Applied Probability and Statistics for Computer Science

Jun Kong, PhD

Associate Professor, Dept. of Mathematics and Statistics

Announcement

- Quiz 8 will be available on Apr 07, 2021 5:00 PM until Apr 09, 2021 11:30 PMăon iCollege (Assessments -> Quizzes). You will have 90 minutes to complete it. Quiz 8 covers the sections 9.1, and 9.2.
- Grades and Solutions to Test3 are posted to iCollege under Content -> Quiz/Test Solutions. If you have any problem with grades, please kindly contact class TA by 4/7/2021. After that, TA will not respond to the grading related request.
- Solutions to exercise problems for chapter 9 are posted to iCollege under Content -> Exercise Problems and Solutions.
- 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
- Your comments about our class are also welcome.

Outline for Chapter 9: Statistical Inference

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

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

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

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

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

Confidence Interval for population mean with known std

* 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 $(1-\alpha)100\%$?

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

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

* 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 $\overline{X}$

unknown

Population 2: μ_Y, σ_Y , a Sample $S_Y = (Y_1, \cdots, Y_m)$ with Sample mean $\overline{Y}$

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

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

In 9.2, we estimate μ with σ known.

In 9.3, " " " " " unknown.

9.3. Unknown Standard Deviation:

* Large Sample
We can use standard normal distribution to construct a C.I. when:

1) Sample size is large (i.e. $n \geq 30$)

or

2) Population has a normal distribution

Procedure to construct a $(1-\alpha)$ level CI for θ.

1) find an unbiased estimator $\hat{\theta}$ of θ (i.e. $E\{\hat{\theta}\} = \theta$.

2) check if $\hat{\theta}$ has a normal distribution.

3) find $Z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$

4) find $\sigma(\hat{\theta})$

5) a $(1-\alpha)$ level CI: $\hat{\theta} \mp ME = \hat{\theta} \mp Z_{\frac{\alpha}{2}}\, \sigma(\hat{\theta})$

Note: when $\hat{\theta} = \bar{X}$, $\sigma(\hat{\theta}) = \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

When the true standard error is unknown, we'll replace it by its estimator $S(\hat{\theta})$.

E.g. A $(1-\alpha)100\%$ CI for μ from a Sample of size n:

$$\bar{X} \mp ME = \bar{X} \mp Z_{\frac{\alpha}{2}}\, \sigma(\bar{X}) = \bar{X} \mp Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

when σ is unknown, we use $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ to estimate σ

population Standard deviation

Sample standard deviation.

$S(\bar{X})$ is an estimator of $\sigma(\bar{X})$

$S(\bar{X}) = \frac{S}{\sqrt{n}}$, $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Ex. 1) Between 6pm – 7pm, send 500 packets with sample mean delay time 0.8 sec.
Std. 0.1 sec.

2) " 10pm ~ 11pm, " 300 " " " " " " 0.5 sec
Std 0.08 sec.

Does the delay time increase during 6pm ~ 7pm?

* Confidence Intervals for proportions (p).

In some cases, we don't know variance when we estimate a population proportion.

Def. Assuming that there is a subpopulation A of items that have a certain attributes. By the population proportion, we estimate the prob. $P\{i \in A\}$ for a randomly selected item i to have this attribute, i.e. to belong to the subpopulation A.

A sample proportion: $$\hat{p} = \frac{\#\text{ of sampled items in } A}{\text{total }\#\text{ of sampled items } n}$$
is used to estimate the population proportion p.

i.e. Let X be an element in Ω

$P(X \in A) = P(A) = p$ is called population proportion.
(unknown)

Let S be a sample drawn from Ω

$$\hat{p} = \frac{\#\text{ of items in } S \text{ from } A}{\text{size of } S}$$ is sample proportion

$\hat{p}$ is used to estimate p.

Let $X_i = \begin{cases} 1, & \text{if item } i \in A \\ 0 & \text{if item } i \notin A \end{cases}$ where X_i is a R.V. $\sim$ Bernoulli distribution with parameter p.

PMF of X_i:

$$f(X_i) = \begin{cases} P(X_i=1) = P(i \in A) = p \\ P(X_i=0) = P(i \notin A) = 1-p \end{cases} \Rightarrow \begin{array}{l} E\{X_i\} = p \\ Var\{X_i\} = p(1-p) \end{array}$$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad : \quad E\{\hat{p}\} = p$$
$$Var\{\hat{p}\} = \frac{Var\{X_i\}}{n} = \frac{p(1-p)}{n}$$

↑ Sample mean

Note. 1) $\hat{p}$ is unbiased estimator of p

2) $\hat{p}$ has a form of sample mean. when n is large, $\hat{p}$ has approximately normal distribution

3) when n is large, $S(\hat{\theta}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx \sigma(\hat{\theta}) = \sqrt{\frac{p(1-p)}{n}}$

$\Rightarrow$ For a large n, an approximate $(1-\alpha)$ level CI for p:

$$\hat{p} \mp ME = \hat{p} \mp Z_{\frac{\alpha}{2}} \cdot \underbrace{\sigma(\hat{p})}_{\frac{p(1-p)}{n}} \approx \hat{p} \mp Z_{\frac{\alpha}{2}} \cdot S(\hat{p}) = \hat{p} \mp Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

An approximate $(1-\alpha)$ level CI for difference between two population proportions $p_1 - p_2$ based on a large # of samples is: $(\hat{p}_1 - \hat{p}_2) \mp Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$

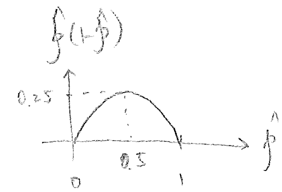
Ex. 9.17. 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B show they'd vote for a candidate. Estimate difference in support that this candidate is getting in towns A and B with 95% confidence. Can we state firmly that the candidate gets a stronger support in town A?

* Sample Size n for a Margin of Error at most Δ.

$$M.E. = Z_{\frac{\alpha}{2}} \cdot \sigma(\hat{p}) \le \Delta$$

$$\Rightarrow Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \Delta$$

$$\Rightarrow n \geq \frac{Z^2_{\frac{\alpha}{2}} \hat{p}(1-\hat{p})}{\Delta^2}$$



As the max value of $\hat{p}(1-\hat{p})$ is 0.25 when $\hat{p}$.

$$\Rightarrow \underbrace{n \geqslant 0.25 \left(\frac{Z_{\frac{\alpha}{2}}}{\Delta}\right)^2}_{\text{provide a sample size corresponding to a desired Margin of Error } \Delta \text{ at } (1-\alpha) \text{ Confidence Level.}} \geqslant \frac{Z^2_{\frac{\alpha}{2}} \hat{p}(1-\hat{p})}{\Delta^2}$$

* Small Samples: Student's T distribution.

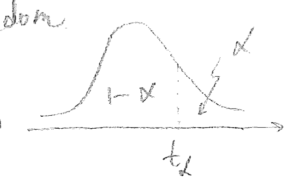
Note:

1) If n is large, (≥ 30) $\sigma(\hat{\theta}) \approx S(\hat{\theta})$ $\Rightarrow$ $(1-\alpha)$ Level CI: $\hat{\theta} \mp Z_{\frac{\alpha}{2}} \sigma(\hat{\theta}) \approx \hat{\theta} \mp Z_{\frac{\alpha}{2}} S(\hat{\theta})$

2) If n is not large (i.e. $n < 30$), $\sigma(\hat{\theta}) \approx S(\hat{\theta})$ may NOT be true.

We need to adjust $(1-\alpha)$ level CI by: $\hat{\theta} \mp t_{\frac{\alpha}{2}} S(\hat{\theta})$

where $t_{\frac{\alpha}{2}}$ (critical value) comes from T-distribution table (Table A5) with $(n-1)$ degrees of freedom (d.f.)



3) T-distribution Table shows values of the T-ratio. $t = \frac{\hat{\theta}-\theta}{S(\hat{\theta})}$ with $(n-1)$ d.f. $\sim$

When sample size is small, t is no longer normally distributed.

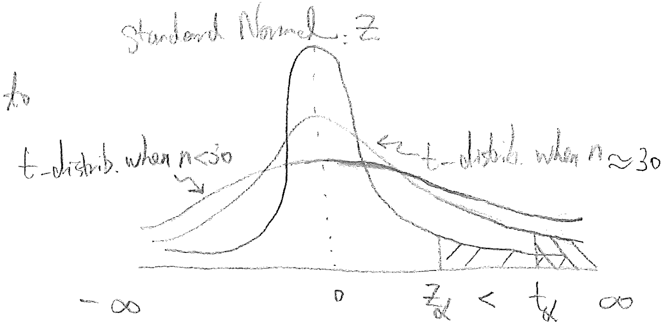
Note: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})$ $X = (X_1, \cdots, X_n)$ has dimension n.

As $\sum_{i=1}^{n} (X_i - \bar{X}) = 0$, there is a linear relationship among elements $X' = (X_1 - \bar{X}, \cdots, X_n - \bar{X})$

$\Rightarrow$ X' has $(n-1)$ dimension.

* T – distribution:

1) PDF has a bell-shaped curve that is symmetric to mean.



2) $Z_\alpha < t_\alpha$

3) As $n \to \infty$, $t_\alpha \to Z_\alpha$

$(1-\alpha)$ level CI for μ based on a sample of size $n < 30$:

$$\bar{X} \mp ME \cong \bar{X} \mp t_{\frac{\alpha}{2}}\, \sigma(\bar{X})$$

$$= \bar{X} \mp t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\cong \bar{X} \mp t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

where $t_{\frac{\alpha}{2}}$ is the critical value from T-distrib. with $(n-1)$ d.f.

9.19.

Ex: Can we detect an unauthorized person accessing an account with a stolen password? The following time (in seconds) have been recorded when a user typed a username and password.

0.24	0.33	0.17
0.22	0.29	0.28
0.26	0.19	0.38
0.34	0.36	0.40
0.35	0.30	0.37
0.32	0.15	0.27

Construct a 99% CI for the mean time between the keystrokes assuming normal distribution of these times.

Comparison of two populations with unknown variances: (Sample size < 30)

Population 1: μ_X, σ_X^2 (unknown)

$S_1 = (X_1, \cdots, X_n)$

Population 2: μ_Y, σ_Y^2 (unknown)

$S_2 = (Y_1, \cdots, Y_m)$

Case 1: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (unknown)

We use data in both samples to estimate σ

Pooled Sample variance: $$S_p^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

$\Rightarrow$ $(1-\alpha)$ level CI for the difference of means $\mu_X - \mu_Y$ with $\sigma_X^2 = \sigma_Y^2$ unknown is:

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

$$\approx (\bar{X}-\bar{Y}) \mp t_{\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n} + \frac{S_p^2}{m}}$$

$$= (\bar{X}-\bar{Y}) \mp t_{\frac{\alpha}{2}}\, S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

where: $S_p^2 = \dfrac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$, $t_{\frac{\alpha}{2}}$ is a critical value from T-distribution with $(n+m-2)$ degrees of freedom.

Ex: 9.20. CD writing affects battery lifetime on laptops. To estimate the effect of CD writing, 30 users are asked to work on their laptops until the "low battery" sign comes on. 18 users without CD writers worked an average of 5.3 hours, with a standard deviation of 1.4 hours. 12 used CD writers and worked an average of 4.8 hours with a standard deviation of 1.6 hours. Assuming Normal distributions with equal population variances ($\sigma_X^2 = \sigma_Y^2$), construct a 95% confidence interval for the battery life reduction caused by CD writer.

Case 2: $\sigma_X^2 \neq \sigma_Y^2$

T-ratio: $t = \dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{S_X^2}{n}+\dfrac{S_Y^2}{m}}}$ does Not follow a T-distrib.

(*) $\nu = \dfrac{\left(\dfrac{S_X^2}{n}+\dfrac{S_Y^2}{m}\right)^2}{\dfrac{S_X^4}{n^2(n-1)}+\dfrac{S_Y^4}{m^2(m-1)}}$ is the degree of freedom of a T-distrib. that is "closest" to the T-raio t.

$\Rightarrow$ A $(1-\alpha)$ level CI for $\mu_X-\mu_Y$ is:

$$(\bar{X}-\bar{Y}) \mp \underbrace{t_{\frac{\alpha}{2}}}_{\text{with d.f. } \nu \text{ defined in } (*)} \sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}$$

where $t_{\frac{\alpha}{2}}$ is a critical value from a T-distribution with d.f. ν

Ex: 9.21. An account on Server A is more expensive, but faster than an account on Server B. A certain Algorithm is executed 30 times on Server A and 20 times on B with the following results:

	Server A	Server B
Sample mean	$\bar{X}=6.7$ min	$\bar{Y}=7.5$ min
Sample std.	$S_X=0.6$ min	$S_Y=1.2$ min

Construct a 95% CI for $\mu_X-\mu_Y$ between the mean execution times on Server A and B, Assuming observed time are approximately Normal.