Statistical Methods MATH2715 info

Teaching material is all online!

- On Minerva `http://minerva.leeds.ac.uk`
- On GitHub
  https://github.com/luisacutillo78/Statistical-Methods-Lecture-Notes

Resources

- Mathematical Statistics and Data Analysis - 3rd ed. (by J. A. Rice);
- http://www1.maths.leeds.ac.uk/statistics/R/Rintro.pdf;
- https://www.datacamp.com/courses/free-introduction-to-r.

Where We've Been, Where We're Going

In the previous Lecture

- Random Samples from Normal Distributions
- Socrative Quiz

Today

- Confidence intervals
- examples and exercises at the whiteboard

Why do we need confidence intervals?

EXAMPLE

- Assume we are given a set of data from a normal distribution
- we wish to find a point estimate of the mean μ.
- We have seen that $\overline{X}$ is an obvious candidate

Questions

We also need to know what is the likely error range. What If we had a different set of data? How reliable is our estimate, can we trust it? To within what error bounds? We need some theory, making use of the previous lectures!

Confidence Intervals

Definition

A $100(1-\alpha)\%$ confidence interval for an unknown parameter θ is defined as the random interval

$$(\hat{\theta}_1, \hat{\theta}_2),$$

where $\hat{\theta}_1 = g_1(\underline{X})$ and $\hat{\theta}_2 = g_2(\underline{X})$ are statistics (random variables) such that

$$p(\hat{\theta}_1 < \theta < \hat{\theta}_2) \;\; = \;\; 1-\alpha.$$

Note 1: CI are not unique, since there are infinitely many choices for these random variables.
Note 2: θ is the true parameter value, and is not random. $\hat{\theta}_1 = g_1(\underline{X})$ and $\hat{\theta}_2 = g_2(\underline{X})$ are random variables.
Note 3: Usual value $\alpha = 0.05$; that is, 95% confidence intervals.

Interpretation of a confidence interval

If we have a 95% (*i.e.*, $\alpha = 0.05$) confidence interval for a parameter θ, the interpretation is:

> *If we do many samplings, and for each observed random sample $\underline{x}$ we construct $(g_1(\underline{x}), g_2(\underline{x}))$, we should expect to have the true value θ within this interval* 95% *of the times.*

Usually statistics $\hat{\theta}_1$ and $\hat{\theta}_1$ are both obtained as a function of a point estimator $\hat{\theta}$ of θ.

CI for μ, σ known, using Z

Recall

A *Z-statistic* is a statistic with a standard normal distribution. The main use of Z-statistics stems from the facts that, for a general distribution, the Central Limit Theorem implies asymptotically that

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

and that the standard normal distribution involves **no unknown parameters**.

Confidence interval for μ with σ^2 known

We can use the Z-statistic to calculate a range of plausible values for μ, under the assumption that σ^2 is known!

CI for μ, σ known, using Z

Remembering that $Z \sim N(0,1)$, choose $z_{\alpha/2}$ such that

$$P\left(Z \leq z_{\alpha/2}\right) = 1 - \frac{\alpha}{2} \Longrightarrow P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha.$$

If $Z = \dfrac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}$ as above, then

$$\begin{aligned} P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \leq z_{\alpha/2}\right) &= 1 - \alpha \\ \Longrightarrow \quad P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) &= 1 - \alpha. \end{aligned}$$

Hence the $100\,(1-\alpha)\,\%$ confidence interval is

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

The most common value of α in use is 0.05, in which case $z_{\alpha/2} = z_{0.025} = 1.960$.

CI for μ, σ known, using Z

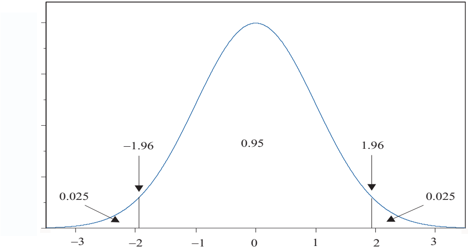

Figure 1 95% interval for $Z \sim N(0,1)$

Note: We could also go backwards, and try to compute the minimum n in order to ensure that the width of the CI is lower than a maximum threshold.

Whiteboard: Examples 1 and 2

CI for μ, σ unknown, using t

We know that if X_1, $X_2, \ldots,$ X_n is iid $N(\mu, \sigma^2)$, then

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} \sim t\,(n-1)$$

We can now look for a confidence interval by replacing the Z-statistic with the t-statistic. Writing $t_{\alpha/2}\,(n-1)$ for the $1 - \frac{\alpha}{2}$ quantile from the distribution $t(n-1)$,

$$P\left(-t_{\alpha/2}\,(n-1) < \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} < t_{\alpha/2}\,(n-1)\right) = 1 - \alpha.$$

Re-arranging gives the random interval

$$\left(\overline{X} - \frac{t_{\alpha/2}}{\sqrt{n}}\,S,\ \overline{X} + \frac{t_{\alpha/2}}{\sqrt{n}}\,S\right),$$

and the $100\,(1-\alpha)\,\%$ confidence interval is the realisation of this interval.

σ^2 unknown

If $X_i \sim N(\mu, \sigma^2)$ with both μ and σ^2 unknown:

- 95% CI for μ: $\bar{X} \pm t_{0.975,n-1} \frac{S}{\sqrt{n}}$

$$T \sim t_{n-1} : \qquad p(T < t_{0.975,n-1}) = 0.975$$

- 95% CI for σ^2: $\left(\frac{(n-1)S^2}{\chi^2_{0.975,n-1}}, \frac{(n-1)S^2}{\chi^2_{0.025,n-1}} \right)$

$$Y \sim \chi^2_{n-1} : \qquad p(Y < \chi^2_{0.975,n-1}) = 0.975$$

Whiteboard: Explain why, and Example 3.

Two Sample Problems

We consider two populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, with two independent random samples. Thus,

$$Var[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

We are interested in inferring how μ_1 and μ_2 compare.

- **Two Means:** Consider the two sample means $\bar{X}_1$ and $\bar{X}_2$:
  - **If σ_1^2 and σ_2^2 known:** Then, a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

    $$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad \text{where} \quad p(Z < z_{1-\frac{\alpha}{2}}) = \left(1 - \frac{\alpha}{2}\right).$$

  - **If $\sigma_1^2 = \sigma_2^2$ unknown:** Then, a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

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

    where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ is the *Pooled Variance*.

Whiteboard: Example 4.