CHAPTER III
CONFIDENCE INTERVAL

PHOK Ponna

Institute of Technology of Cambodia
Department of Applied Mathematics and Statistics (AMS)

2021–2022

Definition 1

Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a population X with density $f(x;\theta)$, where θ is an unknown parameter. The **interval estimator** of θ is a pair of statistics $L = L(X_1, X_2, \dots, X_n)$ and $U = U(X_1, X_2, \dots, X_n)$ with $L \leq U$ such that if $x_1, x_2, \dots, x_n$ is a set of sample data, then θ belongs to the interval $[L(x_1, x_2, \dots, x_n), U(x_1, x_2, \dots, x_n)]$. The interval $[l, u]$ will be denoted as an interval estimate of θ whereas the random interval $[L, U]$ will denote the interval estimator of θ.

The interval estimator of θ is called a $100(1-\alpha)\%$ confidence interval for θ if

$$\boxed{P(L \leq \theta \leq U) = 1 - \alpha.}$$

The random variable L is called the lower confidence limit and U is called the upper confidence limit. The number $(1-\alpha)$ is called the **confidence level** or degree of confidence.

$CI(\theta) = [L, U]$ or $CI(\theta) = (L, U)$.

Ex: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$\Rightarrow Z$ is a pivot.

Definition 2

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population X with probability density function $f(x;\theta)$, where θ is an unknown parameter. A **pivotal quantity** Q is a function of $X_1, X_2, \ldots, X_n$ and θ whose probability distribution is independent of the parameter θ.

Procedure to find a confidence interval for θ using the pivot method

If $Q = Q(X_1, X_2, \ldots, X_n, \theta)$ is a pivot, then a $100(1-\alpha)\%$ confidence interval for θ may be constructed as follows:

1. Find two values a and b such that
$$P(a \leq Q \leq b) = 1 - \alpha$$
Choose a and b such that $P(Q \leq a) = \alpha/2$ and $P(Q \geq b) = \alpha/2$.
2. Convert the inequality $a \leq Q \leq b$ into the form $L \leq \theta \leq U$.

Example 1

Suppose we have a random sample $X_1, \ldots, X_n$ from $N(\mu, 1)$. Construct a 95% confidence interval for μ.

Example 2

Suppose the random sample $X_1, \ldots, X_n$ has $U(0, \theta)$ distribution. Construct a 90% confidence interval for θ and interpret. Identify the upper and lower confidence limits.

Example 1

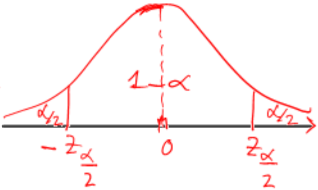
Suppose we have a random sample $X_1, ..., X_n$ from $N(\mu, 1)$. Construct a 95% confidence interval for μ.

Solution

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$, then

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

So Z is a pivot quantity. Then



$$1 - \alpha = P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right)$$

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

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

$a = -z_{\frac{\alpha}{2}}$

$b = z_{\frac{\alpha}{2}}$

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

Thus a $100(1-\alpha)\%$ CI for μ is

$$CI(\mu) = \left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n}},\ \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n}}\right].$$

Example 2

Suppose the random sample $X_1, \ldots, X_n$ has $U(0, \theta)$ distribution. Construct a 90% confidence interval for θ and interpret. Identify the upper and lower confidence limits.

Solution

We know that if $X_1, X_2, \ldots, X_n \overset{iid}{\sim} U[0,\theta]$, then

$\hat{\theta} = \max(X_i)$ is the MLE of θ.

. Let $Y = \hat{\theta} = \max(X_i)$, then the cdf of Y is

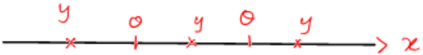
$$G(y) = P(Y \le y) = P(\max(X_i) \le y)$$

$$= P(X_1 \le y, \ldots, X_n \le y) = [P(X \le y)]^n, \quad X \sim U[0,\theta]$$

Since $X \sim U[0,\theta] \Rightarrow f(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$$

$$P(X \le y) = \int_{-\infty}^{y} f(x;\theta)\,dx$$



- If $y < 0 \Rightarrow P(X \le y) = 0$
- If $0 \le y < \theta \Rightarrow P(X \le y) = \int_0^y \frac{1}{\theta}\,dx = \frac{y}{\theta}$
- If $y \ge \theta \Rightarrow P(X \le y) = \int_0^\theta \frac{1}{\theta}\,dx = 1$

Thus, $G(y) = \left[P(X \le y)\right]^n = \begin{cases} 0, & y < 0 \\ \frac{y^n}{\theta^n}, & 0 \le y \le \theta \\ 1, & y > \theta \end{cases}$

So the pdf of Y is

$$g(y) = \frac{d}{dy} G(y) = \begin{cases} \frac{n y^{n-1}}{\theta^n}, & 0 \le y \le \theta \\ 0, & \text{otherwise.} \end{cases}$$

- $E(\hat{\theta}) = E(Y) = \int_{-\infty}^{\infty} y\, g(y)\, dy = \int_0^{\theta} \frac{n}{\theta^n} y^n\, dy$

$$= \frac{n}{\theta^n}\left[\frac{1}{n+1} y^{n+1}\right]_0^{\theta} = \frac{n}{n+1}\theta < \theta$$

So $\hat{\theta} = \max(\bar{x})$ is not an unbiased estimator of θ.

But if we let $M_n = \frac{n+1}{n}\hat{\theta}$, then we have

$$E(M_n) = \frac{n+1}{n} E(\hat{\theta}) = \frac{n+1}{n}\cdot\frac{n}{n+1}\theta = \theta$$

So $M_n = \frac{n+1}{n}\hat{\theta}$ is an unbiased estimator of θ.

Let $Z = \frac{Y}{\theta}$, then the pdf of Z is

$$h(z) = g(\theta z)\cdot|J|, \text{ where } J = \frac{dy}{dz} = \theta$$

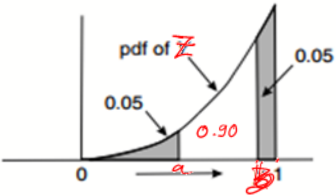
$$\Rightarrow h(z) = g(\theta z)\cdot\theta = \begin{cases} \frac{n}{\theta^n}(\theta z)^{n-1}\cdot\theta, & 0 \le \theta z \le \theta \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} n z^{n-1}, & 0 \le z \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

So $Z = \frac{Y}{\theta}$ is a pivot quantity.

Now, we have to find a and b such that

$$P\left(a \le Z = \frac{Y}{\theta} \le b\right) = 0.90$$



We use the cdf of Z to find a and b

$$\begin{cases} P(Z \le a) = 0.05 \\ P(Z \le b) = 0.95 \end{cases} \Leftrightarrow \begin{cases} P(Y \le \theta a) = 0.05 \\ P(Y \le \theta b) = 0.95 \end{cases}$$

$$\Leftrightarrow \begin{cases} G(\theta a) = 0.05 \\ G(\theta b) = 0.95 \end{cases}$$

But $G(y) = \begin{cases} 0 & , y < 0 \\ \dfrac{y^n}{\theta^n} & , 0 \leq y \leq \theta \\ 1 & , y \geq \theta \end{cases} \Rightarrow \begin{cases} \dfrac{(\theta a)^n}{\theta^n} = 0.05 \\ \dfrac{(\theta b)^n}{\theta^n} = 0.95 \end{cases}$

$$\Rightarrow \begin{cases} a^n = 0.05 \\ b^n = 0.95 \end{cases} \Rightarrow \begin{cases} a = (0.05)^{1/n} \\ b = (0.95)^{1/n} \end{cases}$$

Thus $0.90 = P\left(a \leq \frac{Y}{\theta} \leq b\right) = P\left(\frac{Y}{b} \leq \theta \leq \frac{Y}{a}\right)$

Hence a 90% CI for θ is

$$CI(\theta) = \left(\frac{\max(x_i)}{\sqrt[n]{0.95}} , \frac{\max(x_i)}{\sqrt[n]{0.05}} \right).$$

Large-sample confidence intervals

If the sample size $n \geq 30$, then by the CLT, certain sampling distributions can be assumed to be approximately normal.

1. Find an estimator (such as the MLE) of θ, say $\hat{\theta}$.
2. Obtain the standard error, $\sigma_{\hat{\theta}}$ of $\hat{\theta}$.
3. Find the z-transform $\dfrac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$. Then z has an approximately standard normal distribution.
4. Using the normal table, find two tail values $-z_{\alpha/2}$ and $z_{\alpha/2}$.
5. An approximate $(1 - \alpha)100\%$ for θ is
$$\boxed{\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}.}$$
6. If $\sigma_{\hat{\theta}}$ involve unknown parameters, then we replace $\sigma_{\hat{\theta}}$ by $\hat{\sigma}_{\hat{\theta}}$.

Theorem 1

The large-sample $(1-\alpha)100\%$ CI for the population mean μ is

$$\left[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Example 3

$\nearrow \mu$

Two statistics professors want to estimate average scores for an elementary statistics course that has two sections. Each professor teaches one section and each section has a large number of students. A random sample of 50 scores from each section produced the following results:

(a) Section I: $\bar{x}_1 = 77.01, s_1 = 10.32$.

(b) Section II: $\bar{x}_2 = 72.22, s_2 = 11.02$

Calculate 95% CIs for each of these two samples.

Solution

(a) $n_1 = 50 \geq 30$, $\bar{x}_1 = 77.01$, $s_1 = 10.32$, then

a 95% CI for μ_1 is

$$CI(\mu_1) = \left[\bar{x}_1 \pm z_{\frac{\alpha}{2}} \cdot \frac{s_1}{\sqrt{n_1}}\right] = \left[77.01 \pm 1.96 \times \frac{10.32}{\sqrt{50}}\right]$$

```
A 95 % CI for mu is [ 74.14949 , 79.87051 ]
```

(b) $n_2 = 50 \geq 30$, $\bar{x}_2 = 72.22$, $s_2 = 11.02$, then

a 95% CI for μ_2 is

$$CI(\mu_2) = \left[\bar{x}_2 \pm z_{\frac{\alpha}{2}} \cdot \frac{s_2}{\sqrt{n_2}}\right] = \left[72.22 \pm 1.96 \frac{11.02}{\sqrt{50}}\right]$$

```
A 95 % CI for mu is [ 69.16547 , 75.27453 ]
```

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, $X_i \in \{0, 1\}$

Let $X = \sum_{i=1}^{n} X_i$, then $X \sim Bin(n, p)$ $\quad Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} \approx N(0,1)$

Theorem 2

Let $X \sim Bin(n, p)$, where $p \in (0, 1)$ is the population proportion, and $\hat{p} = \frac{X}{n}$ be the MLE of p. Then an approximate large sample $(1 - \alpha)100\%$ CI for p is

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

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

$\Rightarrow \hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

There are various rules of thumb that are used to determine the adequacy of the sample size for normal approximation. One of the popular rules are that np and $n(1-p)$ should be greater than 10 (or > 5).

Example 4

An auto manufacturer gives a bumper-to-bumper warranty for 3 years or 36,000 miles for its new vehicles. In a random sample of 60 of its vehicles, 20 of them needed five or more major warranty repairs within the warranty period. Estimate the true proportion of vehicles from this manufacturer that need five or more major repairs during the warranty period, with confidence coefficient 0.95. Interpret.

$n = 60, \quad x = 20 \quad \Rightarrow \hat{p} = \frac{x}{n} = \frac{20}{60} = \frac{1}{3}$

$\alpha = 5\% \quad \Rightarrow z_{\frac{\alpha}{2}} = 1.96$

$CI(p) = \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right] = [\, 0.2140537 \,, 0.4526129 \,]$

Theorem 3

Sample size needed for interval estimate of a population proportion is given by

$$\boxed{n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \tilde{p}(1-\tilde{p})}$$ rounded up

where E is the maximum error of estimate and $\tilde{p}$ is the initial estimate of p.

$E = \frac{w}{2}$

Example 5

A researcher wishes to estimate, with 95% confidence, the number of people who own a home computer. A previous study shows that 40% of those interviewed had a computer at home. The researcher wishes to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

$\alpha = 5\%$, $\tilde{p} = 40\%$, $E = 2\%$ $\Rightarrow z_{\frac{\alpha}{2}} = 1.96$

$$\boxed{n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \tilde{p}(1-\tilde{p})}$$ rounded up

$$= \left(\frac{1.96}{0.02}\right)^2 (0.4)(0.6)$$

$$= 2305$$

A sample necessary is 2305

Definition 1

Let X be a crv on $\mathbb{R}$ with cdf f_X. We say that X has Student's t-distribution or t-distribution with degree of freedom $\nu > 0$, written by $X \sim t_\nu$ or $X \sim t(\nu)$, if

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}.$$

Theorem 4

Let $X \sim t_\nu$, $\nu > 0$. Then

1. $E(X) = 0$ for $\nu > 1$, otherwise undefined.
2. $V(X) = \dfrac{\nu}{\nu - 2}$ for $\nu > 2$, $V(X) = \infty$ for $1 < \nu \leq 2$, otherwise undefined.

Theorem 5

Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(\nu)$, $\nu > 0$, and define $T = \dfrac{Z}{\sqrt{V/\nu}}$. Suppose that Z and V are independent. Then, $T \sim t_\nu$.

Theorem 6

Let $X_1, \ldots, X_n$ be a random sample drawn from a normal distribution with mean μ and variance σ^2. Then the random variable

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

has a t distribution with $n-1$ degrees of freedom.

Theorem 7

Let $T \sim t_n$, $n \in \mathbb{N}$, and $Z \sim \mathcal{N}(0,1)$. Then, $T \xrightarrow{p} Z$.

Theorem 8

If $\bar{X}$ and S are the sample mean and the sample standard deviation of a random sample of size n from a normal population, then:

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

is a $(1-\alpha)100\%$ CI for the population mean μ.

Theorem 9

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error $|\bar{x}-\mu|$ will not exceed a specified amount E when the sample size is

$$\boxed{n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.}$$ rounded up.

E is called the maximum error of estimate. $E = \frac{W}{2}$

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where $n < 30$

- If σ is known, then a $100(1-\alpha)\%$ CI for μ is

$$CI(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

- If σ is unknown, then a $100(1-\alpha)\%$ CI for μ is

$$CI(\mu) = \left[\bar{X} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right]$$

Example 6

Assume that the helium porosity (in percentage) of coal samples taken from any particular seam is normally distributed with true standard deviation 0.75. $\sigma = 0.75$

a. Compute a 95% CI for the true average porosity of a certain seam if the average porosity for 20 specimens from the seam was 4.85.

b. Compute a 98% CI for true average porosity of another seam based on 16 specimens with a sample average porosity of 4.56.

c. How large a sample size is necessary if the width of the 95% interval is to be 0.40?

d. What sample size is necessary to estimate true average porosity to within 0.2 with 99% confidence?

Solution

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\sigma = 0.75$

(a) Find a 95% CI for μ

$n = 20$, $\bar{x} = 4.85$

$$CI(\mu) = \left[\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] = \left[4.85 \pm 1.96 \times \frac{0.75}{\sqrt{20}}\right]$$

A 95 % CI for mu is [4.521304 , 5.178696]

(b) Find a 98% CI for μ

$n = 16$, $\bar{x} = 4.56$

$$CI(\mu) = \left[\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] = \left[4.56 \pm 2.33 \times \frac{0.75}{\sqrt{16}}\right]$$

A 98 % CI for mu is [4.12381 , 4.99619]

(c) Find n when $\alpha = 5\%$, $w = 0.4$

$$\boxed{n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.}$$ rounded up.

$E = \frac{w}{2} = 0.2$

$$n = \left(\frac{1.96 \times 0.75}{0.2}\right)^2 = \underline{55}$$

(d) Find n when $\alpha = 1\%$, $E = 0.2$

$$n = \left(\frac{z_{\alpha/2} \times \sigma}{E}\right)^2 = \left(\frac{2.58 \times 0.75}{0.2}\right)^2 = \underline{94}$$

Example 7

The following is a random data from a normal population.

$$7.2 \quad 5.7 \quad 4.9 \quad 6.2 \quad 8.5 \quad 2.8$$

Construct a 95% confidence interval for the population mean μ. Interpret.

Example 8

The scores of a random sample of 16 people who took the TOEFL (Test of English as a Foreign Language) had a mean of 540 and a standard deviation of 50. Construct a 95% CI for the population mean m of the TOEFL score, assuming that the scores are normally distributed.

Example 7

The following is a random data from a normal population.

$$7.2 \quad 5.7 \quad 4.9 \quad 6.2 \quad 8.5 \quad 2.8$$

Construct a 95% confidence interval for the population mean μ. Interpret.

Solution

$x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\sigma = ?$, $n = 6$

$\alpha = 5\% \Rightarrow CI(\mu) = \left[\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right]$

$t_{\frac{\alpha}{2}, n-1} = t_{0.025, 5} = 2.571$, $\bar{x} = 5.88$, $s = 1.96$

A 95 % CI for mu is [3.827493 , 7.939174]

We are 95% confident that μ is between 3.83 and 7.94.

Example 8

The scores of a random sample of 16 people who took the TOEFL (Test of English as a Foreign Language) had a mean of 540 and a standard deviation of 50. Construct a 95% CI for the population mean μ of the TOEFL score, assuming that the scores are normally distributed.

Solution

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2), \sigma = ?$

$n = 16, \bar{x} = 540, s = 50$

$CI(\mu) = \left[\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right]$

$t_{\frac{\alpha}{2}, n-1} = t_{0.025, 15} = 2.131$

A 95 % CI for mu is [513.3569 , 566.6431]

We are 95% confident that the mean score is between 513.36 and 566.64.

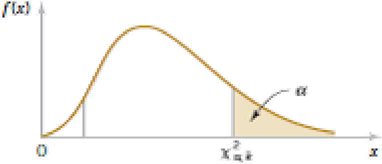
Contents

Theorem 10

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance. Then the random variable

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

has a chi-square (χ^2) distribution with $n-1$ degrees of freedom.



Theorem 11

Let $X_1, \ldots, X_n$ be a random sample drawn from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Then, a $100(1-\alpha)\%$ CI for the population variance is

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

let $a = \chi^2_{1-\frac{\alpha}{2},n-1}$

$b = \chi^2_{\frac{\alpha}{2},n-1}$

A **confidence interval for** σ has lower and upper limits that are the square roots of the corresponding limits in the interval for σ^2. An upper or a lower confidence bound results from replacing $\alpha/2$ with α in the corresponding limit of the CI.

Example 9

Find a 95% confidence interval for the variance and standard deviation of the nicotine content of cigarettes manufactured if a sample of 20 cigarettes has a standard deviation of 1.6 milligrams. Asusme that the variable is approximately normally distributed.

Example 10

Find a 90% confidence interval for the variance and standard deviation for the price in dollars of an adult single-day ski lift ticket. The data represent a selected sample of nationwide ski resorts. Assume the variable is normally distributed.

59	54	53	52	51
39	49	46	49	48

Source: USA TODAY.

Example 9

Find a 95% confidence interval for the variance and standard deviation of the nicotine content of cigarettes manufactured if a sample of 20 cigarettes has a standard deviation of 1.6 milligrams. Asusme that the variable is approximately normally distributed.

Solution

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$n = 20, \; s = 1.6$

$$CI(\sigma^2) = \left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a} \right]$$

$$a = \chi^2_{1-\frac{\alpha}{2}, n-1} = \chi^2_{0.975, 19} = 8.9065$$

$$b = \chi^2_{\frac{\alpha}{2}, n-1} = \chi^2_{0.025, 19} = 32.2583$$

A 95 % CI for variance is [1.480565 , 5.46117]

A 95% CI for σ is $CI(\sigma) = [\, 1.216785, \quad 2.336915 \,]$.

Example 10

Find a 90% confidence interval for the variance and standard deviation for the price in dollars of an adult single-day ski lift ticket. The data represent a selected sample of nationwide ski resorts. Assume the variable is normally distributed.

59	54	53	52	51
39	49	46	49	48

Source: USA TODAY.

Solution

$$CI(\sigma^2) = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}, n-1}} , \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}} \right]$$

A 90 % CI for variance is [15.01273 , 76.38838]

A 90 % CI for standard deviation is [3.874626 , 8.740045]

Contents

Theorem 12

Let $X_{11}, \dots, X_{1n_1}$ be a random sample from a normal distribution with mean μ_1 and variance σ_1^2, and let $X_{21}, \dots, X_{2n_2}$ be a random sample from a normal distribution with mean μ_2 and variance σ_2^2. Let $\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}$ and $\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$. We assume that the two samples are independent. Then $\bar{X}_1$ and $\bar{X}_2$ are independent and the distribution of $\bar{X}_1 - \bar{X}_2$ is $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

$$\Rightarrow Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$\Rightarrow Z$ is a pivot.

Large-sample confidence interval for the difference of two means

(i) If σ_1, σ_2 are known, then a $100(1-\alpha)\%$ large sample CI for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

(ii) If σ_1 and σ_2 are not known, s_1 and s_2 can be replaced respective by sample standard deviations S_1 and S_2 when $n_i \geq 30, i = 1, 2$, then a $100(1-\alpha)\%$ large sample CI for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

Assumptions: The population is normal, and the samples are independent.

Example 11

A study of two kinds of machine failures shows that 58 failures of the first kind took an average of 79.7 minutes to repair with a standard deviation of 18.4 minutes, whereas 71 failures of the second kind took on average 87.3 minutes to repair with a standard deviation of 19.5 minutes. Find a 99% CI for the difference between the true average amounts of time it takes to repair failures of the two kinds of machines.

Sample1: (First kind) $n = 58$, $\bar{x}_1 = 79.7$, $s_1 = 18.4$

Sample2: (Second kind) $n = 71$, $\bar{x}_2 = 87.3$, $s_2 = 19.5$

Find a 99% CI for $\mu_1 - \mu_2$

$$CI(\mu_1-\mu_2) = \left(\bar{x}_1 - \bar{x}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

$$\alpha = 1\% \Rightarrow z_{\frac{\alpha}{2}} = 2.58$$

$$CI(\mu_1-\mu_2) = \left(79.7 - 87.3 \pm 2.58 \sqrt{\frac{18.4^2}{58} + \frac{19.5^2}{71}}\right)$$

A 99 % CI for Mu1-Mu2 is [-16.21763 , 1.017631]

Small-sample confidence interval for the difference of two means

(i) If σ_1 and σ_2 are unknown but $\sigma_1^2 = \sigma_2^2$, then the small sample $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $S_p = \sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$. pooled standard deviation.

(ii) If σ_1 and σ_2 are unknown and $\sigma_1^2 \neq \sigma_2^2$, then the small sample $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,\nu} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where
$\nu = \left(s_1^2/n_1 + s_2^2/n_2\right)^2 / \left[(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)\right]$.

rounded down

Assumption: The samples are independent from two normal populations.

Example 12

Independent random samples from two normal populations with equal variances produced the following data:

Sample 1:	1.2	3.1	1.7	2.8	3
Sample 2:	4.2	2.7	3.6	3.9	

Obtain a 90% CI for $\mu_1 - \mu_2$.

Example 13

Assume that two populations are normally distributed with unknown and unequal variances. Two independent samples are taken with the following summary statistics:

$$\begin{array}{lll} n_1 = 16 & x_1 = 20.17 & s_1 = 4.3 \\ n_2 = 11 & x_2 = 19.23 & s_2 = 3.8 \end{array}$$

Construct a 95% CI for $\mu_1 - \mu_2$.

Example 12

Independent random samples from two normal populations with equal variances produced the following data:

Sample 1:	1.2	3.1	1.7	2.8	3
Sample 2:	4.2	2.7	3.6	3.9	

Obtain a 90% CI for $\mu_1 - \mu_2$. $m = 5, n = 4$

Since $\sigma_1^2 = \sigma_2^2$ are unknown, then

$$CI(\mu_1 - \mu_2) = \left[\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2}, m+n-2} \cdot S_p \cdot \sqrt{\frac{1}{m} + \frac{1}{n}} \right]$$

Degree of freedom is 7

A pooled standard deviation is 0.7738586

A 90 % CI for Mu1-Mu2 is [-2.223514 , -0.2564861]

Example 13

Assume that two populations are normally distributed with unknown and unequal variances. Two independent samples are taken with the following summary statistics:

$$n_1 = 16 \quad \bar{x}_1 = 20.17 \quad s_1 = 4.3$$
$$n_2 = 11 \quad \bar{x}_2 = 19.23 \quad s_2 = 3.8$$

Construct a 95% CI for $\mu_1 - \mu_2$. $\sigma_1^2 \neq \sigma_2^2$ are unknown

Solution

$$CI(\mu_1 - \mu_2) = \left[\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2}, v} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

$\alpha = 5\%$, $t_{\frac{\alpha}{2}, v} = t_{0.025, 23} = 2.069$

Degree of freedom is 23

A 95 % CI for Mu1-Mu2 is [-2.310066 , 4.190066]

Contents

Large-sample confidence interval for $p_1 - p_2$

The $(1-\alpha)100\%$ large-sample CI for $p_1 - p_2$ is given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $\hat{p}_1$ and $\hat{p}_2$ are the point estimators of p_1 and p_2. This approximation is applicable if $\hat{p}_i n_i \geq 5, i = 1, 2$ and $(1-\hat{p}_i)n_i \geq 5, i = 1, 2$ .The two samples are independent.

Example 14

Iron deficiency, the most common nutritional deficiency worldwide, has negative effects on work capacity and on motor and mental development. In a 1999–2000 survey by the National Health and Nutrition Examination Survey, iron deficiency was detected in 58 of 573 white, non-Hispanic females (10% rounded to whole number) and 95 of 498 (19% rounded to whole number) black, non-Hispanic females (source: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5140a1.htm). Let p_1 be the proportion of black, non-Hispanic females with iron deficiency and let p_2 be the proportion of white, non-Hispanic females with iron deficiency. Obtain a 95% CI for $p_1 - p_2$.

Solution

Find a 95% CI for $p_1 - p_2$

$$CI(p_1 - p_2) = \left[(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

Sample 1 : $\hat{p}_1 = \frac{x_1}{n_1} = \frac{95}{498}$

Sample 2 : $\hat{p}_2 = \frac{x_2}{n_2} = \frac{58}{573}$

$\alpha = 5\% \rightarrow z_{\frac{\alpha}{2}} = 1.96$

A 95 % CI for p1-p2 is [0.04710667 , 0.1319761]

Theorem 13

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent samples of size n_1 and n_2 from two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. Let S_1^2 and S_2^2 be the variances of the two random samples. Then the random variable

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

Theorem 14

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent samples of size n_1 and n_2 from two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. Let S_1^2 and S_2^2 be the variances of the two random samples. The a $100(1-\alpha)\%$ CI for $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$\left(\frac{S_1^2}{S_2^2} \frac{1}{F_{n_1-1,n_2-1,\cancel{1-}\alpha/2}}, \frac{S_1^2}{S_2^2} \frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}} \right)$$

Confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$

Example 15

Assuming that two populations are normally distributed, two independent random samples are taken with the following summary statistics:

$$\begin{array}{lll} n_1 = 21 & \bar{x}_1 = 20.17 & s_1 = 4.3 \\ n_2 = 16 & \bar{x}_2 = 19.23 & s_2 = 3.8 \end{array}$$

Construct a 95% CI for $\frac{\sigma_1^2}{\sigma_2^2}$.

Solution

Find a 95% CI for $\frac{\sigma_1^2}{\sigma_2^2}$

- Sample 1 : $n_1 = 21$, $S_1 = 4.3$ $\alpha = 5\% = 0.05$
- Sample 2 : $n_2 = 16$, $S_2 = 3.8$

$$CI\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left(\frac{S_1^2}{S_2^2} \cdot \frac{1}{b}, \frac{S_1^2}{S_2^2} \cdot \frac{1}{a}\right)$$

where $a = F_{1-\frac{\alpha}{2}, n_1-1, n_2-1} = F_{0.975, 20, 15} = 0.3886$

$b = F_{\frac{\alpha}{2}, n_1-1, n_2-1} = F_{0.025, 20, 15} = 2.7559$

A 95 % CI for the ratio of two variances is [0.4646286 , 3.294775]