Maximum Likelihood Estimator of a Poisson Parameter

Suppose $X_1, \ldots, X_n$ are independent Poisson random variables each having mean λ. The likelihood function is given by

$$
\begin{aligned}
f(x_1, \ldots, x_n | \lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\
&= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{x_1! \ldots x_n!}
\end{aligned}
$$

Maximum Likelihood Estimator of a Poisson Parameter

Suppose $X_1, \ldots, X_n$ are independent Poisson random variables each having mean λ. The likelihood function is given by

$$\begin{aligned} f(x_1, \ldots, x_n|\lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1! \ldots x_n!} \end{aligned}$$

Thus

$$log\ f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_{i=1}^{n} x_i log(\lambda) - log(c)$$

where $c = \prod_{i=1}^n x_i!$.

Maximum Likelihood Estimator of a Poisson Parameter

Suppose $X_1, \ldots, X_n$ are independent Poisson random variables each having mean λ. The likelihood function is given by

$$\begin{aligned} f(x_1, \ldots, x_n|\lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{x_1! \ldots x_n!} \end{aligned}$$

Thus

$$log\ f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_{i=1}^{n} x_i log(\lambda) - log(c)$$

where $c = \prod_{i=1}^{n} x_i!$. Differentiating

$$\frac{d}{d\lambda} log\ f(x_1, \ldots, x_n|\lambda) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$

Therefore by equating to zero

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{or} \quad d(X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example

The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:

$$4, 0, 6, 5, 2, 1, 2, 0, 4, 3$$

Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

Example

The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:

$$4, 0, 6, 5, 2, 1, 2, 0, 4, 3$$

Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

Solution: Since there are a large number of drivers, each of whom has a small probability of being involved in an accident in a given day, it seems reasonable to assume that the daily number of traffic accidents is a Poisson random variable. Since

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i = 2.7$$

Example

The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:

$$4, 0, 6, 5, 2, 1, 2, 0, 4, 3$$

Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

Solution: Since there are a large number of drivers, each of whom has a small probability of being involved in an accident in a given day, it seems reasonable to assume that the daily number of traffic accidents is a Poisson random variable. Since

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 2.7$$

Since the long–run proportion of nonrainy days that have 2 or fewer accidents is equal to $P\{X \leq 2\}$, where X is the random number of accidents in a day, it follows that the desired estimate is

$$e^{-2.7}(1 + 2.7 + (2.7)^2/2) = 0.4936$$

Example 3: Hospital Emergency Arrivals

Scenario

A hospital wants to staff their ER appropriately.

Data (patients arriving per hour over 20 hours):

3	5	2	4	6	3	4	2	5	3
4	3	7	2	4	5	3	4	6	2

Model: Patient arrivals follow a Poisson distribution
MLE for Poisson:

$$\hat{\lambda} = \bar{x} = \frac{3 + 5 + 2 + \cdots + 2}{20} = \frac{78}{20} = 3.9 \text{ patients/hour}$$

Hospital Planning: Using the Estimate

With $\hat{\lambda} = 3.9$ patients/hour, we can answer:

Q1: Probability of very busy hour (8 patients)?
Using Poisson formula: $P(X \geq 8) \approx 0.056 = 5.6\%$

Q2: Probability of quiet hour (2 patients)?
$P(X \leq 2) \approx 0.269 = 26.9\%$

Staffing Decision

- Expect about 4 patients per hour on average
- About 1 in 20 hours will have 8+ patients (need backup staff)
- About 1 in 4 hours will be slow (2 patients)

Maximum Likelihood Estimator in a Normal Population

The joint density of n independent, normal random variables each with unknown mean μ and unknown standard deviation σ, $X_1, \ldots, X_n$ is given as

$$f(x_1,\ldots,x_n|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left[\frac{-(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sigma^n} exp\left[\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right]$$

The logarithm of the likelihood is thus given by

$$log\ f(x_1,\ldots,x_n|\mu,\sigma) = -\frac{n}{2}log(2\pi) - nlog(\sigma) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

In order to find the value of μ and σ maximizing the foregoing, we compute

$$\frac{\partial}{\partial\mu} log\ f(x_1,\ldots,x_n|\mu,\sigma) = \frac{\sum_{i=1}(x_i-\mu)}{\sigma^2}$$

$$\frac{\partial}{\partial\sigma} log\ f(x_1,\ldots,x_n|\mu,\sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}$$

Equating to zero,

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

$$\hat{\sigma} = \left[\frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n} \right]^{1/2}$$

Hence, the maximum likelihood estimators of μ and σ are given, respectively, by

$$\bar{X} \quad \text{and} \quad \left[\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \right]^{1/2}$$

Example

Kolmogorov's law of fragmentation states that the size of an individual particle in a large collection of particles resulting from the fragmentation of a mineral compound will have an approximate lognormal distribution, where a random variable X is said to have a **lognormal** distribution if $log(X)$ has a normal distribution. The law, which was first noted empirically and then later given a theoretical basis by Kolmogorov, has been applied to a variety of engineering studies.
Suppose that a sample of 10 grains of metallic sand taken from a large sand pile have respective lengths (in millimeters):

$$2.2, 3.4, 1.6, 0.8, 2.7, 3.3, 1.6, 2.8, 2.5, 1.9$$

Estimate the percentage of sand grains in the entire pile whose length is between 2 and 3 mm.

Example

Kolmogorov's law of fragmentation states that the size of an individual particle in a large collection of particles resulting from the fragmentation of a mineral compound will have an approximate lognormal distribution, where a random variable X is said to have a **lognormal** distribution if $log(X)$ has a normal distribution. The law, which was first noted empirically and then later given a theoretical basis by Kolmogorov, has been applied to a variety of engineering studies.
Suppose that a sample of 10 grains of metallic sand taken from a large sand pile have respective lengths (in millimeters):

$$2.2, 3.4, 1.6, 0.8, 2.7, 3.3, 1.6, 2.8, 2.5, 1.9$$

Estimate the percentage of sand grains in the entire pile whose length is between 2 and 3 mm.

Solution: Taking the natural logarithm of these 10 data values, the following transformed data set results

$$.7885, 1.2238, .4700, ?.2231, .9933, 1.1939, .4700, 1.0296, .9163, .6419$$

Because the sample mean and sample standard deviation of these data are

$$\bar{x} = 0.7504, \quad s = 0.4531$$

Tt follows that the logarithm of the length of a randomly chosen grain has a normal distribution with mean approximately equal to 0.7504 and with standard deviation approximately equal to 0.4351. Hence, if X is the length of the grain, then

$$\begin{aligned}
P\{2 < X < 3\} &= P\{log(2) < log(X) < log(3)\} \\
&= P\left\{\frac{log(2) - 0.7504}{0.4531} < \frac{log(X) - 0.7504}{0.4531} < \frac{log(3) - 0.7504}{0.4531}\right\} \\
&= P\left\{-0.1316 < \frac{log(X) - 0.7504}{0.4531} < 0.8003\right\} \\
&\approx \Phi(0.8003) - \Phi(-0.1316) \\
&\approx 0.3405
\end{aligned}$$

What is a Confidence Interval?

Definition

A confidence interval provides a range of plausible values for an unknown population parameter with an associated confidence level.

Key Concepts:

- Point estimate gives single value
- CI quantifies uncertainty in estimate
- Confidence level: long-run proportion of intervals containing true parameter

Interpretation

95% CI: If we repeated sampling many times, 95% of constructed intervals would contain the true parameter

Interval Estimates

Sometimes it is more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within. To obtain such an interval estimator, we make use of the probability distribution of the point estimator.
since the point estimator $\bar{X}$ is normal with mean μ and variance σ^2/n,it follows that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$$

has a standard normal distribution.

Interval Estimates

Sometimes it is more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within. To obtain such an interval estimator, we make use of the probability distribution of the point estimator.
since the point estimator $\bar{X}$ is normal with mean μ and variance σ^2/n,it follows that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$$

has a standard normal distribution. Therefore,

$$P\left\{-1.96 < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < 1.96\right\} = 0.95$$

$$P\left\{-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

or

Interval Estimates

Sometimes it is more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within. To obtain such an interval estimator, we make use of the probability distribution of the point estimator.
since the point estimator $\bar{X}$ is normal with mean μ and variance σ^2/n,it follows that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$$

has a standard normal distribution. Therefore,

$$P\left\{-1.96 < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < 1.96\right\} = 0.95$$

$$P\left\{-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

or

$$P\left\{\bar{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

That is, with 95 percent confidence we assert that the true mean lies within $1.96\sigma/\sqrt{n}$ of the observed sample meant. The interval

$$\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}}, \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

is called a 95 percent confidence interval estimate of μ.

Example

Suppose that when a signal having value μ is transmitted from location A the value received at location B is normally distributed with mean μ and variance 4. That is, if μ is sent, then the value received is $\mu + N$ where N , representing noise, is normal with mean 0 and variance 4. To reduce error, suppose the same value is sent 9 times. If the successive values received are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, let us construct a 95 percent confidence interval for *mu*

Example

Suppose that when a signal having value μ is transmitted from location A the value received at location B is normally distributed with mean μ and variance 4. That is, if μ is sent, then the value received is $\mu + N$ where N , representing noise, is normal with mean 0 and variance 4. To reduce error, suppose the same value is sent 9 times. If the successive values received are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, let us construct a 95 percent confidence interval for *mu*

Solution:

$$\bar{x} = \frac{81}{9} = 9$$

It follows, under the assumption that the values received are independent, that a 95 percent confidence interval for μ is

$$\left(9?1.96\frac{\sigma}{3}, 9 + 1.96\frac{\sigma}{3}\right) = (7.69, 10.31)$$

Hence, we are 95 percent confident that the true message value lies between 7.69 and 10.31. The interval is called a two-sided confidence interval

Sometimes, however, we are interested in determining a value so that we can assert with, say, 95 percent confidence, that μ is at least as large as that value.
To determine such a value, note that if Z is a standard normal random variable then

$$P\{Z < 1.645\} = 0.95$$

As a result,

$$P\left\{\sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} < 1.645\right\} = 0.95$$

$$P\left\{\bar{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu\right\} = 0.95$$

Confidence intervals of any specified level of confidence

$$P\{Z > z_\alpha\} = \alpha$$

when Z is a standard normal random variable. This implies for any α

$$P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha$$

As a result

$$P\left\{-z_{\alpha/2} < \sqrt{n}\frac{\bar{X} - \mu}{\sigma} < z_{\alpha/2}\right\} = 1 - \alpha$$

$$P\left\{-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu - \bar{X} < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

Hence

$$P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

Hence, a $100(1-\alpha)$ percent two-sided confidence interval for μ is

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

Similarly, knowing that $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ is a standard normal random variable results in one-sided confidence intervals of any desired level of confidence.

One-sided upper confidence interval – $100(1-\alpha)\%$

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \infty\right)$$

One-sided Lower confidence interval – $100(1-\alpha)\%$

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

Sometimes, however, we are interested in determining a value so that we can assert with, say, 95 percent confidence, that μ is at least as large as that value.
To determine such a value, note that if Z is a standard normal random variable then

$$P\{Z < 1.645\} = 0.95$$

As a result,

$$P\left\{\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} < 1.645\right\} = 0.95$$

$$P\left\{\bar{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu\right\} = 0.95$$

Thus a 95 percent one-sided upper confidence interval for μ is

$$\left(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right)$$

A one-sided lower confidence interval is

$$\left(\infty, \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}\right)$$

Example: Salmon Hatchery

Original Problem

Salmon weights are normal with varying seasonal mean but fixed $\sigma = 0.3$ pounds. Want estimate within ± 0.1 pounds with 95% confidence.

Solution: Need $1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq 0.1$

$$1.96 \cdot \frac{0.3}{\sqrt{n}} \leq 0.1$$

$$\frac{0.588}{\sqrt{n}} \leq 0.1$$

$$\sqrt{n} \geq 5.88$$

$$n \geq 34.57$$

Sample Size Required

Need at least 35 salmon for desired precision.

Sample Size Formula Derivation

Given Requirements

- Margin of error: E
- Confidence level: $(1 - \alpha)$
- Population standard deviation: σ

Since margin of error $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Important

Always round UP to ensure adequate precision!

Trade-off: Higher precision (smaller E) requires larger n (higher cost)

Medical Research Planning

Blood Pressure Study

Estimate mean systolic BP within ± 2 mmHg with 95% confidence.
Historical $\sigma = 12$ mmHg.

Calculation:

$$n \geq \left(\frac{1.96 \times 12}{2}\right)^2$$
$$= (11.76)^2$$
$$= 138.3$$

Required: 139 participants

Cost-Benefit Analysis

- Cost per participant: \$100
- Total study cost: \$13,900
- Alternative: $E = \pm 3$ mmHg needs only 62 participants (\$6,200)
- Decision depends on required precision vs. budget

Real-World Application: Quality Control

Battery Manufacturing

A battery plant knows $\sigma = 20$ hours from years of data. Testing 64 batteries: $\bar{x} = 485$ hours. Find 95% CI.

Calculation:

$$\begin{aligned} CI &= 485 \pm 1.96 \cdot \frac{20}{\sqrt{64}} \\ &= 485 \pm 1.96 \cdot 2.5 \\ &= 485 \pm 4.9 \\ &= (480.1, 489.9) \text{ hours} \end{aligned}$$

Business Impact

- Warranty setting: Can guarantee 480 hours
- Quality assurance: Process meets 475-hour standard
- Marketing claim: "Lasts over 480 hours"

Example

Determine the upper and lower 95 percent confidence interval estimates of μ in previous example

Example

Determine the upper and lower 95 percent confidence interval estimates of μ in previous example

Solution:

$$1.645 \frac{\sigma}{\sqrt{n}} = 1.097$$

the 95 percent upper confidence interval is

$$(9 - 1.097, \infty) = (7.903, \infty)$$

and the 95 percent lower confidence interval is

$$(-\infty, 9 + 1.097) = (-\infty, 10.097)$$

Example

From past experience it is known that the weights of salmon grown at a commercial hatchery are normal with a mean that varies from season to season but with a standard deviation that remains fixed at 0.3 pounds. If we want to be 95 percent certain that our estimate of the present seasons mean weight of a salmon is correct to within ± 0.1 pounds, how large a sample is needed?

Example

From past experience it is known that the weights of salmon grown at a commercial hatchery are normal with a mean that varies from season to season but with a standard deviation that remains fixed at 0.3 pounds. If we want to be 95 percent certain that our estimate of the present seasons mean weight of a salmon is correct to within ± 0.1 pounds, how large a sample is needed?

Solution: A 95 percent confidence interval estimate for the unknown mean μ, based on a sample of size n, is

$$\mu \in \left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Because the estimate $\bar{x}$ is within $1.96(\sigma/\sqrt{n}) = 0.588/\sqrt{n}$ of any point in the interval, it follows that we can be 95 percent that $\bar{x}$ is within 0.1 of μ provided that

$$\frac{0.588}{\sqrt{n}} \leq 0.1$$

That is

$$\sqrt{n} \geq 5.88$$

or

$$n \geq 34.57$$

Hence sample size of 35 or larger will suffice.

Confidence Interval for a Normal Mean When the Variance Is Unknown

Suppose now that $X_1, \ldots, X_n$ is a sample from a normal distribution with unknown mean μ and unknown variance σ^2, and that we wish to construct a $100(1-\alpha)$ percent confidence interval for μ.

We no longer can assume the fact that $\sqrt{n}(\bar{X}-\mu)/\sigma$ is a standard normal variable since σ is unknown. But we can compute the sample variance.

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{2} n - 1$$

And it is known that

$$\sqrt{n}\frac{\bar{X}-\mu}{S}$$

is a t–random variable with $n-1$ degrees of freedom. Hence , we have that for any $\alpha \in (0, 0.5)$

$$P\left\{-t_{\alpha/2,n-1} < \sqrt{n}\frac{\bar{X}-\mu}{S} < t_{\alpha/2,n-1}\right\} = 1-\alpha$$

$$P\left\{\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right\} = 1-\alpha$$

Thus, if $\bar{X} = \bar{x}$ and $S = s$, we can say with $100(1-\alpha)\%$ confidence

$$\mu \in \left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

A one–sided upper confidence interval can be obtained by noting that

$$P\left\{\sqrt{n}\frac{\bar{X}-\mu}{S} < t_{\alpha,n-1}\right\} = 1-\alpha$$

or

$$P\left\{\mu > \bar{X} - \frac{S}{\sqrt{n}}t_{\alpha,n-1}\right\} = 1-\alpha$$

Hence if $\bar{X} = \bar{x}$ and $S = s$, then we can assert with $100(1-\alpha)\%$ confidence that upper confidence interval is,

$$\mu \in \left(\bar{x} - \frac{s}{\sqrt{n}}t_{\alpha,n-1}, \infty\right)$$

similarly lower confidence interval would be

$$\mu \in \left(-\infty, \bar{x} - \frac{s}{\sqrt{n}}t_{\alpha,n-1}\right)$$

Example

When the value μ is transmitted at location A then the value received at location B is normal with mean μ and variance σ^2 but with σ^2 being unknown. If 9 successive values are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, compute a 95 percent confidence interval for μ.

Example

When the value μ is transmitted at location A then the value received at location B is normal with mean μ and variance σ^2 but with σ^2 being unknown. If 9 successive values are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, compute a 95 percent confidence interval for μ. **Solution:**

$$\bar{x} = 9$$

$$s^2 = \frac{\sum x_i^2 - 9(\bar{x})^2}{8} = 9.5$$

$$s = 3.082$$

Hence, as $t_{0.025,8} = 2.306$, a 95 percent confidence interval for μ is

$$\left[9 - 2.306\frac{3.082}{3}, 9 + 2.306\frac{3.082}{3}\right] = (6.63, 11.37)$$

Confidence Intervals for the Variance of a Normal Distribution

If $X_1, \ldots, X_n$ is a sample from a normal distribution having unknown parameters μ and σ^2, then we can construct a confidence interval for σ^2 by using the fact that

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hence,

$$P\left\{\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha$$

or

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\} = 1-\alpha$$

Hence when $S^2 = s^2$, a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left\{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right\}$$

Example

A standardized procedure is expected to produce washers with very small deviation in their thicknesses. Suppose that 10 such washers were chosen and measured. If the thicknesses of these washers were, in inches,

$$0.123, 0.133, 0.124, 0.125, 0.126, 0.128, 0.120, 0.124, 0.130 \text{ and } 0.126$$

what is a 90 percent confidence interval for the standard deviation of the thickness of a washer produced by this procedure?

Example

A standardized procedure is expected to produce washers with very small deviation in their thicknesses. Suppose that 10 such washers were chosen and measured. If the thicknesses of these washers were, in inches,

$$0.123, 0.133, 0.124, 0.125, 0.126, 0.128, 0.120, 0.124, 0.130 \text{ and } 0.126$$

what is a 90 percent confidence interval for the standard deviation of the thickness of a washer produced by this procedure?

Solution:

$$S^2 = 1.366 \times 10^{-5}$$

Since, $\chi^2_{0.05,9} = 16.917$ and $\chi^2_{0.95,9} = 3.334$

$$\frac{9 \times 1.366 \times 10^{-5}}{16.917} = 7.267 \times 10^{-6}, \qquad \frac{9 \times 1.366 \times 10^{-5}}{3.334} = 36.875 \times 10^{-6}$$

with confidence of 90 percent

$$\sigma^2 \in (7.267 \times 10^{-6}, 36.875 \times 10^{-6})$$

$$\sigma \in (2.696 \times 10^{-3}, 6.072 \times 10^{-3})$$

$100(1-\alpha)\%$ Confidence Intervals

$$X_1, \ldots, X_n \in \mathcal{N}(\mu, \sigma^2)$$

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$$

Assumption	Parameter	Confidence Interval	Lower Interval	Upper Interval
σ^2 known	μ	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\left(-\infty, \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}\right)$	$\left(\bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \infty\right)$
σ^2 unknown	μ	$\bar{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$	$\left(-\infty, \bar{X} + t_{\alpha, n-1} \frac{S}{\sqrt{n}}\right)$	$\left(\bar{X} - t_{\alpha, n-1} \frac{S}{\sqrt{n}}, \infty\right)$
μ unknown	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}\right)$	$\left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}\right)$	$\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}, \infty\right)$

Estimating the difference in means of two normal populations

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a normal population having mean μ_1 and variance σ_1^2 and let $Y_1, \ldots, Y_m$ be a sample of size m from a different normal population having mean μ_2 and variance σ_2^2 and suppose that the two samples are independent of each other. We are interested in estimating $\mu_1 - \mu_2$.

Estimating the difference in means of two normal populations

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a normal population having mean μ_1 and variance σ_1^2 and let $Y_1, \ldots, Y_m$ be a sample of size m from a different normal population having mean μ_2 and variance σ_2^2 and suppose that the two samples are independent of each other. We are interested in estimating $\mu_1 - \mu_2$.

To obtain a confidence interval estimator, we need the distribution of $\bar{X} - \bar{Y}$. Because

$$\bar{X} \sim \mathcal{N}(\mu_1, \sigma_1^2/n)$$

$$\bar{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2/m)$$

it follows from the fact that the sum of independent normal random variables is also normal, that

$$\bar{X} - \bar{Y} \sim \left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Hence assuming σ_1^2 and σ_2^2 are known

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1)$$

Hence

$$P\left\{-z_{\alpha/2} < \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} < z_{\alpha/2}\right\} = 1-\alpha$$

equivalently

$$P\left\{\bar{X}-\bar{Y}-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}} < (\mu_1-\mu_2) < \bar{X}-\bar{Y}+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right\} = 1-\alpha$$

Therefore

$$\mu_1-\mu_2 \in \left(\bar{x}-\bar{y}-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}, \bar{x}-\bar{y}+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right)$$

One sided interval can be given as

$$\mu_1-\mu_2 \in \left(-\infty, \bar{x}-\bar{y}+z_{\alpha}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right)$$

If the population variances are unknown, the sample variances need to be used

$$S_1^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

$$S_2^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{m-1}$$

If the population variances are unknown, the sample variances need to be used

$$S_1^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

$$S_2^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{m-1}$$

Assuming $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$(n-1)\frac{S_1^2}{\sigma^2} \sim \chi^2_{n-1}$$

and

$$(m-1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{m-1}$$

If the population variances are unknown, the sample variances need to be used

$$S_1^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

$$S_2^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{m-1}$$

Assuming $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$(n-1)\frac{S_1^2}{\sigma^2} \sim \chi^2_{n-1}$$

and

$$(m-1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{m-1}$$

It follows that,

$$(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

Since

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

we see that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0, 1)$$

Since

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

we see that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0, 1)$$

Now it follows from the fundamental result that in normal sampling $\bar{X}$ and S^2 are independent that $\bar{X}, S_1^2, \bar{Y}, S_2^2$ are independent random variables. Hence, using the definition of a t–random variable (as the ratio of two independent random variables, the numerator being a standard normal and the denominator being the square root of a chi-square random variable divided by its degree of freedom parameter), it follows that if we let

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

then

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \Big/ \sqrt{\frac{S_p^2}{\sigma^2}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

has a t–distribution with $n + m - 2$ degrees of freedom.

Consequently,

$$P\left\{-t_{\alpha/2,n+m-2} \leq \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} \leq t_{\alpha/2,n+m-2}\right\} = 1-\alpha$$

we obtain the following $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$: