İST 292 STATISTICS

Sections: 05-06

For Department of Computer Engineering

LESSON 5 CONFIDENCE INTERVAL-PART I

Dr. Ayten Yiğiter and Dr. Esra Polat Lecture Notes

INTERVAL ESTIMATION

In statistics, point estimation involves the use of sample data to calculate a single value (known as a point estimate since it identifies a point in some parameter space) of an unknown population parameter (for example, the population mean, μ). https://en.wikipedia.org/wiki/Point_estimation

How can we assess the accuracy of this point estimation/estimator.

The Central Limit Theorem (CLT) and sampling distributions help us at this point. For large samples, according to the CLT, the distribution of the sample mean, is approximately normal with μ and variance σ^2/n, $\boldsymbol{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$. Using the distribution of $\bar{X}$ we construct an interval estimation of the population parameter μ.

Basically approximately 95% of all values fall between 2σ away to the mean μ. Hence the interval will contain the mean μ with a probability approximately equal to 0.95.

In other words, approximately 95% of intervals would contain $\bar{x} \pm 2\sigma_{\bar{x}}$ if 100 repeated random samples were drawn from this population. Since there is now way of knowing whether our sample interval is one of the 95% that contain μ or one of the 5% that does not, but the odds (ihtimal, olasılık) certainly favor its containing μ.

An importing point here is that the interval estimation is associated with confidence level such as 90%, 95%, 99%. That’s why we prefer an interval estimation of a parameter to the point estimation of the parameter.

The confidence level is referred by 1-α where α is called as significance level.

- Confidence interval for a population mean (μ)
- Confidence interval for a population variance (σ^2)
- Confidence interval for a population ratio (p)

Confidence Interval of a Population Mean (μ)

Suppose a random sample of size n from a normal population with mean μ and variance σ², $X_1, X_2 \ldots, X_n$, and thus sample mean $\bar{X}$ is a normal distributed random variable with μ and variance σ²/n. Hence,

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

❑ When the Population Variance σ² is Known

To confidence interval of μ for given confidence level (1-α)100%.

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

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

We assert (idda etmek) with (1- α)100% confidence that the interval from $\bar{X} - Z_{\alpha/2\frac{\sigma}{\sqrt{n}}}$ to $\bar{X} + Z_{\alpha/2\frac{\sigma}{\sqrt{n}}}$ contains the true mean of the population

- **NOTE:** If we have a sample from non-normal population with known population variance, by virtue of the Central Limit Theorem, this result can be also used for random samples from non-normal populations provided that n is sufficiently large; that is, n ≥ 30.

Normal Population with known σ^2

Non-normal Population with known σ^2 and $n \geq 30$

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\textit{maximum error}}) = 1 - \alpha$$

Example 1: A team of efficiency experts intends to use the mean of random sample of size n=150 to estimate the average mechanical aptitude (uygunluk, kabiliyet) of assembly-line (montaj hattı) workers in a large industry and suppose that they get $\bar{x} = 69.5$. If, based on experience, the efficiency experts can assume that $\sigma=6.2$ for such data, what can they assert (iddia etmek, ileri sürmek) with probability 0.99 about maximum error of their estimate?

Non-normal population with known variance $\sigma^2=6.2^2$ and n=150 (n≥30) , 99% confidence interval of μ,

$$P(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

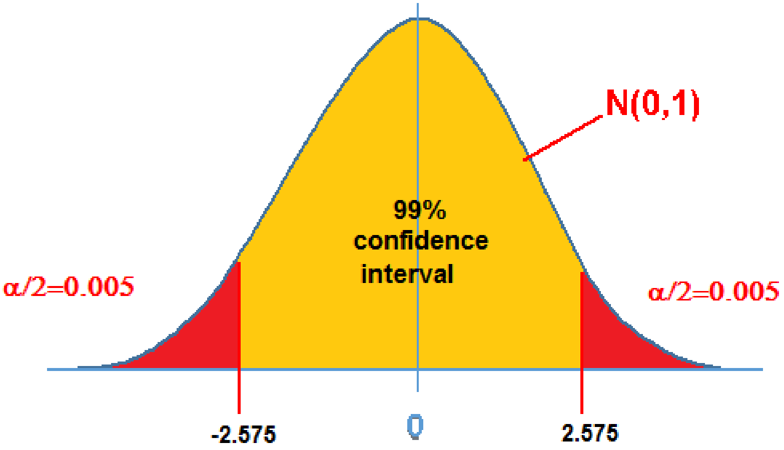
$$z_{\alpha/2} = z_{0.005} = 2.575$$

$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 2.575 \times \frac{6.2}{\sqrt{150}} = 1.3$ as the maximum error of the estimate of μ.

$$P\left(69.5 - 2.575 \times \frac{6.2}{\sqrt{150}} < \mu < 69.5 + 2.575 \times \frac{6.2}{\sqrt{150}}\right) = 0.99$$

$$P(68.2 < \mu < 70.8) = 0.99$$

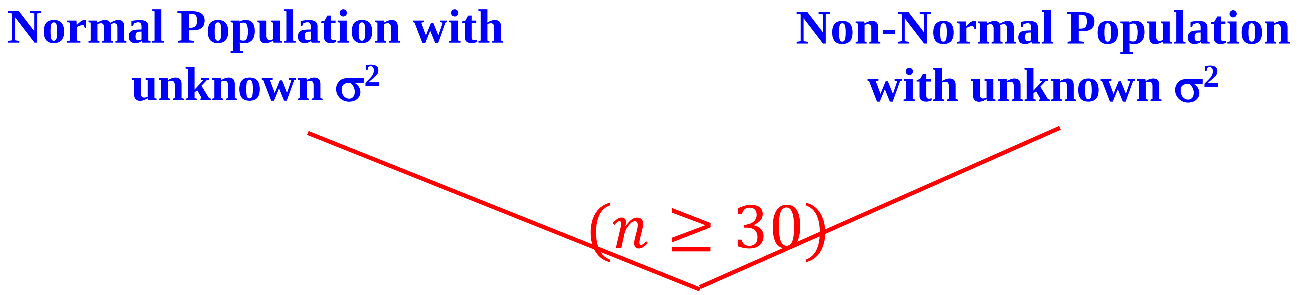
The 99% confidence interval of μ is equal to (68.2, 70.8)



When the Population Variance σ^2 is Unknown

If sample size n enough large, $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim N(0,1)$ for n≥30, where S is sample standard deviation.

To confidence interval of μ for given confidence level (1-α)100%:



Normal Population with unknown σ^2

Non-Normal Population with unknown σ^2

$(n \geq 30)$

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

Example 2: Suppose a large hospital wants to estimate the average length of time patients remain in the hospital. It is assumed that the length of time patients staying in the hospital has normal distribution. To accomplish this objective, the hospital administrators plan to sample 100 of all previous patients records. Find a point estimation and a confidence interval of the mean stay, μ, of all patients' visits using given a $\sum_{i=1}^{100} x_i = 465$ days and $\sum_{i=1}^{100}(x_i - \bar{x})^2 = 2387$ for α=0.05.

$$\bar{x} = 4.65 \qquad s^2 = \frac{\sum_{i=1}^{100}(x_i-\bar{x})^2}{n-1} = \frac{2387}{100-1} = 24.11$$

$$z_{\alpha/2} = z_{0.025} = 1.96 \;\Rightarrow\; P(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$$

$$P(4.65 - 1.96 \times \frac{\sqrt{24.11}}{\sqrt{100}} < \mu < 4.65 + 1.96 \times \frac{\sqrt{24.11}}{\sqrt{100}})=0.95$$

$P(3.69 < \mu < 5.61) = 0.95$ That is, we estimate the mean length stay in the hospital for all patients to fall in the interval 3.69 and 5.61 days with 95 % confidence level.

	The Distribution of Population is Normal – N(μ, σ^2)		**The Distribution of Population is Nonnormal with mean μ and variance σ^2**	
Sample Size	with known population variance σ^2	with unknown population variance σ^2	with known population variance σ^2	with unknown population variance σ^2
$n \geq 30$	z statistic (we use σ^2 in formula)	z statistic (we use s^2 in formula)	As a result of central limit theorem, z statistic (we use σ^2 in formula)	As a result of central limit theorem, z statistic (we use s^2 in formula)
$n<30$	z statistic (we use σ^2 in formula)	t statistic (we use s^2 in formula)	In that case n must be made larger we do not know a special statistic for this case	In that case n must be made larger we do not know a special statistic for this case

- **When Normal Population Variance σ² is Unknown, and n < 30 (small sample)**

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

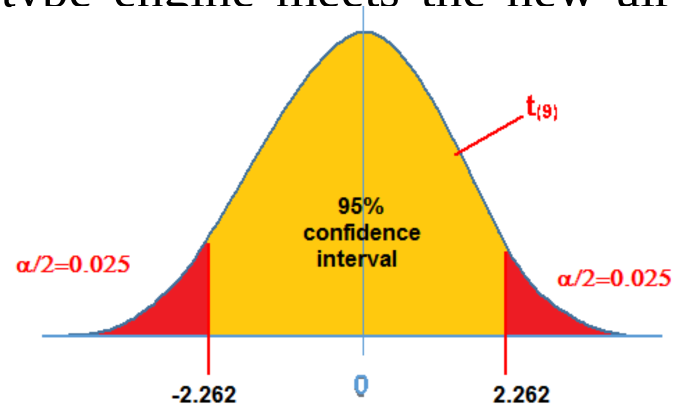
The statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has Student's t distribution with (n-1) degrees of freedom. Based on Student's t distribution, (1-α)100% confidence interval of μ is:

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

Example 3: A major car manufacture wants to test a new engine to determine whether it meets new air pollution standards. The mean emission μ of all engines of this type must be less than 20 parts per million carbon. It is assumed that the emission measures are normally distributed. Ten engines are manufactured for testing purposes and the mean and the standard deviation of the emission for this sample of engines are determined to be $\bar{x} = 17.1$ parts per million of carbon and s=3.0 parts per million of carbon. Find the 95% confidence interval of the true mean emission. Could you decide whether a new type engine meets the new air pollution standards.

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

$$t_{\alpha/2,(n-1)} = t_{0.025,(9)} = 2.262$$



$$P\left(\underbrace{17.1 - 2.262 \times \frac{3}{\sqrt{10}}}_{14.95} < \mu < \underbrace{17.1 + 2.262 \times \frac{3}{\sqrt{10}}}_{19.25}\right) = 0.95$$

That is, the interval 14.95 parts per million to 19.25 parts per million contains the true mean emission with 95% confidence. Hence, a new type engine meets the new air pollution standards.

Confidence Interval of a Population Variance (σ^2)

- **Given a random sample of size n from a normal population**

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

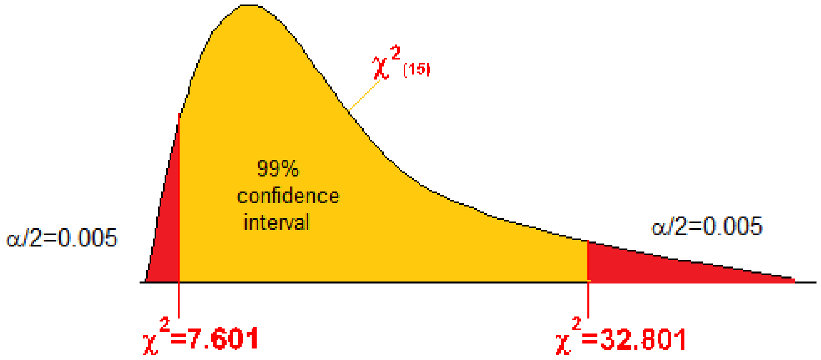
Based on chi-squared distribution, (1-α)100% confidence interval of σ^2 is:

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

Example 4: In 16 tests runs the gasoline consumption of an experimental engine had a standard deviation of 2.2 of gallons. It is assumed that the gasoline consumption of engine has normal distribution. Construct a 99% confidence interval for σ², which measures the true variability of the gasoline consumption of the engine.

$$\chi^2_{\frac{\alpha}{2},(n-1)} = \chi^2_{0.005,(15)} = 32.801$$

$$\chi^2_{1-\frac{\alpha}{2},(n-1)} = \chi^2_{0.995,(15)} = 7.601$$



$$P\left(\underbrace{\frac{(16-1)(2.2)^2}{32.801}}_{2.21} < \sigma^2 < \underbrace{\frac{(16-1)(2.2)^2}{7.601}}_{15.78}\right) = 0.95$$

The true variability of the gasoline consumption of the engine, σ², falls in the interval 2.21 to 15.78 gallons with 99% confidence.

Confidence Interval of a Population Ratio (*p*)

In many problems we must estimate proportions, probabilities or rates such as the proportion of defectives in a large shipment (yük, sevkiyat) of transistors, the probability that a car will be written traffic ticket in a particular day, the mortality rate of a disease.

Assume that each observation (unit) is classified into two groups with respect to whether it has particular property referred as «success» or not, such as defective/non-defective units, yes/no replies, like/unlike tweets, spam/normal mails etc. Under some assumtions, each data of them is considered as Bernoulli population with ratio (*p*), and the distribution of the number of success in a sample from the Bernoulli population has Binomial distribution with parameters (*n*, *p*).

- **Given a random sample of size n from Bernoulli (p) population**
- **X:** The number of success in a sample of size n

$$X \sim Binom(n,p)$$

$$E(X) = np \quad V(X) = np(1-p)$$

As a result of the Central Limit Theorem, for large sample size (n≥30)

$$\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1) \qquad \text{or} \qquad \frac{\frac{X}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

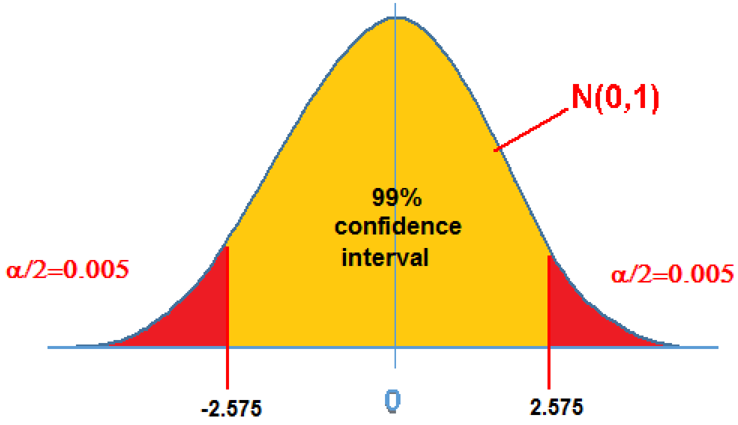
Based on normal distribution, (1-α)100% confidence interval of p is:

$$P\left(\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

$\hat{p} = \frac{x}{n}$ shows sample ratio.

Example 5: A study is made to determine the proportion of the voters in a sizable community who favor the construction of a nuclear power plant (tesis). If 140 of 400 voters selected at random favor the project and we use $\hat{p} = \frac{140}{400} = 0.35$ as an estimate of the actual proportion of all voters in the community who favor the project, what can we say with 99% confidence about the maximum error.

$$z_{\alpha/2} = z_{0.005} = 2.575$$



$$P\left(\underbrace{0.35 - 2.575\sqrt{\frac{0.35(1-0.35)}{400}}}_{0.29} < \mu < \underbrace{0.35 - 2.575\sqrt{\frac{0.35(1-0.35)}{400}}}_{0.41}\right) = 0.99$$

The actual proportion of voters in the community who favor the project falls in the interval (0.29 , 0.41) with 99% confidence.