20210 Summer 2025: Problem Set 1
Econometric
Instructor(s): Murilo Ramos

Agustín Esteva
aesteva@uchicago.edu
Due Date: July 29, 2025

Problem 1.1

Let Y be a Bernoulli random variable with success probability $\Pr(Y = 1) = p$, and let $Y_1, \ldots, Y_n$ be i.i.d. draws from this distribution. Let $\hat{p}$ be the fraction of successes (1s) in the sample.

(a) Show that $\hat{p} = \overline{Y}$.

> SOLUTION: By definition,
>
> $$\hat{p} = \frac{\sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\}}{n} = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}$$
>
> ♡

(b) Show that $\hat{p}$ is an unbiased estimator of p.

> SOLUTION:
>
> $$\mathbb{E}[\hat{p}] - p = \frac{1}{n}\sum \mathbb{E}[Y] - p = \frac{n}{n}p - p = 0$$
>
> ♡

(c) Show that $\mathrm{Var}(\hat{p}) = \dfrac{p(1-p)}{n}$.

> SOLUTION:
>
> $$\mathrm{Var}(\hat{p}) = \mathrm{Var}(\frac{1}{n}\sum \mathrm{Var}[Y_i]) = \frac{1}{n^2}n\mathrm{Var}[Y] = \frac{p(1-p)}{n}$$
>
> ♡

Problem 1.2

Now suppose a survey of 400 likely voters finds that 215 support the incumbent and 185 support the challenger. Let p denote the fraction of all likely voters who support the incumbent, and let $\hat{p}$ be the sample proportion supporting the incumbent.

(a) Use the survey data to estimate p.

Solution: Lett Y_i be one for support and zero for not, we estimate

$$\hat{p} = \frac{215}{400} = 0.5375$$

♡

(b) Use the plug-in estimator of the variance of $\hat{p}$, namely $\dfrac{\hat{p}(1-\hat{p})}{n}$, to compute the standard error.

Solution:

$$\hat{\sigma}_Y = \sqrt{\frac{\frac{215}{400}\frac{185}{400}}{400}} = 0.02494$$

♡

(c) What is the p-value for testing $H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$?

Solution: Using CLT, we have that

$$\frac{\sqrt{n}(\hat{p} - \mu_Y)}{\sigma_Y} \to N(0,1)$$

Thus, under the null,

$$p = 2\Phi\{\frac{\sqrt{400}(0.5375 - \frac{1}{2})}{0.025}\} = 2\Phi(-1,5) = 0.1336$$

♡

(d) What is the p-value for testing $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$?

Solution:

$$p = \frac{0.1336}{2} = 0.0668$$

♡

(e) Explain why the results of parts (c) and (d) differ.

Solution: Because we in part (d) we are just worried about one possible source of error, and is thus the chance of rejecting the null is half as small. ♡

(f) Based on the results above, does the survey provide statistically significant evidence (at the 5% level) that the incumbent was ahead?

Solution: No. ♡

Problem 1.3

Using the same data:

(a) Construct a 95% confidence interval for p.

Solution: Letting $\alpha = 0.05$, we consider that we want

$$\mathbb{P}\{\left|\frac{\sqrt{n}(\hat{p}-\mu_Y)}{\hat{\sigma}_Y}\right| \leq z_{\frac{\alpha}{2}}\} = 0.95$$

by the CLT we have this is equivalent to finding

$$\left|\frac{\sqrt{n}(\hat{p}-\mu_Y)}{\hat{\sigma}_Y}\right| \leq 1.96.$$

Rearranging, we find our confidence interval to be

$$\mu_Y \in [\hat{p} \pm 1.96\frac{\hat{\sigma_Y}}{\sqrt{400}}] = [0.488, 0.586]$$

♡

(b) Construct a 99% confidence interval for p.

Solution: Using now $z_{\frac{\alpha}{2}} = 2.56$, we find that

$$\mu_y \in [0.4732, 0.6018]$$

♡

(c) Why is the interval in part (b) wider than the interval in part (a)?

Solution: Because in order to be more confident about where the true value lies, we need to expand the ‘net’ we are trying to catch it with. ♡

(d) Without recalculating, use your results to test the hypothesis

$$H_0 : p = 0.50 \quad \text{vs.} \quad H_1 : p \neq 0.50$$

at the 5% significance level.

Solution: Using the 95% confidence interval, we see that we are 95% confident that the true mean is wiith in $[0.488, 0.586]$, and since 0.5 is in there, then we fail to reject the null. ♡

Problem 2

Let $X_1, \dots, X_n$ be i.i.d. $\sim X$. Consider the estimator

$$\hat{\theta}_n = \sum_{i=1}^{n} a_i X_i$$

for some constants $a_1, \dots, a_n$.

(a) Show that if $\hat{\theta}_n$ is an unbiased estimator of $\mathbb{E}[X]$, then $\sum_{i=1}^n a_i = 1$.

Solution: Suppose $\hat{\theta}_n$ is an unbiased estimator of μ. Then

$$\begin{aligned} 0 &= \mathbb{E}[\hat{\theta}_n] - \mu \\ &= \mathbb{E}[\sum a_i X_i] - \mu \\ &= \sum a_i \mathbb{E}[X_i] - \mu \\ &= \mu \sum a_i - \mu \end{aligned}$$

which is true iff $\sum a_i = 1$. ♡

(b) Show that $\text{Var}[\hat{\theta}_n] = \text{Var}[X] \sum_{i=1}^n a_i^2$.

Solution: Computing,

$$\begin{aligned} \text{Var}(\hat{\theta}_n) &= \text{Var}(\sum a_i X_i) \\ &= \sum a_i^2 \text{Var}(X_i) \\ &= \sum a_i^2 \text{Var}(X) \\ &= \text{Var}(X) \sum a_i^2 \end{aligned}$$

♡

(c) Find $a_1, \dots, a_n$ that minimize $\text{Var}[\hat{\theta}_n]$ subject to the constraint that $\hat{\theta}_n$ is an unbiased estimator of $\mathbb{E}[X]$.

Solution: Noting that $\sum a_i = 1$, we seek $\min_{a_1,\dots,a_n} \sum_{i=1}^n a_i^2$. We claim that $a_i = \frac{1}{n}$ for all i minimizes this. To see this, we can solve using a Lagrangian:

$$\mathcal{L}(a_1, \dots, a_n, \lambda) = \sum a_i^2 - \lambda(\sum a_i - 1)$$

$$\frac{\partial \mathcal{L}}{\partial a_j} = 2a_j - \lambda = 0 \implies a_j = \frac{\lambda}{2}$$

With the boundary condition we find that

$$\sum a_i = 1 \implies \sum \frac{\lambda}{2} = 1 \iff \frac{\lambda n}{2} = 1 \iff \lambda = \frac{2}{n}$$

Hence, $a_i = \frac{1}{n}$ for all i. ♡

Problem 3

Let $X_1, \ldots, X_n$ be i.i.d. $\sim X$. Suppose $\mathrm{Var}[X] < \infty$. For $1 \leq i \leq n$, define $Z_i = a + bX_i$ and $Z = a + bX$ for some constants a and b.

(a) Show that $\bar{Z}_n = a + b\bar{X}_n$ and $\hat{\sigma}_Z^2 = b^2\hat{\sigma}_X^2$.

Solution: Computing,

$$\overline{Z}_n = \frac{1}{n}(\sum_{i=1}^{n} Z_i) = \frac{1}{n}(\sum_{i=1}^{n} a + bX_i) = \frac{n}{n}a + \frac{b}{n}\sum X_i = a + b\overline{X}_n$$

We calculate the variance to be

$$\mathrm{Var}(Z) = \mathrm{Var}(a + bX) = \mathrm{Var}(bX) = b^2\mathrm{Var}(X)$$

♡

(b) Prove that $\bar{Z}_n$ is an unbiased estimator of $\mathbb{E}[Z]$.

Solution: We simply compute

$$\begin{aligned}
\mathbb{E}[\overline{Z}_n] - \mathbb{E}[Z] &= \mathbb{E}[\frac{1}{n}\sum Z_i] - \mathbb{E}[Z] \\
&= \frac{1}{n}\sum \mathbb{E}[Z_i] - \mathbb{E}[Z] \\
&= \frac{1}{n}\sum \mathbb{E}[Z] - \mathbb{E}[Z] \\
&= \frac{n}{n}\mathbb{E}[Z] - \mathbb{E}[Z] \\
&= 0
\end{aligned}$$

♡

(c) Prove that $\bar{Z}_n$ is a consistent estimator of $\mathbb{E}[Z]$. (Hint: Is the function $g(t) = a + bt$ continuous?)

Solution: Since $\mathrm{Var}(X) < \infty$, we have that $\mathbb{E}[X^2] < \infty$ since $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Thus, we apply the weak law of large numbers to find that $\bar{X}_n \to \mathbb{E}[X]$ in probability. Since $g(t) = a + bt$ is continuous, we apply the continuous mapping theorem to see that $g(\overline{X}_n) \to g(\mathbb{E}[X])$ in probability. In other words,

$$a + b\overline{X}_n \to a + b\mathbb{E}[X] = \mathbb{E}[Z]$$

which by part (a) implies then that

$$\overline{Z_n} \to \mathbb{E}[Z]$$

in probability, and thus we are done. ♡

Problem 4

Let $X_1, \dots, X_n$ be i.i.d. $\sim X$, where X is the SAT score of a high school senior in Chicago. A researcher is interested in θ, the fraction of high school seniors in Chicago with an SAT score higher than 1200.

(a) Write θ as the expected value of a suitable random variable.

Solution: We note that this fraction we are interested in is the same as

$$\theta = \mathbb{P}\{X > 1200\} = \mathbb{E}[\mathbb{1}\{X > 1200\}]$$

$\heartsuit$

(b) Propose an estimator $\hat{\theta}_n$ of θ that is unbiased and consistent. Prove unbiased ness and consistency.

Solution: Using the analogy principle, we propose

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > 1200\}$$

TO show (un)bias:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_n] - \theta &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > 1200\}] - \theta \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{X_i > 1200\}] - \theta \\
&= \frac{1}{n}\sum_{i-1} \mathbb{P}\{X > 1200\} - \theta \\
&= \mathbb{P}\{X > 1200\} - \theta \\
&= 0
\end{aligned}$$

by definition of θ. To show that $\hat{\theta}_n$ is consistent, we use the weak law of large numbers. We claim that $\mathbb{E}[\theta^2] < \infty$. This is obvious since $\theta \leq 1$ almost surely since it is a probability, and thus $\mathbb{E}[\theta^2] \leq 1^2$ almost surely. Thus, we apply the law of large numbers to the i.i.d. r.v.s of $\mathbb{1}\{X_1 > 1200\}, \dots \mathbb{1}\{X_n > 1200\} \sim \mathbb{1}\{X > 1200\}$ and find that

$$\frac{1}{n}\sum_{i=1}^{n} \nVdash\{X_i > 1200\} \to \mathbb{E}[X > 1200] = \mathbb{P}\{X > 1200\} = p$$

where the convergence is in probability. $\heartsuit$

(c) What is $\text{Var}[\hat{\theta}_n]$?

Solution: We just compute bro

$$\begin{aligned}
\text{Var}(\hat{\theta}_n) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > 1200\}) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(\mathbb{1}\{X_i > 1200\}) \\
&= \frac{\mathbb{E}[\mathbb{1}\{X_i > 1200\}^2] - (\mathbb{P}\{X > 1200\})^2}{n} \\
&= \frac{\mathbb{P}\{X > 1200\}(1 - \mathbb{P}\{X > 1200\})}{n} \\
&= \frac{\theta(1-\theta)}{n}
\end{aligned}$$

♡

(d) Propose a consistent estimator for $n \cdot \text{Var}[\hat{\theta}_n]$. Justify your answer.

Solution: We propose

$$\hat{\sigma}_n^2 = \hat{\theta}_n(1 - \hat{\theta}_n).$$

To show $\hat{\theta}_n$ is consistent, we note that by part b $\hat{\theta}_n \to \theta$ in probability. Letting $g(t) = t(1-t)$ be the continuous function, we use the continuous mapping theorem and part (c) to say that

$$g(\hat{\theta}_n) \to g(\theta) \implies \hat{\theta}_n(1-\hat{\theta}_n) \to \theta(1-\theta) \iff \hat{\sigma}_n^2 \to n\text{Var}(\hat{\theta}_n)$$

♡

(e) A researcher wishes to test the null hypothesis that at least $\frac{1}{4}$ of high school seniors in Chicago scored higher than 1200 in SAT at significance level α. In what follows, suppose that it is known that $0 < \mathbb{P}\{X > 1200\} < 1$.

(i) Formally state the null and alternative hypotheses.

Solution: $H_0 : \theta \geq \frac{1}{4}$ and $H_a : \theta < \frac{1}{4}$ ♡

(ii) Suppose your sample size is large. How would you perform the test? Write down your test statistic, critical value, and the rule you would use to determine whether or not to reject the null hypothesis.

Solution: Test statistic is just the Z score. I would the sample proportion, $\hat{\theta}_n$ defined in part (b). Since the estimator is Bernoulli and n is large, then by the

CLT

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\frac{\theta(1-\theta)}{n}}} = \frac{n(\hat{\theta}_n - \theta)}{\sqrt{\theta(1-\theta)}} \sim N(0,1)$$

Under the null,

$$Z = \frac{n(\hat{\theta}_n - \frac{1}{4})}{\frac{3}{16}}$$

and reject when ($\alpha = 0.05$ as our significance level implying critical value $z_\alpha = -1.65$)

$$Z < -1.65 \iff \frac{n(\hat{\theta}_n - \frac{1}{4})}{\sqrt{\frac{3}{16}}} < -1.65 \iff \hat{\theta}_n < \frac{1}{4} - \frac{1.65}{n}\frac{\sqrt{3}}{4}$$

♡

(iii) State in words the definition of p-value. What is the p-value for your test?

SOLUTION: The p value is the smallest $\alpha-$value for which we reject the null hypothesis. In other words,

$$1 - \mathbb{P}\{Z \leq \text{observed}\} = 1 - \Phi(\frac{n(\hat{\theta}_n - \frac{1}{4})}{\frac{3}{16}})$$

♡

Problem 5 (Optional)

The following question involves the California Test Score dataset. The dataset is described in Appendix 4.1 of Stock and Watson and can be downloaded from the course website.

(a) Load the California Test Score dataset into R. How many observations do you have in the dataset?

(b) The variable `avginc` is average district income measured in 1000s of dollars. Define a new variable, `income`, which is the variable `avginc` multiplied by 1000.

   (i) What does the variable `income` measure?

   (ii) What is the mean and standard deviation of `avginc`?

   (iii) What is the mean and standard deviation of `income`? Given your result to part (ii), are the mean and standard deviation for `income` what you expected? Why?

(c) (i) What is the mean math score across all districts?

   (ii) What fraction of districts have an average class size of 20 or fewer students? What is the mean math score in districts with average class size of 20 or fewer students?

   (iii) What fraction of districts have an average class size of more than 20 students? What is the mean math score in districts with average class size greater than 20?

   (iv) What is the connection between your answer in (i) and your answers in (ii) and (iii)?

   (v) Calculate a test at the 10% level of whether the mean math score in districts with average class size of 20 or fewer students is equal to the mean math score in districts with average class size greater than 20. Formally state your null hypothesis in terms of population-level conditional expectations. Describe your testing procedure. Can you reject the null hypothesis?

   (vi) What is the covariance between `avginc` and mean math score? What is the covariance between `income` and mean math score? Are the two covariances the same or different? Explain.

   (vii) What is the correlation between `avginc` and mean math score? What is the correlation between `income` and mean math score? Are the two correlations the same or different? Explain.