innovate achieve lead

BITS Pilani
Pilani Campus

Course No: MATH F113

Probability and Statistics

Chapter 7: Statistical Intervals Based on a Single Sample

Sumanta Pasari

BITS Pilani
Pilani Campus

sumanta.pasari@pilani.bits-pilani.ac.in

Interval Estimation

- A point estimate cannot be expected to provide the exact value (close value) of the population parameter.

- **Usually**, an interval estimate can be obtained by adding and subtracting a margin of error to the point estimate. Then,

Interval Estimate = Point Estimate + / – Margin of Error

- Interval estimation provides us information about **how close** the point estimate is to the value of the parameter.
- Why we use the term ***confidence interval?***

Interval (CI) Estimation

– Instead of considering a statistic as a point estimator, we may use *random intervals* to trap the parameter.

– In this case, the end points of the interval are RVs and we can talk about the probability that it traps the parameter value.

Confidence Interval : A $100(1-\alpha)\%$ confidence interval for a parameter is a **random interval** $[L_1, L_2]$ such that

$$P[L_1 \leq \theta \leq L_2] = 1 - \alpha$$, regardless the value of θ.

Theorem 7.1: Interval estimation for μ: σ known

Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal population with mean μ (unknown) and the variance σ^2 (known). Then, we know,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0,1)$$

Taking two points $\pm z_{\alpha/2}$ symmetrically about the origin, we get

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

Here $(1-\alpha)$ is known as confidence level, and α is the level of significance.

Interval estimation for μ: σ known

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

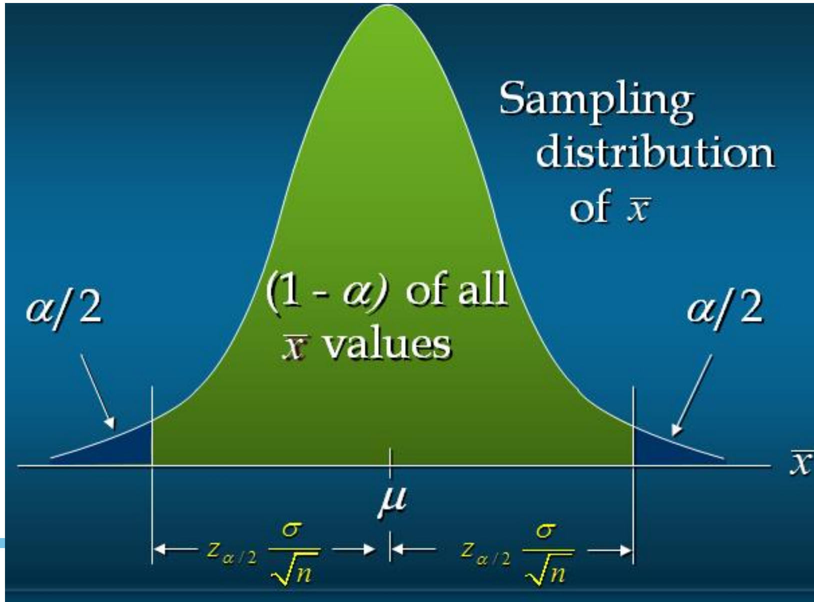
Hence, the confidence interval for population mean μ having confidence level $100\times(1-\alpha)\%$ is given as $\left(\bar{x}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\ \bar{x}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right)$.

The endpoints of the confidence interval is called confidence limits.

Interval Estimate of μ

$$\bar{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where:
- $\bar{x}$ is the sample mean
- $1-\alpha$ is the confidence coefficient
- $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution
- σ is the population standard deviation
- n is the sample size



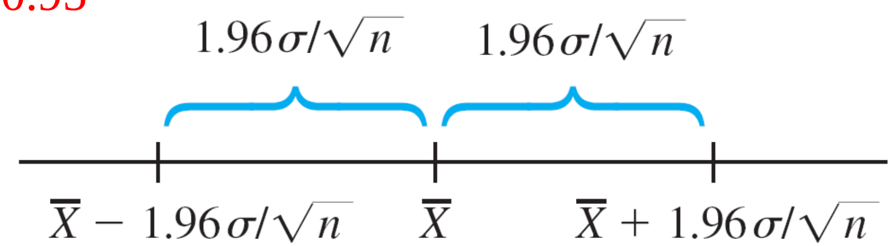
Interval estimation for μ: σ known

Most commonly used confidence levels:

Confidence Level	α	$\alpha/2$	Table Look-up Area	$z_{\alpha/2}$
90%	.10	.05	.9500	1.645
95%	.05	.025	.9750	1.960
99%	.01	.005	.9950	2.576

Hence, 95% CI for μ is given as $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$.

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



Practice Problems

Ex.1. The mean of a sample size 50 from a normal population is observed to be 15.68. If the s.d. of the population is 3.27, find (a) 80% (b) 95%, (c) 99% confidence interval for the population mean. Can you find out the respective margin of errors? What is the length of CI for each case?

Sol. (b) First check the two assumptions : (i) normality (ii) σ known

Step 1: Here $n = 50$, $\bar{x} = 15.68$, $\sigma = 3.27$, and $\alpha = 0.05$. We need CI for μ.

Step 2: As $\alpha = 0.05$, we need to find $z_{\alpha/2}$ such that $P\left(Z < z_{\alpha/2}\right) = 0.975$.

From cumulative normal distribution table, we see $z_{\alpha/2} = 1.96$.

Step 3: The CI for $\mu\left(\sigma \text{ known}\right)$ is $\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{0.025}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{0.025}\right) = \left(14.77, 16.59\right)$

Interval estimation for μ: σ known

Confusions and confusions? ☹ ☹

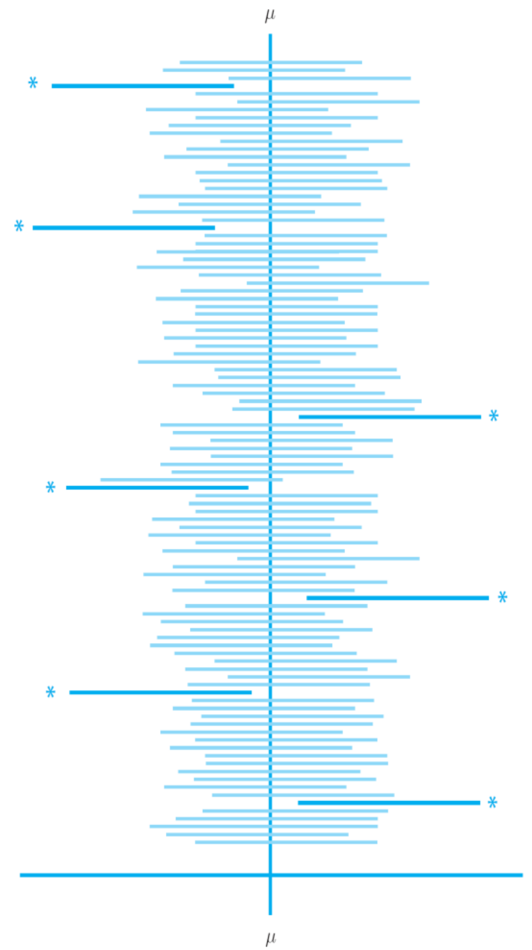
In the above example, we found 95% CI for μ is (14.77, 16.59). This means, the unknown μ lies within the fixed interval with probability 0.95. That is,

P [μ lies in (14.77, 16.59)]= 0.95 **– right?**

If not, then what is the interpretation of “95% confidence”?

- Long run relative frequency?
- A single replication/realization of random interval is not enough! Not satisfactory, at least.

95% CIs for population mean



One hundred 95% CIs (asterisks identify intervals that do not include μ).

Practice Problems

HW 1. Studies have shown that the random variable X, the processing time required to do a multiplication on a new 3-D computer, is normally distributed with mean μ and standard deviation 2 microseconds. A random sample of 16 observations is to be taken

(a) These data are obtained

42.65	45.15	39.32	44.44
41.63	41.54	41.59	45.68
46.50	41.35	44.37	40.27
43.87	43.79	43.28	40.70

Based on these data, find an unbiased estimate for μ.

(b) Find a 95% confidence interval for μ.

Confidence Level, Precision, and Sample Size

Ex.2. Extensive monitoring of a computer time-sharing system has suggested that response time to a particular editing command is normally distributed with standard deviation 25 millisec.

A new operating system has been installed, and we wish to estimate the true average response time μ for the new environment.

Assuming that response times are still normally distributed with $\sigma = 25$, what sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10?

Confidence Level, Precision, and Sample Size

The sample size *n* must satisfy

$$10 = 2 \cdot (1.96)(25/\sqrt{n})$$

Rearranging this equation gives

$$\sqrt{n} = 2 \cdot (1.96)(25)/10 = 9.80$$

So

$$n = (9.80)^2 = 96.04$$

Since *n* must be an integer, a sample size of 97 is required.

Deriving a CI: Example 7.5

A theoretical model suggests that the time to breakdown of an insulating fluid between electrodes at a particular voltage has an exponential distribution with parameter λ.

A random sample of n = 10 breakdown times yields the following sample data (in min): $x_1 = 41.53$, $x_2 = 18.73$, $x_3 = 2.99$, $x_4 = 30.34$, $x_5 = 12.33$, $x_6 = 117.52$, $x_7 = 73.02$, $x_8 = 223.63$, $x_9 = 4.00$, $x_{10} = 26.78$.

A 95% CI for λ and for the true average breakdown time are desired.

Example 7.5

Let $h(X_1, X_2, \ldots, X_n; \lambda) = 2\lambda\Sigma X_i$. It can be shown that this random variable has a probability distribution called a chi-squared distribution with $2n$ degrees of freedom (df) ($\nu = 2n$, where ν is the parameter of a chi-squared distribution).

(https://en.wikipedia.org/wiki/Relationships_among_probability_distributions)

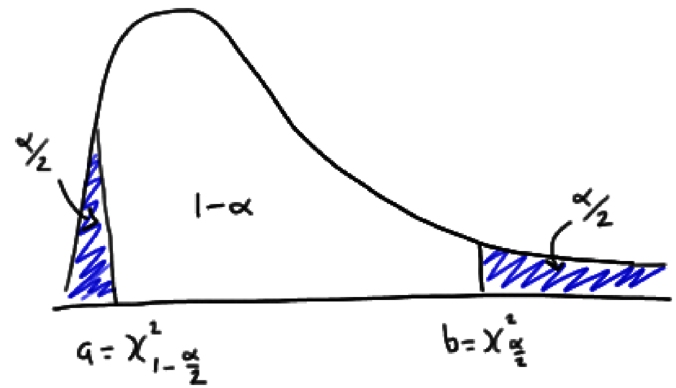
Appendix Table A.7 pictures a typical chi-squared density curve and tabulates critical values that capture specified tail areas. The relevant number of df here is 2(10) = 20.

The $\nu = 20$ row of the table shows that 34.170 captures upper-tail area .025 and 9.591 captures lower-tail area .025 (upper-tail area .975).

Example 7.5

Thus for n = 10,

$$P(9.591 < 2\lambda\Sigma X_i,\ 34.170) = .95$$



Division by

$2\Sigma X_i$ isolates λ, yielding

$$P(9.591/(2\Sigma X_i) < \lambda < (34.170/(2\Sigma X_i)) = .95$$

The lower limit of the 95% CI for λ is $9.591/(2\Sigma X_i)$, and the upper limit is $34.170/(2\Sigma X_i)$.

Example 7.5

For the given data, $\Sigma x_i = 550.87$, giving the interval (.00871, .03101).

The expected value of an exponential rv is $\mu = 1/\lambda$.

Since

$$P((2\Sigma X_i/34.170 < 1/\lambda < (2\Sigma X_i/9.591) = .95$$

the 95% CI for true average breakdown time is

$$(2\Sigma x_i/34.170,\ 2\Sigma x_i/9.591) = (32.24,\ 114.87).$$

This interval is obviously quite wide, reflecting substantial variability in breakdown times and a small sample size.

Impracticality of Assumptions in CI

In practice, we usually face mainly two problems in application of previous C.I. formula.

- What if the population is not normal? **(large sample size is needed) - *Can we take help from CLT?***

- What if the population variance is unknown?

7.2: Large Sample CI for μ

innovate achieve lead

Let $X_1,\ X_2,\ \cdots, X_n$ (large sample) be a random sample from a with mean μ (unknown) and the variance σ^2 (known). Then, using CLT,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X}-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0,1)$$

$100\times(1-\alpha)\%$ CI for μ is given as $\left(\bar{x}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\ \bar{x}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right)$.

7.2: Large Sample CI for μ

Let $X_1,\ X_2,\ \cdots, X_n$ be a random sample from a large sample $(n \geq 40)$ with mean μ (unknown) and sample variance S^2. Then,

$$\frac{\bar{X}-\mu}{\left(\frac{S}{\sqrt{n}}\right)} \sim N(0,1) \quad \left[\text{approximately, a standard normal distribution}\right]$$

$$\Rightarrow \quad P\left(\bar{X}-\frac{S}{\sqrt{n}}z_{\alpha/2} < \mu < \bar{X}+\frac{S}{\sqrt{n}}z_{\alpha/2}\right)=1-\alpha$$

Hence, the large-sample confidence interval for population mean μ having confidence level $100\times(1-\alpha)\%$ is approximately given as $\left(\bar{x}-\frac{s}{\sqrt{n}}z_{\alpha/2},\ \bar{x}+\frac{s}{\sqrt{n}}z_{\alpha/2}\right)$, $n \geq 40$ is needed.

HW 2

Udaipur in Rajasthan is one of the most attractive tourist destinations for foreigners. The average cost per night of a deluxe hotel room in Udaipur city is Rs. 2500. Assuming that this estimate is based on a sample of 144 hotels and that the sample standard deviation is Rs. 600, what is the 95% confidence interval estimate of the population mean? You may use $P(Z < -1.96) = 0.025$. **[4]**

$$\left(\overline{x} - \frac{s}{\sqrt{n}} z_{0.025}, \overline{x} + \frac{s}{\sqrt{n}} z_{0.025} \right)$$

$$= (2402, 5098), \text{as } z_{0.025} = 1.96.$$

One Sided CI for μ

The confidence intervals discussed thus far give both a lower confidence bound *and* an upper confidence bound fc the parameter being estimated.

In some circumstances, an investigator will want only one of these two types of bounds.

For example, a psychologist may wish to calculate a 95% upper confidence bound for true average reaction time to a particular stimulus, or a reliability engineer may want only a lower confidence bound for true average lifetime of components of a certain type.

Sample/Population Proportion

Let us draw a random sample $X_1, X_2, \ldots, X_n$ of size n from population, where

$X_i = 1$, if the i-th member of the sample has the trait

$\quad = 0$, if i-th sample does not have the trait

Then, $X = \sum_{i=1}^{n} X_i$ gives the number of objects in the sample with the trait and the **statistic X/n** gives the proportion of the sample with the trait. Note that $\boldsymbol{X}$ is a binomial RV with parameters $\boldsymbol{n}$ (known) and $\boldsymbol{p}$.

Sample Proportion

The statistic that estimates the parameter p, a proportion of a population that has some property, is the sample proportion

$$\hat{p} = \frac{\text{number in sample with the trait (success)}}{\text{sample size}} = \frac{X}{n}$$

Properties:

(i) As the sample size increases (n large), the sampling distribution of $\hat{p}$ becomes approximately normal (WHY?)

(ii) The mean of $\hat{p}$ is p, and variance of $\hat{p}$ is $\dfrac{p(1-p)}{n}$ (WHY?)

(iii) Can we get estimators of p? Point and interval estimator

Sample Proportion

Note that $\hat{p} = \dfrac{X}{n} = \dfrac{\sum_{1}^{n} X_i}{n} = \bar{X}$,

where, each X_i is an independent point binomial (Bernoulli RV),

that is, $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$

x_i	1	0
$f(x_i)$	p	$1-p$

$$E[X_i] = 1(p)+0(1-p)=p$$

$$Var(X_i) = E[X_i^2]-(E[X_i])^2 = p(1-p)$$

Form of Sampling Distribution of Sample Proportion

If $np \geq 10$ and $n(1-p) \geq 10$, sampling distribution of $\bar{p}$ can be approximated by a normal distribution with mean p and s.d. $\sqrt{\dfrac{p(1-p)}{n}}$.

Knowing, $E(\hat{p}) = 0.60$ and $\sigma_{\hat{p}} = 0.0894$, can we find

(i) $P(0.55 < \hat{p} < 0.65) = ?$

(ii) $P(0.50 < \hat{p} < 0.65) = ?$

Confidence Interval on p

Interval Estimate = Point Estimate + / – Margin of Error

Note that for large n (using CLT),

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \Rightarrow \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Taking two points $\pm z_{\alpha/2}$ symmetrically about the origin, we get

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

Here $(1-\alpha)$ is known as confidence level.

Confidence Interval on p

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

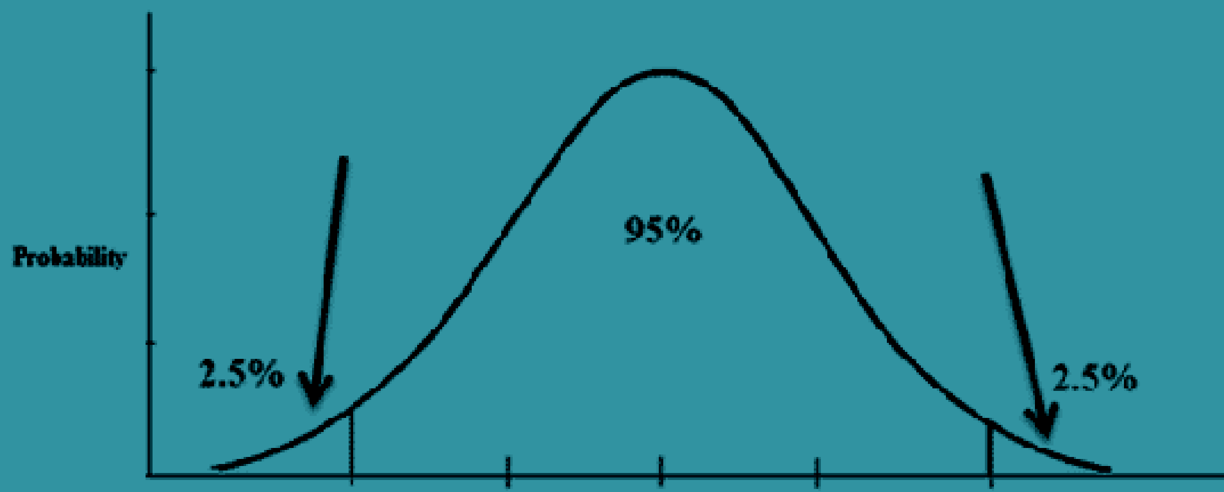
As p is unknown, above confidence bounds are not statistics. So replace p by unbiased estimator $\hat{p}$, and then the CI on p having confidence level $(1-\alpha)$ is

$$\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

The endpoints of the confidence interval is called confidence limits.

Confidence Interval on p

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



$\alpha = 0.05$ $\qquad$ $z_{0.05/2} = z_{0.025}$

Sample Size for Estimating p

innovate achieve lead

We can be 100(1-α)% sure that $\hat{p}$ and p differ by at most d , where d is given by

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

Thus, sample size for estimating p, when prior estimate available is

$$n = z_{\alpha/2}^{2}\frac{\hat{p}(1-\hat{p})}{d^{2}}$$

Sample Size for Estimating p

It can be shown that the value of $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$.

Thus, sample size for estimating p, when

prior estimate is not available is $n = \frac{z_{\alpha/2}^{2}}{4d^{2}}$.

Problem Solving

Ex 3

A study of electromechanical protection devices used in electrical power systems showed that of 193 devices that failed when tested, 75 were due to mechanical part failures.

a) Find a point estimate for p, the proportion of failures that are due to mechanical failures.

b) Find a 95% confidence interval on p.

c) How large a sample is required to estimate p to within 0.03 with 95% confidence.

Problem Solving

innovate achieve lead

Random variable X= number of failed devices which were due to mechanical failure among 193 failed devices.

X has approx. normal dist with mean = 193p, variance=193p(1-p).

a) Point estimate for p = $\hat{p}_{obs}$ = *x/n* = 75/193 = 0.3886.

b) 95% confidence interval on p is

$$\frac{75}{193} \pm z_{0.025}\sqrt{\frac{75}{193}\left(1-\frac{75}{193}\right)\Big/193} = (0.3198, 0.4574)$$

$$\text{(c) (with using prior estimate)} \quad n = \frac{1.96^2(0.389)(0.611)}{(0.03^2)} \sim 1015$$

$$\text{(without using prior estimate)} \ n = \frac{z_{\alpha/2}^2}{4d^2} = \frac{1.96^2}{(4)(0.03^2)} \sim 1068.$$