Date: October 05,07,08,…,2020
MA 2302: Introduction to Probability and Statistics

Statistical Inference

Instructor
Prof. Gopal Krishna Panda
Department of Mathematics
NIT Rourkela

Statistical Inference

Statistical inference consists of two parts:

- Estimation of parameters
- Testing of statistical hypotheses

Usually, the population parameters remain unknown. For example, if the B. Tech. students of our institute constitute a population, and we want to get information about the mean mark and standard deviation in Math I, of course with the assumption that the marks of all students is not accessible and we can estimate it by choosing a random sample of students of size say 100 and asking them about their Math I marks. Once we get a sample, we can get many things.

1. We can estimate the mean mark and standard deviation of marks of the population
2. We can test if the mean population mark is equal to say 55 with some allowed error.
3. We can determine an interval in which the population mark lie with a very high probability.

Point Estimation

Estimating the mean mark or variance or any other population parameter is known as point estimation. We assume that the population is governed by certain probability distribution with one or more parameters. By saying that $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2 means that $X_1, X_2, \ldots, X_n$ are independent random variables each with mean μ and variance σ^2. This is an example and in place of normal distribution, the population distribution can be binomial, Poisson, exponential, gamma or any other distribution. If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with probability mass function $f(x)$ and say one parameter θ, then we write the pmf as $f(x, \theta)$. Observe that because of independence,

$$\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \Pr\{X_1 = x_1\} \Pr\{X_2 = x_2\} \cdots \Pr\{X_n = x_n\}$$

$$= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

The product $L = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$ is known as the likelihood function. L is the probability that $x_1, x_2, \ldots, x_n$ belong to the population under consideration.

Maximum Likelihood Estimation (MLE)

In case of a continuous distribution, though we cannot have an equation like

$$\Pr\left\{x_1 - \frac{dx_1}{2} \le X_1 \le x_1 + \frac{dx_1}{2}, x_2 - \frac{dx_2}{2} \le X_2 \le x_2 + \frac{dx_2}{2}, \dots, x_n - \frac{dx_n}{2} \le X_n \le x_n + \frac{dx_n}{2}\right\}$$
$$= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) dx_1 dx_2 \cdots dx_n$$

and the product $L = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$ is known as the likelihood function. In either case, the method of maximum likelihood of estimating parameter(s) consists of maximizing the likelihood function with respect to the parameter(s). To find MLE, we need to solve

$$\frac{\partial L}{\partial \theta} = 0$$

for θ and get the estimate as a function of the sample observations, i.e. $x_1, x_2, \dots, x_n$, that is the observed values of $X_1, X_2, \dots, X_n$. If the population has more parameters to estimate, we may consider solving the system of equations

$$\frac{\partial L}{\partial \theta_1} = 0, \qquad \frac{\partial L}{\partial \theta_2} = 0, \dots$$

Maximum Likelihood Estimation (MLE)

Since, the functions $f(x)$ and $\log f(x)$ attains their extreme values at the same point, instead of solving

$$\frac{\partial L}{\partial \theta} = 0$$

we prefer solving the equation

$$\frac{\partial \log L}{\partial \theta} = 0$$

for θ and in case of more than one parameters, we consider solving

$$\frac{\partial \log L}{\partial \theta_1} = 0, \qquad \frac{\partial \log L}{\partial \theta_2} = 0, \ldots$$

for $\theta_1, \theta_2, \ldots.$

***Example 1**: Using a sample of size n, find the maximum-likelihood estimate for the parameter λ of a population governed by the probability density function $f(x,\lambda) = \lambda e^{-\lambda x}$ if $x > 0$ and $f(x) = 0$ otherwise.*

Ans. The likelihood function is given by

$$L = \prod_{i=1}^{n} f(x_i, \lambda) = \prod_{i=1}^{n} \lambda\, e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

$$\Rightarrow \log L = n \log \lambda - \lambda \sum_{i=1}^{n} x_i\,.$$

For MLE of λ, we have solve

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

where $\bar{x}$ is the sample mean. In the final step $\hat{\lambda}$ is used in place of λ to differentiate between the parameter and its estimate.

***Example 2**: Using a sample of size n, find the maximum-likelihood estimate for the mean/variance of a Poisson distribution. (Actually, a population governed by the Poisson probability law.)*

Ans. The likelihood function is given by

$$L = \prod_{i=1}^{n} f(x_i, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

$$\Rightarrow \log L = -n\lambda + \log\lambda \sum_{i=1}^{n} x_i - \log \prod_{i=1}^{n} x_i!\,.$$

For MLE of λ, we have solve

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

where $\bar{x}$ is the sample mean.

***Example 3**: Using a sample of size n, find the maximum-likelihood estimate for the mean and variance of a normal population (distribution) if (a) both mean and variance are unknown, (b) mean is unknown, the variance is known, (c) the mean is known and the variance is unknown.*

Ans. The likelihood function of normal distribution is given by

$$L = \prod_{i=1}^{n} f(x_i, \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\sqrt{2\pi}\right)^{-n} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2},$$

$$\Rightarrow \log L = C - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 .$$

Case I: Both μ and σ^2 are unknown. In this case to find the MLEs of μ and σ^2, we need to solve

$$\frac{\partial \log L}{\partial \mu} = 0, \frac{\partial \log L}{\partial \sigma^2} = 0,$$

$$\Rightarrow -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)=0 \text{ and } -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

$$\Rightarrow \sum_{i=1}^{n}(x_i-\mu)=0 \text{ and } n\sigma^2=\sum_{i=1}^{n}(x_i-\mu)^2 .$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu \text{ and } n\sigma^2=\sum_{i=1}^{n}(x_i-\mu)^2 .$$

$$\Rightarrow \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

and

$$\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 .$$

Case II: μ is unknown, σ^2 is known. In this case, we need to solve only

$$\frac{\partial \log L}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Case II: μ is known, σ^2 is unknown. In this case, we need to solve only

$$\frac{\partial \log L}{\partial \sigma^2} = 0 \Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

$$\Rightarrow n\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \Rightarrow \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 .$$

Observe that in case of a binomial distribution, a sample of size 1 is sufficient since if out of n trials, if it is known that the number of successes is x, that is enough to estimate the only parameter p of binomial distribution. Hence, in case of a binomial distribution, the pmf is the likelihood function.

***Example 4**: Using a sample of size* 1*, find the maximum-likelihood estimate for the parameter p of a binomial distribution.*

Ans. Since the sample size is 1, the likelihood function is given by

$$L = f(x,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\Rightarrow \log L = \log \binom{n}{x} + x \log p + (n-x)\log(1-p)\,.$$

To find MLE of p, we need to solve

$$\frac{\partial \log L}{\partial p} = 0.$$

But

$$\frac{\partial \log L}{\partial p} = 0 \Rightarrow \frac{x}{p} - \frac{n-x}{1-p} = 0 \Rightarrow x(1-p) = (n-x)p$$

$$\Rightarrow np = x \Rightarrow \hat{p} = \frac{x}{n}.$$

What is the implication of this MLE? If n tosses of a coin results in x heads, then the MLE for the probability of head is x/n. Similarly, if n balls are drawn from box with replacement out of which x balls are white then the MLE for the proportion of white ball is x/n. If a student fails x tests out of n math tests, then than the MLE for his probability of his passing a test is $\frac{n-x}{n}$ and so on.

Example 5*: Using a sample of size n, find the maximum-likelihood estimate for the parameter λ of a gamma distribution.*

Ans. The pdf of a gamma distribution with parameters λ *and* α are given by

$$f(x, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}.$$

Hence, the likelihood function is given by

$$L = \prod_{i=1}^{n} f(x_i, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} e^{-\lambda x_i} x_i^{\alpha-1} = \frac{\lambda^{n\alpha}}{\{\Gamma(\alpha)\}^n} e^{-\lambda \sum_{i=1}^{n} x_i} \left\{ \prod_{i=1}^{n} x_i \right\}^{\alpha-1}$$

Hence,

$$\log L = n\alpha \log \lambda + \log \left\{ \frac{1}{\{\Gamma(\alpha)\}^n} \right\} - \lambda \sum_{i=1}^{n} x_i + \log \left\{ \prod_{i=1}^{n} x_i \right\}^{\alpha-1} .$$

For MLE of λ

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\lambda} = \frac{n\alpha}{\sum_{i=1}^{n} x_i} = \frac{\alpha}{\bar{x}}.$$

***Example 5**: Using a sample of size n, find the maximum-likelihood estimate for the parameter θ of the uniform distribution $f(x,\theta) = \frac{1}{\theta}$ if $0 < x < \theta$ and $f(x,\theta) = 0$ otherwise.*

Ans: Let $x_1, x_2, \ldots, x_n$ be the observed value of the sample. Then the likelihood function is given by

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

which is not a function of the sample observation. However, since the observations form the uniform distribution on the interval $(0, \theta)$ cannot exceed θ, it follows that

$$\hat{\theta} = \max_{1 \le i \le n} x_i \, .$$

Similarly, the MLEs for a and b for the uniform distributions on (a, b) are respectively,

$$a = \min_{1 \le i \le n} x_i \quad \text{and} \quad b = \max_{1 \le i \le n} x_i \, .$$

***Example 6**: Using the sample* $\{67,75,79,52,49,75,69,70,64\}$*, find the maximum-likelihood estimate for the* $mean$ and variance of $N(\mu,\sigma^2)$. Also, find the MLE of σ^2if it is known that $\mu = 65$.

Ans: In this case $n = 9$. It is known that if both μ and σ^2 are unknown, their MLEs are

$$\hat{\mu} = \bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = (You\ calculate).$$

$$\widehat{\sigma^2} = \frac{1}{9}\sum_{i=1}^{9}(x_i - \hat{\mu})^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i - \bar{x})^2 = (You\ calculate).$$

If μ is know, then

$$\widehat{\sigma^2} = \frac{1}{9}\sum_{i=1}^{9}(x_i - \mu)^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i - 65)^2 = (You\ calculate).$$

Method of moments.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population governed by the probability law (pmf or pdf) $f(x, \theta)$, where θ may a scalar or vector. The r-th moment of the distribution is given by

$$\mu_r' = E(X^r) = \sum_x x^r f(x)$$

if X is continuous and

$$\mu_r' = E(X^r) = \int_{-\infty}^{\infty} x^r f(x) dx$$

if X is discrete. If we define

$$M_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r$$

then, it is easy to see that

$$E(M_r) = \frac{1}{n} \sum_{i=1}^{n} E(X_i^r) = \frac{1}{n} \sum_{i=1}^{n} \mu_r{}' = \mu_r{}'.$$

Recall that $X_1, X_2, \ldots, X_n$ is a random sample from a population governed by the probability law (pmf or pdf) $f(x, \theta)$ means that $X_1, X_2, \ldots, X_n$ are independent random variables governed by the probability law $f(x, \theta)$. In Statistics, such random variables are called identically and independently distributed (i.i.d) random variables and $E(M_r) = \mu_r'$ means that M_r is an unbiased estimate of μ_r'. This means that each sample of size n corresponds to a M_r and if the population size is N, then there are $\binom{N}{n}$ possible samples of size n and hence, there are $\binom{N}{n}$ possible values of M_r and the average of all these M_r's is equal to μ_r'.

In the method of moments, solve the equations $\mu_r' = M_r, r = 1,2,3, \ldots$ till all the parameters of the population are estimated. This concept can be better clarified by means of the following examples:

Example 7: *Using the method of moments and the random sample* $\{1.37, 2.31, 0.06, 3.21, 1.92, 3.02\}$ *estimate for the parameter* θ *of the uniform distribution* $f(x,\theta) = \frac{1}{\theta}$ *if* $0 < x < \theta$ *and* $f(x,\theta) = 0$ *otherwise.*

Ans: Observe that

$$\mu_1' = E(X) = \int_{-\infty}^{\infty} x\, f(x) dx = \int_0^{\theta} x \cdot \frac{1}{\theta} dx = \frac{\theta}{2}.$$

In the present case, $n = 6$ and

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{6}(1.37 + 2.31 + 0.06 + 3.21 + 1.92 + 3.02) = 1.9817.$$

To estimate θ we need to solve $\mu_1' = M_1$. Hence,

$$\frac{\theta}{2} = 1.9817, \Rightarrow \hat{\theta} = 2 \times 1.9817 = 3.9633.$$

Example 8: *Using the method of moments and the random sample* $\{1.1, 4.4, 3.1, 7.1, 5.7, 4.8, 3.9, 5.8, 1.6, 0.7, 1.3, 7.0\}$ *estimate for the parameter* θ *of the distribution* $f(x,\theta) = \frac{1}{\theta^2} e^{-\frac{x}{\theta^2}}$ *if* $x > 0$ *and* $f(x,\theta) = 0$ *otherwise.*

Ans: Notice that

$$\mu_1' = E(X) = \int_{-\infty}^{\infty} x\, f(x) dx = \frac{1}{\theta^2} \int_0^{\infty} x\, e^{-\frac{x}{\theta^2}} dx = \theta^2.$$

In the present case, $n = 12$ and

$$M_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$= \frac{1}{12} (1.1 + 4.4 + 3.1 + 7.1 + 5.7 + 4.8 + 3.9 + 5.8 + 1.6 + 0.7 + 1.3 + 7.0) = 3.9$$

To estimate θ we need to solve $\mu_1' = M_1$. Hence,

$$\theta^2 = 3.9 \Rightarrow \hat{\theta} = 1.97484.$$

Example 9: *Using the method of moments and the random sample* $\{51,76,84,31,55,12,21,66,82,92\}$ *from* $N(\mu,\sigma^2)$ *estimate both* μ *and* σ^2.

Ans: Here,

$$\mu_1' = E(X) = \mu, \qquad \mu_2' = E(X^2) = E(X^2) - \mu^2 + \mu^2 = \sigma^2 + \mu^2.$$

In the present case, $n = 10$ and

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}(51 + 76 + 84 + 31 + 55 + 12 + 21 + 66 + 82 + 92) = 57$$

$$M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{10}(51^2 + 76^2 + 84^2 + 31^2 + 55^2 + 12^2 + 21^2 + 66^2 + 82^2 + 92^2)$$

$$= 3954.8$$

For method of moments estimates of μ and σ^2 are

$$\mu_1' = M_1 \Rightarrow \hat{\mu} = M_1 = 57,$$

$$\mu_2' = M_2 \Rightarrow \sigma^2 + \mu^2 = M_2 \Rightarrow \widehat{\sigma^2} = M_2 - M_1^2 = 3954.8 - 57^2 = 705.8$$

Example 10: *Using the method of moments and the random sample* $\{1.2, 3.5, 4.6, 5.2, 7.5, 8.7\}$ *estimate the parameter* β *of the distribution with pdf* $f(x) = \frac{\beta}{x^{\beta+1}}$ *if* $x > 1$ *and* $f(x) = 0$ *otherwise* $(\beta > 0)$. (*This distribution is known as the Pareto distribution*.)

Ans: In this case,

$$\mu_1' = E(X) = \int_{-\infty}^{\infty} x\, f(x) dx = \beta \int_1^{\infty} x \cdot \frac{1}{x^{\beta+1}} dx$$

$$= \beta \int_1^{\infty} x^{-\beta}\, dx = \frac{\beta}{\beta - 1}.$$

From the given sample,

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{6}(1.2 + 3.5 + 4.6 + 5.2 + 7.5 + 8.7) = 5.1167.$$

Hence, the method of moments estimate of β can be obtained by solving $\mu_1' = M_1$.

$$\Rightarrow \frac{\beta}{\beta - 1} = 5.1167 \Rightarrow \hat{\beta} = 1.243.$$

Confidence Interval Estimation

The method of maximum likelihood estimation and the method of moments estimations are point estimations for the parameter(s) of a population (distribution). However, sometimes point estimates may not provide adequate information about the population parameter(s) and estimating an interval for a parameter is more informative. The interval is so estimated that the population parameter lies in this interval with a very high preassigned probability. This interval is known as the confidence interval and the high preassigned probability is known as the confidence coefficient. In this section, our discussion is limited to setting up confidence intervals for the mean and variance of a normal distribution only.

Confidence interval for the mean of a normal population

Let us assume that $X_1, X_2, \dots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, where one or more of the parameters μ and σ^2 are unknown. This means that $X_1, X_2, \dots, X_n$ are independent random variables and $X_i \sim N(\mu, \sigma^2), i = 1,2, \dots, n$. Thus, $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for each i. Now let

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

be the sample mean. By linearity of expectation,

$$E(\bar{X}) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

and

$$Var(\bar{X}) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}.$$

Consequently,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and the standardized variable corresponding to $\bar{X}$ is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Now let γ be such that $0 < \gamma < 1$ and is close to 1, say $\gamma = 0.95, 0.98, 0.99$ etc. Then we can find a positive number c such that

$$\Pr\{|Z| \leq c\} = \gamma \Rightarrow \Phi(c) - \Phi(-c) = \gamma$$

$$\Rightarrow 2\Phi(c) = 1 + \gamma \Rightarrow \Phi(c) = \frac{1+\gamma}{2}$$

and for a given value of γ, such a c can be obtained from Table A8. For example, if $\gamma = 0.95$, then $\Phi(c) = \frac{1+\gamma}{2} = 0.975$ and then $c = 1.96$. Similarly, if $\gamma = 0.98$, then $\Phi(c) = 0.99$ and $c = 2.326$ and for if $\gamma = 0.99$, then $\Phi(c) = 0.995$ and $c = 2.576$.

Table A8 Normal Distribution

Values of z for given values of $\Phi(z)$ [see (3), Sec. 24.8] and $D(z) = \Phi(z) - \Phi(-z)$
Example: $z = 0.279$ if $\Phi(z) = 61\%$; $z = 0.860$ if $D(z) = 61\%$.

%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
1	−2.326	0.013	41	−0.228	0.539	81	0.878	1.311
2	−2.054	0.025	42	−0.202	0.553	82	0.915	1.341
3	−1.881	0.038	43	−0.176	0.568	83	0.954	1.372
4	−1.751	0.050	44	−0.151	0.583	84	0.994	1.405
5	−1.645	0.063	45	−0.126	0.598	85	1.036	1.440
6	−1.555	0.075	46	−0.100	0.613	86	1.080	1.476
7	−1.476	0.088	47	−0.075	0.628	87	1.126	1.514
8	−1.405	0.100	48	−0.050	0.643	88	1.175	1.555
9	−1.341	0.113	49	−0.025	0.659	89	1.227	1.598
10	−1.282	0.126	50	0.000	0.674	90	1.282	1.645
11	−1.227	0.138	51	0.025	0.690	91	1.341	1.695
12	−1.175	0.151	52	0.050	0.706	92	1.405	1.751
13	−1.126	0.164	53	0.075	0.722	93	1.476	1.812
14	−1.080	0.176	54	0.100	0.739	94	1.555	1.881
15	−1.036	0.189	55	0.126	0.755	95	1.645	1.960
16	−0.994	0.202	56	0.151	0.772	96	1.751	2.054
17	−0.954	0.215	57	0.176	0.789	97	1.881	2.170
18	−0.915	0.228	58	0.202	0.806	97.5	1.960	2.241
19	−0.878	0.240	59	0.228	0.824	98	2.054	2.326
20	−0.842	0.253	60	0.253	0.842	99	2.326	2.576
21	−0.806	0.266	61	0.279	0.860	99.1	2.366	2.612
22	−0.772	0.279	62	0.305	0.878	99.2	2.409	2.652
23	−0.739	0.292	63	0.332	0.896	99.3	2.457	2.697
24	−0.706	0.305	64	0.358	0.915	99.4	2.512	2.748
25	−0.674	0.319	65	0.385	0.935	99.5	2.576	2.807
26	−0.643	0.332	66	0.412	0.954	99.6	2.652	2.878
27	−0.613	0.345	67	0.440	0.974	99.7	2.748	2.968
28	−0.583	0.358	68	0.468	0.994	99.8	2.878	3.090
29	−0.553	0.372	69	0.496	1.015	99.9	3.090	3.291
30	−0.524	0.385	70	0.524	1.036			
31	−0.496	0.399	71	0.553	1.058	99.91	3.121	3.320
32	−0.468	0.412	72	0.583	1.080	99.92	3.156	3.353
33	−0.440	0.426	73	0.613	1.103	99.93	3.195	3.390
34	−0.412	0.440	74	0.643	1.126	99.94	3.239	3.432
35	−0.385	0.454	75	0.674	1.150	99.95	3.291	3.481
36	−0.358	0.468	76	0.706	1.175	99.96	3.353	3.540
37	−0.332	0.482	77	0.739	1.200	99.97	3.432	3.615
38	−0.305	0.496	78	0.772	1.227	99.98	3.540	3.719
39	−0.279	0.510	79	0.806	1.254	99.99	3.719	3.891
40	−0.253	0.524	80	0.842	1.282			

Having obtained such a c, one can see that

$$\Pr\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \le c\right\} = \gamma \Rightarrow \Pr\left\{\left|\frac{\mu-\bar{X}}{\sigma/\sqrt{n}}\right| \le c\right\} = \gamma$$

$$\Rightarrow \Pr\left\{-c \cdot \frac{\sigma}{\sqrt{n}} \le \mu - \bar{X} \le c \cdot \frac{\sigma}{\sqrt{n}}\right\} = \gamma$$

$$\Rightarrow \Pr\left\{\bar{X} - c \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right\} = \gamma$$

Thus, the population mean μ lies in the interval

$$\left[\bar{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right]$$

with probability γ and is called a $100\gamma\%$ confidence interval for the population mean μ. Writing $k = c \cdot \frac{\sigma}{\sqrt{n}}$, the confidence interval can be written as $[\bar{X} - k, \bar{X} + k]$. $\bar{X} - k$ and$\bar{X} + k$ are respectively called the lower and upper confidence limits LCL and UCL.

Example 11: *Using a sample of size 40 with sample mean 57 and known population variance* $\sigma^2 = 225$ *from a normal distribution, find a 98% confidence interval for the population mean* μ.

Ans. Given that $n = 40$, $\sigma^2 = 225 \Rightarrow \sigma = 15, \bar{X} = 57$ and $\gamma = 0.98$. To find a 98% confidence interval, we need to find c such that

$$\Phi(c) = \frac{1 + 0.98}{2} = 0.99 \Rightarrow c = 2.326.$$

Thus,

$$k = c \cdot \frac{\sigma}{\sqrt{n}} = 2.326 \times \frac{15}{\sqrt{40}} = 5.52,$$

$$LCL = \bar{X} - k = 57 - 5.52 = 51.48, \qquad UCL = \bar{X} + k = 57 + 5.52 = 62.52$$

and a 98% confidence interval for μ is $[51.48, 62.52\ ]$.

When σ^2 is unknown, we cannot find the confidence interval for μ as

$$\left[\bar{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right]$$

and σ^2 needs to be replaced by its sample unbiased estimate

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right],$$

called the sample variance and then, of course, the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is no more $N(0,1)$, rather, it follows a t-distribution with $n-1$ degrees of freedom (d.f.). The t-distribution approaches to the standard normal distribution as its d.f.$\to \infty$. Hence, for large degrees of freedom, say ≥ 30, the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is approximately $N(0,1)$ and we can set up a $100\gamma\%$ confidence interval for μ as

$$\left[\bar{X} - c \cdot \frac{S}{\sqrt{n}}, \bar{X} + c \cdot \frac{S}{\sqrt{n}}\right]$$

where as usual c: $\Phi(c) = \frac{1+\gamma}{2}$.

Example 12: *The CGPA obtained by the final year students of NIT Rourkela in the year 2016 was approximately normally distributed. The CGPA of randomly chosen 45 students are* $X_1, X_2, \dots, X_{45}$ such that $\sum_{i=1}^{45} X_i = 301.7$, $\sum_{i=1}^{45} X_i^2 = 2801.67$. *Find a 95% confidence interval for the mean CGPA of final year students of 2016.*

Ans. In this case $n = 45$, hence the sample is large.

$$\bar{X} = \frac{1}{45}\sum_{i=1}^{45} X_i = \frac{301.7}{45} = 6.704,$$

$$S^2 = \frac{1}{45-1}\sum_{i=1}^{45}(X_i - \bar{X})^2 = \frac{1}{44}\left[\sum_{i=1}^{45} X_i^2 - 45 \times \bar{X}^2\right]$$

$$= \frac{2801.67 - 45 \times 6.704^2}{44} = 17.709 \Rightarrow S = 4.21$$

Since $\gamma = 0.95$,

$$c: \Phi(c) = \frac{1 + 0.95}{2} = 0.975$$

and from Table A8, $c = 1.96$.

Thus,

$$k = c \cdot \frac{S}{\sqrt{n}} = 1.96 \times \frac{4.21}{\sqrt{45}} = 1.23,$$

$LCL = \bar{X} - k = 6.704 - 1.23 = 5.474,\ UCL = \bar{X} + k$ =6.704+1.23=**7.934**

and a 95% confidence interval for μ is $[5.474, 7.934\ ]$.

IMP: When the sample size is small, say $n \leq 30$, the confidence interval for the population mean still remains $\left[\bar{X} - c \cdot \frac{S}{\sqrt{n}}, \bar{X} + c \cdot \frac{S}{\sqrt{n}}\right]$, but the value of c will be different. In this case, $c: F(c) = \frac{1+\gamma}{2}$, where F is the distribution function of a t-distribution with $n - 1$ degrees of freedom.

Example 13: *A manufacturing unit of a company produces 56Ω resisters and it is known that the resistance of the resistors is not exactly equal to 56Ω, but normally distributed with some unknown μ mean and unknown variance σ^2. The resistance of a sample of 11 resisters are* $55.6, 57.9, 56.1, 58.2, 56.9, 55.3, 55.1, 57.6, 54.8, 56.1$ *and* 55.7 *Ohms. Find a 98% confidence interval for the mean value of resisters produced by this unit.*

Ans. In this case $n = 11$, hence the sample is small.

$$\sum_{i=1}^{11} X_i = 619.3 \Rightarrow \bar{X} = \frac{1}{11}\sum_{i=1}^{11} X_i = 56.3\,, \sum_{i=1}^{11} X_i^2 = 34880.43$$

$$S^2 = \frac{1}{11-1}\sum_{i=1}^{11}(X_i - \bar{X})^2 = \frac{1}{10}\left[\sum_{i=1}^{11} X_i^2 - 11 \times \bar{X}^2\right]$$

$$= \frac{34880.43 - 11 \times 56.3^2}{10} = 1.384 \Rightarrow S = 1.1764$$

Thus,

$$n = 11, \bar{X} = 56.3, S = 1.1764.$$

Here $\gamma = 0.98$ and degrees of freedom is $n - 1 = 10$. To find c, we need to solve

$$F(c) = \frac{1 + 0.98}{2} = 0.99$$

where F is the distribution function of a t-distribution with 10 degrees of freedom. From table A9, the value of c is 2.72. Hence

$$k = c \cdot \frac{S}{\sqrt{n}} = 2.76 \times \frac{1.1764}{\sqrt{11}} = 0.979.$$

Hence,

$$LCL = \bar{X} - k = 56.3 - 0.979 = 55.321,$$
$$UCL = \bar{X} + k = 56.3 + 0.979 = 57.279$$

and a 98% confidence interval for the mean value of resistance produced by the unit is $[\mathbf{55.321, 57.279}]$.

Table A9 t-Distribution

Values of z for given values of the distribution function $F(z)$ (see (8) in Sec. 25.3). Example: For 9 degrees of freedom, $z = 1.83$ when $F(z) = 0.95$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.32	0.29	0.28	0.27	0.27	0.26	0.26	0.26	0.26	0.26
0.7	0.73	0.62	0.58	0.57	0.56	0.55	0.55	0.55	0.54	0.54
0.8	1.38	1.06	0.98	0.94	0.92	0.91	0.90	0.89	0.88	0.88
0.9	3.08	1.89	1.64	1.53	1.48	1.44	1.41	1.40	1.38	1.37
0.95	6.31	2.92	2.35	2.13	2.02	1.94	1.89	1.86	1.83	1.81
0.975	12.7	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23
0.99	31.8	6.96	4.54	3.75	3.36	3.14	3.00	2.90	2.82	2.76
0.995	63.7	9.92	5.84	4.60	4.03	3.71	3.50	3.36	3.25	3.17
0.999	318.3	22.3	10.2	7.17	5.89	5.21	4.79	4.50	4.30	4.14

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
0.7	0.54	0.54	0.54	0.54	0.54	0.54	0.53	0.53	0.53	0.53
0.8	0.88	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86
0.9	1.36	1.36	1.35	1.35	1.34	1.34	1.33	1.33	1.33	1.33
0.95	1.80	1.78	1.77	1.76	1.75	1.75	1.74	1.73	1.73	1.72
0.975	2.20	2.18	2.16	2.14	2.13	2.12	2.11	2.10	2.09	2.09
0.99	2.72	2.68	2.65	2.62	2.60	2.58	2.57	2.55	2.54	2.53
0.995	3.11	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.85
0.999	4.02	3.93	3.85	3.79	3.73	3.69	3.65	3.61	3.58	3.55

$F(z)$	Number of Degrees of Freedom									
	22	24	26	28	30	40	50	100	200	∞
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.25	0.25	0.25
0.7	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.52
0.8	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.85	0.84	0.84
0.9	1.32	1.32	1.31	1.31	1.31	1.30	1.30	1.29	1.29	1.28
0.95	1.72	1.71	1.71	1.70	1.70	1.68	1.68	1.66	1.65	1.65
0.975	2.07	2.06	2.06	2.05	2.04	2.02	2.01	1.98	1.97	1.96
0.99	2.51	2.49	2.48	2.47	2.46	2.42	2.40	2.36	2.35	2.33
0.995	2.82	2.80	2.78	2.76	2.75	2.70	2.68	2.63	2.60	2.58
0.999	3.50	3.47	3.43	3.41	3.39	3.31	3.26	3.17	3.13	3.09

Example 14: *A recent survey shows that the heights of 8 B. Tech. boy students of NIT Rourkela in the year 2020 are* $69.1, 72.5, 61.7, 66.4, 71.5, 59.3, 67.2, 64.3$ *inches. Find a 99% confidence interval for the population mean height of B. Tech. boys of NIT Rourkela in the year* 2020.

Ans. In this case $n = 8$, hence the sample is small.

$$\sum_{i=1}^{8} X_i = 532 \Rightarrow \bar{X} = \frac{1}{8}\sum_{i=1}^{8} X_i = 66.5\,, \sum_{i=1}^{8} X_i^2 = 35525.98$$

$$S^2 = \frac{1}{8-1}\sum_{i=1}^{8}(X_i - \bar{X})^2 = \frac{1}{7}\left[\sum_{i=1}^{8} X_i^2 - 8 \times \bar{X}^2\right]$$

$$= \frac{35525.98 - 8 \times 66.5^2}{7} = 21.14 \Rightarrow S = 4.5978.$$

Thus,

$$n = 11, \bar{X} = 66.5, S = 4.5978.$$

Here $\gamma = 0.99$ and degrees of freedom is $n - 1 = 7$. To find c, we need to solve

$$F(c) = \frac{1 + 0.99}{2} = 0.995$$

where F is the distribution function of a t-distribution with 7 degrees of freedom. From table A9, the value of c is 3.50. Hence,

$$k = c \cdot \frac{S}{\sqrt{n}} = 3.50 \times \frac{4.5978}{\sqrt{8}} = 5.69.$$

Hence,

$$LCL = \bar{X} - k = 66.5 - 5.69 = 60.81,$$
$$UCL = \bar{X} + k = 66.5 + 5.69 = 72.19$$

and a 99% confidence interval for the mean height of B. Tech. boys for the year 2020 is $[\mathbf{60.8}", \mathbf{72.2}"]$.

Confidence interval for the variance of a normal population

By definition, the square of a standard normal variable has a χ^2-distribution (chi square) with 1 degree of freedom. Similarly, the sum of squares of n standard normal variables has a χ^2-distribution with n degrees of freedom. Thus, if $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

has a χ^2-distribution with n degrees of freedom. But, the random variable

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \{(X_i - \mu) - (\bar{X} - \mu)\}^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 - \frac{1}{\sigma^2} \sum_{i=1}^{n} (\bar{X} - \mu)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

Sum of squares of n SNVs

Square of 1 SNV

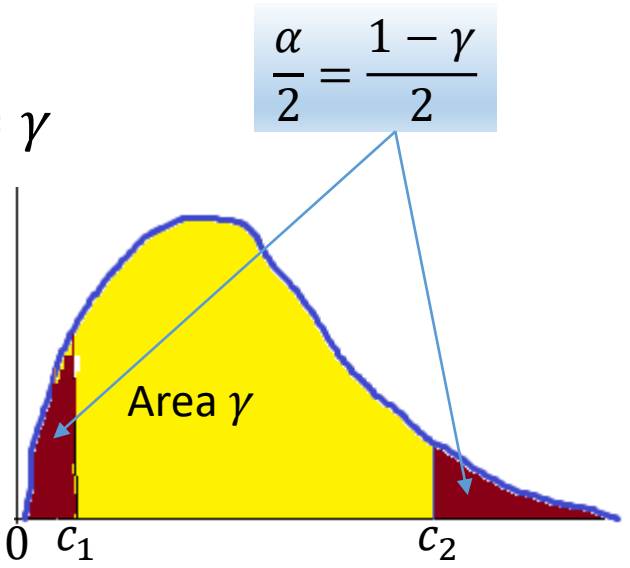
and has a χ^2-distribution with $n - 1$ degrees of freedom.

A χ^2-distribution is not a symmetric distribution like a $N(0,1)$ distribution;, it is a positively skewed distribution—a longer tail towards the right side and has the range the positive x-axis. If we need to set up a confidence interval with confidence coefficient γ using the random sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$, we need to find two points c_1 and c_2 from a χ^2-distribution with $n-1$ degrees of freedom such that

$$\Pr\{c_1 \le \chi^2 \le c_2\} = \gamma, i.e. F(c_1) - F(c_2) = \gamma$$

where,

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



If we set $\alpha = 1 - \gamma$, our choice of c_1 and c_2 are such that

$$\boldsymbol{F(c_1) = \frac{\alpha}{2} = \frac{1-\gamma}{2}, F(c_2) = 1 - \frac{\alpha}{2} = \frac{1+\gamma}{2}}$$

and the values of c_1 and c_2 from a **χ^2-distribution with $n-1$ degrees of freedom** can be obtained from **Table A10.**

Now, having obtained the values of c_1 and c_2 with the property that $\Pr\{c_1 \leq \chi^2 \leq c_2\} = \gamma$, we can replace χ^2 by $\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$ and then

$$\left\{c_1 \leq \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \leq c_2\right\} = \gamma$$

$$\Rightarrow \left\{c_1 \leq \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \leq c_2\right\} = \gamma$$

$$\Rightarrow \left\{\frac{1}{c_2} \leq \frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \leq \frac{1}{c_1}\right\} = \gamma$$

$$\Rightarrow \left\{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{c_2} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{c_1}\right\} = \gamma$$

Hence a $100\gamma\%$ confidence interval for the variance of a normal population is given by

$$\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{c_2}, \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{c_1}\right].$$

Observe that $\sum_{i=1}^{n}(X_i-\bar{X})^2=(n-1)S^2$ and the LCL and UCL can be written as

$$LCL=\frac{(n-1)S^2}{c_2}, UCL=\frac{(n-1)S^2}{c_1}.$$

The summary of the entire discussion is that to find a $100\gamma\%$ confidence interval for the variance of a normal population from a sample of size n, compute $\sum_{i=1}^{n}(X_i-\bar{X})^2$ and two constants from Chi-square table with $n-1$ d.f. such that $F(c_1)=\frac{\mathbf{1-\gamma}}{\mathbf{2}}$ and $F(c_2)=\frac{\mathbf{1+\gamma}}{\mathbf{2}}$ (Table A10), then calculate LCL and UCL by the above formula and the desired confidence interval is $[LCL, UCL]$. You are now done.

Table A10 Chi-square Distribution

Values of x for given values of the distribution function $F(z)$ (see Sec. 25.3 before (17)).
Example: For 3 degrees of freedom, $z = 11.34$ when $F(z) = 0.99$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.005	0.00	0.01	0.07	0.21	0.41	0.68	0.99	1.34	1.73	2.16
0.01	0.00	0.02	0.11	0.30	0.55	0.87	1.24	1.65	2.09	2.56
0.025	0.00	0.05	0.22	0.48	0.83	1.24	1.69	2.18	2.70	3.25
0.05	0.00	0.10	0.35	0.71	1.15	1.64	2.17	2.73	3.33	3.94
0.95	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31
0.975	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48
0.99	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21
0.995	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.005	2.60	3.07	3.57	4.07	4.60	5.14	5.70	6.26	6.84	7.43
0.01	3.05	3.57	4.11	4.66	5.23	5.81	6.41	7.01	7.63	8.26
0.025	3.82	4.40	5.01	5.63	6.26	6.91	7.56	8.23	8.91	9.59
0.05	4.57	5.23	5.89	6.57	7.26	7.96	8.67	9.39	10.12	10.85
0.95	19.68	21.03	22.36	23.68	25.00	26.30	27.59	28.87	30.14	31.41
0.975	21.92	23.34	24.74	26.12	27.49	28.85	30.19	31.53	32.85	34.17
0.99	24.72	26.22	27.69	29.14	30.58	32.00	33.41	34.81	36.19	37.57
0.995	26.76	28.30	29.82	31.32	32.80	34.27	35.72	37.16	38.58	40.00

$F(z)$	Number of Degrees of Freedom									
	21	22	23	24	25	26	27	28	29	30
0.005	8.0	8.6	9.3	9.9	10.5	11.2	11.8	12.5	13.1	13.8
0.01	8.9	9.5	10.2	10.9	11.5	12.2	12.9	13.6	14.3	15.0
0.025	10.3	11.0	11.7	12.4	13.1	13.8	14.6	15.3	16.0	16.8
0.05	11.6	12.3	13.1	13.8	14.6	15.4	16.2	16.9	17.7	18.5
0.95	32.7	33.9	35.2	36.4	37.7	38.9	40.1	41.3	42.6	43.8
0.975	35.5	36.8	38.1	39.4	40.6	41.9	43.2	44.5	45.7	47.0
0.99	38.9	40.3	41.6	43.0	44.3	45.6	47.0	48.3	49.6	50.9
0.995	41.4	42.8	44.2	45.6	46.9	48.3	49.6	51.0	52.3	53.7

$F(z)$	Number of Degrees of Freedom							
	40	50	60	70	80	90	100	> 100 (Approximation)
0.005	20.7	28.0	35.5	43.3	51.2	59.2	67.3	$\frac{1}{2}(h - 2.58)^2$
0.01	22.2	29.7	37.5	45.4	53.5	61.8	70.1	$\frac{1}{2}(h - 2.33)^2$
0.025	24.4	32.4	40.5	48.8	57.2	65.6	74.2	$\frac{1}{2}(h - 1.96)^2$
0.05	26.5	34.8	43.2	51.7	60.4	69.1	77.9	$\frac{1}{2}(h - 1.64)^2$
0.95	55.8	67.5	79.1	90.5	101.9	113.1	124.3	$\frac{1}{2}(h + 1.64)^2$
0.975	59.3	71.4	83.3	95.0	106.6	118.1	129.6	$\frac{1}{2}(h + 1.96)^2$
0.99	63.7	76.2	88.4	100.4	112.3	124.1	135.8	$\frac{1}{2}(h + 2.33)^2$
0.995	66.8	79.5	92.0	104.2	116.3	128.3	140.2	$\frac{1}{2}(h + 2.58)^2$

In the last column, $h = \sqrt{2m - 1}$, where m is the number of degrees of freedom.

Example 15: *A manufacturing unit of a company produces 56Ω resisters and it is known that the resistance of the resistors is not exactly equal to 56Ω, but normally distributed with some unknown μ mean and unknown variance σ^2. The resistance of a sample of 11 resisters are* $55.6, 57.9, 56.1, 58.2, 56.9, 55.3, 55.1, 57.6, 54.8, 56.1$ *and* 55.7 *Ohms. Find a 98% confidence interval for the variance of resisters produced by this unit.*

Ans. In this case,

$$\sum_{i=1}^{11} X_i = 619.3 \Rightarrow \bar{X} = \frac{1}{11}\sum_{i=1}^{11} X_i = 56.3\,, \sum_{i=1}^{11} X_i^2 = 34880.43$$

$$\sum_{i=1}^{11} (X_i - \bar{X})^2 = \sum_{i=1}^{11} X_i^2 - 11 \times \bar{X}^2 = 34880.43 - 11 \times 56.3^2 = 13.84.$$

Here $\gamma = 0.98$ and hence

$$\frac{1-\gamma}{2} = 0.01, \frac{1+\gamma}{2} = 0.99.$$

Degrees of freedom is $n - 1 = 11 - 1 = 10$. Moreover, c_1 and c_2 are such that

$F(c_1) = \frac{1-\gamma}{2} = 0.01, F(c_2) = \frac{1+\gamma}{2} = 0.99 \Rightarrow c_1 = 2.56$, $c_2 = 23.21$. Hence,

$$LCL = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{c_2} = \frac{13.84}{23.21} = 0.60,$$

$$UCL = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{c_1} = \frac{13.84}{2.56} = 5.41.$$

Hence, a 98% confidence interval for the population variance σ^2 is $[0.60, 5.41]$.

Example 16: *A recent survey shows that the heights of 8 B. Tech. boy students of NIT Rourkela in the year 2020 are* $69.1, 72.5, 61.7, 66.4, 71.5, 59.3, 67.2, 64.3$ *inches. Find a 99% confidence interval for the population variance of B. Tech. boys of NIT Rourkela in the year* 2020.

Ans. In this case $n = 8$,

$$\sum_{i=1}^{8} X_i = 532\,, \sum_{i=1}^{8} X_i^2 = 35525.98$$

$$\sum_{i=1}^{8} (X_i - \bar{X})^2 = \left[\sum_{i=1}^{8} X_i^2 - \frac{\left(\sum_{i=1}^{8} X_i\right)^2}{8}\right]$$

$$= 35525.98 - \frac{532^2}{8} = 147.98.$$

Here $\gamma = 0.99$ and hence

$$\frac{1-\gamma}{2} = 0.005, \frac{1+\gamma}{2} = 0.995.$$

Degrees of freedom is $n - 1 = 8 - 1 = 7$.

$F(c_1) = \frac{1-\gamma}{2} = 0.005, F(c_2) = \frac{1+\gamma}{2} = 0.995 \Rightarrow c_1 = 0.99$, $c_2 = 20.48$. Hence,

$$LCL = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{c_2} = \frac{147.98}{20.48} = 7.2256,$$

$$UCL = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{c_1} = \frac{147.98}{0.99} = 149.4747.$$

Hence, a 99% confidence interval for the population variance σ^2 is $[7.2256, 149.4747]$.

Example 17: *Find a* 95% *confidence interval for the population variance of length of bolts if the mean length of a sample of* 20 *bolts is* $20.2\ cm$ *and the sample variance is* $0.04\ cm^2$.

Ans. In this case $n = 20$, $\bar{X} = 20.2$ and $S^2 = 0.04$, $\gamma = 0.95$. Hence, the degrees of freedom is $n - 1 = 20 - 1 = 19$. Next, to find c_1 and c_2, we need to solve

$$F(c_1) = \frac{1-\gamma}{2} = 0.025, F(c_2) = \frac{1+\gamma}{2} = 0.975$$

and from Table A10, the values of c_1 and c_2 are c_1 =8.91, $c_2 = 32.85$. Hence,

$$LCL = \frac{(n-1)S^2}{c_2} = \frac{19 \times 0.04}{32.85} = 0.0231,$$

$$UCL = \frac{(n-1)S^2}{c_1} = \frac{19 \times 0.04}{8.91} = 0.0852,$$

and a 95% confidence interval for the population variance is $[0.0231, 0.0852]$.

Test for goodness of fit (Chi-square test)

Chi-Square test for goodness of fit is a non-parametric test used to test whether or not the observed values of a given phenomena is consistent with the expected value. The term goodness of fit is used to compare the observed sample frequency distribution with the expected or theoretical frequency distribution. For example, if it is known that male and female births are equally like. In a given locality, if the number of male and female child born during a given period is known, then it is important to test whether male and female births are equally likely in that locality or not. Here, the observed values are the actual number of male and female child born during the given period and the corresponding expected or theoretical values are half the total number of births each during the given period.

Let $O_1, O_2, \dots, O_n$ be the observed (experimental) values of a particular phenomenon, and theoretically these should be $E_1, E_2, \dots, E_n$ (may be due to some physical or probabilistic law). Here, we need to compare the consistency of these n observed valued with the corresponding expected values.

To do so, we need to calculate

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{n} \frac{O_i^2}{E_i} - N$$

where $N = \sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$ is the total number of observations. If the observed and the expected (theoretical) values do not differ significantly, then χ_0^2 remains small and is the amount error associated with the observed values. For a given level of significance α(Type I error, $\alpha = 1 - \gamma$) and with $n - 1$ degrees freedom there is a corresponding tabulated value of χ^2, say c which can be determined from Table A10, and which is the maximum permissible error. If the observed value of χ^2 is less than this tabulated values, then we conclude that the observed or experimental values agrees with the theoretical values, otherwise, we conclude that, the observed values are not consistent with the theoretical or expected values. Actually, we set a null hypothesis as Ho: The observed values are consistent with the expected values. Finally, we accept this hypothesis if $\chi_0^2 \leq c$, else we reject the null hypothesis Ho.

Example 18: A coin was tossed 50 times and 32 heads were obtained. Can you conclude that the coin is fair? *Test at* 5% *level of significance.*

Ans: We set up the null hypothesis that the coin is fair. Symbolically, Ho: The coin is fair. Under the assumption of this null hypothesis, the expected number of heads and tails should be 25 each. We next proceed to calculate χ_0^2.

Thus,

Outcome	O	E	$(O-E)^2/E$
Head	32	25	$49/25$
Tail	18	25	49/25

$$\chi_0^2 = \frac{49}{25} + \frac{49}{25} = 3.92.$$

The degree of freedom is $2-1=1$ and $c: F(c) = 1-\alpha = 0.95$. From table A10, $c = 3.84$. Since $\chi_0^2 > c$, the null hypothesis is rejected (at 5% level of significance) and we conclude that the coin is not fair. Observe that we we let $\alpha = 0.01$ (1%), then $c = 6.63$ and the hypothesis can be accepted. Hence, the level of significance is very important in a statistical testing of hypothesis.

Example 19: *The following figures show the distribution of digits in numbers chosen at random from a telephone directory*.

Didits	1	2	3	4	5	6	7	8	9	0	Total
Freq.	1107	997	966	1075	933	1107	972	964	853	1026	10,000

Test whether the digits occur equally frequently in the directory. Use 5% *level of significance .*

Ans: Null hypothesis Ho: the digits occurs equally frequently in the directory. Under the assumption of the null hypothesis, the expected number of times each number should appear is 1000. The following table is needed to calculate χ_0^2.

O	1107	997	966	1075	933	1107	972	964	853	1026
E	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
$(O-E)^2/E$	11.449	0.009	1.056	5.625	4.489	11.449	0.784	1.296	21.609	0.676

From the above table, it is clear that the observed value of Chi-square is

$$\chi_0^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = 58.542.$$

Degrees of freedom is $10 - 1 = 9$. Level of significance is $\alpha = 0.05$. The tabulated value of Chi-square corresponds to $c: F(c) = 1 - \alpha = 0.95$. From table A10, $c = 16.92$. Since, $\chi_0^2 > c$, the null hypothesis is rejected and we conclud that the digits 0,1,…,9 do not occur equally frequently in the directory.

Example 20: *The following table gives the number of car accidents on different days of a week in a city:*

Days	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
No. of acc.	24	26	18	22	21	19	24

Test if the accidents are uniformly distributed over the days of the week. Use $\alpha = 5\%$.

Ans:

YOU DO IT.

Example 21: *The theory predicts that the proportion of beans in the four groups A, B,* C *and D should be in the ratio* 9 : 3 : 3 : 1. *In an experiment with 1600 beans, the numbers in the four groups were* 882, 313, 287 *and* 118. *Does the experimental result support the theory? (Test at 1% level of significance)*

Ans: Ho: The experimental result support the theory. Assuming the null hypothesis, the expected number of beans in each group are 900,300,300 and 100. D.F. $= 4 - 1 = 3$. c: $F(c) = 1 - \alpha = 0.95$.

O	882	313	287	118
E	900	300	300	100

Rest you do.

Example 21: *A survey of 320 joint families with 5 children each revealed the following distribution. Is this result consistent with the hypothesis that male and female births are equally probable* ? Use $\alpha = 5\%$.

No. of boys	5	4	3	2	1	0
No. of girls	0	1	2	3	4	5
No. of families	14	56	110	88	40	12

Ans: We set up the null hypothesis that mail and female births are equally likely. Symbolically, Ho:$p = 1/2$, where p is the probability of a female birth. Under the null hypothesis, the probability of x girls in a family of 5 children is given by

$$f(x) = \binom{5}{x}\left(\frac{1}{2}\right)^{x}\left(1-\frac{1}{2}\right)^{5-x} = \binom{5}{x} \times \frac{1}{32}, x = 0,1,2,3,4,5.$$

Hence, $f(0) = \frac{1}{32}, f(1) = \frac{5}{32}, f(2) = \frac{10}{32}, f(3) = \frac{10}{32}, f(4) = \frac{5}{32}$ and $f(5) = \frac{1}{32}$.

Since, there are 320 joint families in total, the expected number of joint families with $0,1,2,3,4,5$ girls are $10, 50, 100, 100, 50, 10$ respectively. We can now proceed to calculate χ_0^2 as follows

O	14	56	110	88	40	12
E	10	50	100	100	50	10
$(O-E)^2/E$	1.60	0.72	1.00	1.44	2.00	0.40

Thus,

$$\chi_0^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 7.16.$$

Degrees of freedom is $6 - 1 = 5$. Level of significance is $\alpha = 0.05$. Thus,

$$c: F(c) = 1 - \alpha = 0.95 \Rightarrow c = 11.07.$$

Since, $\chi_0^2 < c$, the null hypothesis is accepted and we conclude that male and female births are equally probable.

Example 22: *A bag contains* 47Ω *and* 57Ω *radio resistors.* 50 *resistors are drawn at random without replacement. Find the minimum number of* 47Ω *resisters in the sample that rejects the hypothesis that* 47Ω *and* 57Ω *resistors are in the ratio* $1:1$*. Similarly, find the maximum numbers of* 47Ω *resisters in the sample that rejects the same hypothesis. Do it for both* $\alpha = 2\%$ *and* $\alpha = 5\%$*. Modify your answer if the hypothesis is* 47Ω *and* 57Ω *resistors are in the ratio* $3:2$*?*

Ans. Let x be the 47Ω resisters in the sample. Let the null hypothesis be Ho: The resistors are in the ratio $1:1$. Under the null hypothesis, we proceed to calculate χ_0^2 as follows:

Type of Reg.	$\boldsymbol{O}$	$\boldsymbol{E}$	$(\boldsymbol{O}-\boldsymbol{E})^2/\boldsymbol{E}$
47Ω	x	25	$(x-25)^2/25$
56Ω	$50-x$	25	$(25-x)^2/25$

Thus,

$$\chi_0^2 = \frac{2(25-x)^2}{25}.$$

Degrees of freedom is $2-1=1$. c: $F(c)=1-\alpha=0.95$. Hence, from Table A10, $c=3.84$. The null hypothesis is accepted if $\chi_0^2 < c$. But this happens if and only if

$$\chi_0^2 = \frac{2(x-25)^2}{25} < 3.84 \Leftrightarrow (x-25)^2 < 48$$

$$\Leftrightarrow (x-31.93)(x-18.07) < 0$$

$$\Leftrightarrow (x-31.93) > 0, (x-18.07) < 0,$$

$$or, (x-31.93) < 0, (x-18.07) > 0$$

$$\Leftrightarrow x > 31.93, x < 18.07 \;\; or, x < 31.93, x > 18.07$$

$$\Leftrightarrow x \leq 31, x \geq 19 \Leftrightarrow 19 \leq x \leq 31$$

Thus, the maximum numbers of 47Ω resisters in the sample that rejects the null hypothesis is 18 and the minimum numbers of 47Ω resisters in the sample that rejects the null hypothesis is 32. (Notice that is $x=31$, the hypothesis is accepted and if $x=19$, then also the hypothesis is accepted. The acceptance zone is $19 \leq x \leq 31$ and the rejection zone is $x \leq 18$ and $x \geq 32$.)