Confidence Intervals

Anastasiia Kim

April 27/29, 2020

Idea

- the point estimate $\hat{\theta}$ alone does not give much information about parameter θ
- Without additional information, we do not know how close $\hat{\theta}$ is to real θ
- Instead of giving just one value $\hat{\theta}$ as the estimate for θ, we may produce an interval that is likely to include the true value of θ

$$\hat{\theta} = 5.2$$

  A plausible range of values for the population parameter is called a confidence interval (CI):

$$[L, U] = [4.9, 5.45]$$

- Two important factors: the length of the interval and the confidence interval
- The length of the interval $C_U - C_L$ shows the precision with which we can estimate θ

CI on the Mean of a Normal Distribution, Variance known

For i.i.d. r.v.s $X_1, X_2, ..., X_n$ with unknown expected value $E(X_i) = \mu$ and known variance $Var(X_i) = \sigma^2$ the sample mean is approximately $Normal(\mu, \sigma^2/n)$.

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

- a confidence interval estimate for μ is an interval of the form $l \leq \mu \leq u$
- different samples will produce different values of l and u, these end-points are values of random variables L and U, respectively

CI on the Mean of a Normal Distribution, Variance known

$$P(L \leq \mu \leq U) = 1 - \alpha, \quad 0 \leq \alpha \leq 1$$

- There is a probability of $1 - \alpha$ of selecting a sample for which the CI will contain the true value of μ.
- Once we have selected the sample and computed l and u, the resulting confidence interval for is $l \leq \mu \leq u$
  - l and u are the lower- and upper-confidence limits (bounds)
  - $1 - \alpha$ is the confidence coefficient

CI on the Mean of a Normal Distribution, Variance known

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

We can write

$$P(L \leq \mu \leq U) = 1 - \alpha, \quad 0 \leq \alpha \leq 1$$

as

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha, \quad 0 \leq \alpha \leq 1$$

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

This is a random interval because the end-points

$$\bar{X} \pm Z_{\alpha/2}\sigma/\sqrt{n}$$

involve the random variable $\bar{X}$.

CI on the Mean of a Normal Distribution, Variance known

If $\bar{x}$ is the sample mean of a random sample of size n from a Normal population with known variance σ^2, a $100(1-\alpha)\%$ CI on μ is given by

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

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution.
Z-scores for commonly used confidence intervals:
Confidence level and Z score

$$90\% \quad z_{0.10/2} = 1.645$$

$$95\% \quad z_{0.05/2} = 1.96$$

$$99\% \quad z_{0.01/2} = 2.576$$

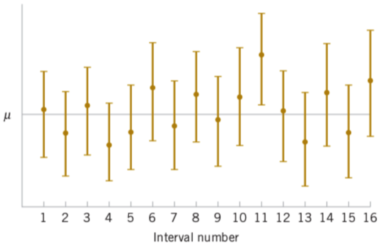
Interpreting a Confidence Interval

Say, the 95% CI for the mean boiling temperature of a certain liquid is $102.3 \le \mu \le 104.2$. Does it mean that μ is within this interval with probability 0.95?

- the true value of μ is unknown and the obtained CI above might be either correct or wrong
- a CI is a random interval because in the probability statement defining the endpoints of the interval L and U are random variables
- the correct interpretation of a $100(1-\alpha)\%$ CI depends on the relative frequency view of probability
- if an infinite number of random samples are collected and a $100(1-\alpha)\%$ CI for μ is computed from each sample, $100(1-\alpha)\%$ of these intervals will contain the true value of μ

Interpreting a Confidence Interval

Repeated construction of a confidence interval for μ:



- the dots at the center of the intervals indicate the point estimate of μ (that is, $\bar{x}$)
- one of the intervals fails to contain the true value of μ
- if this were a 95% confidence interval, in the long run only 5% of the intervals would fail to contain μ.

Interpreting a Confidence Interval

- In practice, we obtain only one random sample and calculate one confidence interval
- We can't talk about the probability that the given CI estimate contains μ
- The appropriate statement is that the observed interval $[l, u]$ brackets the true value of μ with confidence $100(1-\alpha)$.

CI on the Mean of a Normal Distribution, Variance known

Suppose a number of weekly hours of internet use among 9-11 y.o. Australian children is normaly distributed with variance of 36 hours. Suppose the mean number of hours of internet use per week is 5.75 hours obtained from the data of 2500 children. Calculate a 95% confidence interval for the mean number of hours of internet use per week.

$$95\% \qquad z_{0.05/2} = 1.96$$

$$\bar{x} = 5.75,\ n = 2500, \sigma^2 = 36$$

The 95% CI is

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

$$5.75 - (1.96)6/\sqrt{2500} \leq \mu \leq 5.75 + (1.96)6/\sqrt{2500}$$

$$[5.515, 5.985] \text{ hous per week}$$

Confidence Level and Precision of Estimation

The 99% CI is longer than the 95% CI $\rightarrow$ we have a higher level of confidence in the 99% confidence interval

- For a fixed sample size n and standard deviation σ, the higher the confidence level, the longer the resulting CI
- The length of a confidence interval is a measure of the precision of estimation
- Obtain a confidence interval that is short enough for decision-making purposes
- Choose the sample size n to be large enough to give a CI of specified length or precision with prescribed confidence.

Choice of sample size

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

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

if n is not an integer, it must be rounded up.

Large-Sample Confidence Interval

When n is large ($n > 30$, better to have $n > 40$), the quantity

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

has an approximate standard normal distribution. Consequently,

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

is a large-sample confidence interval for μ, with confidence level of approximately $100(1 - \alpha)\%$

Large-Sample Confidence Interval. Example

Mercury contamination in a certain fish. Note that the distribution of mercury concentration is not normal. A sample of fish was selected from 53 Florida lakes, and mercury concentration in the muscle tissue was measured (ppm):

$$n = 53 > 40, \bar{x} = 0.525, s = 0.3486$$

The approximate 95% CI on μ is:

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

$$0.525 - 1.96(0.3486)/\sqrt{53} \leq \mu \leq 0.525 + 1.96(0.3486)/\sqrt{53}$$

$$0.4311 \leq \mu \leq 0.6189$$

CI on the Mean of a Normal Distribution, Variance is Unknown

If $\bar{x}$ is the sample mean of a random sample of size n from a Normal population with unknown variance σ^2. The random variable

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

has a t distribution with n - 1 degrees of freedom.
The general appearance of the t distribution is similar to the standard normal distribution in that both distributions are symmetric and unimodal, and the maximum ordinate value is reached when the mean $\mu = 0$. The t distribution has heavier tails than the normal; that is, it has more probability in the tails than does the normal distribution.

t confidence interval on μ

If $\bar{x}$ and s are the mean and standard deviation of a random sample of size n from a Normal population with unknown variance σ^2, a $100(1-\alpha)\%$ CI on μ is given by

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n-1$ degrees of freedom.
$t_{\alpha/2,n-1}$ depends on desired confidence level and degrees of freedom.
R (90% confidence level , sample of size n=20): qt(1-0.10/2, 20-1) = 1.729

t confidence interval on μ. Example

A farmer weighs 10 randomly chosen watermelons from his farm. Data $\bar{x} = 9.26$ and $s = 1.99$. Assuming that the weight is normally distributed with mean μ and variance σ^2, find a 95% confidence interval for μ.
R (95% confidence level, sample of size n=10): qt(1-0.05/2, 10-1) = 2.262

$$\bar{x} - t_{\alpha/2,n-1} s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1} s/\sqrt{n}$$

$$9.26 - 2.262(1.99)/\sqrt{10} \leq \mu \leq 9.26 + 2.262(1.99)/\sqrt{10}$$

[7.84, 10.68] is a 95% confidence interval for μ.

The Pivotal Method

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a parameter θ that is to be estimated. The random variable Q is said to be a pivot or a pivotal quantity, if it has the following properties:

- It is a function of the observed data $X_1, X_2, \ldots, X_n$ and the unknown parameter θ but it does not depend on any other unknown parameters:

$$Q = Q(X_1, X_2, \ldots, X_n; \theta)$$

- The probability distribution of Q does not depend on θ or any other unknown parameters.

Pivotal quantities allow the construction of exact confidence intervals, meaning they have exactly the stated confidence level, as opposed to so-called 'large-sample' (asymptotic) confidence intervals.

- an exact CI is valid for any sample size
- an asymptotic confidence interval is valid only for sufficiently large sample size

Exact intervals. The Pivotal Method

- Find a pivotal quantity $Q = Q(X_1, X_2, ..., X_n; \theta)$
- Find upper and lower confidence limits on the pivotal quantity, that is, l and u such that
$$P(q_1 \leq Q \leq q_2) = 1 - \alpha, \quad 0 \leq \alpha \leq 1$$
- The constants q_1 and q_2 are called critical values. They are obtained from a table for the distribution of the pivotal quantity or from a computer program.

The example of a pivotal quantity is

$$Q = Q(X_1, X_2, ..., X_n; \theta) = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$P(q_1 \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq q_2) = 0.95$$

which is equivalent to

$$P(\bar{X} - q_1 \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + q_2 \frac{S}{\sqrt{n}}) = 0.95$$

The Pivotal Method. Exponential

Suppose $X_1, X_2, ..., X_n$ are i.i.d. *Exponential*(λ). Then

$$\sum_{i=1}^{n} X_i \sim Gamma(n, \lambda)$$

It can be shown that

$$\lambda \bar{X} \sim Gamma(n, n)$$

Since the distribution here does not depend on the parameter λ, we see that

$$Q = \lambda \bar{X}$$

is a pivotal quantity.
We choose q_1 and q_2 to be the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of the pivotal quantity. In R (alpha $= 0.05$ corresponds to the confidence level 95%):

$$qgamma(alpha/2, shape = n, rate = n)$$

$$qgamma(1 - alpha/2, shape = n, rate = n)$$

Example: $\bar{x} = 20, n = 10$, the 95% CI is $0.479 \leq \lambda/\bar{x} \leq 1.708$.

The Pivotal Method. Uniform

Suppose $X_1, X_2, ..., X_n$ are i.i.d. $Uniform(0, \theta)$. Then

$$\frac{X_i}{\theta} \sim Uniform(0, 1)$$

It can be shown that

$$max\left(\frac{X_1}{\theta}, \frac{X_2}{\theta}, ..., \frac{X_n}{\theta}\right) \sim Uniform(0, 1)$$

Since the distribution here does not depend on the parameter θ, we see that

$$Q = max\left(\frac{X_1}{\theta}, \frac{X_2}{\theta}, ..., \frac{X_n}{\theta}\right) = \frac{X_{(n)}}{\theta}$$

is a pivotal quantity.

The Pivotal Method. Uniform

To find 95% CI, we need to choose q_1 and q_2 such that

$$P(q_1 \leq Q = \frac{X_{(n)}}{\theta} \leq q_2) = 0.95$$

$$P(q_1 \leq \frac{X_{(n)}}{\theta} \leq q_2) = \int_{q_1}^{q_2} ny^{n-1}dy = q_2^n - q_1^n = 0.95$$

So $q_2^n - q_1^n = 0.95$ must hold and $0 < q_1, q_2 < 1$ because $X_{(n)} \sim Uniform(0,1)$

$$P\left(\frac{X_{(n)}}{q_2} \leq \theta \leq \frac{X_{(n)}}{q_1}\right) = 0.95$$

The length of the interval is $X_{(n)}\left(\frac{1}{q_1} - \frac{1}{q_2}\right)$. We can do anything with $X_{(n)}$ but we can minimize $\left(\frac{1}{q_1} - \frac{1}{q_2}\right)$ subject to the constraint $q_2^n - q_1^n = 0.95$. The solution is $q_2 = 1, q_1 = 0.05^{1/n}$. Among 95% CIs the shortest one is $\theta \in [X_{(n)}, X_{(n)}/0.05^{1/n}]$.

Confidence Intervals for the Variance of Normal Random Variables

Suppose $X_1, X_2, ..., X_n$ are i.i.d. $Normal(\mu, \sigma^2)$. Then find an interval estimator for σ^2. Assume that μ is also unknown.

The random variable Q

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

has a chi-squared distribution with n-1 degrees of freedom, i.e., $Q \sim \chi^2_{n-1}$. A chi-squared distribution is a special case of Gamma distribution, $\chi^2_n \sim Gamma(n/2, 2)$. Q is a pivotal quantity because its distribution does not depend on σ^2 or any other unknown parameters. The $100(1-\alpha)\%$CI can be found by solving

$$P(\chi^2_{1-\alpha/2,n-1} \leq Q \leq \chi^2_{\alpha/2,n-1}) = 1 - \alpha$$

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

One-Sided Confidence Bounds on the Variance

Two-sided: $\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$
One-sided confidence bounds on σ^2 are

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

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

One-Sided Confidence Bounds on the Variance

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.01532$ (fluid ounce). If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. We will assume that the fill volume is approximately normally distributed. A 95% upper confidence bound is found from

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

$$\sigma^2 \leq \frac{(19)0.01532}{10.117} = 0.0287$$

The standard deviation is $\sigma = 0.17$. Therefore, at the 95% level of confidence, the data indicate that the process standard deviation could be as large as 0.17 fluid ounce.
In R: qchisq(.95, df=19,lower.tail=FALSE) = 10.117 gives the right tail probability.