Classical statistical inference

1 Properties of estimators

1. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $\hat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n}$. Find the bias, standard error, and MSE of this estimator.

   a. Bias

   Recall that $\text{bias}(\hat{\lambda}) = \text{E}_\lambda[\hat{\lambda}] - \lambda$. The expected value of the estimator $\hat{\lambda} = \text{E}[X] = \lambda$. Because the expected value of the estimator $\hat{\lambda}$ is the true value λ, bias $= 0$.

   b. Standard error

$$\text{Var}(\hat{\lambda}) = \frac{\text{Var}[X]}{n} = \frac{\lambda}{n}$$

$$\text{s.e.} = \sqrt{\frac{\lambda}{n}}$$

   c. MSE

$$\begin{aligned} \text{MSE} &= \text{bias}^2(\hat{\lambda}) + \text{Var}_\lambda(\hat{\lambda}) \\ &= 0^2 + \frac{\lambda}{n} \\ &= \frac{\lambda}{n} \end{aligned}$$

2. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\hat{\theta} = 2\bar{X}_n$. Find the bias, standard error, and MSE of this estimator.

   Recall that for the uniform distribution

$$\text{E}(X_i) = \frac{\theta}{2}, \text{Var}(X_i) = \frac{\theta^2}{12}$$

   a. Bias

$$\text{E}_\theta(2\bar{X}) = 2\text{E}_\theta(\bar{X}) = 2 \times \frac{\theta}{2} = \theta$$

$$\text{bias}(\hat{\theta}) = \text{E}_\theta[\hat{\theta}] - \theta = \theta - \theta = 0$$

   b. Standard error

$$\text{Var}_\theta(2\bar{X}) = 4\text{Var}_\theta(\bar{X}) = \frac{4\sigma^2}{12n} = \frac{4\sigma^2}{12n} = \frac{\sigma^2}{3n}$$

$$\text{s.e.} = \sqrt{\text{Var}_\theta(2\bar{X})} = \sqrt{\frac{\sigma^2}{3n}}$$

c. MSE

$$\begin{aligned}\text{MSE} &= \text{bias}^2(\hat{\theta}) + \text{Var}_\theta(\hat{\theta}) \\ &= 0^2 + \frac{\sigma^2}{3n} \\ &= \frac{\sigma^2}{3n}\end{aligned}$$

2 Birds of a feather get their news on Twitter

A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion.

a. Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

Solution: Recall the general formula is Point estimate $\pm\, z^* \times SE$. First, identify the three different values. The point estimate is 52%, $z^* = 2.58$ for a 99% confidence level, and $SE = 2.4\%$. Then plug the values into the formula:

$$52\% \pm 2.58 \times 2.4\% \rightarrow (45.8\%, 58.4\%)$$

b. Identify the follow statements as true or false. Provide an explanation to justify each of your answers.

a. The data provide statistically significant evidence that more than half of U.S. adult Twitter users get some news through Twitter. Use a significance level of $\alpha = 0.01$.

Solution: False. 50% is included in the 99% confidence interval, hence a null hypothesis of $p = 0.50$ would not be rejected at this level.

b. Since the standard error is 2.4%, we can conclude that 97.6% of all U.S. adult Twitter users were included in the study.

Solution: False. The standard error measures the variability of the sample proportion, and is unrelated to the proportion of the population included in the study.

c. If we want to reduce the standard error of the estimate, we should collect less data.

Solution: False. We need to increase the sample size to decrease the standard error.

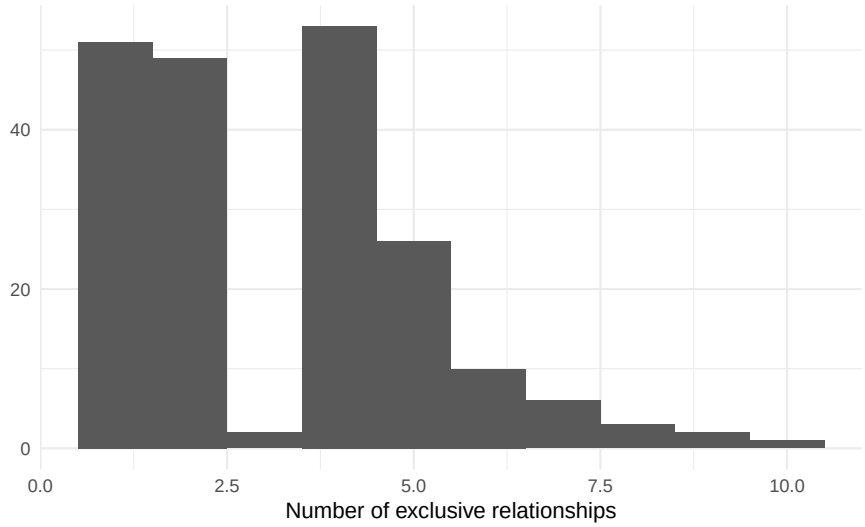
d. If we construct a 90% confidence interval for the percentage of U.S. adults Twitter users who get some news through Twitter, this confidence interval will be wider than a corresponding 99% confidence interval.

Solution: False. As the confidence level decreases so does the margin of error, and hence the width of the confidence interval.

3 Dating on college campuses

A survey conducted on a reasonably random sample of 203 undergraduates asked, among many other questions, about the number of exclusive relationships these students have been in. The histogram below shows the distribution of the data from this sample.

The sample average is 3.2 with a standard deviation of 1.97.



Estimate the average number of exclusive relationships undergraduate students have been in using the Normal distribution and a 90% confidence interval and interpret this interval in context.

Solution: The confidence interval can be calculated as follows:

$$\begin{aligned}\bar{x} \pm Z^* SE_{\bar{x}} &= 3.2 \pm 1.65 \times \frac{1.97}{\sqrt{203}} \\ &= 3.2 \pm 0.23 \\ &= (2.97, 3.43)\end{aligned}$$

4 Choose your own death

There is a theory that people can postpone their death until after an important event. To test the theory, Phillips and King (1988) collected data on deaths around the Jewish holiday Passover. Of 1919 deaths, 922 died the week before the holiday and 997 died the week after. Think of this as a binomial and test the null hypothesis that $\theta = \frac{1}{2}$. Report and interpret the p-value. Also construct a confidence interval for θ.

Solution: $\hat{\theta} = \frac{922}{1919} = 0.48$ with estimated standard error $\widehat{\text{s.e.}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.011$. The Wald test statistic is

$$\frac{0.48 - 0.50}{0.011} = -1.818$$

and the p-value is $\Pr(|Z| > 1.818) = 2\Pr(Z > 1.818) = 0.07$ which is moderate evidence against the null $H_0 : p = \frac{1}{2}$. A 95% confidence interval is

$$\hat{p} \pm 2 \times \widehat{\text{s.e.}} = (0.458, 0.503)$$

The results are rather equivocal. I would suggest collecting further data.

5 Evaluating eyesight in children

It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted.

a. Construct hypotheses appropriate for the following question: do these data provide evidence that the 8% value is inaccurate?

Solution:

- $H_0 : p = 0.08$: Proportion of children who are near sighted is 8%
- $H_A : p \neq 0.08$: Proportion of children who are near sighted is not 8%

b. What proportion of children in this sample are nearsighted?

Solution: $\frac{21}{194} = 0.108 \rightarrow 10.8\%$ of children in this sample are near sighted.

c. Given that the standard error of the sample proportion is 0.0195 and the point estimate follows a nearly normal distribution, calculate the test statistic (the Z-statistic).

Solution:

$$Z = \frac{\hat{p} - p}{SE} = \frac{0.108 - 0.08}{0.0195} = 1.44$$

d. What is the p-value for this hypothesis test?

Solution:

$$\text{p-value} = 2 \times \Pr(Z > 1.44) = 2 \times (1 - 0.9251) = 0.1498$$

e. What is the conclusion of the hypothesis test?

Solution: Since the p-value is large we fail to reject H_0. The data do not provide convincing evidence that the proportion of children who are nearsighted is different than 8%.

6 Statistical significance

Determine whether the following statement is true or false, and explain your reasoning: "With large sample sizes, even small differences between the null value and the point estimate can be statistically significant."

Solution: True. If the sample size is large, then the standard error will be small, meaning even relatively small differences between the null value and point estimate can be statistically significant. **This is kind of a big deal when analyzing big data.**

7 Sleep deprivation

New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. Do these data provide strong evidence that New Yorkers sleep less than 8 hours a night on average?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

a. Write the hypotheses in symbols and in words.

Solution:

$H_0 : \mu = 8$ (New Yorkers sleep 8 hrs per night on average.)

$H_A : \mu < 8$ (New Yorkers sleep less than 8 hrs per night on average.)

b. Calculate the test statistic, T, and the associated degrees of freedom.

Solution: The test statistic and degrees of freedom can be calculated as follows:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{7.73 - 8}{\frac{0.77}{\sqrt{25}}} = \frac{-0.27}{0.154} = -1.75$$

$$df = 25 - 1 = 24$$

c. Find and interpret the p-value in this context.

Solution: p-value $= \Pr(T_{24} < -1.75) \rightarrow 0.025 <$ p-value < 0.05. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hrs per night or less is between 0.025 and 0.05.

d. What is the conclusion of the hypothesis test?

Solution: Since p-value < 0.05, reject H_0. The data provide convincing evidence that New Yorkers sleep less than 8 hours per night on average.

e. If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

Solution: No, the hypothesis test suggests that the average amount of sleep New Yorkers get is significantly lower than 8 hours per night, therefore we wouldn't expect 8 hours to be in the interval. Note that the confidence level corresponding to this test is 90%, since the test is one-sided and uses a significance level of 5%.

8 Interpreting public opinion polls

On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.

a. We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

Solution: False. A confidence interval is constructed to estimate the population proportion, not the sample proportion.

b. We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

Solution: True. This is the correct interpretation of the confidence interval, which can be calculated as $0.46 \pm 0.03 = (0.43, 0.49)$.

c. If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

Solution: False. The confidence interval does not tell us what we might expect to see in another random sample.

d. The margin of error at a 90% confidence level would be higher than 3%.

False. As the confidence level decreases, the margin of error decreases as well.

9 Approval of marijuana

The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.

a. Is 48% a sample statistic or a population parameter? Explain.

Solution: 48% is a sample statistic, it's the observed sample proportion.

b. Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

Solution: A 95% confidence interval can be calculated as follows:

$$\begin{aligned}\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.48 \pm 1.96 \sqrt{\frac{0.48(1-0.48)}{1259}} \\ &= 0.48 \pm 1.96 \times 0.014 \\ &= 0.48 \pm 0.0274 \\ &= (0.4526, 0.5074)\end{aligned}$$

We are 95% confident that approximately 45% to 51% of Americans think marijuana should be legalized.

c. A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

Solution: No, the interval contains 50%, suggesting that the true population proportion could be 50%, or even lower. Using this interval we wouldn't reject a null hypothesis where $p = 0.50$.

10 Adopting open-source textbooks

A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

a. State the hypotheses for testing if the professor's predictions were inaccurate.

Solution: The hypotheses are as follows:

- H_0: The distribution of the format of the book used by the students follows the professor's predictions.
- H_A: The distribution of the format of the book used by the students does not follow the professor's predictions.

b. How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

Solution:

$$\begin{aligned} E_{hardcopy} &= 126 \times 0.60 &&= 75.6 \\ E_{print} &= 126 \times 0.25 &&= 31.5 \\ E_{online} &= 126 \times 0.15 &&= 18.9 \end{aligned}$$

c. Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

Solution: The χ^2 statistic, the degrees of freedom associated with it, and the p-value can be calculated as follows:

$$\begin{aligned} \chi^2 &= \sum \frac{O-E)^2}{E} = \frac{(71-75.6)^2}{75.6} + \frac{(30-31.5)^2}{31.5} + \frac{(25-18.9)^2}{18.9} = 2.32 \\ df &= 2 \\ \text{p-value} &> 0.3 \end{aligned}$$

d. Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

Solution: Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.