What to do today (Jan 24, 2023)?

Part 2. Basic Statistical Inference (Chp 6-9)

Reminder: Homework 3 is due on Monday 5:00pm.

§2.2 Interval Estimation (Chp 7)

§2.2.1 Some General Concepts (Chp 7.1)

Goal: Suppose $X \sim F(\cdot;\theta)$ and $X_1,\ldots,X_n$ iid observations from the population. To obtain a 'good' interval estimator of θ.

Suppose $X \sim F(\cdot;\theta)$ and $X_1,\ldots,X_n$ iid observations from the population. What is a 'good' interval estimator?

$100(1-\alpha)\%$ **confidence interval of** θ:

- **Definition**. $\hat{\theta}_L$ and $\hat{\theta}_U$ are two statistics. The random interval $\big(\hat{\theta}_L, \hat{\theta}_U\big)$ is a $100(1-\alpha)\%$ CI of θ is

$$P\big(\theta \in (\hat{\theta}_L, \hat{\theta}_U)\big) = 100(1-\alpha)\%.$$

  Here, $(1-\alpha)$ is called the confidence level of the CI.

  - eg, $\alpha = 0.05$, a $100(1-\alpha)\%$ CI of θ is a CI with confidence level of 95%.
  - eg, in **Example 3.1**, $\mu = \theta$ and $\alpha = 5\%$:

$$\begin{array}{cc} \text{lower limit } \hat{\theta}_L & \text{upper limit } \hat{\theta}_U \\ \Big(\bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, & \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) \end{array}$$

- **Interpretation**. (frequentist)
  With 100 experiments' outcomes, there're at least $100(1-\alpha)$ out of the 100 CI realizations containing the true value of θ.

  Bayesian interpretation: different!

- **Confidence Level, Precision, and Sample Size**.
  - $100(1-\alpha)\%$ CI $(\hat{\theta}_L, \hat{\theta}_U)$: the confidence level is $1-\alpha$.

    $$P\big(\theta \in (\hat{\theta}_L, \hat{\theta}_U)\big) = 1-\alpha$$

  - Length (Width) of CI: $\hat{\theta}_U - \hat{\theta}_L$, about CI's **precision/accuracy**.

  - Often to determine the sample size n such that a $1-\alpha$ CI has a desired precision $\Rightarrow$ **Study Design** eg, recall **Example 3.1**, $(1-\alpha)$ CI of μ with $\sigma = 2$ in $N(\mu, \sigma^2)$:

$$\left(\bar{X} - Z_{1-\alpha/2}\frac{2}{\sqrt{n}}, \quad \bar{X} + Z_{1-\alpha/2}\frac{2}{\sqrt{n}}\right)$$

- If $1-\alpha = 95\%$, the CI is $\left(\bar{X} - 1.96\frac{2}{\sqrt{n}}, \bar{X} + 1.96\frac{2}{\sqrt{n}}\right)$ with length $2(1.96)(2/\sqrt{n})$.
- If n increases, the length decreases.
- If the desired length is A, the required sample size should be $\{(7.84)/A\}^2$

Example 3.2 (p282)

- **Study.** To determine the true average response time of a new operating system. What sample size is necessary to ensure the resulting 95% CI has a width of (at most) 10?
- **Formulation.** Assuming a response time $X \sim N(\mu, \sigma^2)$ with $\sigma = 25$. To obtain a 95% CI of μ with length ≤ 10
- **Interval estimator.** $\left(\bar{X} - 1.96\frac{25}{\sqrt{n}}, \bar{X} + 1.96\frac{25}{\sqrt{n}}\right)$.
- **Sample size determination.** The length $2(1.96)(25/\sqrt{n})$ is to be at most 10:

$$2(1.96)(25/\sqrt{n}) \leq 10.$$

Thus $\sqrt{n} \geq 2(1.96)(25)/10 = 9.80$. So, n should be at least 97 ($9.80^2 = 96.04$).

§2.2.2 Confidence Intervals Based on a Normal Population

Framework.

- **Population**. $X \sim N(\mu, \sigma^2)$;
- **Data**. a random sample $\{X_1, \ldots, X_n\}$
- **Goal**. **to construct CI of μ and σ^2**

§2.2.2A CI of μ when σ^2 is known:
Recall from **Example 3.1** that a $1 - \alpha$ CI of μ with known σ^2 is

$$\left(\bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

with $P(Z < Z_{1-\alpha/2}) = 1 - \alpha/2$ if $Z \sim N(0, 1)$.

- **Rationale.** $\bar{X} \sim N(\mu, \sigma^2/n)$ and

$$\begin{aligned} & P\big(\bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\big) \\ = \quad & P\big(|\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}| \le Z_{1-\alpha/2}\big) = 1-\alpha \end{aligned}$$

- **Evaluation.** Plug in the CI with (a) $\bar{X}$, (b) $Z_{1-\alpha/2}$, (c) n and (d) σ^2.

- **Other** $1-\alpha$ **CI of** μ**.**
  - Is $\big(\bar{X} - Z_{1-\alpha/3}\frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_{1-2\alpha/3}\frac{\sigma}{\sqrt{n}}\big)$ another $1-\alpha$ CI? If so, which one is better?
  - How about $\big(-\infty,\ \bar{X} + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\big)$ or $\big(\bar{X} - Z_{1-\alpha}\frac{\sigma}{\sqrt{n}},\ \infty)$? $1-\alpha$ *confidence upper/lower bound of* μ

§2.2.2B CI of μ when σ^2 is unknown

Recall a $1-\alpha$ CI of μ with known σ^2

$$\left(\bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

It is not applicable when σ^2 is unknown!

- *Using an estimator to substitute σ^2 in the CI?*
- *If so, is the resulting estimator still $1-\alpha$ CI?*

$$\begin{aligned} & P\left(\bar{X} - Z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + Z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right) \\ = \quad & P\left(\left|\frac{\bar{X}-\mu}{\sqrt{\hat{\sigma}^2/n}}\right| \le Z_{1-\alpha/2}\right) = 1-\alpha \quad ??? \end{aligned}$$

- *How about using $\hat{\sigma}^2 = S^2$, and replacing $Z_{1-\alpha/2}$ with $t_{\alpha/2}(n-1)$?*

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

for $(\bar{X}-\mu)/\sqrt{S^2/n} \sim t(n-1)$.

A $1-\alpha$ CI of μ:

$$\left(\bar{X} - t_{\alpha/2}(n-1)\sqrt{\frac{S^2}{n}}, \quad \bar{X} + t_{\alpha/2}(n-1)\sqrt{\frac{S^2}{n}}\right)$$

Notation.

S^2 – the sample variance;

$t_{\alpha/2}(n-1)$ – the $\alpha/2$-right tail of the Student t-distribution with $\mathrm{df} = n-1$.

Why?

Any theoretical justification?

Review 2 in Part 1: Sampling Distributions

- **Some distributions derived from** $N(\mu, \sigma^2)$

Chi-Square Distribution. Suppose $Z_1, \ldots, Z_K$ are i.i.d. with $N(0, 1)$. Let W be $W = Z_1^2 + \ldots + Z_K^2$. The distribution of W is the chi-square distribution with the degrees of freedom (df) K, denoted by $W \sim \chi^2(K)$.

Properties:

(i) $\mathrm{E}(W) = K$.

(ii) $\mathrm{V}(W) = 2K$.

(iii) If $W_1 \sim \chi^2(K_1)$, $W_2 \sim \chi^2(K_2)$ and W_1 and W_2 are independent, then $W_1 + W_2 \sim \chi^2(K_1 + K_2)$. (*why?*)

How to obtain relevant values of $\chi^2(\cdot)$?

- χ^2*-distribution table:* Table A.7 Critical Values of Chi-Square Distribution
- *Alternatively, using* `R` *function*:

```
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p =
FALSE)
```

Student's t-Distribution. Suppose $Z \sim N(0,1)$ and $W \sim \chi^2(K)$, and $Z \perp W$. Let T be

$$T = \frac{Z}{\sqrt{W/K}}.$$

The distribution of T is the t-distribution with K degrees of freedom (df): $T \sim t(K)$. *It was initially derived by Gosset (1908).*

Properties:

(i) $\mathrm{E}(T) = 0$.

(ii) $\mathrm{V}(T) = K\big/(K-2)$, if $K > 2$.

(iii) If $T \sim t(K)$ with $K >> 1$, T's distribution is approximately $N(0,1)$. That is $t(\infty) = N(0,1)$. (*why?*)

How to obtain relevant values of $t(\cdot)$?

- *Student's t-distribution table.* Table A.5 Critical Values for t-Distributions
- *Alternatively, using* `R` *function*:

```
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
```

$\Longrightarrow$ **the theoretical support to the CI contruction:**

$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1)$

Example 3.3 (p299)

- **Study**. Consider hot dog's fat content (in %): what is the average hot dog fat content? what is the fat content of the next hot dog?

- **Data**. The fat contents of $n = 10$ randomly selected hot dogs and $\bar{x} = 21.90$.

- **Formulation**. The fat content of a hot dog, r.v. $X \sim N(\mu, \sigma^2)$. What is μ? What is x_{10+1}?

- **Estimation**. $\bar{X}$: the best unbiased estimator of μ; $\bar{x} = 21.90$

  95% CI of μ: $\left(\bar{X} - t_{\alpha/2}(n-1)\sqrt{\frac{S^2}{n}}, \quad \bar{X} + t_{\alpha/2}(n-1)\sqrt{\frac{S^2}{n}}\right)$.
  Its evaluation is $(18.94, 24.86)$ because of (i) $n = 10$ (ii) $\alpha = 0.05$ (iii) $t_{0.025}(9) = 2.262$ and $s = 4.134$.

- **Prediction**.

Thinking ... $X_{n+1} \sim N(\mu, \sigma^2)$ and

$$\bar{X} - X_{n+1} \sim N\big(0, \sigma^2(1 + 1/n)\big)$$

$$Z = \frac{\bar{X} - X_{n+1}}{\sqrt{\sigma^2(1 + 1/n)}} \sim N(0, 1)$$

$$T = \frac{\bar{X} - X_{n+1}}{\sqrt{S^2(1 + 1/n)}} \sim t(n - 1).$$

Here comes ... ... **a prediction interval (PI) with level of $1 - \alpha$:**

$$\Big(\bar{X} - t_{\alpha/2}(n-1)S\sqrt{1 + \frac{1}{n}}, \quad \bar{X} + t_{\alpha/2}(n-1)S\sqrt{1 + \frac{1}{n}}\Big)$$

$\Rightarrow$ a 95% PI for the next hot dog's fat content is $(12.09, 31.71)$.
much wider than the 95% *CI of* μ*! Why?*

2.2.2C Confidence interval of σ^2

Recall we constructed CI of μ based on $\bar{X}$'s distribution.

- What is S^2's distribution?

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

- That is

$$P\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha$$

Notation:
If $W \sim \chi^2(\nu)$, the critical value $\chi^2_\beta(\nu)$ satisfies
$P(W > \chi^2_\beta(\nu)) = \beta$.

- So that

$$P\Big(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\Big) = 1-\alpha$$

Here come ...

a 95% **CI of** σ^2:

$$\Big(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\quad \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\Big)$$

a 95% **CI of** σ:

$$\Big(\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}},\quad \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}}\Big)$$

Example 3.4 (p305-306)

- **Study**. on breakdown voltage: what is its variation?
- **Data**. $n = 17$ observations collected, with $s^2 = 137324.3$
- **Stats formulation**. A breakdown voltage $X \sim N(\mu, \sigma^2)$. What is σ^2?
- **Estimation**. 95% CI:

$$\Big(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \quad \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\Big)$$

Its evaluation with the given s^2 and
$n - 1 = 16$, $\chi^2_{0.025}(16) = 28.845$, $\chi^2_{0.975}(16) = 6.908$
is $(76172.3, 318064.4)$.

Rather wide because of a small sample size!

What will we do next?

Part 1. Introduction and Review (Chp 1-5)

Part 2. Basic Statistical Inference (Chp 6-9)

2.1 Point Estimation (Chp 6)

2.2 Confidence Interval (Chp 7)

- *2.2.1 Some General Concepts*
- *2.2.2 Confidence Intervals with Normal Population*
- **2.2.3 Approximate Confidence Intervals**

2.3 One-Sample Test (Chp 8)

2.4 Inference Based on Two-Samples

Part 3. Important Topics in Statistics (Chp 10-13)

Part 4. Further Topics (Selected from Chp 14-16)

Homework 3 is due on Monday 5:00pm.