Statistical Inference

HW Author: Ali Mikaeili
Instructor: Mohammadreza A. Dehaqani

Spring 2024

HW 2: Part II

- These are short answers, so we encourage you to engage and ask questions from the TAs if you need more in-depth and complete solutions.

Problem 1: The Truth of Confidence

For each following statements, explain if it is true and determine if there is any problem and fix it.

1. A confidence interval is an estimate for which there is a specified degree of certainty that the population parameter will fall within the range of the interval. **True**
   - Explanation: A confidence interval provides a range of values, derived from the sample data, that is likely to contain the value of an unknown population parameter. The specified degree of certainty (confidence level) indicates the reliability of the interval.
2. A smaller sample size is recommended when generating resampled samples. **False**
   - Explanation: A larger sample size is generally recommended for generating resampled samples, such as in bootstrapping, to ensure that the resampled data adequately represents the population and provides more reliable estimates.
3. The difference between each pair of observations is the starting point for a paired analysis. Then we use these differences to make inferences. **True**
   - Explanation: In a paired analysis, we focus on the differences between paired observations (e.g., before and after measurements) to control for variability between subjects and to make more precise inferences about the population.
4. In the CLT we are looking for a confidence interval for the sample mean. **False**
   - Explanation: The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean will be approximately normally distributed, allowing us to make inferences about the population mean. However, the CLT itself is not specifically about constructing confidence intervals.
5. With large sample size, even small differences between the null value and the true value of the parameter, a difference often called the effect size, will be identified as statistically significant. **True**
   - Explanation: Large sample sizes increase the power of a statistical test, making it easier to detect small effect sizes as statistically significant. However, the practical significance of such small differences should also be considered.
6. For a positively skewed distribution, the mean usually has a larger value than either the median or the mode. **True**
   - Explanation: In a positively skewed distribution, the mean is pulled in the direction of the skew (to the right) and tends to be greater than the median, which in turn is greater than the mode.

7. A 98% confidence interval obtained from a random sample of 1000 people has a better chance of containing the population percentage than a 98% confidence interval obtained from a random sample of 500 people. **False**

    - Explanation: A 98% confidence interval from a random sample of 1000 people would typically be narrower, but it does not necessarily have a better chance of containing the population percentage compared to a 98% confidence interval from a smaller sample size. The confidence level (98%) already dictates the probability of containing the true parameter.

8. Suppose the null hypothesis is $\mu = 5$ and we fail to reject H_0. Under this scenario, the true population mean is 5. **False**

    - Explanation: Failing to reject the null hypothesis means that there is not enough evidence to conclude that the population mean is different from 5, but it does not prove that the population mean is exactly 5.

9. Suppose we have constructed the following confidence interval about the mean age of freshmen college students in a State: $17.3 \leq \mu \leq 22.6$. The proper interpretation is that we are 95% confident that the sample mean is in the range 17.3 and 22.6 years. **False**

    - Explanation: The correct interpretation of a confidence interval is that we are 95% confident that the true population mean, not the sample mean, lies within the interval 17.3 to 22.6 years.

Problem 2: Where Bernouli and Gauss Meet

Consider $X_1, X_2, \dots, X_n$ as independent and identically distributed (i.i.d.) Bernoulli random variables with parameter p . These variables represent the outcomes of n binomial trials. Let the sample mean be denoted as $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Discuss how the Central Limit Theorem (CLT) is applicable in this context:

1. Using the CLT, justify why: $\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{n\to\infty} N(0,1)$

2. Explain why the following adjustment using the sample proportion, $\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{n\to\infty} N(0,1)$

(1) Sol

2.1

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{n\to\infty} N(0,1)$$

$$\mu = E[X_i] = p$$

$$\sigma^2 = \text{Var}(X_i) = p(1-p)$$

- $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $E[\overline{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = p$
- $\text{Var}(\overline{X}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{p(1-p)}{n}$
- The standardized form: $\frac{\overline{X}_n - p}{\sqrt{\text{Var}(\overline{X}_n)}} = \frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}}$

2.2

1. **Central Limit Theorem:**
   - For large n, $\sqrt{n}(\overline{X}_n - p) \xrightarrow{d} N(0, p(1-p))$
   - Thus, $\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$
2. **Variance Estimation:**
   - Use $\overline{X}_n(1-\overline{X}_n)$ as an estimator for $p(1-p)$
   - By the Law of Large Numbers, $\overline{X}_n(1-\overline{X}_n) \to p(1-p)$
   - Replace $p(1-p)$ with $\overline{X}_n(1-\overline{X}_n)$
   - $\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{d} N(0,1)$

Problem 3: Sample Size and Significance

The daily electricity consumption in the homes of a specific population can be assumed normally distributed with a standard deviation (α) of 16.8 kWh. After an intense campaign to reduce electricity use, we want to estimate the average daily consumption per home in that population.

1. What sample size is required to obtain a 95% confidence interval (CI) for the mean μ with a $2kWh$ margin of error?
2. If the value of σ increases, would the needed sample size be bigger or smaller than that required in the first part?
3. If the significance level is $\alpha = 0.10$, would the needed sample size be bigger or smaller than that required in the first part?

(1) Solution

3.1

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

$$n = \left(\frac{1.96 \cdot 16.8}{2}\right)^2 = \left(\frac{32.928}{2}\right)^2 = (16.464)^2 \approx 271.03$$

$$n \approx 272$$

3.2

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

3.3 If α changes to 0.10 (90% confidence level):

$$n = \left(\frac{1.645 \cdot 16.8}{2}\right)^2 = \left(\frac{27.636}{2}\right)^2 = (13.818)^2 \approx 190.91$$

$$n \approx 191$$

Problem 4: Bonus: Confidence Interval of Insurance Confidence

A survey was conducted in the fall of 1979 by a national health organization regarding health insurance coverage in the USA. The survey revealed that a large proportion of citizens were very pessimistic about their health insurance coverage. When asked if they believed that their health insurance would be sufficient, 62.9% of the 6100 surveyed citizens answered negatively. Compute a 95%-confidence interval for the proportion of citizens that believed their health insurance would not be sufficient.

(1) Solution

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$ME = z_{\alpha/2} \times SE = 1.96 \times SE$$

Problem 5: A Happy Exponential Family

Analyze whether the specified distributions can be classified as members of the exponential family by appropriately transforming their probability density functions. For each distribution, identify and list the sufficient statistics and the canonical parameters involved:

1. A shape parameter k and a scale parameter θ characterize the Gamma distribution.
2. The central chi-squared (χ^2) distribution with k degrees of freedom, is commonly used in statistical testing.

Solution

5.1

$$f(x \mid k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}, \quad x > 0$$

$$f(x \mid k, \theta) = \frac{1}{\Gamma(k)} \left(\frac{1}{\theta}\right)^k x^{k-1} e^{-x/\theta}$$

$$f(x \mid k, \theta) = \frac{1}{\Gamma(k)} \exp\left((k-1)\log x - \frac{x}{\theta} - k\log\theta\right)$$

5.2

$$f(x \mid k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}, \quad x > 0$$

Problem 6: MLE vs MoM

Let $X_1, \dots, X_n$ be iid with pdf

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta, \theta > 0.$$

Estimate θ using both the method of moments and maximum likelihood. Calculate the means and variances of the two estimators. Which one should be preferred and why?

(1) Solution

Method of Moments

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

$$\bar{X} = \frac{\theta}{2} \Rightarrow \hat{\theta}_{\text{MM}} = 2\bar{X}$$

Maximum Likelihood Estimation (MLE)

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}$$

$$\ell(\theta) = -n\log\theta \quad \text{subject to } \theta \geq \max(X_i)$$

$$\hat{\theta}_{\text{MLE}} = \max(X_1, X_2, \ldots, X_n)$$

$$\hat{\theta}_{\text{MM}} = 2\bar{X}$$

$$E(\hat{\theta}_{\text{MM}}) = \theta$$

$$Var(\hat{\theta}_{\text{MM}}) = \frac{\theta^2}{3n}$$

$$\hat{\theta}_{\text{MLE}} = \max(X_1, X_2, \ldots, X_n)$$

$$E(\hat{\theta}_{\text{MLE}}) = \frac{n}{n+1}\theta$$

$$Var(\hat{\theta}_{\text{MLE}}) = \frac{n\theta^2}{(n+1)^2(n+2)}$$

Problem 7: The War of Estimators

Given that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from $N(\mu, \sigma^2)$:

1. Show that the sample mean $\overline{X}$ and the sample variance S^2 are unbiased estimators of μ and σ^2 respectively.

2. Compare the mean squared error (MSE) of S^2 with that of the alternative estimator $\hat{\sigma}^2 = \frac{n-1}{n}S^2$ derived from the method of moments or maximum likelihood estimation.

   Discuss whether the estimator S^2 or $\hat{\sigma}^2$ has a lower MSE, and elaborate on the implications of this result. Is one always better than the other?

Solution

7.1

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

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n\mu = \mu$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

$$E\left[\sum_{i=1}^{n} X_i^2\right] = \sum_{i=1}^{n} E[X_i^2] = n(\sigma^2 + \mu^2)$$

$$E[n\bar{X}^2] = nE[\bar{X}^2] = n\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2 + n\mu^2$$

$$E[S^2] = \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - (\sigma^2 + n\mu^2)\right) = \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) = \frac{1}{n-1}\left((n-1)\sigma^2\right) = \sigma^2$$

7.2

MSE of S^2

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta}, \theta)]^2$$

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

MSE of $\hat{\sigma}^2 = \frac{n-1}{n}S^2$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$$

$$E(\sigma^2) = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2$$

$$Var(\sigma^2) = Var\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 Var(S^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)^4}{n^2}.$$

$$\begin{aligned} E(\sigma^2 - \sigma^2)^2 &= Var(\sigma^2) + (Bias)^2 \\ &= \frac{2(n-1)^4\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

$$MSE_{\sigma^2} = \frac{2n-1}{n^2}\sigma^4 < \frac{2n}{n^2}\sigma^4 = \frac{2\sigma^4}{n} < \frac{2\sigma^4}{n-1} = MSE_{S^2}.$$

Problem 8: Chi-Square

Suppose U is distributed normally with mean 0 and variance 1 ($U \sim N(0,1)$), and V follows a chi-squared distribution with p degrees of freedom ($V \sim \chi_p^2$). Assume U and V are independent.

1. Determine the joint probability density function of U and V.
2. Define the variable T as follows:

   $$T = \frac{U}{\sqrt{\frac{V}{p}}}$$

   What is the probability distribution of T, according to basic statistical theory?
3. Determine the joint distribution of T and W, where $W = V$.

Solution

8.1

$$f_{U,V}(u,v) = f_U(u) \cdot f_V(v) = \left(\frac{1}{\sqrt{2\pi}} e^{-u^2/2} \right) \left(\frac{v^{(p/2-1)} e^{-v/2}}{2^{p/2}\Gamma(p/2)} \right)$$

8.2

$$T = \frac{U}{\sqrt{V/p}}$$

$$T \sim t_p$$

8.3

$$W = V$$

$$T = \frac{U}{\sqrt{W/p}}, \quad W = V$$

$$U = T\sqrt{W/p}$$

The Jacobian matrix J:

$$J = \begin{vmatrix} \frac{\partial U}{\partial T} & \frac{\partial U}{\partial W} \\ \frac{\partial V}{\partial T} & \frac{\partial V}{\partial W} \end{vmatrix} = \begin{vmatrix} \sqrt{\frac{W}{p}} & \frac{T}{2\sqrt{Wp}} \\ 0 & 1 \end{vmatrix}$$

$$\det(J) = \sqrt{\frac{W}{p}}$$

$$f_{T,W}(t,w) = f_{U,V}(u,v) \cdot |\det(J)|$$

Problem 9: Coding Confidence Intervals

(Use R or Python) The following are the systolic blood pressure measurements for two groups of men and women randomly selected for a heart study:

Men:

128.35, 160.34, 133.74, 138.12, 91.00, 97.43, 128.58, 148.78, 150.65, 110.96, 135.7, 118.77, 147.1, 107.2, 122.46, 129.36, 158.14, 102.72, 136.59, 146.02, 105.88, 111.24, 131.22, 124.6, 137.85, 136.46, 145.31, 166.71, 158.66, 108.63, 103.11, 149.29

Women:

116.62, 137.15, 106.07, 172.58, 151.33, 98.73, 136.11, 149.9, 140.8, 98.58, 158.4, 97.97, 117.99, 126.53, 128.67, 126.57, 124.3, 120.39, 150.08, 143.05, 130.18, 108.04, 136.39, 124.94, 136.86, 143.03, 128.58, 142.51, 151.68, 120.94

1. **Visualize the Data:**
   - Create histograms to view the distribution of systolic blood pressures for both groups.
   - Generate boxplots to compare the central tendencies and variability of the two groups.
2. **Statistical Analysis:**
   - Calculate basic statistics such as mean, variance, and range for each group.
   - Interpret these statistics to discuss potential differences in blood pressure variability and central tendency between the groups.
3. **Confidence Interval Construction:**

(a) Construct a 95% confidence interval for the difference in mean systolic blood pressures between men and women. Interpret the result.

(b) Is there any difference in mean systolic blood pressures between men and women? Explain the conclusion.

4. **Interpretation:**
   - Interpret the results of the histograms, boxplots, and confidence interval.
   - Discuss the implications of these findings in the context of the heart study.

Problem 10: Bonus: Learn about the t-dist before too late

(Use R or Python) This problem examines a t-distribution with 20 degrees of freedom.

1. Plot a graph of a t-distribution with 20 degrees of freedom and a standard normal curve from -4 to 4.
2. Find the area to the right of 2 under each curve.
3. Find the 0.90 quantile of each curve.
4. Find the 0.95 quantile of each curve.
5. Find the 0.975 quantile of each curve.

Problem 11: Bonus: True Coverage of Different CI

(Use R or Python) This problem examines the true coverage probability for three different methods of making confidence intervals for a sample proportion.

Given: Sample size $n = 60$ and $p = 0.4$.

For each of the following methods, find which outcomes x result in confidence intervals that capture p and compute the coverage probability.

1. Normal from maximum likelihood estimate, $\hat{p} = \frac{X}{n}$, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, with the interval $\hat{p} \pm 1.96SE$.
2. Normal from adjusted maximum likelihood estimate, $\hat{p} = \frac{X+2}{n+4}$, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{(n+4)}}$, with the interval $\hat{p} \pm 1.96SE$.
3. Within $z^2/2$ of the maximum likelihood loglikelihood. (For this method, use the provided function `logl.ci.p()` which returns the lower and upper bounds of a 95% confidence interval given n and x. You can graph the loglikelihood using `glogl.p()` for n, x, and $z = 1.96$ to see if the returned values make sense.)