Applied Probability and Statistics for Computer Science

Jun Kong, PhD

Associate Professor, Dept. of Mathematics and Statistics

Announcement

- Quiz 7 will be available from 5:00 PM of Mar 31, 2021 to 11:30 PM of Apr 02, 2021 on iCollege (Assessments -> Quizzes). You will have 90 minutes to complete it. Quiz 7 covers the section 9.1.
- You are encouraged to visit my and TA's online office hours.
- In addition to visiting my and TA's online office hours, you can visit the online STEM Tutoring center (1/19-5/3/2021): `gsu-as.tutorocean.com`. Tutors will be waiting online to serve students. MAC Business Hours: 9 am - 8 pm, Mon - Fri; 11 am - 6 pm, Sat; 12 pm - 6 pm, Sun
- You are always welcome to send me your comments or suggestions about our class.

Outline for Chapter 9: Statistical Inference

- Parameter estimation
- Confidence intervals
- Unknown standard deviation
- Hypothesis testing

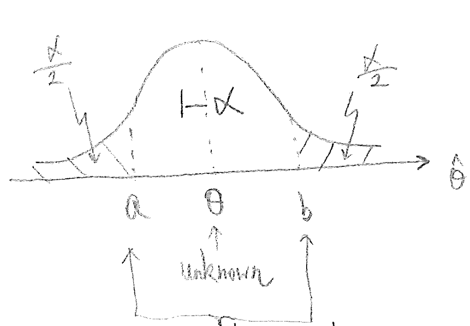
9.2 Confidence Interval: (CI)

Def: An interval $[a, b]$ is $(1-\alpha)100\%$ confidence interval for the parameter θ if $[a, b]$ contains the parameter θ with probability $(1-\alpha)$

i.e. $P(a \le \theta \le b) = \underbrace{1-\alpha}_{\text{confidence level}}$.

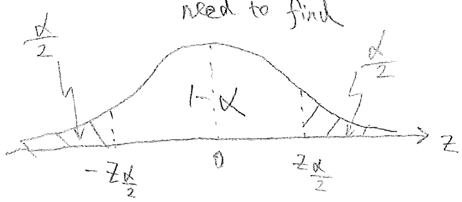
* Question. Given a sample data and a desired confidence level $(1-\alpha)$, how can we construct a confidence interval $[a, b]$ that satisfies $P(a \le \theta \le b) = 1-\alpha$?

1) find an estimator $\hat{\theta}$ of θ using the sample data
2) Suppose $\hat{\theta}$ is an unbiased estimator of θ: i.e. $E\{\hat{\theta}\} = \theta$
3) Suppose $\hat{\theta}$ follows a normal distribution with $\mu = E\{\hat{\theta}\} = \theta$ and $\sigma(\hat{\theta}) = \sqrt{Var\{\hat{\theta}\}}$



$\hat{\theta} \sim Normal(\theta, \sigma(\hat{\theta}))$

$$Z = \frac{\hat{\theta} - E\{\hat{\theta}\}}{\sigma(\hat{\theta})} = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \sim Normal(0, 1)$$



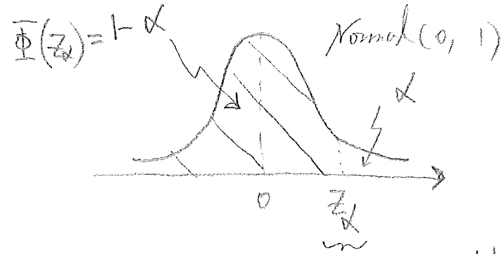
$$P\left(-z_{\frac{\alpha}{2}} \le z = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\underbrace{\hat{\theta} - z_{\frac{\alpha}{2}}\sigma(\hat{\theta})}_{a} \le \theta \le \underbrace{\hat{\theta} + z_{\frac{\alpha}{2}}\sigma(\hat{\theta})}_{b}\right) = 1-\alpha$$

$$\Rightarrow \begin{cases} a = \hat{\theta} - Z_{\frac{\alpha}{2}}\, \sigma(\hat{\theta}) \\ b = \hat{\theta} + Z_{\frac{\alpha}{2}}\, \sigma(\hat{\theta}) \end{cases} \Rightarrow [a,\ b] \text{ is a } (1-\alpha)100\% \text{ confidence interval for } \theta.$$

* If parameter θ has an unbiased, normally distributed estimator $\hat{\theta}$, a $(1-\alpha)100\%$ confidence interval for θ is: $\hat{\theta} \mp Z_{\frac{\alpha}{2}}\, \sigma(\hat{\theta}) = [\hat{\theta} - Z_{\frac{\alpha}{2}}\, \sigma(\hat{\theta}),\ \hat{\theta} + Z_{\frac{\alpha}{2}}\, \sigma(\hat{\theta})]$

* Margin of Error: $\underbrace{Z_{\frac{\alpha}{2}}}_{\text{quantile}}\ \underbrace{\sigma(\hat{\theta})}_{\substack{\text{Standard error} \\ \text{(SE)}}}$
(M.E.)



$(1-\alpha)$ quantile

or $Z_\alpha = \Phi^{-1}(1-\alpha)$

* Confidence interval for the population mean μ. (with σ known)

Given a Sample $X = (X_1, \cdots, X_n)$ from a random variable X, let's construct a confidence interval for the population mean $\mu = E\{X\}$.

1) $\bar{X}$ is unbiase estimator of μ (as $E\{\bar{X}\} = \mu$).

2) If $S = (X_1 \cdots X_n)$ comes from a normal distribution, $\Rightarrow \bar{X}$ is normally distributed.

3) If $S = (X_1, \cdots, X_n)$ comes from any distribution, but sample size n is large,

$\Rightarrow \bar{X}$ is approximately normally distributed due to Central Limit Theorem.

Suppose n is large (i.e. $n \geq 30$)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$E\{S_n\} = n E\{\bar{X}\} = n\mu$

$\sigma^2(S_n) = n^2 \operatorname{Var}\{\bar{X}\}$

$= n^2 \cdot \frac{1}{n^2} \cdot n\sigma^2$

$= n\sigma^2$

$$\Rightarrow \bar{X} \text{ is approximately } \sim \text{Normal}\Big(\mu,\ \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{std}}\Big)$$

$\Rightarrow \bar{X} \mp Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ is a $(1-\alpha)100\%$ C.I. for μ.

$\downarrow \qquad \downarrow$

$\hat{\theta} \mp Z_{\frac{\alpha}{2}}\, \sigma(\hat{\theta})$

$\Rightarrow$ Margin Error (ME) $= Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Ex.[9.13] Construct a 95% CI for the population mean μ based on a sample of measurements

$S = (2.5,\ 7.4,\ 8.0,\ 4.5,\ 7.4,\ 9.2)$

if measurement errors have normal distribution and measurement device guarantees a std. of $\sigma = 2.2$

Ex: Assuming that individual SAT math scores in a state consistently have a normal distribution with $\sigma = 100$. Researchers choose a random sample of 667 exams. $\bar{X} = 488$. Find a 99% CI to estimate the mean SAT math score.

* Selection of a Sample Size:

How large should a sample be so that Margin of Error (M.E.) is at most Δ with a confidence level $(1-\alpha)100\%$?

$$ME = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \Delta \quad \Rightarrow \quad \left(\frac{Z_{\frac{\alpha}{2}}\, \sigma}{\Delta}\right)^2 \le n$$

$\Rightarrow$ In order to have a margin of error Δ for estimating a population mean with a CI level $(1-\alpha)$, a sample size $n \ge \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{\Delta}\right)^2$ is required.

Ex. Suppose that data in a population are normally distributed with $\sigma = 2.2$ and an unknown mean μ. How large a sample do we need to estimate the population mean μ with a margin of error at most 0.4 with 95% confidence?

* Confidence Interval for Difference between Two means.

Population 1: μ_X, σ_X , a Sample $S_X=(X_1,\cdots,X_n)$ with sample mean $\bar{X}$
(μ_X, μ_Y unknown)
Population 2: μ_Y, σ_Y , a Sample $S_Y=(Y_1,\cdots,Y_m)$ with Sample mean $\bar{Y}$

To construct a CI for the difference between two means:

1) a sample is collected from each population.

2) $\hat{\theta} = \bar{X} - \bar{Y}$ is an estimator of $\theta = \mu_X - \mu_Y$

$$E\{\hat{\theta}\} = E\{\bar{X}-\bar{Y}\} = E\{\bar{X}\} - E\{\bar{Y}\} = \mu_X - \mu_Y$$

$\Rightarrow \hat{\theta} = \bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_X - \mu_Y$

3) Suppose that populations are normal or Sample sizes are large.

$\Rightarrow \hat{\theta} = \bar{X} - \bar{Y}$ is normally distributed or approximately normally distributed.

4) $$\sigma(\hat{\theta}) = \sigma(\bar{X}-\bar{Y}) = \sqrt{Var\{\bar{X}-\bar{Y}\}} \overset{\bar{X},\bar{Y}\text{ indep}}{=} \sqrt{Var\{\bar{X}\} + Var\{\bar{Y}\}} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

5) find the quantile $Z_{\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$

$\Rightarrow$ A $(1-\alpha)100\%$ CI for $\theta = \mu_X - \mu_Y$: $\hat{\theta} \mp Z_{\frac{\alpha}{2}}\,\sigma(\hat{\theta})$

$$= (\bar{X}-\bar{Y}) \mp Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

Ex: 9.14. A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 mins before the upgrade, 7.2 mins after it. Standard deviation is 1.8 mins before and after upgrade. Please construct a 90% CI showing how much the mean running time reduced due to the hardware upgrade.