Statistical Inference

HW Author:
Instructor: Mohammadreza A. Dehaqani

Spring 2024

HW 2: Part II

- Please note that this is the 2nd half of a homework. But we understand that it still takes time (Maybe as much as a full homework) so we will weight the grade such that it benefits you better.
- If you have any questions about the homework, don't hesitate to use the class group to ask questions or drop an email to the HW Authors.
- Please consult the course page for important information on submission guidelines and delay policies to ensure your homework is turned in correctly and on time.
- This course aims to equip you with the skills to tackle all problems in this domain and encourages you to engage in independent research. Utilize your learnings to extend beyond the classroom teachings where necessary.

Problem 1: The Truth of Confidence

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.
2. A smaller sample size is recommended when generating resampled samples.
3. The difference between each pair of observations is the starting point for a paired analysis. Then we use these differences to make inferences.
4. In the CLT we are looking for a confidence interval for the sample mean.
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.
6. For a positively skewed distribution, the mean usually has a larger value than either the median or 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.
8. Suppose the null hypothesis is $\mu = 5$ and we fail to reject H_0. Under this scenario, the true population mean is 5.

i. 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.

Problem 2: Where Bernouli and Gauss Meet

Consider $X_1, X_2, \ldots, X_n$ 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, $\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{n\to\infty} N(0,1)$

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?

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.

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 characterize the Gamma distribution.
2. The central chi-squared (χ^2) distribution with k degrees of freedom, is commonly used in statistical testing.

Problem 6: MLE vs MoM

Let $X_1, \ldots, X_n$ 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?

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?

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 ($V \sim \chi^2_p$). 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 $W = V$.

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.)