# Homework 4

S1201, Summer 2022

100 points

Due Aug 9, 11:59PM EST

**Notes:** Please read the following instructions.

(i) Please submit a **single** PDF file on courseworks. You may scan your written solutions or directly write the solutions on your tablet.

(ii) Late submission will lead to some penalty: 20 points off for within 24 hours late, 40 points off for over 24 hours late, all points off for submission after the solutions are posted on courseworks.

(iii) You can discuss the problems with others. But a direct plagiarism will lead to zero point for this assignment, and this will be reported to the university.

(iv) There is one extra problem with 9 bonus points, which is optional. Therefore all points add up to 109, but the maximum score of this homework is 100.

1. **$2 \times 5 = 10$ points** TRUE/FALSE questions. No explanations are needed.

   (a) If $X_1, \ldots, X_n$ is a random sample from some distribution with population mean $\mu$, then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is not an unbiased estimator of $\mu$. F

   (b) If $X_1, \ldots, X_n$ is a random sample from some distribution with population variance $\sigma^2$, then the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of $\sigma^2$. T

   (c) Suppose we observe some data from a normal distribution with unknown mean $\mu$, and we have calculated that a 95% confidence interval of $\mu$ is $[5.04, 6.16]$. Then it means the probability that $\mu$ falls into $[5.04, 6.16]$ equals 0.95. F

   (d) For the same problem, the same parameter of interest, and the same confidence level, we can construct more than one confidence interval. T

   (e) For the same problem, with the same confidence interval construction and a fixed confidence level, it is possible to decrease the sample size and the CI width at the same time. F

# Homework 4

100 points S1201, Summer 2022

Due Aug 9, 11:59PM EST

2. **$2 \times 5 = 10$ points** Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with $\sigma > 0$ known to us. Let $\alpha_1 > 0$, $\alpha_2 > 0$, with $\alpha_1 + \alpha_2 = \alpha$. Then we know

$$\mathbb{P}\left(-z_{\alpha_1} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha_2}\right) = 1 - \alpha.$$

   (a) Use this equation to derive a more general expression for a $100(1-\alpha)\%$ CI for $\mu$ of which the interval $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$ (which we covered in class) is a special case.

   **Hint:** In other words, find a $100(1-\alpha)\%$ CI via the equation above (may depend on $z_{\alpha_1}$, $z_{\alpha_2}$, $\sigma$, ...)

   (b) Let $\alpha = 0.05$ and $\alpha_1 = \alpha/4$, $\alpha_2 = 3\alpha/4$. Does this result in a narrower or wider interval than the interval the interval $\underline{\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]}$? (**)

   **Hint:** Plug in the $\alpha$, $\alpha_1$, and $\alpha_2$, calculate and compare the width of two CIs.

Solution: (a) $1-\alpha = \mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha_2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha_1}\right)$

$\Rightarrow 100(1-\alpha)\%$ CI : $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha_2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha_2}\right]$ (*)

(b) Width of (*) $= \frac{\sigma}{\sqrt{n}} z_{0.0375} + \frac{\sigma}{\sqrt{n}} z_{0.0125}$

$\approx \frac{\sigma}{\sqrt{n}} (1.78 + 2.24)$

$= \frac{\sigma}{\sqrt{n}} \times 4.02.$

Width of (**) $= \frac{2\sigma}{\sqrt{n}} z_{0.025}$

$= \frac{\sigma}{\sqrt{n}} \times 3.92$

$<$ width of (*).

$\Rightarrow$ (*) is wider than the "symmetric" CI (**)

$\square$

3. **$2 \times 5 = 10$ points** A state legislator wishes to survey residents of her district to see what proportion of the electorate is aware of her position on using state funds to pay for abortions.

   (a) What sample size is necessary if the 95% CI for $p$ is to have a width of at most 0.10 irrespective of $\hat{p}$?

   (b) If the legislator has strong reason to believe that at least 2/3 of the electorate know of her position (i.e. the sample mean $\hat{p} \geq 2/3$), how large a sample size would you recommend?

   **Hint**:

   (1) See the example "The college dean ..." on slide 23 of Lecture 13.

   (2) Comment from Ye: Part (b) looks very weird... You can plug $\hat{p} = 2/3$ into the inequality in part (a). The take-home message is the sample size you get from (b) is less than the number in part (a), because we are considering the worst case in part (a).

Solution: (a) $95\%\ CI = \left[\hat{p} - z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

$\text{width} = 2 z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.1$

$\Rightarrow n \geq 39.2^2 \times \hat{p}(1-\hat{p})$ for $\forall \hat{p} \in [0, 1]$

By quadratic function properties: $\max_{\hat{p} \in [0,1]} \hat{p}(1-\hat{p}) = 0.5 \times 0.5 = 0.25$ (when $\hat{p} = 0.5$)

$\Rightarrow n \geq 39.2^2 \times 0.25 = 384.16$

$\Rightarrow n$ needs to be at least 385.

(b) $n \geq 39.2^2 \times \frac{2}{3} \times (1 - \frac{2}{3}) = 341.48$

$\Rightarrow n$ needs to be at least 342. $\square$

4. **$3 \times 4 = 12$ points** A CI is desired for the true average stray-load loss $\mu$ (watts) for a certain type of induction motor when the line current is held at 10 amps for a speed of 1500 rpm. Assume stray-load loss is normally distributed.

   (a) Suppose we know the true standard deviation of stray-load loss $\sigma = 3$. Compute a 95% confidence interval when sample size $n = 25$ and sample mean $\bar{x} = 58.3$.

   (b) Suppose we know the true standard deviation of stray-load loss $\sigma = 3$. Compute a 95% confidence interval when sample size $n = 100$ and sample mean $\bar{x} = 58.3$.

   (c) Suppose the true standard deviation is unknown. Compute a 95% confidence interval when sample size $n = 100$, sample mean $\bar{x} = 58.3$ and the sample variance $s^2 = 10.1$.

   **Hint:** You shall use different CI formulas in part (a)-(b) and part (c), which depend on whether $\sigma$ is known or not.

Solution:

(a) $\left[\bar{x} - \frac{\sigma}{\sqrt{n}} z_{0.025},\ \bar{x} + \frac{\sigma}{\sqrt{n}} z_{0.025}\right]$

$= \left[58.3 - \frac{3}{5} \times 1.96,\ 58.3 + \frac{3}{5} \times 1.96\right]$

$= [57.124,\ 59.476]$

(b) $\left[\bar{x} - \frac{\sigma}{\sqrt{n}} z_{0.025},\ \bar{x} + \frac{\sigma}{\sqrt{n}} z_{0.025}\right]$

$= \left[58.3 - \frac{3}{10} \times 1.96,\ 58.3 + \frac{3}{10} \times 1.96\right]$

$= [57.712,\ 58.888]$

(c) $n = 100 > 30 \Rightarrow$ We can replace t-values with z-values.

$\left[\bar{x} - \frac{s}{\sqrt{n}} z_{0.025},\ \bar{x} + \frac{s}{\sqrt{n}} z_{0.025}\right]$

$= \left[58.3 - \frac{3.178}{10} \times 1.96,\ 58.3 + \frac{3.178}{10} \times 1.96\right]$

$= [57.677,\ 58.923]$ $\square$.

5. **$4 \times 2 = 8$ points** Determine the $t$ critical value (i.e. the $t$-distribution quantile used to construct CI) for a confidence interval in each of the following situations:

   (a) Confidence level = 95%, degree of freedom = 10

   (b) Confidence level = 95%, sample size $n = 10$

   (c) Confidence level = 99%, degree of freedom = 6

   (d) Confidence level = 99%, sample size $n = 39$

   **Hint:** In other words, find the $t_{n-1,\alpha/2}$ value for each part. Note that $t_{n-1,\alpha/2}$ is the upper-tail quantile that satisfies $\mathbb{P}(T > t_{n-1,\alpha/2}) = \alpha/2$, where $T \sim t_{n-1}$ distribution. Look up the t-table on courseworks.

Solution: (a) $t_{10,\,0.025} = 2.228$

(b) $t_{9,\,0.025} = 2.262$

(c) $t_{6,\,0.005} = 3.707$

(d) $t_{38,\,0.005} = 2.712$ (or use $z_{0.005} \approx 2.575$)

6. **6 × 2 = 12 points** As we mentioned in class, in a hypothesis testing problem, we tend to protect the null and require strong evidence to reject it. Therefore, standing at different position, we you may want to formulate the hypothesis test differently (i.e. changing $H_0$ and $H_a$). See the following example.

   A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.

   (a) **As a diner in this restaurant**, how would you like to formulate the hypothesis testing problem? i.e. What are the null and alternative hypotheses?

   (b) What is a Type 1 Error in (a)?

   (c) What is a Type 2 Error in (a)?

   (d) Which error is more problematic for the diners? Why?

   (e) Which error is more problematic for the restaurant owner? Why?

   (f) So, **as the restaurant owner**, how would you like to formulate the hypothesis testing problem? (suppose you only want to pursue profit and continue running it!)

   **Hint:** For parts (b) and (c), you are expected to explain the Type 1 and Type 2 Errors under the context of this "food safety inspector" example.

Solution: (a) $H_0$: regulations are not met.

$H_a$: regulations are met.

(b) Regulations are not met, but the inspector give the restaurant a pass.

(c) Regulations are met, but the inspector thinks the restaurant is in gross violation.

(d) Type-I error.

(e) Type-II error.

(f) $H_0$: regulations are met.

$H_a$: regulations are not met. □.

Maybe too challenging.
Give students full marks as long as they write something.



7. **$3 \times 5 = 15$ points** Answer the following two questions.

   (a) Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Unif}(0, \theta)$. What's the method of moment estimator of $\theta$?

   (b) Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Unif}(1, 2\theta)$ $(\theta > 1/2)$. What's the method of moment estimator of $\theta$?

   (c) Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$. What's the maximum likelihood estimator?
   **Hint:** Recall the Columbia email example we covered in Lecture 11.

Solution: (a) $\mathbb{E}X = \frac{\theta}{2}$

$\Rightarrow$ Let $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\hat{\theta}}{2}$

$\Rightarrow \hat{\theta} = \frac{1}{2n}\sum_{i=1}^{n} X_i$

(b) $\mathbb{E}X = \frac{1+2\theta}{2}$

$\Rightarrow$ Let $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1+2\theta}{2}$

$\Rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2}$.

(c) Joint pmf $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} \cdot (1-p)^{1-x_i}$

$= p^{\sum x_i} \cdot (1-p)^{n-\sum x_i}$

$\Rightarrow$ log-likelihood function $\ell(p) = \sum_{i=1}^{n} x_i \cdot \log p + (n - \sum x_i) \cdot \log p$

$\Rightarrow$ Let $\ell'(p) = \sum_{i=1}^{n} x_i \cdot \frac{1}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0$

$\Rightarrow \hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$.  □

8. **6 + 7 = 13 points** Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Exp}(\lambda)$ with an unknown $\lambda > 0$. In this problem, we will construct a $100(1-\alpha)\%$ CI for $\lambda$. Recall the following steps to construct a general CI (Lecture 13).

   (a) **Step 1:** Find a pivot random variable $V$ which is a function of both $X_1, \ldots, X_n$ and $\theta$.

   **Step 2:** Verify that the distribution of $V$ does NOT depend on $\theta$ or any other unknown parameters.

   Recall that in Lecture 13, we mentioned the connection between Chi-squared distribution and exponential distribution, which is $2\lambda \sum_{i=1}^n X_i \sim \chi^2_{2n}$. Given this fact, do you think $V = 2\lambda \sum_{i=1}^n X_i$ is a good choice? Why?

   (b) By (a), we know that

$$\mathbb{P}\left(\chi^2_{2n,1-\alpha/2} \le 2\lambda \sum_{i=1}^n X_i \le \chi^2_{2n,\alpha/2}\right) = 1-\alpha.$$

   Construct a $100(1-\alpha)\%$ CI for $\lambda$ based on the equation above.

Solution: (a) Yes.

① It's a function of data $x_1, \ldots, x_n$ and parameter $\lambda$.

② Its distribution is known (unrelated to $\theta$).

(b) $$1-\alpha = \mathbb{P}\left(\frac{\chi^2_{2n,1-\alpha/2}}{2\sum_{i=1}^n X_i} \le \lambda \le \frac{\chi^2_{2n,\alpha/2}}{2\sum_{i=1}^n X_i}\right)$$

$\Rightarrow$ $100(1-\alpha)\%$ CI: $$\left[\frac{\chi^2_{2n,1-\alpha/2}}{2\sum_{i=1}^n X_i}, \frac{\chi^2_{2n,\alpha/2}}{2\sum_{i=1}^n X_i}\right]$$ $\square$

9. **10 points** Suppose that you wish to estimate the percent of students at Columbia who have a pet. You survey 500 randomly selected students around campus and find that 421 of them have a pet. Create a 95% confidence interval for the true proportion of Columbia students who have pets.

Solution: sample proportion $\hat{p} = \frac{421}{500} = 0.842$

By formula:

$$(\text{"approximate"})\ 95\%\ CI = \left[\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot z_{0.025},\ \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\, z_{0.025}\right]$$

$$= \left[0.842 - \sqrt{\frac{0.842 \times 0.158}{500}} \times 1.96,\ 0.842 + \sqrt{\frac{0.842 \times 0.158}{500}} \times 1.96\right]$$

$$= [0.8100,\ 0.8740].$$

$\square$.

# Homework 4

S1201, Summer 2022

100 points

Due Aug 9, 11:59PM EST

10. **(Extra problem, not required) $3 \times 3 = 9$ bonus points** Suppose some mechanists are testing the quality of light bulbs and they record whether each light bulb can work for over 500 hours (denoted as 1) or not (denoted as 0). They observe 100 Bernoulli samples $X_1, \ldots, X_{100}$, each of which follows Bernoulli$(\log(\alpha) + 0.9)$, where $\alpha$ is a physical parameter and is believed to be large than 1. Suppose the sample mean or sample success proportion $\hat{p} = \bar{x} = 0.898$.

    (a) Construct a 95% CI for $\alpha$;

    **Hint:** Recall 95% CI for the proportion or success probability $p$, then replace $p$ by $\log(\alpha) + 0.9$ and solve $\alpha$ (i.e., rewrite it into the form $A \leq \alpha \leq B$ for some numbers $A$ and $B$).

    (b) Construct the method of moment estimator of $\alpha$ and calculate its value under the current data;

    **Hint:** What's the population mean $\mathbb{E}X$ and how does it relate to $\alpha$? Then replace $\mathbb{E}X$ by sample proportion $\hat{p}$ and solve $\alpha$.

    (c) Do you think the **estimate** calculated based on the current data is reliable? Why?

    **Hint:** Is the $\alpha$ estimate you get larger than 1?

Solution: (a) By the CI of $p = \log\alpha + 0.9$:

$$95\% \approx \mathbb{P}\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \times z_{0.025} \leq p \leq \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \times z_{0.025}\right)$$

$$= \mathbb{P}\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \times z_{0.025} \leq \log\alpha + 0.9 \leq \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \times z_{0.025}\right)$$

$$= \mathbb{P}\left(0.898 - \sqrt{\frac{0.898 \times 0.102}{100}} \times 1.96 \leq \log\alpha + 0.9 \leq 0.898 - \sqrt{\frac{0.898 \times 0.102}{100}} \times 1.96\right)$$

$$= \mathbb{P}(-0.0613 \leq \log\alpha \leq 0.0573)$$

$$= \mathbb{P}(0.9405 \leq \alpha \leq 1.0590)$$

$\Rightarrow$ 95% CI: $[0.9405, 1.0590]$

(b) $\mathbb{E}X = p = \log\alpha + 0.9 \Rightarrow$ Let $\hat{p} = \bar{X} = \log\alpha + 0.9$

$\Rightarrow \hat{\alpha} = \exp(\hat{p} - 0.9)$

(c) Based on current data, $\hat{p} = 0.898$

$\Rightarrow \hat{\alpha} = 0.998 < 1 \Rightarrow$ NOT reliable! $\square$