STA255 Week-9 (Day-1)

Shahriar Shams

09/03/2020

Review of Week-8

- Idea of interval estimation
- Definition of confidence interval
  - 100*$(1-\alpha)$% CI for $\mu \implies P[l(X_1, ...X_n) \leq \mu \leq u(X_1, ...X_n)] = 1-\alpha$
- Idea of Pivotal Quantity
- Confidence interval(CI) for μ with
  - population variance (σ^2) known (using Z dist.)
  - population variance (σ^2) unknown (using t dist.)
- Two sided vs one sided CI
- Interpretation of confidence interval
  - Interpretation does not involve the two numeric numbers that we calculate.

Learning goals

- Confidence interval for population proportion
- Sample size calculation

Confidence interval for population proportion

- Last week we constructed confidence interval for population mean (μ).
- We used either $\bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ or $\bar{X} \pm t_{1-\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$
- One way to look at this:
  - we use sample mean ($\bar{X}$) as a point estimator
  - WE calculated the standard error of the estimator, $SE[\bar{X}] = \sqrt{var[\bar{X}]} = \frac{\sigma}{n}$
  - We wrote 100*$(1-\alpha)$% CI as

$$point\ estimator \pm z_{1-\frac{\alpha}{2}}\ standard\ error$$

- Suppose p is the proportion of "success" in a population.
  - "success" represent one of the outcomes of a Bernoulli trial.
- The goal is to construct a confidence interval for p.

- Suppose we take a sample of n observation and find X number of success out of n.
- Then $\hat{p} = \frac{X}{n}$ is a natural estimate/estimator of p.

- Since X is the number of success in n trials with probability of success being p in each trial,

$$X \sim Bin(n, p)$$

- We know, $E[X] = np$ and $V[X] = np(1-p)$
- Therefore, $E[X/n] = p$ and $V[X/n] = \frac{p(1-p)}{n} \implies SE[X/n] = \sqrt{\frac{p(1-p)}{n}}$
- Since, p is unknown, it's not possible to calculate the standard error of X/n
- One way to estimate the standard error is to replace p by the sample proportion ($\hat{p}$)
- This gives us $estimated\, standard\, error = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- We can write,

$$P[z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}] \leq z_{1-\frac{\alpha}{2}}] = 1 - \alpha$$

- Re-arranging the above expression we get the 100*$(1-\alpha)$% CI for p as

$$\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \; \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sample Size calculation

Width of an interval

- Let l and u be the lower and upper value of an interval.
- Width of an interval $= u - l$

- For CI for μ with σ known, $\text{width} = 2 * z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Margin of error

- Keeping the point estimator in the middle adding or subtracting a quantity gives us confidence intervals.
- The quantity that we add or subtract is called the margin of error.
- For CI for μ with σ known, $\text{margin of error} = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Sample size as a function of width or margin or error

- We can see, width = 2* margin of error
- If margin of error increases, width will also increase.

- If we look at the expression of the margin of error(ME) we will see
  - $\sigma \uparrow \implies ME \uparrow$
  - $(1-\alpha) \uparrow \implies ME \uparrow$
  - $n \uparrow \implies ME \downarrow$

- A typical sample size calculation involves an statement like, "If we want the 95% CI for μ to have a width no longer 10, how many samples should we collect given that the population standard deviation is known to be 25".
- From this statement, $\sigma = 25$, $z_{0.975} = 1.96$ and $width < 10$
- We get an inequality,

$$2 * z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < 10$$

- Re-arranging this we get, $n > (\frac{2*1.96*25}{10})^2 \implies n > 96.04 \implies n \approx 97$

Sample size calculation for population proportion

- Sample size calculation involves plugging in a value of the population standard deviation.
- When it's unknown, we put a reasonably large value of sigma which is also known as "conservative" approach as this gives us the widest interval.

- As we have seen on page-2 of this document, CI for proportion involves the unknown parameter p in the standard error expression.
- Standard error involves $p(1-p)$. Since p is bounded between 0 and 1, it is possible to get a sense of the maximum value of $p(1-p)$.
- It can be shown that the maximum value of $p(1-p) = 1/4$ which is resulted when $p = 1/2$
- This will give us the widest interval.
- A typical sample size calculation for proportion involves a statement like, "If we want the 95% CI for the true proportion to have a width no longer than 0.1, how many samples we should collect?
- For this we get the inequality,

$$2 * z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} < 0.1$$

- Replacing $p(1-p)$ by 1/4 we get,

$$2 * z_{1-\frac{\alpha}{2}} \sqrt{\frac{1/4}{n}} < 0.1 \implies n > (\frac{z_{1-\frac{\alpha}{2}}}{0.1})^2 \implies n > (\frac{1.96}{0.1})^2 \implies n \approx 385$$

Chapter 8.4 and 8.5 not needed.

Homework

Chapter 8.1

4(e), 5(c,d), 7

Chapter 8.2

20, 21, 23, 25

Chapter 8.3

33, 38