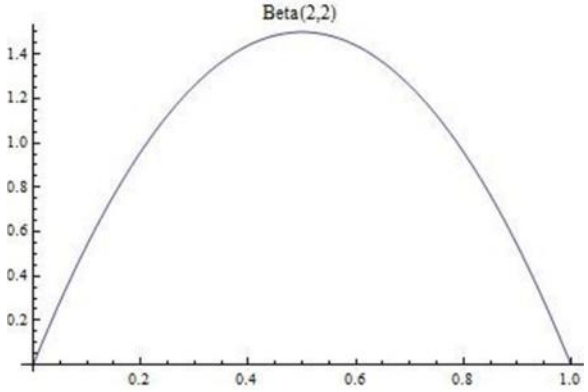
$$f(x) = \frac{\Gamma(\sim + \ )}{\Gamma(\sim)\Gamma(\ )} x^{\sim-1}(1-x)^{\ -1}$$

for $0 \leq x \leq 1$.

The following are examples of beta distribution:







Example 1.6.11 The proportion of car owners who service their car by the due date is given by a beta$(\sim\, = 2, \quad = 3)$ random variable X . Find the probability that this proportion is less than or equal to 0.7.
The density function of X is given by

$$f(x) = \frac{\Gamma(5)}{\Gamma(2)\Gamma(3)} x(1-x)^2 = 12x(1-x)^2$$

for $0 \le x \le 1$.

$$\begin{aligned} P(X \le 0.7) &= \int_0^{.7} 12x(1-x)^2\,dx \\ &= 12\int_0^{.7} x(1-2x+x^2)dx \\ &= 12\int_0^{.7} x - 2x^2 + x^3 dx \\ &= 12\left[\frac{x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{4}\right]_0^{.7} \\ &= 12\left[\frac{.7^2}{2} - 2\left(\frac{.7^3}{3}\right) + \frac{.7^4}{4}\right] \\ &= 0.9163 \end{aligned}$$

Proposition 1.6.12 Let X be a random variable with beta$(\sim, \quad)$ distribution. Then

1. $E(X) = \dfrac{\sim}{\sim + \quad}$
2. $V(X) = \dfrac{\sim}{(\sim + \quad)^2(\sim + \quad + 1)}$

Example 1.6.13 Let X be the random variable in Example 1.6.11. Find the mean and standard deviation of X.

$E(X)=\frac{2}{5}$ and $V(X)=\frac{6}{(25)6}=\frac{1}{25}$. So $SD(X)=\frac{1}{5}$.

1.6.5 Weibull

A random variable X has $\text{Weibull}(\text{~},\text{ª})$ distribution if its pdf is given by

$$f(x)=\text{~ª}(\text{ª}x)^{\text{~}-1}e^{-(\text{ª}x)^{\text{~}}}$$

for $x>0$. Note that when $\text{~}=1$, $\text{Weibull}(\text{~},\text{ª})$ reduces to $\texttt{exponential}(\text{ª})$.

Weibull distribution is often used to model the size of reinsurance claims. It is also used in survival analysis, reliability, failure analysis, in industrial engineering to represent manufacturing and delivery times, in weather forecasting to describe wind speed distributions, and the cumulative development of asbestosis losses, among many others.

Proposition 1.6.14 Let X be a random variable with Weibull $(\text{~},\ )$ distribution. Then

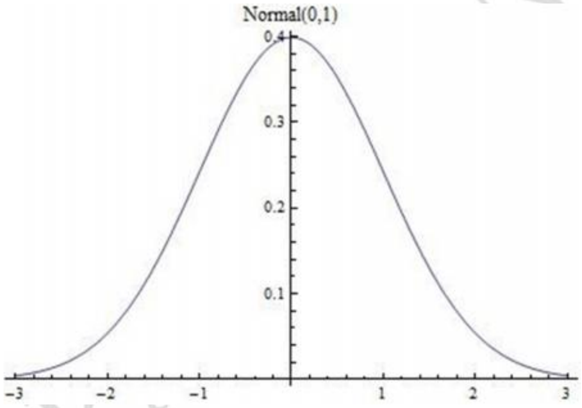
1. $E(X)=\frac{1}{\text{ª}}\Gamma\left(1+\frac{1}{\text{~}}\right)$

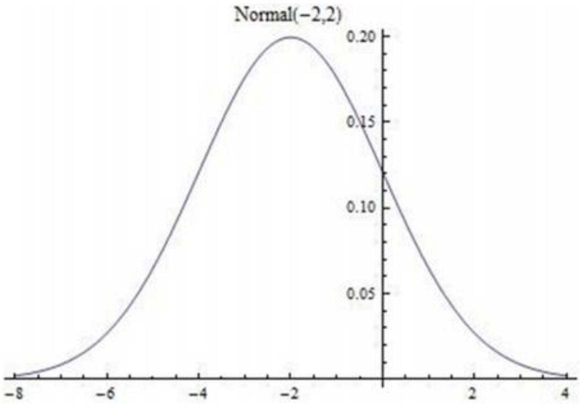
2. $V(X)=\frac{1}{\text{ª}^2}\left[\Gamma\left(1+\frac{2}{\text{~}}\right)-\left(\Gamma\left(1+\frac{1}{\text{~}}\right)\right)^2\right]$

1.6.6 Normal

A random variable X has *normal* distribution or *Gaussian* distribution with parameters « and ± if its pdf is given by

$$\frac{1}{\sqrt{2®}±}e^{-\frac{1}{2}\left(\frac{x-«}{±}\right)^2}.$$





Proposition 1.6.15 Let X be a normal random variable with parameters « and ±. Then $E(X)=\mu$ and $V(X)=±^2$

Normal distribution is by far the most important distribution. It is widely used in many practical applications. We get a very special case of normal called *standard normal* when the mean μ is zero and the standard deviation ± is 1. A standard normal random variable is usually denoted by Z and has density function given by

$$\frac{1}{\sqrt{2®}}e^{-\frac{1}{2}x^2}.$$

For computation of probabilities for normal distribution, we have to use tables. Tables are available only for standard normal, but we can convert non-standard normal to standard normal by a procedure called *standardization*, which is, subtracting the mean and dividing by the standard deviation.

Example 1.6.16 Let $Z \sim N\left(0,1\right)$. Find

1. $P(Z \le 1.2)$
2. $P(1.5 \le Z \le 2.3)$
3. $P(Z \ge 2)$
4. a such that $P(Z \le a) = .9251$
5. b such that $P(Z > b) = .1230$

Solution:

1. $P(Z \le 1.2) = .8849$
2. $P(1.5 \le Z \le 2.3) = .9893 - .9332 = .0561$
3. $P(Z \ge 2) = 1 - P(Z \le 2) = 1 - .9772 = .0228$
4. From the body of the table, we see that $a = 1.44$.
5. $P(Z > b) = .1230 \Rightarrow P(Z \le b) = .8770$. So $b = 1.16$.

Example 1.6.17 Let $X \sim N(2.3,1.5)$. Find

1. $P(X \le 1.2)$
2. $P(3.8 \le X \le 5.3)$
3. $P(X \ge 2)$
4. a such that $P(X \le a) = .9251$
5. b such that $P\left(X > b\right) = .1230$

Solution:

1. $P(X \le 1.2) = P\left(Z \le \dfrac{1.2-2.3}{1.5}\right) = P(Z \le - .73) = P\left(Z > .73\right) = 1 - .7673 = .232$
2. $P(3.8 \le X \le 5.3) = P\left(\dfrac{3.8-2.3}{1.5} \le Z \le \dfrac{5.3-2.3}{1.5}\right) = P(1 \le Z \le 2) = .9772 - .8413 = .135$
3. $P(X \ge 2) = P\left(Z > \dfrac{2-2.3}{1.5}\right) = P(Z > - .2) = P(Z \le .2) = .579$
4. $P(X \le a) = .9251 \Rightarrow P\left(Z \le \dfrac{a-2.3}{1.5}\right) = .9251 \Rightarrow \dfrac{a-2.3}{1.5} = 1.44 \Rightarrow a = 4.4$
5. $P(X > b) = .1230 \Rightarrow P(X \le b) = .8770 \Rightarrow P\left(Z \le \dfrac{b-2.3}{1.5}\right) = .8770 \Rightarrow \dfrac{b-2.3}{1.5} = 1.16 \Rightarrow b = 4.$

Note: The remaining continuous distributions, t, ¡ ^{2}, and F are important for statistical estimation and inference, but it is not necessary for us to learn their density functions, expected values etc. In the subsequent chapters we will learn how these distributions are used to construct confidence intervals and perform tests of statistical hypotheses.

1.7 Jointly Distributed Random Variables

Let X and Y be two random variables defined on the same sample space (that is, both are numerical outcomes of the same random experiment). Knowing the probability distributions of X and Y separately is not sufficient to know their behaviour completely. For instance, from knowing $P(X \le 1)$ and $P(Y \le 2)$, we will not normally be able to compute $P(X \le 1, Y \le 2)$. We need to know their *joint* distribution. Study of joint distributions can be complicated, particularly if it involves many random variables that that are interdependent.

On the other hand, if the random variables are known to behave independent of each other, then the joint probabilities can be calculated from the respective individual probabilities by multiplying them together. In the previous example, $P(X \le 1, Y \le 2) = P(X \le \ )P(Y \le \ )$ if X and Y are independent. If you decide that it is reasonable to assume that the numbers of car insurance claims in California(X) and Pennsylvania(Y) are independent of each other, then the probability that there are at least 200 claims in California and at least 100 claims in Pennsylvania in a given year can be calculated by multiplying the two individual probabilities. In other words, $P(X \ge 200, Y \ge 10) = P(X \ge 200)P(Y \ge 10)$.

Definition 1.7.1 The Covariance between X and Y is given by

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

The Correlation between X and Y is given by

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

1.8 Properties of Covariance and Correlation

Theorem 1.8.1

1. $Cov\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov\left(X_i, Y_j\right)$
2. $V\left(\sum_{i=1}^{m} a_i X_i\right) = \sum_{i=1}^{m} a_i^2 V\left(X_i\right) + \sum_{i<j} a_i a_j Cov\left(X_i, Y_j\right)$
3. $Cov(X,X) = V(X)$
4. $Cov(aX+b, cY+d) = acCov(X,Y)$
5. $V(aX+bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$
6. $V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$
7. $SD(X+Y) = \sqrt{(SD(X))^2 + (SD(Y))^2 + 2Cov(X,Y)}$
8. $-1 \le Corr(X,Y) \le 1$
9. $Corr(aX+b, cY+d) = Corr(X,Y)$ if $ac > 0$, $= -Corr(X,Y)$ if $ac < 0$, $= 0$ if $ac = 0$.
10. If X and Y are independent,
    (a) $Cov(X,Y) = 0$
    (b) $Corr(X,Y) = 0$
    (c) $V(X+Y) = V(X) + V(Y)$
    (d) $SD(X+Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$

Note: The converses of (10a) and (10b) above are false. Even though independence implies that covariance and correlation are zero, two dependent random variables can have zero covariance, and hence zero correlation.

Example 1.8.2 Let X and Y be independent with $SD(X) = 4$, $SD(Y) = 5$. Find

1. $V(X+Y)$
2. $Cov(X+Y, X-Y)$
3. $Corr(X+Y, X-Y)$

Solution:

1. $V\left(X+Y\right)=V\left(X\right)+V(Y)=16+25=41$, so $SD(X+Y)=\sqrt{41}$
2. $Cov(X+Y,X-Y)=Cov\left(X,X\right)+Cov(Y,X)-Cov\left(X,Y\right)-Cov(Y,Y)=V\left(X\right)-V(Y)=-9$
3. $V(X-Y)=V\left(X\right)+V(Y)=16+25=41$, so

$$Corr(X+Y,X-Y)=\frac{Cov(X+Y,X-Y)}{SD(X+Y)SD(X-Y)}=\frac{-9}{41}.$$

Example 1.8.3 Let X and Y be two random variables with $V\left(X\right)=4$, $V(Y)=9$, $Cov\left(X,Y\right)=-2$. Find

1. $Cov(2X+3,-4Y+2)$
2. $Cov(2X-1,4-X)$
3. $V(2X-Y)$
4. $Corr(X,Y)$
5. $Corr(3X+1,-2Y+2)$

Solution:

1. $Cov(2X+3,-4Y+2)=2(-4)Cov\left(X,Y\right)=2(-4)(-2)=16$
2. $Cov(2X-1,4-X)=2(-1)Cov\left(X,X\right)=(-2)V\left(X\right)=-8$
3. $V(2X-Y)=4V\left(X\right)+V(Y)-4Cov\left(X,Y\right)=16+9+8=33$
4. $Corr(X,Y)=\dfrac{Cov(X,Y)}{SD(X)SD(Y)}=\dfrac{-2}{(2)(3)}=\dfrac{-1}{3}$
5. $Corr(3X+1,-2Y+2)=(-1)\dfrac{-1}{3}=\dfrac{1}{3}$

1.9 Central Limit Theorem

Theorem 1.9.1 Let $X_1,X_2,...,X_n$ be independent identically distributed random variables with mean « and standard deviation ± . Let $\overline{X}_n=\dfrac{X_1+X_2+\cdots+X_n}{n}$. Then

$$E\left(\overline{X}_n\right)=\text{«} \text{ and } SD\left(\overline{X}_n\right)=\frac{\pm}{\sqrt{n}}.$$

Theorem 1.9.2 Central Limit Theorem
Le t $X_1,X_2,...,X_n$ be independent identically distributed random variables with mean « and standard deviation ± . Let $\overline{X}_n=\dfrac{X_1+X_2+\cdots+X_n}{n}$. Then

$$\frac{\overline{X}_n-m}{\frac{\pm}{\sqrt{n}}}\rightarrow^{d} N(0,1)$$

where $\rightarrow^{d}$ indicates distributional convergence.

Example 1.9.3 The content of 500ml bottles of Coca-Cola has normal distribution mean 503ml and stan-dard deviation 5ml.

1. Find the probability that a randomly chosen bottle has less than 500ml cola in it.
2. Find the probability that a randomly chosen 6-pack of bottles has an average of less than 500ml.

Solution:

1. $P\left(X < 500\right) = P\left(Z < \frac{500-503}{5}\right) = P(Z < -0.6) = 0.2743$.
2. $P\left(\overline{X} < 500\right) = P\left(Z < \frac{500-503}{\frac{5}{\sqrt{6}}}\right) = P(Z < -1.47) = 0.0708$.

Example 1.9.4 The waiting time for a particular service for a person is exponentially distributed with mean 5 minutes. Find the probability that a total service time for hundred people is no more than 8 hours.

Solution: For exponential distribution the mean is the same as the standard deviation, so $\pm = 5$. From the CLT we know that the sample mean is approximately normal. Let T be the total service time in minutes.

$$P(T \leq 480) = P\left(\overline{X} < 4.8\right) \approx P\left(Z < \frac{4.8-5}{\frac{5}{\sqrt{100}}}\right) = P(Z < -0.4) = 0.3446 .$$

Chapter 2

Confidence Intervals and Hypothesis Testing

2.1 Confidence Intervals

As mentioned before, if we specify a range or an interval around the point estimator within which the parameter being estimated is expected to lie, then that is called an *interval estimator*. Usually it will be an interval such that the point estimator is in the center, but not always. First we decide on what is called a *confidence coefficient* $\sim$, often chosen to be 0.05. Then the interval around the point estimator is so chosen that the probability that this random interval actually contains the unknown parameter is $1-\sim$. Such an interval is called a $100(1-\sim)\%$ *confidence interval*. If $\sim$ is 0.05, then we get a 95% confidence interval. In general, a $100(1-\sim)\%$ confidence interval for a parameter $^{-}$ based on a sample X is a random interval of the form $(L(X), R(X))$ such that

$$P[^{-} \in (L(X), R(X))] = 1-\sim$$

For the mean « of a normal population whose standard deviation is known to be $\pm$, a $100(1-\sim)\%$ confidence interval is given by

$$\bar{x} \pm z_{\frac{\sim}{2}} \frac{\pm}{\sqrt{n}}$$

where z_{ϵ} is the value found in the standard normal table (also known as the Z -table) so that the probability of Z exceeding it is (that is, $P(Z > z_{\epsilon}) = \epsilon$).

When $\pm$ is unknown, as is the case in most situations, we need to use its estimator s in its place. In this case, you substitute s in place of $\pm$ and t in place of z to get

$$\bar{x} \pm t_{n-1,\frac{\sim}{2}} \frac{s}{\sqrt{n}}$$

where $t_{n-1,\frac{\sim}{2}}$ is from the t -distribution with $n-1$ degrees of freedom. We no longer use the values from the Z -table in this case unless n is very large. If the sample is large, then the requirement that the underlying population is normal is not necessary. All we need is finiteness of the variance.

The following table summarizes the methods for different cases.

Table 2.1: Confidence Intervals

Distribution	Pop. St. Deviation	Sample Size	Conf. Interval
Normal	Known	Small/Large	$\overline{x} \pm z_{\frac{\sim}{2}} \frac{\pm}{\sqrt{n}}$
Normal	Unknown	Small/Large	$\overline{x} \pm t_{n-1,\frac{\sim}{2}} \frac{s}{\sqrt{n}}$
Non-normal	Known	Large	$\overline{x} \pm z_{\frac{\sim}{2}} \frac{\pm}{\sqrt{n}}$
Non-normal	Unknown	Large	$\overline{x} \pm t_{n-1,\frac{\sim}{2}} \frac{s}{\sqrt{n}}$
Non-normal		Small	No formula

Example 2.1.1 A sample of size 100 taken from a population with $\pm = 12$ yielded a sample mean of $\overline{x} = 31.6$. Construct a 99% confidence interval for the population mean.

Solution: Here the population standard deviation is known, so we use the z confidence interval. As $\sim = 0.01$, we use the formula $\overline{x} \pm z_{.005} \frac{\pm}{\sqrt{n}}$. Substituting, we get

$$31.6 \pm 2.576\left(\frac{12}{10}\right) = 31.6 \pm 3.091 = (28.509, 34.691) .$$

Example 2.1.2 A sample of size 16 was taken from a normal population with $\pm = 3$. The sample mean and the sample standard deviation were $\overline{x} = 17.6$ and $s = 3.7$. Construct a 95% confidence interval for the population mean.

Solution: Here the population standard deviation is known, so we use the z confidence interval. As $\sim = 0.05$, we use the formula $\overline{x} \pm z_{.025} \frac{\pm}{\sqrt{n}}$. Substituting, we get

$$17.6 \pm 1.96\left(\frac{3}{4}\right) = 17.6 \pm 1.47 = (16.13, 19.07)$$

Note that here we ignored the value of s because when we know the value of $\pm$, we don’t need s.

Example 2.1.3 A sample of size 36 was taken from a population and the sample mean and the sample variance obtained were $\overline{x} = 19.9$ and $s^2 = 32.8329$. Construct a 90% confidence interval for the population mean.

Solution: Here the population standard deviation is unknown, so we use the t confidence interval. As $\sim = 0.1$ and the $d.f. = 35$, we use the formula $\overline{x} \pm t_{35,.05} \dfrac{s}{\sqrt{n}}$.

Substituting, we get:

$$19.9 \pm 1.690\left(\frac{5.73}{6}\right) = (18.286, 21.514)\ .$$

2.2 Testing Statistical Hypotheses

2.2.1 The Null and Alternative Hypotheses

A statistical hypothesis is an assertion or conjecture about one or more population parameters or the nature of the population. The idea is first to formulate a hypothesis, which we shall call *null hypothesis*, then to check if there is enough statistical evidence to refute it. The null hypothesis is usually denoted by H_0.

The hypothesis which we use as an alternative to null hypothesis, which we accept if null hypothesis is rejected, is called the *alternative hypothesis*. The alternative hypothesis is usually denoted by H_1 or H_A or H_a. It covers all or part of the situations not covered by the null hypothesis. For example, we may have

$$H_0:\ \mu = \mu_0 \text{ vs. } H_1:\ \mu < \mu_0$$

or

$$H_0:\ \mu = \mu_0 \text{ vs. } H_1:\ \mu > \mu_0$$

or

$$H_0:\ \mu \neq \mu_0 \text{ vs. } H_1:\ \mu = \mu_0$$

The first two alternatives are called one-sided alternatives and the third alternative is called a *two-sided* alternative.

Whether we reject a null hypothesis or not depends on the value of the statistic we use for the test. This is called a *test statistic*. If the test statistic falls in a specific region, we reject H_0 and accept it otherwise. This region is called the *critical region* or the *rejection region*. Its complement is called the *acceptance region*.

It is possible that we make an error in deciding whether the null hypothesis is true or not. If H_0 is true and you decided that it is not, you made what is known as the *type I error*. On the other hand, if you made the wrong decision that H_0 is true when it really is not, then you are making *type II error*. Obviously when you try to reduce one of them, the other will increase. The type I error is deemed more serious of the two. Because of this, the statistician tries to keep the probability of making that type of error no more than a specified value $\sim$, known also as the *level of significance*, and make decision that would minimize the type II error. Most of the times $\sim$ is taken to be 0.05. The probability of type II error is usually denoted by , and $1-$ is called the *power*. Naturally, the higher the power of the test is, the better.

Table 2.2: Hypothesis Testing

	Accept H_0	Reject H_0
H_0 is true	Correct$(1-\sim)$	Type I error$(\sim)$
H_0 is false	Type II error$(\ )$	Correct$(1-\ )$

The alternative hypotheses are usually composite hypotheses (hypotheses that cover more than possibil-ities) as opposed to the null hypotheses which are usually simple hypotheses (covering only one possibility). So when we talk about the probability of type II error, it will strongly depend on the value of the parameter. Thus , and hence the power, are functions of the parameter. The latter is called the *power function*.

Suppose we want to test a hypothesis with a two-sided alternative about the population mean. Formally, we are testing

$$H_0 : \mu = «_0 \text{ vs. } H_1 : « \neq «_0$$

where we reject the null hypothesis if the value of the test statistic is too large or too small. If ± is known, then the test statistic is

$$Z = \frac{\overline{x} - «_0}{\frac{±}{\sqrt{n}}}$$

and we reject the null hypothesis H_0 if $Z \geq z_{\frac{\sim}{2}}$ or $Z \leq -z_{\frac{\sim}{2}}$ (in other words, we reject H_0 if $|Z| \geq z_{\frac{\sim}{2}}$). If $-z_{\frac{\sim}{2}} \leq Z \leq z_{\frac{\sim}{2}}$, we do not reject the null hypothesis.

In case of a one-sided alternative of $H_1 : \mu > \mu_0$, we reject H_0 if $Z \geq z_{\sim}$ and in case of a one-sided alternative of $H_1 : \mu < \mu_0$, we reject if $Z \leq -z_{\sim}$. These cut-off values beyond which we reject the null hypothesis are called *critical values*.

If ± is unknown, use s in the place of ± and t instead of z. The test statistic used in this case is thus $t = \frac{\overline{x} - «_0}{\frac{s}{\sqrt{n}}}$. The critical value would then be $t_{n-1,\sim}$ or $t_{n-1,\frac{\sim}{2}}$ depending on whether we have a one-sided or a two-sided alternative. To use either the Z-procedure or the t-procedure for small samples, the underlying population distribution has to be normal.

Example 2.2.1 The manufacturer of a certain brand of cigarettes claim that the average nicotine content does not exceed 2.5 milligrams. Suppose that when a sample of size 100 was taken, it was found that the sample mean and the sample standard deviation are 2.55 and 0.5 respectively. Decide whether there is enough statistical evidence to reject the manufacturer's claim based on a 0.05 level test.

The manufacturer's claim is to be rejected only if μ is greater than 2.5 milligrams and accepted if it is less than or equal to 2.5 milligrams. As we will always specify the null hypothesis as a single number, this is how we will formulate the null and alternative hypotheses:

$$H_0 : \mu = 2.5 \text{ vs. } H_1 : \mu > 2.5$$

The t-statistic is equal to $\frac{2.55 - 2.5}{\frac{0.5}{\sqrt{100}}} = 1$ which is to be compared to $t_{99,.05} \approx t_{100,.05} = 1.66$. As the t value is smaller than 1.66, we do not reject the null hypothesis.

Note: The right terminology is to say that we not reject H_0 rather than *accept* H_0. Failing to reject the null hypothesis does not mean that we have concluded it to be true. It is just that we do not have sufficient evidence to reject it.

Table 2.3: Rejection Regions for Z - and t -Tests

±	**Distribution**	**Sample Size**	**Alternative**	**Rejection Region**
Known	Normal	Small/large	$« > «_0$	$Z \ge z_{\sim}$
			$« < «_0$	$Z \le -z_{\sim}$
			$« \ne «_0$	$\lvert Z \rvert \ge z_{\sim}$
	Non-normal	Large	$« > «_0$	$Z \ge z_{\sim}$
			$« < «_0$	$Z \le -z_{\sim}$
			$« \ne «_0$	$\lvert Z \rvert \ge z_{\sim}$
	Non-normal	Small		No formula
Unknown	Normal	Small/large	$« > «_0$	$t \ge t_{n-1,\sim}$
			$« < «_0$	$t \le -t_{n-1,\sim}$
			$« \ne «_0$	$\lvert t \rvert \ge t_{n-1,\sim}$
	Non-normal	Large	$« > «_0$	$t \ge t_{n-1,\sim}$
			$« < «_0$	$t \le -t_{n-1,\sim}$
			$« \ne «_0$	$\lvert t \rvert \ge t_{n-1,\sim}$
	Non-normal	Small		No formula

Example 2.2.2 An economist wants to test whether the average annual household income of a small town is below \$25000. Assume that the population is normal with a known population standard deviation 2000. A sample of size 25 was taken and the mean of this sample was 24000. Formulate the hypotheses and test the null hypothesis at 5% level of significance.

$$H_0:\ \mu = 25000 \text{ vs. } H_1:\ \mu < 25000$$

$$Z = \frac{24000 - 25000}{\frac{2000}{\sqrt{25}}} = -2.5$$

As the calculated value of z is smaller than $-z_{.05} = -1.645$, we reject the null hypothesis in favor of the alternative that $\mu < 25000$.

Example 2.2.3 A sample of size 20 was taken from a normal population. The sample mean and sample variance are 12.1 and 12.25 respectively. Test the null hypothesis $\mu = 10$ against a two-sided alternative. Use 1% level of significance.

$$H_0:\ \mu = 10 \text{ vs. } H_1:\ \mu \ne 10$$

$$t = \frac{12.1}{\frac{\sqrt{12.25}}{20}} = 2.683$$

As the calculated value of t is smaller than $t_{19,.005} = 2.861$, we do not reject the null hypothesis.

2.2.2 P-Values

There is another way of looking at the hypothesis-testing problem – through so-called p-values. We now discuss how this is done.

Suppose we suspect that a certain cereal company's 500g packets of cereal are being under filled. We cannot accuse the company unless we feel quite confident that we are correct in our claim. The alternative hypothesis would be $H_1:\ \mu < 500$ where μ is the population mean of the net weight of the company's 500g cereal packets. All calculations are carried out assuming that the null hypothesis is true, so the null hypothesis is always in the form '$=$'. Thus the hypotheses are

$$H_0:\ \mu = 500 \text{ vs. } H_1:\ \mu < 500 .$$

To test the hypothesis, we take a simple random sample and calculate the sample mean. We now want to know whether the sample value provides enough evidence to suggest $\mu < 500$. Even if μ is indeed 500, we expect some random fluctuations about 500, but how far from 500 does the sample mean need to be before we can say it is unlikely that the sample came from a population with mean 500?

We first calculate the test statistic $Z = \dfrac{\overline{x} - «_0}{\dfrac{±}{\sqrt{n}}}$ and then the probability of obtaining such an extreme value *in the direction of the alternative hypothesis*. If this probability, called *p-value*, is small, then we say that we have enough evidence to reject the null hypothesis (ie. there is only a small chance of this event if H_0 is correct, so we are willing to back the alternative hypothesis).

In the above problem of cereal boxes, suppose a sample size of 30 gives a sample mean of 498g. Given that the population standard deviation is 5, is there enough evidence to suggest that the packages are underweight? Here,

$$H_0:\ \mu = 500 \text{ vs. } H_1:\ \mu < 500 .$$

We calculate $P\left(\overline{X} \le 498\right)$ assuming that H_0 is true, that is, $N\left(500, \dfrac{5}{\sqrt{30}}\right)$.

$$P\left(\overline{X} \le 498\right) = P\left(Z \le \frac{498-500}{\dfrac{5}{\sqrt{30}}}\right) = P(Z \le -2.19) = 0.0143$$

So if « is indeed 500, the probability of obtaining a sample mean less than 498 is 0.0143. Is this a small enough probability to say that we think the null hypothesis is wrong? (It is a value that would occur 1.43% of the time by chance alone when in fact H_0 is true). If the p-value is smaller than the significance level, we reject the null hypothesis. If we are to use 5% level of significance, we will reject the null hypothesis that $\mu = 500$ and conclude that the population mean is less than 500. We will not reject H_0 if we are using 1% significance level.

Note: If the alternative hypothesis is two-sided, then we need to add the two tail probabilities (or double the right-tail probability) to get the p-value.

Statistical significance:
If the p-value is as small or smaller than ~ , we say the data are statistically significant at level ~ and we reject the null hypothesis.

The method of testing hypotheses using p-values is equivalent to the previous method we learned using critical values.

There is a relation between confidence intervals and hypothesis tests with two-sided alternatives. The hypothesis would be rejected at significance level $\sim$ if and only if the hypothesized value μ_0 falls outside the $100(1-\sim)\%$ confidence interval. If the test is one-sided, we cannot make a decision based on the confidence Intervals.

Example 2.2.4 A manufacturer of small appliances employs a market research firm to estimate retail sales of its products by gathering information from a sample of retail stores. This month an SRS (simple random sample) of 75 stores in the Midwest sales region finds that these stores sold an average of 24 of the manufacturer's hand mixers, with standard deviation 11.

(a) Give a 95% confidence interval for the mean number of mixers sold by all stores in the region.

(b) The distribution of sales is strongly right skewed, because there are many smaller stores and a few very large stores. The use of t in (a) is reasonably safe despite this violation of the normality assumption. Why?

(c) Would you reject the null hypothesis $H_0: \ \mu = 22$ against the alternative $H_1: \ \mu \neq 22$ at significance level 0.05? What about at 0.01?

(d) What is the approximate p-value for the test in (c)?

Solution:

(a) As $d.f. = n-1 = 74$ is not available in the t table, take the closest available, which is 70. The confidence interval given by

$$\begin{aligned} \overline{x} \pm t_{n-1,\frac{\sim}{2}} \frac{s}{\sqrt{n}} &= 24 \pm 1.994 \frac{11}{\sqrt{75}} \\ &= 24 \pm 2.53 \\ &= (21.47, 26.53) \end{aligned}$$

(b) As sample size is large, the sample mean will be approximately normal even though the parent population is not. Thus it is safe to use t-distribution.

(c) Since 22 is inside the 95% confidence interval, we will accept H_0 against a two-sided alternative at 0.05 level. If we accept at 0.05 level, we certainly will accept at 0.01 level. Another way of looking at this is that if 22 is in the 95% confidence interval, then it is certainly inside the bigger 99% interval. If the question is about level 0.1, then we cannot answer it based on the fact that it is inside the 95% interval. We will then need to recalculate.

(d) First we calculate the test statistic $t = \frac{24-22}{\frac{11}{\sqrt{75}}} = 1.575$. $P(t_{74} > 1.575) \approx P(Z > 1.575) = 0.0576$. Thus p-value is $2 \times 0.0576 = 0.1152$

2.3 Selection of Sample Size

Consider the formula for the confidence interval for the population mean for the case of known population standard deviation. Note how the margin of error changes as n increases. To obtain higher confidence from the same data you have to accept a higher margin of error. To obtain higher confidence without increasing the margin of error we need more observations. But sometimes we may want to select a sample size that will guarantee a desired confidence level for a fixed margin of error m. A formula for the sample size is derived as follows:

$$m = \frac{z_{\frac{\sim}{2}} \pm}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{z_{\frac{\sim}{2}} \pm}{m} \Rightarrow n = \left(\frac{z_{\frac{\sim}{2}} \pm}{m} \right)^2$$

Here we select the confidence level and the error margin that you are willing to tolerate and find the sample size we need. As the formula will not usually give a whole number, we need to round it to an integer. **Always round up** because rounding down will give you a sample size that would not guarantee the required precision.

Example 2.3.1 Suppose we need to construct a 98% confidence interval for μ with a margin of error no more than .0001. It is known that $\pm = 0.0002$. The required sample size is calculated as follows:

$$n = \left(\frac{z_{\frac{\sim}{2}}\pm}{m}\right)^2 = \left(\frac{2.326(.0002)}{.0001}\right)^2 = 21.64$$

So we choose n to be 22.

In many situations, we would not know the value of the population standard deviation. In such cases, we may use an estimated value from past experience, possibly a prior survey. If no such information is available, one can conduct a pilot study with a smaller sample to estimate the standard deviation.

2.4 Inference for variance and standard deviation

2.4.1 Confidence intervals

Given a random sample of size n from a normal population, we can get a $100(1-\sim)\%$ confidence interval for the population variance $\pm^2$ as follows:

$$\left(\frac{(n-1)s^2}{\text{¡}^2_{n-1,\frac{\sim}{2}}}, \frac{(n-1)s^2}{\text{¡}^2_{n-1,1-\frac{\sim}{2}}}\right)$$

You need to look up the Chi-square distribution table for the critical values. Confidence intervals for the standard deviation can be found by taking square-roots.

Example 2.4.1 A food inspector examined 12 jars of a certain brand of peanut butter and determined the percentages of impurities. The result is given below:

$$2.3,\ 1.9,\ 2.1,\ 2.8,\ 2.3,\ 3.6,\ 1.4,\ 1.8,\ 2.1,\ 3.2,\ 2.0,\ 1.9$$

Construct a 90% confidence interval for the standard deviation.

Solution: From the data, we get $s = 0.625$. From the table, the 11 d.f. ¡^2 values for 0.05 and 0.95 are 19.675 and 4.575. So the confidence interval for the variance is given by

$$\left(\frac{11\times 0.625^2}{19.675}, \frac{11\times 0.625^2}{4.575}\right) = \left(0.2184, 0.939\right).$$

Thus the confidence interval for the standard deviation is (.47, .97).

2.4.2 Hypothesis Testing

Testing the hypothesis $H_0 :\ \pm^2 = \pm_0^2$ (which us the same as $H_0 :\ \pm = \pm_0$) is done as follows. First we calculate the test statistic

$$\text{¡}^2 = \frac{\left(n-1\right)s^2}{\pm_0^2}.$$

We reject the null hypothesis if

$$\text{¡}^{2} \le \text{¡}^{2}_{n-1,1-\sim} \text{ when } H_1 \text{ is } \pm^{2} < \pm^{2}_{0}$$

$$\text{¡}^{2} \ge \text{¡}^{2}_{n-1,\sim} \text{ when } H_1 \text{ is } \pm^{2} > \pm^{2}_{0}$$

$$\text{¡}^{2} \le \text{¡}^{2}_{n-1,1-\frac{\sim}{2}} \text{ or } \text{¡}^{2} \ge \text{¡}^{2}_{n-1,\frac{\sim}{2}} \text{ when } H_1 \text{ is } \pm^{2} \ne \pm^{2}_{0}.$$

Example 2.4.2 An experiment was conducted to determine the specific heat of iron and a random sample of size 9 resulted in a standard deviation of 0.0086. Assuming normality of the underlying population, test the hypothesis that $\pm = 0.01$ against the alternative hypothesis that $\pm < 0.01$. Use the 0.05 level of significance.

Solution: $H_0 : \pm = 0.01$ vs. $H_1 : \pm < 0.01$. $n = 9$, $s = 0.0086$, $\sim = 0.05$. We reject the null hypothesis if $\text{¡}^{2} \le \text{¡}^{2}_{8,.95} = 2.733$. $\text{¡}^{2} = \frac{(n-1)s^2}{\pm^{2}_{0}} = \frac{8(.0086)^2}{.01^2} = 5.92 > 2.733$, so do not reject H_0.

Example 2.4.3 From past experience, it is assumed that standard deviation of measurements on sheet metal stampings is 0.41. A new set of 30 stampings are used to test the accuracy of this assumption, and a sample standard deviation of 0.49 was obtained. Test the hypothesis that $\pm = 0.41$ against the alternative hypothesis that $\pm > 0.41$. Use the 0.05 level of significance.

Solution: $H_0 : \pm = 0.41$ vs. $H_1 : \pm > 0.41$. $n = 30$, $s = 0.49$, $\sim = 0.05$. We reject the null hypothesis if $\text{¡}^{2} \ge \text{¡}^{2}_{29,.05} = 42.557$. $\text{¡}^{2} = \frac{(n-1)s^2}{\pm^{2}_{0}} = \frac{29(.49)^2}{.41^2} = 41.42 < 42.557$, so do not reject H_0.

Example 2.4.4 Based on a sample of size 10 that gave $s = 1.5$, test the hypothesis at the 0.05 level of significance that $\pm = 1$ against the alternative hypothesis that $\pm \ne 1$.

Solution: $H_0 : \pm = 1$ vs. $H_1 : \pm \ne 1$. $n =$ 10, $s = 1.5$, $\sim = 0.05$. We reject the null hypothesis if $\text{¡}^{2} \le \text{¡}^{2}_{9,.975} = 2.7$ or $\text{¡}^{2} \ge \text{¡}^{2}_{9,.025} = 19.023$. $\text{¡}^{2} = \frac{(n-1)s^2}{\pm^{2}_{0}} = \frac{9(1.5)^2}{1^2} = 20.25$. This exceeds 19.023, so we reject H_0.

Chapter 3

Two-Sample Inference

3.1 Confidence Intervals For Difference of Means

Suppose we have two populations from which we draw samples based on which we want to make inferences about the underlying parameters. Specifically, we are interested in constructing confidence intervals for the difference between their population means.

Assume that the two populations are normal, the first with mean «_1 and standard deviation $\pm_1$, and the second with mean «_2 and standard deviation $\pm_2$. We draw independent samples of size n_1 and n_2 respectively from these populations. Based on these, we construct our confidence interval for $\mu_1 - \mu_2$.

3.1.1 Case 1: the standard deviations are known

If $\pm_1$ and $\pm_2$ are known, then a $100(1-\sim)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\overline{x}_1 - \overline{x}_2 \pm z_{\frac{a}{2}} \sqrt{\frac{\pm_1^2}{n_1} + \frac{\pm_2^2}{n_2}}.$$

If the sample sizes are large, say both at least 30, then the assumption of normality can be dropped. The method will usually work well for non-normal populations too when the sample sizes are large.

Example 3.1.1 Independent random samples of size 16 and 25 are taken from two normal populations with standard deviations 4.8 and 3.5. The sample means obtained were 18.2 and 23.4. Construct a 90% confidence interval for the difference of the population means.

Solution: Here $n_1 = 16$, $n_2 = 25$, $\overline{x}_1 = 18.2$, $\overline{x}_2 = 23.4$, $\pm_1 = 4.8$ and $\pm_2 = 3.5$. $\pm_1 = 4.8$. $\sim = 0.1$, so $z_{\frac{\sim}{2}} = 1.645$.

Therefore the 90% C.I. is given by

$$\overline{x}_1 - \overline{x}_2 \pm z_{\frac{a}{2}} \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2}} = 18.2 - 23.4 \pm 1.645 \sqrt{\frac{4.8^2}{16} + \frac{3.5^2}{25}} = -5.2 \pm 2.285$$

3.1.2 Case 2: the standard deviations are unknown:

If the two normal populations have unknown standard deviations an approximate $100(1-\sim)\%$ confidence interval is given below:

$$\overline{x}_1 - \overline{x}_2 \pm t_{m,\frac{\sim}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the approximate degrees of freedom is given by $m = min(n_1 - 1,\ n_2 - 1)$.

Example 3.1.2 Two independent samples of size 40 and 50 from a normal population yielded sample means 13.5 and 9.3. The corresponding sample variances were 14.4 and 112.5. Find an approximate 99% confidence interval for the difference between the means.

$$13.5-9.3\pm t_{39,.005}\sqrt{\frac{14.4}{40}+\frac{112.5}{50}}=4.2\pm 2.704\left(1.6155\right)=4.2\pm 4.368$$

3.1.3 Case 3: the standard deviations are equal and unknown:

If the variances of two normal populations are unknown but known to be equal, then we estimate the common variance from the individual sample variances by taking a weighted average. This is called the pooled estimator.

$$s_p^2=\frac{\left(n_1-1\right)s_1^2+\left(n_2-1\right)s_2^2}{n_1+n_2-2}.$$

The $100(1-\sim)\%$ confidence interval for $\mu_1-\mu_2$ is then given by

$$\overline{x}_1-\overline{x}_2\pm t_{n_1+n_2-2,\frac{\sim}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Example 3.1.3 Twelve randomly selected mature citrus trees of one variety have a mean height of 13.8 feet with a standard deviation of 1.2 feet. Fifteen randomly selected mature citrus trees of another variety have a mean height of 12.9 feet with a standard deviation of 1.5 feet. Assuming that the samples were taken from normal populations with equal variances, construct a 95% confidence interval for the difference between the true average heights of the two varieties of citrus trees.

Solution: $s_p^2=\frac{(12-1)1.2^2+(15-1)1.5^2}{12+15-2}=1.8936$, so the 95% confidence interval for $\mu_1-\mu_2$ is given by

$$13.8-12.9\pm t_{25,.025}\sqrt{1.8936}\sqrt{\frac{1}{12}+\frac{1}{15}}=.9\pm 2.060\left(1.376\right)\left(.3873\right)=.9\pm 1.098.$$

So we can be 95% certain that the difference between means is between -0.198 and 1.998 feet.

3.2 Hypothesis testing for difference of means

Here we work with the same framework as in the case of confidence intervals. We want to test the hypothesis $H_0:\ \mu_1-\mu_2=\text{¢}$ against one of these alternatives:

$$H_1:\ \mu_1-\mu_2<\text{¢}$$
$$H_1:\ \mu_1-\mu_2>\text{¢}$$
$$H_1:\ \mu_1-\mu_2\neq\text{¢}.$$

The alternative hypothesis is chosen depending on the situation. In most commonly occurring situations, $\text{¢}=0$.

3.2.1 Case 1: the standard deviations are known

We calculate the test statistics

$$Z=\frac{\overline{x}_1-\overline{x}_2-\text{¢}}{\sqrt{\frac{\pm_1^2}{n_1}+\frac{\pm_2^2}{n_2}}}.$$

We reject the null hypothesis in favor of the alternative hypothesis at significance level $\sim$ if

$$Z\geq z_{\sim}\ \text{ for }\ H_1:\ \mu_1-\mu_2<\text{¢}$$

$$Z\geq z_{\sim}\ \text{ for }\ H_1:\ \mu_1-\mu_2>\text{¢}$$

$$\left|Z\right|\geq z_{\frac{\sim}{2}}\ \text{ for }\ H_1:\ \mu_1-\mu_2\neq\text{¢}$$

Example 3.2.1 For the data in Example 3.1.1, test the hypothesis that the difference of means is -7 against a two-sided alternative at 10% level of significance.

Solution: The null and the alternative hypotheses are $H_0: \mu_1 - \mu_2 = -7$ against $H_1: \mu_1 - \mu_2 \neq -7$. We reject the null hypothesis if $|Z| \geq z_{0.05} = 1.645$.

$$Z = \frac{\overline{x}_1 - \overline{x}_2 - \text{¢}}{\sqrt{\frac{\pm_1^2}{n_1} + \frac{\pm_2^2}{n_2}}} = \frac{18.2 - 23.4 - (-7)}{\sqrt{\frac{4.8^2}{16} + \frac{3.5^2}{25}}} = \frac{1.8}{\sqrt{1.93}} = 1.296$$

As this falls in the acceptance region, we do not reject the null hypothesis. We have no reason to suspect that the difference is not -7.

3.2.2 Case 2: the standard deviations are unknown:

If the two normal populations have unknown standard deviations, we reject the null hypothesis in favor of the alternative hypothesis at significance level $\sim$ if

$$t \leq -t_{m,\sim} \text{ for } H_1: \mu_1 - \mu_2 < \text{¢}$$

$$t \leq -t_{m,\sim} \text{ for } H_1: \mu_1 - \mu_2 > \text{¢}$$

$$|t| \geq t_{m,\frac{\sim}{2}} \text{ for } H_1: \mu_1 - \mu_2 \neq \text{¢}$$

where

$$t = \frac{\overline{x}_1 - \overline{x}_2 - \text{¢}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and $m = min(n_1 - 1, n_2 - 1)$.

Example 3.2.2 Consider the situation given in Example 3.1.2. Test, at .01 level of significance, the hypothesis that the means of these populations are same against the alternative hypothesis that the first population has a larger mean.

Solution: The null and the alternative hypotheses are $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 > 0$. We reject the null hypothesis if $t \geq t_{39,.01} \approx 2.423$.

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{13.5 - 9.3}{\sqrt{\frac{14.4}{40} + \frac{112.5}{50}}} = \frac{4.2}{1.616} = 2.599.$$

As the computed t-value is larger than the critical value $t_{39,.01} = 2.423$, we reject the null hypothesis and conclude that the first population has a larger mean.

3.2.3 Case 3: the standard deviations are equal and unknown:

As before, we calculate the test statistics

$$t = \frac{\overline{x}_1 - \overline{x}_2 - \text{¢}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

We reject the null hypothesis in favor of the alternative hypothesis at significance level $\sim$ if

$$t \leq -t_{n_1+n_2-2,\sim} \text{ for } H_1: \mu_1 - \mu_2 < \text{¢}$$

$$t \geq t_{n_1+n_2-2,\sim} \text{ for } H_1: \mu_1 - \mu_2 > \text{¢}$$

$$|t| \geq t_{n_1+n_2-2,\frac{\sim}{2}} \text{ for } H_1: \mu_1 - \mu_2 \neq \text{¢}$$

Example 3.2.3 Two independent random samples of size six each from two normal populations with equal variance yielded sample means 77.4 and 72.2. The corresponding standard deviations were 3.3 and 2.1. At the significance level 0.01, test whether the two population means are the same.

Solution: The null and the alternative hypotheses are $H_0:\ \mu_1-\mu_2=0$ against $H_1:\ \mu_1-\mu_2\neq 0.$ We reject the null hypothesis if $|t|\geq t_{10,.005}=3.169.$

$$s_p=\sqrt{\frac{5\times 3.3^2+5\times 2.1^2}{10}}2.766.$$

$$t=\frac{\overline{x}_1-\overline{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{5.2}{2.766\sqrt{\frac{1}{6}+\frac{1}{6}}}=3.256$$

As this is in the rejection region, we reject the null hypothesis. There is a significant difference between the means.

3.3 Confidence intervals for ratio of variances

Suppose we have two independent samples of size n_1 and n_2 from normal populations. Then a $100(1-\sim)\%$ confidence interval for the ratio of the variances $\frac{\pm_1^2}{\pm_2^2}$ is given by

$$\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{n_1-1,n_2-1,\frac{\sim}{2}}},\frac{s_1^2}{s_2^2}F_{n_2-1,n_1-1,\frac{\sim}{2}}\right)$$

You need to look up the F-distribution table for the critical values. Confidence intervals for ratio of the standard deviations can be found by taking square-roots.

Example 3.3.1 Find a 98% confidence intervals for the ratio of variances and the ratio of standard deviations for the data in Example 3.1.3.

Solution: 98% confidence interval is given by

$$\left(\frac{1.2^2}{1.5^2}\frac{1}{F_{11,14,.01}},\frac{1.2^2}{1.5^2}F_{11,14,.01}\right)=\left(\frac{.64}{3.86},.64\times 4.29\right)=(.166,2.746).$$

Taking square-roots, we get the confidence intervals for the ratio of standard deviations to be

$$\left(\sqrt{.166},\sqrt{2.746}\right)=(.407,1.657).$$

3.4 Hypothesis Testing for equality of two variances

Suppose we want to test the hypothesis $H_0:\ \pm_1=\pm_2$ against a one-sided or two-sided alternative. We reject the null hypothesis if

$$\frac{s_1^2}{s_2^2}\geq F_{n_1-1,n_2-1,\sim}\ \text{ for }\ H_1:\ \pm_1>\pm_2,$$

$$\frac{s_2^2}{s_1^2}\geq F_{n_2-1,n_1-1,\sim}\ \text{ for }\ H_1:\ \pm_1<\pm_2.$$

For the two-sided alternative $H_1:\ \pm_1\neq\pm_2$, we reject the null hypothesis if

$$\frac{s_1^2}{s_2^2}\geq F_{n_1-1,n_2-1,\frac{\sim}{2}}\ \text{ if }\ s_1\geq s_2$$

$$\frac{s_2^2}{s_1^2}\geq F_{n_2-1,n_1-1,\frac{\sim}{2}}\ \text{ if }\ s_1<s_2.$$

Example 3.4.1 In the comparison of two kinds of paint, it was found that four 1-gallon cans of one brand cover 546 square feet on the average with a standard deviation of 31 square feet, while four 1-gallon cans of another brand cover 492 square feet on the average with a standard deviation of 26 square feet. Test the hypothesis that the variances are the same against the alternative that the variance of the first population is larger.

Solution: Here $n_1 = n_2 = 4$, $s_1 = 31$, $s_2 = 26$. The hypothesis is $H_0 : \pm_1 = \pm_2$ and the alternative hypothesis is $H_1 : \pm_1 > \pm_2$. We reject H_0 if $\frac{s_1^2}{s_2^2} \geq F_{3,3,.05} = 9.28$. As $\frac{s_1^2}{s_2^2} = \frac{31^2}{26^2} = 1.42$, do not reject the null hypothesis.

Example 3.4.2 To compare two kinds of bumper guards, six of each kind were mounted on a certain make of car and each car was run into a concrete wall at 5 miles per hour. Costs of repair are given below: Is it reasonable to assume that the two populations have the same variance?

Bumper Guard 1	127	168	143	165	122	139
Bumper Guard 2	154	135	132	171	153	149

Solution: Here $n_1 = n_2 = 6$, $s_1 = 19.06$, $s_2 = 14.21$. We are testing the hypothesis $H_0 : \pm_1 = \pm_2$ against the alternative hypothesis $H_1 : \pm_1 \neq \pm_2$. As $s_1 \geq s_2$, we reject H_0 if $\frac{s_1^2}{s_2^2} \geq F_{5,5,.01} = 1.0$. As $\frac{s_1^2}{s_2^2} = \frac{19.06^2}{14.21^2} = 1.8$, do not reject the null hypothesis. It is reasonable to assume that the two populations have the same variance.

Chapter 4

Analysis of Variance

Statistical Tables

Table 4.1: **Normal Distribution**

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Table 4.2: ***t* distribution**

~

d.f.	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.376	3.078	6.314	12.706	31.821	63.656	318.289
2	1.061	1.886	2.920	4.303	6.965	9.925	22.328
3	0.978	1.638	2.353	3.182	4.541	5.841	10.214
4	0.941	1.533	2.132	2.776	3.747	4.604	7.173
5	0.920	1.476	2.015	2.571	3.365	4.032	5.894
6	0.906	1.440	1.943	2.447	3.143	3.707	5.208
7	0.896	1.415	1.895	2.365	2.998	3.499	4.785
8	0.889	1.397	1.860	2.306	2.896	3.355	4.501
9	0.883	1.383	1.833	2.262	2.821	3.250	4.297
10	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.330	1.734	2.101	2.552	2.878	3.610
19	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.325	1.725	2.086	2.528	2.845	3.552
21	0.859	1.323	1.721	2.080	2.518	2.831	3.527
22	0.858	1.321	1.717	2.074	2.508	2.819	3.505
23	0.858	1.319	1.714	2.069	2.500	2.807	3.485
24	0.857	1.318	1.711	2.064	2.492	2.797	3.467
25	0.856	1.316	1.708	2.060	2.485	2.787	3.450
26	0.856	1.315	1.706	2.056	2.479	2.779	3.435
27	0.855	1.314	1.703	2.052	2.473	2.771	3.421
28	0.855	1.313	1.701	2.048	2.467	2.763	3.408
29	0.854	1.311	1.699	2.045	2.462	2.756	3.396
30	0.854	1.310	1.697	2.042	2.457	2.750	3.385
31	0.853	1.309	1.696	2.040	2.453	2.744	3.375
32	0.853	1.309	1.694	2.037	2.449	2.738	3.365
33	0.853	1.308	1.692	2.035	2.445	2.733	3.356
34	0.852	1.307	1.691	2.032	2.441	2.728	3.348
35	0.852	1.306	1.690	2.030	2.438	2.724	3.340
40	0.851	1.303	1.684	2.021	2.423	2.704	3.307
50	0.849	1.299	1.676	2.009	2.403	2.678	3.261
60	0.848	1.296	1.671	2.000	2.390	2.660	3.232
70	0.847	1.294	1.667	1.994	2.381	2.648	3.211
80	0.846	1.292	1.664	1.990	2.374	2.639	3.195
∞	0.841	1.282	1.645	1.960	2.326	2.576	3.091

Table 4.3: **Chi-square distribution**

~

$d.f.$	0.995	0.990	0.975	0.950	0.050	0.025	0.010	0.005
1	3.9E-05	0.00016	0.00098	0.00393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.170

Table 4.4: ***F*** **distribution witl** ~ = 0.05

$\neg_1$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161	199	216	225	230	234	237	239	241	242	243	244	245	245	246
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.23
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40

Table 4.5: ***F*** **distribution witl** ~ = 0.01

$\neg_1$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106	6126	6143	6157
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1	27.0	26.9	26.9
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.5	14.4	14.3	14.2	14.2
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89	9.82	9.77	9.72
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25
12	9.33	6.93	5.95	4.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82
14	8.86	6.51	7.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66
15	8.68	6.36	7.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52