Dec 8, 2022

Engineering Statistics

Week 7: Confidence intervals

Reminder

An estimator of a population parameter is a random variable that depends on the sample information; its value provides approximations of this unknown parameter. A specific value of that random variable is called an estimate.

Parameter	Estimator (Statistic)	Estimate
μ	$\bar{X}$	$\bar{x}$

Confidence interval estimation

Confidence interval estimator

A confidence interval estimator for a population parameter is a rule for determining (based on sample information) an interval that is likely to include the parameter. The corresponding estimate is called a **confidence interval** estimate.

Confidence interval and confidence level

- Let θ be an unknown parameter. Suppose that on the basis of sample information, random variables A and B are found such that $P(A < \theta < B) = 1 - \alpha$ where α is any number between 0 and 1.
- If the specific sample values of A and B are a and b, then the interval from a to b is called a $100(1 - \alpha)\,\%$ confidence interval of θ.
- The quantity $100(1 - \alpha)\,\%$ is called the confidence level of the interval.

Confidence interval and confidence level

If the population is repeatedly sampled a very large number of times, the true value of the parameter θ will be covered by $100(1-\alpha)\,\%$ of intervals calculated in this way. The confidence interval calculated in this manner is written as,

$$a < \theta < b$$

with $100(1-\alpha)\,\%$ confidence.

Confidence interval for population mean μ

Interval based on the normal distribution

Assumptions

(i) $X_1, X_2, \ldots, X_n$ is a random sample drawn from a normal distribution with mean μ and variance σ^2.

(ii) Population mean μ is unknown but population variance σ^2 is known.

Reminder

Let random variables $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution with mean μ and variance σ^2. Then, sampling distribution of the $\bar{X}$ is normal with mean μ and variance $\frac{\sigma^2}{n}$,I.e.,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Interval based on the normal distribution

$100(1-\alpha)\,\%$ confidence interval of μ

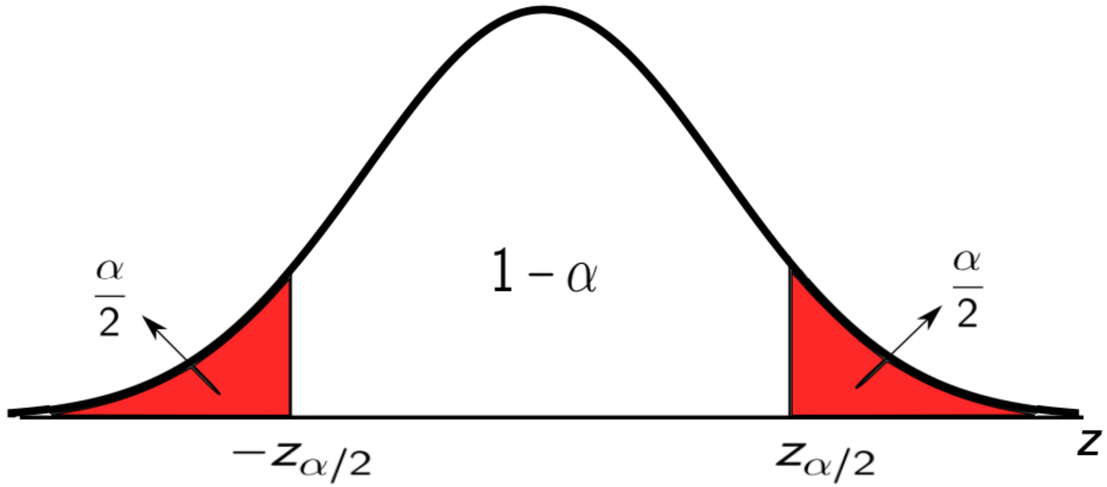
$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the value from the standard normal distribution such that the upper percentile (upper probability) is $\alpha/2$.

Interval based on the normal distribution

$z_{\alpha/2}$ = upper $\alpha/2$ point of standard normal distribution

$$Z \sim N(0,1) \rightarrow P(Z > z_{\alpha/2}) = \alpha/2 \quad \text{or} \quad P(|Z| < z_{\alpha/2}) = 1 - \alpha$$



Interval based on the normal distribution

Selected confidence levels and corresponding values of $z_{\alpha/2}$

$100(1-\alpha)\,\%$	$z_{\alpha/2}$
$90\,\%$	1.64
$95\,\%$	1.96
$99\,\%$	2.58

Example

The daily carbon monoxide (CO) emission from a large production plant will be measured on 25 randomly selected weekdays. The production process is always being modified and the current mean value of daily CO emissions μ is unknown. Data collected over several years confirm that, for each year, the distribution of CO emission is normal with a standard deviation of 0.8 ton. Suppose the sample mean is found to be $\bar{x} = 2.7$ tons.

Construct a $95\,\%$ confidence interval for the current daily mean emission μ.

Example

$\bar{x} = 2.7$ tons

$\sigma = 0.8$ tons

$n = 25$ days

The $95\ \%$ confidence interval for the current daily mean emission μ:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$2.7 - 1.96\frac{0.8}{\sqrt{25}} < \mu < 2.7 + 1.96\frac{0.8}{\sqrt{25}}$$

$$2.3864 < \mu < 3.0136$$

Note

If random samples of n observations are drawn repeatedly and independently from the population and $100(1-\alpha)\,\%$ confidence intervals are calculated, then over a very large number of repeated trials, $100(1-\alpha)\,\%$ of these intervals will contain the true value of the population mean.

Example

How changes the length of confidence interval for the population mean,

(i) if the population standard deviation (variance) gets larger for given $100(1-\alpha)\,\%$ and sample size n?

(ii) if sample size n gets larger for given $100(1-\alpha)\,\%$ and population standard deviation (variance)?

(iii) if $100(1-\alpha)\,\%$ gets larger for given sample size and population standard deviation (variance)?

Example

How changes the length of confidence interval for the population mean,

(i) if the population standard deviation (variance) gets larger for given $100(1-\alpha)\,\%$ and sample size n? The confidence interval gets larger

(ii) if sample size n gets larger for given $100(1-\alpha)\,\%$ and population standard deviation (variance)? The confidence interval gets narrower

(iii) if $100(1-\alpha)\,\%$ gets larger for given sample size and population standard deviation (variance)? The confidence interval gets larger

Interval based on the normal distribution

Assumptions

(i) $X_1, X_2, \ldots, X_n$ is a random sample drawn from a normal distribution with mean μ and variance σ^2.

(ii) Both population mean μ and population variance σ^2 are unknown.

We want to a $100(1-\alpha)\,\%$ confidence interval of the population mean μ.

Interval based on the normal distribution

Since population variance σ^2 is unknown, we cannot use the following fact here:

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

An intuitive approach is to replace σ by its estimator s in this formula.

Student's t distribution

If $X_1, X_2, \ldots, X_n$ be sample drawn from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Then, the statistic defined as,

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

has Student's t distribution with degrees of freedom $n - 1$.

Student's t distribution

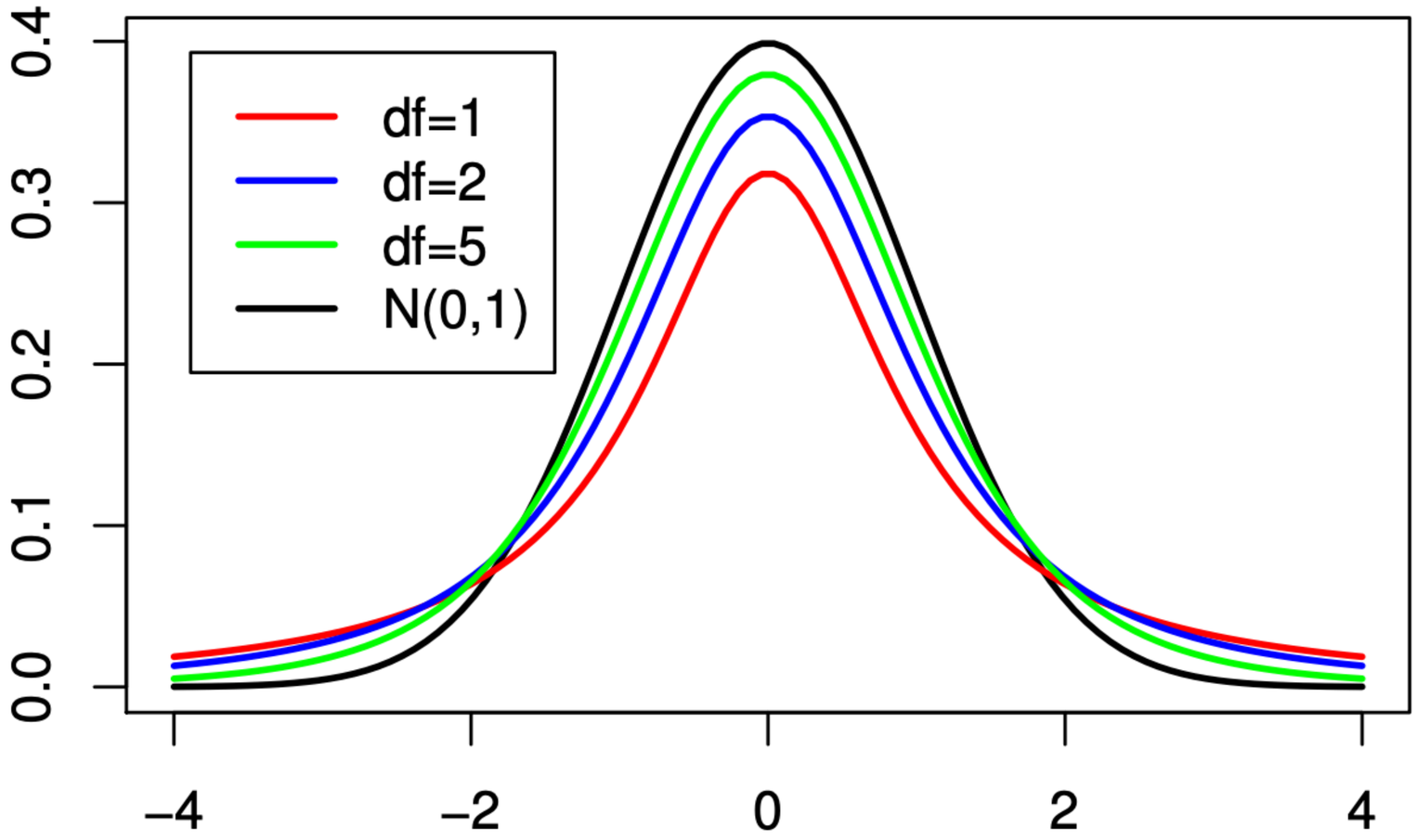
Random variable T having Student's t distribution with degrees of freedom $n-1$ shortly denoted by,

$$T \sim t_{(n-1)}$$

Student’s t distribution

- A specific member of the family of Student’s t distributions is characterized by the number of degrees of freedom associated with the computation of the standard error.
- The shape of the Student’s t distribution is rather similar to that of the standard normal distribution. Both distributions have mean 0, and the probability density functions of both are symmetric about their means.
- However, the density function of the Student’s t distribution has a wider dispersion (reflected in a larger variance) than the standard normal distribution.

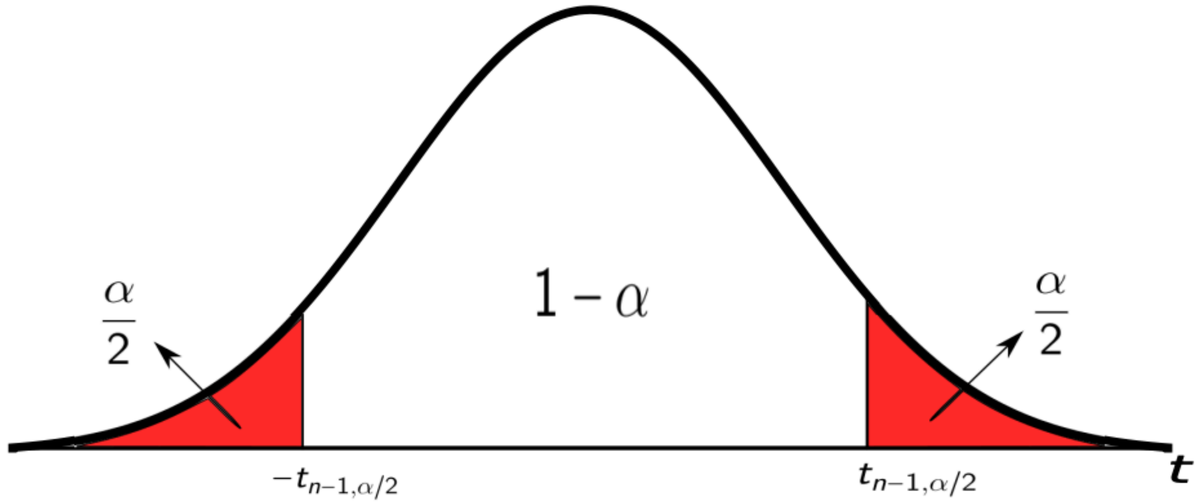
Student's t distribution



Confidence interval

$t_{n-1,\alpha/2}$ = upper $\alpha/2$ point of Student's t distribution

$$T \sim t_{(n-1)} \rightarrow P(T > t_{(n-1,\alpha/2)}) = \alpha/2 \quad \text{or} \quad P(|T| < t_{(n-1,\alpha/2)}) = 1 - \alpha$$



Confidence interval

Suppose that $X_1, X_2, \ldots, X_n$ be sample drawn from $N(\mu, \sigma^2)$ where both μ and σ^2 is unknown.

$100(1-\alpha)\,\%$ confidence interval of μ:

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

where $t_{(n-1,\alpha/2)}$ is the value from Student's t distribution with degrees of freedom $n-1$ such that the upper percentile (upper probability) is $\alpha/2$.

Example

Recently gasoline prices rose drastically. Suppose that a study was conducted using truck drivers with equivalent years of experience to test run 24 trucks of a particular model over the same highway. Estimate the population mean fuel consumption for this truck model with 90% confidence if the fuel consumption, in miles per gallon, for these 24 trucks was as follows:

15.5	21.0	16.5	19.2	18.6	19.1	18.5	19.3
19.7	16.9	18.7	18.2	18.0	17.5	19.8	18.0
19.8	18.2	20.2	14.5	18.5	20.5	20.3	21.8

Example

$n = 24$ trucks

$\bar{x} = 18.68$

$s = 1.70$

No information about the population variance -> means that “use t-distribution”

90% confidence interval for mean fuel consumption:

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

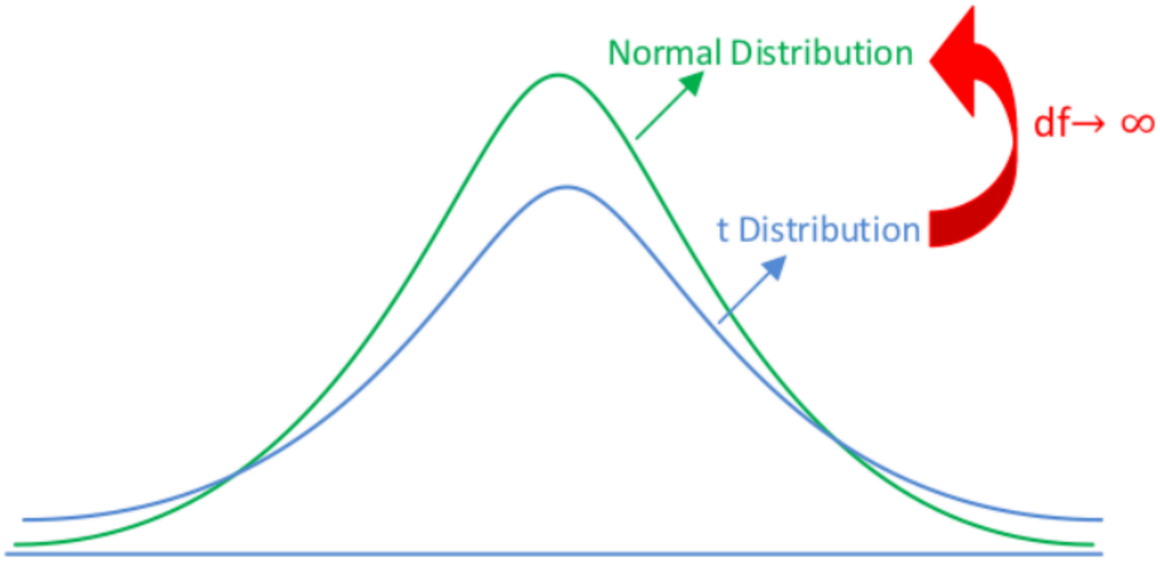
$$18.68 - t_{(23,10/2)}\frac{1.7}{\sqrt{24}} < \mu < 18.68 + t_{(23,10/2)}\frac{1.70}{\sqrt{24}}$$

$$18.68 - 2.069\frac{1.7}{\sqrt{24}} < \mu < 18.68 + 2.069\frac{1.70}{\sqrt{24}}$$

$$17.96 < \mu < 19.39$$

Note

- For larger values of the degrees of freedom, the Student's t distribution approaches to normal distribution.
- In other words, if $n > 30$ then Z table can be used instead of t table for computing the confidence interval of population mean μ.



Note

If n is large enough, the confidence interval for μ can be given as

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

even the population variance is unknown.

Example

How much do students pay, on the average, for textbooks during the first semester of college? From a random sample of 400 students the mean cost was found to be $375.75, and the standard deviation was $37.89. Assuming that the population is normally distributed,

(a) Find a 95% confidence interval for the population mean.

(b) Find a 99% confidence interval for the population mean.

(c) Compare the results obtained in (a) and (b).

(d) Without doing the calculations, state whether an 80% confidence interval for the population mean would be wider than, narrower than, or the same as the answer to part (a).

Example

$n = 400$

$\bar{x} = 375.75$ USD

$s = 37.89$ USD

No information about the population variance -> means that “use t-distribution”, but we can use Z-distribution also because of n > 30

(a) The 95% confidence interval for the population mean:

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$375.75 - 1.96\frac{37.89}{\sqrt{400}} < \mu < 375.75 + 1.96\frac{37.89}{\sqrt{400}}$$

$$372.0367 < \mu < 379.4632$$

Example

(b) The 99% confidence interval for the population mean:

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$375.75 - 2.58\frac{37.89}{\sqrt{400}} < \mu < 375.75 + 2.58\frac{37.89}{\sqrt{400}}$$

$$370.8621 < \mu < 380.6378$$

Example

(c) The CI in (a) is narrower than the CI in (b) because of the confidence level.

The CI of population mean in (a) is $(372.0367, 379.4632)$

The CI of population mean in (b) is $(370.8621, 380.6378)$

Example

(c) Without doing the calculations, state whether an 80% confidence interval for the population mean would be wider than, narrower than, or the same as the answer to part (a).

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

The Z value for 80% will be lower than the Z for 95% confidence level. This means that the CI for 80% confidence level will be narrower.

Confidence interval for population variance σ^2

Intervals based on the normal distribution

Assumptions:

(i) $X_1, X_2, \ldots, X_n$ be a sample drawn from $N(0,\sigma^2)$ distribution.

(ii) Population variance σ^2 is unknown.

We want a $100(1-\alpha)\,\%$ confidence interval of the population variance σ^2.

Chi-square distribution

If $X_1, X_2, \ldots, X_n$ be a sample drawn from normal distribution having variance σ^2. Then,

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

has chi-square distribution with degrees of freedom $n - 1$.

Chi-square distribution

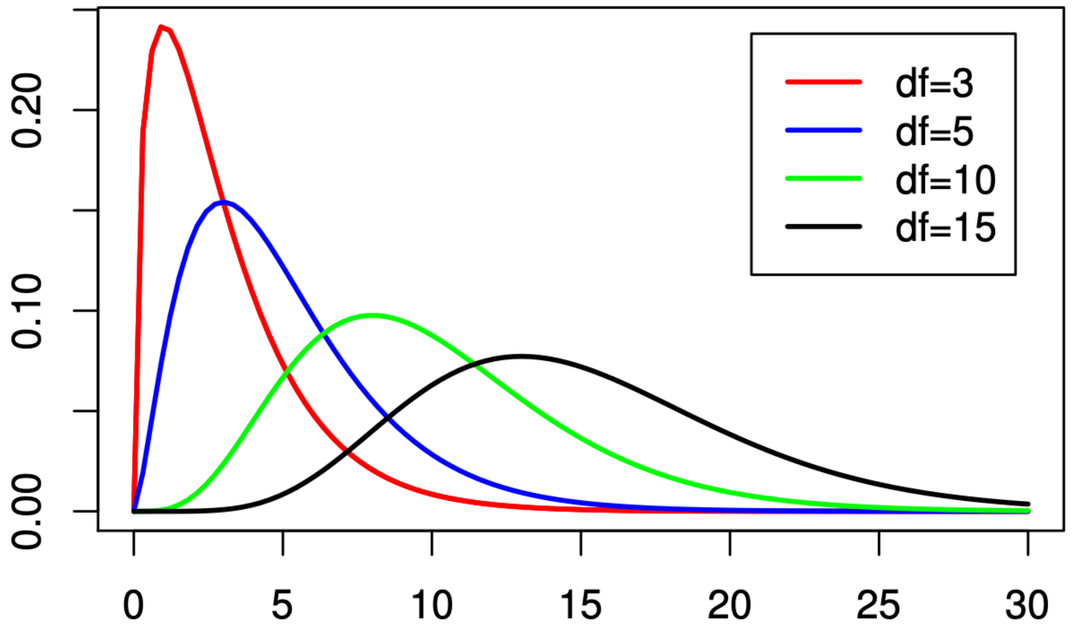
Random variable χ^2 having chi-square distribution with degrees of freedom $n-1$ is shortly denoted by

$$\chi^2 \sim \chi^2_{n-1}$$

Chi-square distribution

- The chi-square distribution, like the t distribution, is characterized by a quantity called the degrees of freedom associated with the distribution. In other words, the parameter of the chi-square distribution is its degree of freedom.

- Unlike normal and t distributions, the chi-square distribution is not symmetric around 0.

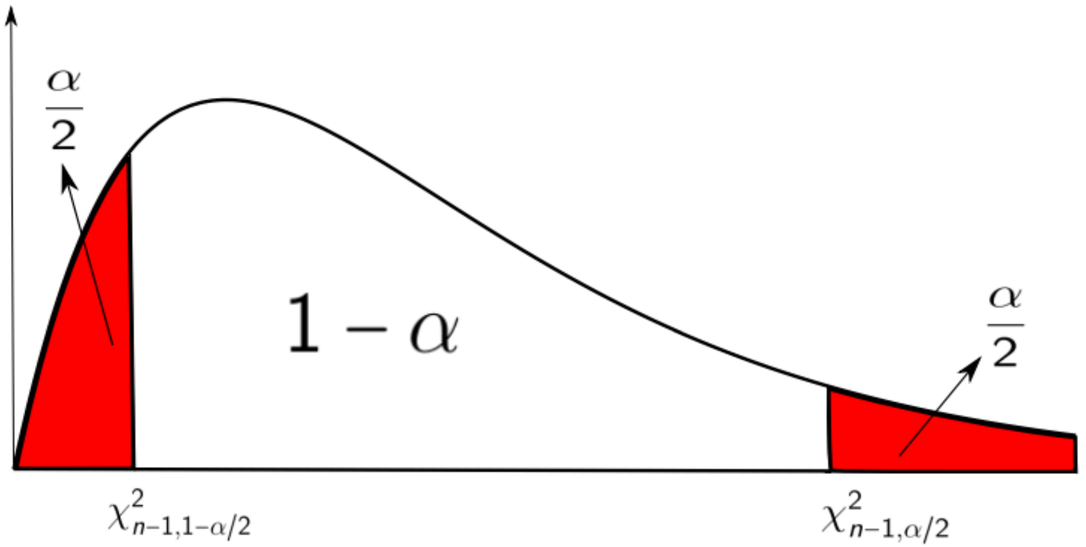


Chi-square distribution

$\chi^2_{n-1,\alpha/2}$ = upper $\alpha/2$ point of chi-square distribution.

$\chi^2_{n-1,1-\alpha/2}$ = lower $\alpha/2$ point of chi-square distribution.

$\chi^2 \sim \chi^2_{n-1} \rightarrow P(\chi^2 > \chi^2_{n-1,\alpha/2}) = \alpha/2$ and $\chi^2 \sim \chi^2_{n-1} \rightarrow P(\chi^2 > \chi^2_{n-1,1-\alpha/2}) = 1 - \alpha/2$



$100(1-\alpha)\,\%$ confidence interval of σ^2

Suppose that $X_1, X_2, \ldots, X_n$ be sample drawn from a normal distribution having variance σ^2 which is unknown.

$100(1-\alpha)\,\%$ confidence interval of σ^2:

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

here $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$ are the values from chi-square distribution with degrees of freedom $n-1$ such that the upper and lower percentiles are both $\alpha/2$.

Example

The accompanying data on breakdown voltage of electrically stressed circuits was read from a normal probability plot that appeared in the article “Damage of Flexible Printed Wiring Boards Associated with Lightning-Included Voltage Surges” (IEEE Transactions on Components, Hybrids, and Manuf. Tech., 1985: 214-220). The straightness of the plot gave strong support to the assumption that breakdown voltage is approximately normally distributed.

1470, 1510, 1690, 1740, 1900, 2000, 2030, 2100, 2190, 2200, 2290, 2380, 2390, 2480, 2500, 2580, 2700

Compute the 95% confidence interval for the variance of the breakdown voltage distribution.

Example

Compute the 95% confidence interval for the variance of the breakdown voltage distribution.

The CI of variance -> means use chi-square distribution

$n = 17$

$s^2 = 137324.3$

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

$$\frac{(17-1)137324.3}{\chi^2_{16,0.975}} < \sigma^2 < \frac{(17-1)137324.3}{\chi^2_{16,0.025}}$$

$$\frac{(17-1)137324.3}{28.845} < \sigma^2 < \frac{(17-1)137324.3}{6.908}$$

$$76172.3 < \sigma^2 < 318064.4$$

Confidence interval for population proportion p

Sampling distribution of $\hat{p}$

- When n elements are randomly sampled from the population, the data will consists of the count X of the number of sampled elements possessing the characteristics. Common sense suggests the sample proportion

$$\hat{p} = \frac{X}{n}$$

as an estimator of p.

- When the sample size n is only a small fraction of the population size, the sample count X has the binomial distribution with mean np and variance $np(1-p)$.

Sampling distribution of $\hat{p}$

- When n is large, the binomial variable X is well approximated by a normal with mean np and variance $np(1-p)$, I.e.

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

  is approximately standard normal distribution.

- This statement can be converted into a statement about proportions by dividing the numerator and the denominator by n. In particular,

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

Sampling distribution of $\hat{p}$

- The last equation shows that $\hat{p}$ is approximately normally distributed with mean p and variance $p(1-p)/n$. Let population proportion be p and X is the number having the characteristic in a random sample of size n. Then,

$$\hat{p} = \frac{X}{n}$$

and

$$E(\hat{p}) = p, Var(\hat{p}) = \frac{p(1-p)}{n}, SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Note

Since the population proportion p is unknown, estimated standard error of $\hat{p}$ formulated by,

$$\text{estimated } SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is used during the calculations.

Confidence interval of p

A random sample of size n is selected, and X is the number of success in the sample. Provided that n is large enough,

$100(1-\alpha)\,\%$ confidence interval of p:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2}$ is the value from the standard normal distribution such that the upper percentile (upper probability) is $\alpha/2$.

Example #1

Many public polling agencies conduct surveys to determine the current consumer sentiment concerning the state of the economy. For example, the Bureau of Economic and Business Research (BEBR) at the University of Florida conducts quarterly to gauge consumer sentiment in the Sunshine State. Suppose that BEBR randomly samples 484 consumers and finds that only 157 are optimistic about the state of the economy.

Use a 90% confidence interval to estimate the proportion of all consumers in Florida who are optimistic about the state of the economy.

Example #1

Suppose that BEBR randomly samples 484 consumers and finds that only 157 are optimistic about the state of the economy:

$$\hat{p} = \frac{X}{n} = \frac{157}{484} = 0.324$$

The 90% confidence interval to estimate the proportion of all consumers in Florida who are optimistic about the state of the economy:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.324 - z_{0.05}\sqrt{\frac{0.324(1-0.324)}{484}} < p < 0.324 + z_{0.05}\sqrt{\frac{0.324(1-0.324)}{484}}$$

$$0.324 - 1.64\sqrt{\frac{0.324(1-0.324)}{484}} < p < 0.324 + 1.64\sqrt{\frac{0.324(1-0.324)}{484}}$$

$$0.289 < p < 0.359$$

Example #2

An automobile club which pays for emergency road services (ERS) requested by its members wishes to estimate the proportions of the different types of ERS requests. Upon examining a sample of 2927 ERS calls, it finds that 1499 calls related to starting problems, 849 calls involved serious mechanical failures requiring towing, 498 calls involved flat tires or lockouts, and 81 calls were for other reasons.

(a) Estimate the true proportion of ERS calls that involved serious mechanical problems requiring towing and determine its 95% margin of error.

(b) Calculate a 98% confidence interval for the true proportion of ERS calls that related to starting problems.

Example #2

Examining a sample of 2927 ERS calls, it finds that 1499 calls related to starting problems, 849 calls involved serious mechanical failures requiring towing, 498 calls involved flat tires or lockouts, and 81 calls were for other reasons.

(a) Estimate the true proportion of ERS calls that involved serious mechanical problems requiring towing and determine its 95% margin of error = Find CI for 95% confidence level

$$\hat{p} = \frac{X}{n} = \frac{849}{2927} = 0.29$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.29 - 1.96\sqrt{\frac{0.29(1-0.29)}{2927}} < p < 0.29 + 1.96\sqrt{\frac{0.29(1-0.29)}{2927}}$$

$$0.2735 < p < 0.3064$$

Example #2

Examining a sample of 2927 ERS calls, it finds that 1499 calls related to starting problems, 849 calls involved serious mechanical failures requiring towing, 498 calls involved flat tires or lockouts, and 81 calls were for other reasons.

(b) Calculate a 98% confidence interval for the true proportion of ERS calls that related to starting problems.

$$\hat{p} = \frac{X}{n} = \frac{1499}{2927} = 0.5121$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.5121 - 2.33\sqrt{\frac{0.5121(1-0.5121)}{2927}} < p < 0.5121 + 2.33\sqrt{\frac{0.5121(1-0.5121)}{2927}}$$

$$0.4905 < p < 0.5336$$

* The Z-value for 98% confidence level is obtained from the standard normal distribution table as 2.33.

Course materials

You can download the notes and codes from:

https://github.com/mcavs/ESTUMatse_2022Fall_EngineeringStatistics

Contact

Do not hesitate to contact me on:

https://twitter.com/mustafa__cavus

https://www.linkedin.com/in/mustafacavusphd/

mustafacavus@eskisehir.edu.tr