Math 4501 - Probability and Statistics II

7.1 - Confidence intervals for means

Normal distribution of unknown variance

iid

Let $X_1, X_2, \dots, X_n$ be a random sample from a normal distribution of unknown variance σ^2.

both unknown

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$

We have seen previously that

$$\boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)}$$

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

Chp 5: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

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

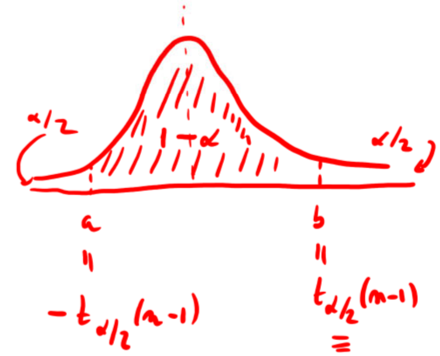
$T = \frac{Z}{\sqrt{U/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

- S^2 is the unbiased estimator of σ^2

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Repeating the previous argument with $z_{\alpha/2}$ replaced by $t_{\alpha/2}(n-1)$:

$$P\left[-t_{\alpha/2}(n-1) \le \overbrace{\frac{\bar{X}-\mu}{S/\sqrt{n}}}^{T} \le t_{\alpha/2}(n-1)\right] = 1 - \alpha$$



solve inequality inside for μ

and so

$$\left[\bar{x} - t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right), \bar{x} + t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right]$$

is a $100(1-\alpha)\%$ confidence interval for μ.

Example

Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves. Assume that the distribution of X is $N(\mu, \sigma^2)$.

$X_1, X_2, \ldots, X_{20}$

To estimate μ, a farmer measured the butterfat production for $n = 20$ cows. The data yielded a sample mean $\bar{x} = 507.50$ and sample ~~variance~~ $s = 89.75$.

standard deviation

Determine an approximate 90% confidence interval for μ.

$1-\alpha = 0.9 \Rightarrow \alpha = 0.1$

Note that:

KEY INFORMATION

- sample taken from a normal distribution
- variance is unknown
- small sample

$n = 20$

$X_1, \ldots, X_n \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

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

$U = \frac{(n-1)s^2}{\sigma^2} = \frac{19 \cdot s^2}{\sigma^2} \sim \chi^2(\overbrace{n-1}^{19})$

$T = \frac{Z}{\sqrt{U/(n-1)}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(\overbrace{n-1}^{19})$

The 90% confidence interval for μ is

$$\left[\bar{x} - t_{\overbrace{0.05}^{\alpha/2}}(\overbrace{19}^{n-1})\left(\frac{s}{\sqrt{n}}\right), \bar{x} + t_{0.05}(19)\left(\frac{s}{\sqrt{n}}\right)\right] =$$

$$= \left[507.50 - 1.729\left(\frac{89.75}{\sqrt{20}}\right), 507.50 + 1.729\left(\frac{89.75}{\sqrt{20}}\right)\right] = \underbrace{[17.94, 20.20]}.$$

Large sample with known variance

ASSUMPTION

not very realistic

Let $X_1, X_2, \ldots, X_n$ be a random sample from a not necessarily normal distribution with known variance σ^2.

By the Central limit theorem, for n large enough:

- $\bar{X}$ is approximately $N\left(\mu, \sigma^2/n\right)$
- $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approximately $N(0, 1)$

approximation improves with increasing sample size

$$\text{cdf of } Z \xrightarrow[\text{as } n \to \infty]{} \text{cdf } N(0,1)$$

Repeating the argument devised previously for samples from a normal distribution, we find that

$$P\left(-z_{\alpha/2} \leq \overbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}^{Z} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

approximately

and so

solve for μ

$$\left[\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for μ.

same formula as the case we saw last time!!

Note: the closeness of the approximate probability $1-\alpha$ to the exact probability depends on both the underlying distribution and the sample size.

1) When the underlying distribution is unimodal (has only one mode), symmetric, and continuous, the approximation is usually quite good even for small n.

distribution is already very close to normal

- For distribution close to normal, $n = 5$ might suffice to obtain reasonably good approximations.

2) As the underlying distribution becomes "less normal" (i.e., badly skewed or discrete), a larger sample size might be required to keep a reasonably accurate approximation.

- In most cases, an n of at least 30 is usually adequate.

↳ 50 if we want to be extra safe

Example

Let X equal the amount of orange juice (in grams per day) consumed by an American.

$\sigma^2 = 96^2$

Suppose it is known that the standard deviation of X is $\sigma = 96$. To estimate the mean μ of X, an orange growers' association took a random sample of $n = 576$ Americans and found that they consumed, on the average, $\bar{x} = 133$ grams of orange juice per day.

Determine an approximate 90% confidence interval for μ.

Note that:

- distribution from which the sample is taken is not necessarily normal
- variance is known
- large sample (n = 576 is large!) $\Rightarrow$ use CLT $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approx. $N(0,1)$

approximate

The 90% confidence interval for μ is

table!

$$\left[\bar{x} - z_{0.05}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{0.05}\left(\frac{\sigma}{\sqrt{n}}\right)\right] =$$

$$= \left[133 - 1.645\left(\frac{96}{\sqrt{576}}\right), 133 + 1.645\left(\frac{96}{\sqrt{576}}\right)\right] = [126.42, 139.58] .$$

Large sample with unknown variance

→ reasonably realistic situation

Let $X_1, \ldots, X_n$ be a random sample from a distribution with unknown variance σ^2.

For n large enough, we use the fact that $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ is approximately $N(0,1)$:

→ HEURISTIC

→ usually a good approximation

→ NOT a convergence result

- S^2 is the unbiased estimator of σ^2.
- statement is true whether or not the underlying distribution is normal.
- n large enough usually means $n \geq 30$. (50 to be safer)
- if the underlying distribution is badly skewed or contaminated with occasional outliers, we would prefer to have a larger sample size ($n \geq 50$).

Repeating the argument devised previously, we find that

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1-\alpha$$

and so

solve for μ

$$\left[\bar{x} - z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right), \bar{x} + z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for μ.

Example

To measure the effect on a lake of salting the streets of a city on its shore, students took 32 samples of water and measured the amount of sodium in parts per million in order to make a statistical inference about the unknown mean μ.

$n = 32$

look for a confidence interval for μ

Their data yielded a sample mean $\bar{x} = 19.07$ and sample variance $s^2 = 10.60$.

Determine an approximate 95% confidence interval for μ.

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \boxed{\frac{\alpha}{2} = 0.025}$

Note that:

- distribution from which the sample is taken is not necessarily normal
- variance is unknown
- large sample (just barely since $n = 32 \geq 30$)

$n = 32$ is NOT very large
just reasonably large

Use $Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ is approx. $N(0,1)$

The 95% confidence interval for μ is

$$\left[\bar{x} - \underbrace{z_{0.025}}_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right), \bar{x} + z_{0.025}\left(\frac{s}{\sqrt{n}}\right)\right] =$$

$$= \left[19.07 - 1.96\sqrt{\frac{10.60}{32}}, 19.07 + 1.96\sqrt{\frac{10.60}{32}}\right] = [17.94, 20.20] \,.$$

Smaller sample of unknown variance

↳ smaller but not necessarily small

Let $X_1, \ldots, X_n$ be a random sample from a distribution of unknown variance σ^2.

↳ not necessarily normal

We can use that

we account for the smaller sample size by using a distribution with larger tails →

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

HEURISTIC STATEMENT

has an approximate t distribution with $r = n - 1$ degrees of freedom.

- S^2 is the unbiased estimator of σ^2.

by replacing $N(0,1)$ with $t(n-1)$ we get a larger interval for a fixed confidence level

↑

because $t(n-1)$ has larger tails than $N(0,1)$

↑

smaller n is, the larger the tails are.

Repeating the previous argument:

$$P\left[-t_{\alpha/2}(n-1) \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right] \approx 1 - \alpha$$

and so

solve for μ

$$\left[\bar{x} - t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right), \bar{x} + t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for μ.

One-sided confidence intervals

Sometimes we might want only a lower (or upper) bound on the unknown μ

The approach developed above can be applied to one-sided confidence intervals.

For instance, let $\bar{X}$ be the mean of a random sample of size n from a normal distribution $N\left(\mu, \sigma^2\right)$ with known variance σ^2. Then:

$$P\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_\alpha\right] = 1-\alpha$$

or equivalently

$$P\left[\bar{X} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu\right] = 1-\alpha\ .$$

Once $\bar{X}$ is observed to be equal to $\bar{x}$, it follows that

$$\left[\bar{x} - z_{\alpha}(\sigma/\sqrt{n}), \infty\right)$$

is a $100(1-\alpha)\%$ one-sided confidence interval for μ.

- $\bar{x} - z_{\alpha}(\sigma/\sqrt{n})$ is a lower bound for μ with confidence $1-\alpha$.

Similarly, we can show that

$$\left(-\infty, \bar{x} + z_{\alpha}(\sigma/\sqrt{n})\right]$$

is a one-sided confidence interval for μ:

- $\bar{x} + z_{\alpha}(\sigma/\sqrt{n})$ is an upper bound for μ with confidence $1-\alpha$.

When σ is unknown, we would use

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

to find the corresponding lower or upper bounds for μ:

$$\bar{x} - t_{\alpha}(n-1)(s/\sqrt{n}) \quad \text{and} \quad \bar{x} + t_{\alpha}(n-1)(s/\sqrt{n}) \; .$$

Summary 1

Mean μ of normal distribution $N(\mu, \sigma^2)$ with known variance σ^2

- $X_1, \ldots, X_n$ random sample of size n
- confidence level $1 - \alpha$
- sample mean $\bar{x}$
- two-sided confidence interval: $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$
- upper bound: $\bar{x} + z_{\alpha}(\sigma/\sqrt{n})$
- lower-bound: $\bar{x} - z_{\alpha}(\sigma/\sqrt{n})$

Mean μ of normal distribution $N(\mu, \sigma^2)$ with unknown variance σ^2

- $X_1, \ldots, X_n$ random sample of size n
- confidence level $1 - \alpha$
- sample mean $\bar{x}$ and sample variance s^2
- two-sided confidence interval: $\bar{x} \pm t_{\alpha/2}(n-1)(s/\sqrt{n})$
- upper bound: $\bar{x} + t_{\alpha}(n-1)(s/\sqrt{n})$
- lower-bound: $\bar{x} - t_{\alpha}(n-1)(s/\sqrt{n})$

Summary 2

Mean μ of distribution of known variance σ^2

- large random sample $X_1, \ldots, X_n$ of size $n \geq 30$
- confidence level $1 - \alpha$
- sample mean $\bar{x}$
- approximate two-sided confidence interval: $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$
- approximate upper-bound: $\bar{x} + z_{\alpha}(\sigma/\sqrt{n})$
- approximate lower-bound: $\bar{x} - z_{\alpha}(\sigma/\sqrt{n})$

Mean μ of distribution of unknown variance σ^2

- large random sample $X_1, \ldots, X_n$ of size $n \geq 30$
- confidence level $1 - \alpha$
- sample mean $\bar{x}$ and sample variance s^2
- approximate two-sided confidence interval: $\bar{x} \pm z_{\alpha/2}(s/\sqrt{n})$
- approximate upper-bound: $\bar{x} + z_{\alpha}(s/\sqrt{n})$
- approximate lower-bound: $\bar{x} - z_{\alpha}(s/\sqrt{n})$