BUSGEN108/CS171/DATASCI161/ECON115 Fall 2025
Causality, Decision Making, and Data Science Homework 0

Instructions: Please complete this problem set *on your own*. You may either type or *very* legibly write your solutions, and upload them on Gradescope by 11:59pm on Monday September 29.

Note: This homework is graded for completeness, not correctness. The point is for us (and you) to assess your statistics background relevant to this course.

For this problem set, consider the following scenario. There are $N = 8000$ undergraduates at Stanford. We'd like to understand how many caffeinated beverages each student drinks before 2pm each day. To do this, we chose $n = 100$ random undergraduates (with replacement) and asked them. We got[1] the following data:

Number of caffeinated bevs	Number of students
0	20
1	15
2	40
3	20
4	5

1. Let X_i, for $i = 1, \ldots, n$, be the number of caffeinated beverages the i'th student in our sample consumed.

   (a) Let μ be the average number of caffeinated beverages consumed before 2pm out of the entire undergraduate population. In terms of the X_i, give an unbiased estimator $\hat{\mu}$ for μ. What is $\hat{\mu}$ with the data above?

   Solution

   As we saw in class, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of μ. Plugging in the data above, this is

   $$\hat{\mu} = \frac{1}{100}(15 \times 1 + 40 \times 2 + 20 \times 3 + 4 \times 5) = 1.75 \text{ caffeinated beverages.}$$

   (b) Let σ^2 be the variance of the number of caffeinated beverages consumed before 2pm out of the entire undergraduate population. In terms of the X_i, give an unbiased estimator for σ^2. What is your estimate of σ^2 with the data above?

[1] Disclaimer: these data are made up. In future problem sets, you'll be working with real data!

Solution

An unbiased estimator of σ^2 is the sample variance, which is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2,$$

where $\hat{\mu}$ is as in part (a). In our case, this is

$$\begin{aligned} s^2 &= \frac{1}{99}\left(20(0-1.75)^2 + 15(1-1.75)^2 + 40(2-1.75)^2 + 20(3-1.75)^2 + 5(4-1.75)^2\right) \\ &\approx 1.3. \end{aligned}$$

2. In this question, we'll compute (analytically) standard errors for $\hat{\mu}$ from the previous problem. One way to do this is (a) work out what the variance of $\hat{\mu}$ should be, and then (b) plug in any estimates we need to do get a number. To get standard errors, we then take the square root of our estimate of $\text{Var}(\hat{\mu})$.

   (a) Explain why $\text{Var}(\hat{\mu}) = \frac{1}{n}\sigma^2$, where σ^2 is as in the previous problem. (We are looking for a mathematical derivation / justification).

   Hint: If X and Y are independent, then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

Solution

We have

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) \\ &= \frac{1}{n}\text{Var}(X), \end{aligned}$$

where X is the number of caffeinated beverages for a random student in the whole population. That's because, individually, each X_i is distributed the same as X. But $\sigma^2 = \text{Var}(X)$ by definition. Above, we used the fact that the X_i's are independent to bring the Variance inside the sum.

   (b) Use your answer to the previous part to obtain a standard error for your estimate of $\hat{\mu}$ in Question 1. (We are looking for a number; show your work).

Solution

From the above, we know that $\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\sigma^2$. We don't know σ^2, but we estimated it by $s^2 \approx 1.3$ in the previous problem. So plugging that in, we get

$$\sqrt{\mathrm{Var}(\hat{\mu})} \approx \sqrt{\frac{1.3}{100}} \approx 0.114.$$

(c) What does your answer to part (b) tell you about the accuracy of your estimate $\hat{\mu}$? In particular, do you think that $\hat{\mu}$ is a "good" estimate for μ in this case?

Solution

It tells us that we would expect our estimate of 1.75 may be off by something in the ballpark of 0.11. (As we'll see in Question 4, plus or minus about twice that is a good confidence interval in this case). Since 0.11 is a lot less than 1.75, perhaps we are happy with this estimate, although it depends on the question we are trying to answer.

3. In this question, we'll discuss another way to estimate standard errors, which is *bootstrapping* (also called *resampling*). The way this works is the following. As above, let $X_1, X_2, \ldots, X_n$ be our samples of number of caffeinated beverages.

Consider the following procedure, which produces an estimate for $\mathrm{Var}(\hat{\mu})$.

- For $t = 1, 2, \ldots, T$ (where T is a large number, say $1,000$ or so):
  - draw $Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_n^{(t)}$ *with replacement* at random from $X_1, X_2, \ldots, X_n$.
  - Compute $\hat{\mu}_t^*$ from the samples $Y_1^{(t)}, \ldots, Y_n^{(t)}$ in the same way that you computed $\hat{\mu}$ from $X_1, \ldots, X_n$.
- Now, you have T quantities $\hat{\mu}_1^*, \hat{\mu}_2^*, \ldots, \hat{\mu}_T^*$.
- Estimate the variance of $\hat{\mu}$ by the sample variance of $\{\hat{\mu}_1^*, \hat{\mu}_2^*, \ldots, \hat{\mu}_T^*\}$.

We aren't going to ask you to implement this on this problem set (that's best done with a computer, not by hand). Instead, answer the following questions.

(a) Have you seen bootstrapping before? ("No" is a perfectly fine answer).

Solution

Yes for us, but "No" can also be correct :)

(b) Whether or not you've seen it before, why do you think this might be a good idea for estimating $\mathrm{Var}(\hat{\mu})$. Intuitively, what's going on here?

Solution

In a perfect world, we'd estimate $\text{Var}(\hat{\mu})$ according to the procedure above, except where $Y_i^{(t)}$ are all chosen iid from the original population of N undergraduates. But of course we don't have that data. So instead we approximate what that might look like, by resampling from the data that we do have. This gives us a sense of how much our estimate of $\hat{\mu}$ would change, if we had selected a slightly different sample of n students.

(c) What should you do to the output of the procedure above to estimate the standard error for $\hat{\mu}$?

Solution

Since the above outputs an estimate of $\text{Var}(\hat{\mu})$, we should take the square root to get an estimated standard error.

(d) *(**BONUS:** This problem isn't required, but might be fun if you know how to program.)* Implement the procedure above and estimate the standard error for $\hat{\mu}$. How does this compare to your estimate of the standard error in Question 2?

Solution

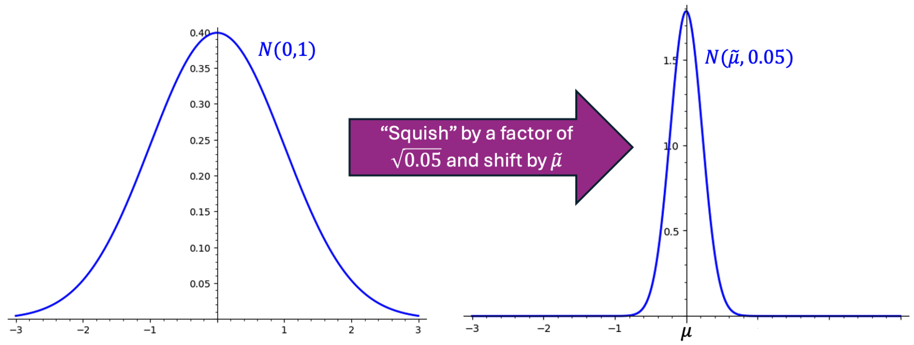
We implemented it and got a standard error of about 0.1. So it's pretty close to question 2.

4. In this question, we'll analytically construct confidence intervals for $\hat{\mu}$. The first observation is that $\hat{\mu}$ is a sum of independent random variables. (If the $\hat{\mu}$ that you got in Question 1 isn't a sum of independent random variables, go back and think about that question again...) The *Central Limit Theorem* implies that the distribution of $\hat{\mu}$ should be approximately Gaussian, with mean μ, and variance $\text{Var}(\hat{\mu})$.

(a) For a standard Gaussian $Z \sim \mathcal{N}(0, 1)$ (that is, Z is a Gaussian random variable with mean zero and variance one), a 95% confidence interval is $[-1.96, +1.96]$. This means that the probability that Z is outside that interval is at most 5%. What is a 95% confidence interval for a random variable $W \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$?

Hint: Think about what the pdf of $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ looks like. To get from the pdf of $\mathcal{N}(0, 1)$ to the pdf of $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, we "squish" the pdf by a factor of $\tilde{\sigma}$, and then shift it by $\tilde{\mu}$ (see the picture below). What does that transformation do the confidence interval?

Example: To get from $N(0,1)$ to $N(\tilde{\mu}, 0.05)$:



Solution

The mean is $\tilde{\mu}$ instead of 0, so the confidence interval should be centered at $\tilde{\mu}$. Following the hint, we need to “squish” the confidence interval by a factor of $\tilde{\sigma}$. So we should have

$$[\tilde{\mu} - 1.96 \cdot \tilde{\sigma}, \tilde{\mu} + 1.96 \cdot \tilde{\sigma}].$$

(b) Based on your answer above, give a 95% confidence interval for $\hat{\mu}$ in our running example. We are looking for an interval with real numbers (like “$[0.26, 23.33]$”, but more correct), along with an explanation.

Solution

Based on the above, we should plug in $\tilde{\mu} \leftarrow \hat{\mu} = 1.75$, and $\tilde{\sigma} \leftarrow 0.11$, our estimate for the standard error of $\hat{\mu}$ from earlier. (If you wanted to plug in an estimated standard error that you got from bootstrapping, that’s okay too). So we get

$$[1.75 - 1.96 \cdot 0.11, 1.75 + 1.96 \cdot 0.11] \approx [1.53, 1.97].$$

5. (**BONUS.** This question is more open-ended. It is optional, but worth thinking about!)

In Question 3 we explored bootstrapping. You can use the same technique to compute confidence intervals! Propose a way to use bootstrapping to come up with confidence intervals, and explain why it should be a good idea. (Note: there are several ways to do this. In the solutions to this problem set we’ll give one way for your reference.)

Solution

The basic idea is to resample with replacement to get $\hat{\mu}_1^*, \ldots, \hat{\mu}_T^*$ as before. But instead of looking at the sample variance to compute standard errors, we can look at the whole distribution of $\{\hat{\mu}_1^*, \ldots, \hat{\mu}_T^*\}$, and see what a 95% (or whatever) confidence interval would have been for that distribution. Here is one way to do it:

- Compute $\hat{\mu}_1^*, \ldots, \hat{\mu}_T^*$ as before.
- For each $t = 1, \ldots, T$, compute $\delta_t = \hat{\mu}_t^* - \hat{\mu}$.
- Sort the values δ_i. Let δ_1^* be the value at the 2.5'th percentile, and δ_2^* be the value at the 97.5'th percentile.
- Return $[\hat{\mu} - \delta_2^*, \hat{\mu} - \delta_1^*]$ as a 95% confidence interval.

This is sometimes called the "pivotal interval" or the "empirical interval" or the "basic interval." Another natural thing to do is just to take a 95% confidence interval directly from the distribution of the $\hat{\mu}_t^*$; that's called the "percentile interval."[a] There are other fancier ways to do it too.

[a] In general, the pivotal interval is a better idea if you think that the sampling distribution is approximately symmetric around the true parameter, but are worried that your estimator may be biased. The percentile interval is a better idea if you expect that the sampling distribution isn't symmetric, and that your estimator is unbiased.

6. You are curious about how caffeine consumption varies by demographic. Some people you talk to think that seniors drink more caffeinated beverages than freshman, on average. Others think that the number of caffeinated beverages for these two groups come from the same distribution. Unfortunately, you didn't record the class year in the data set at the beginning of the problem set, so you do a quick test: you ask five random freshman and five random seniors. Suppose you see:

   - Freshmen: $[1, 1, 0, 1, 0]$
   - Seniors: $[3, 4, 2, 2, 3]$

   Comparing the means of your samples, it does seem that seniors are more caffeinated than freshman, but how can you be sure with only five samples in each group?

   (a) If the groups were larger, so that the central limit theorem might apply, how would you decide if the hypothesis "The number of caffeinated beverages for freshmen and seniors are drawn from the same distribution" is likely true or not?

   Hint: We saw in Problem 4 how to construct confidence intervals for Gaussians. If the samples were bigger, the central limit theorem would imply that the random variable $\hat{\mu}_F - \hat{\mu}_S$ is approximately Gaussian. (Here, $\hat{\mu}_F$ is the sample mean for freshman, and $\hat{\mu}_S$ is the sample mean for seniors; the randomness is over the random choice of students). What is its mean and variance? How could you use that to construct a confidence interval? What would that confidence interval tell you about this hypothesis?

Solution

Let H_0 be the hypothesis that $\mu_F - \mu_S = 0$, where μ_F is the average for frosh and μ_S is the average for seniors. Let $\hat{\mu}_F$ and $\hat{\mu}_S$ be our sample means. Let s_F^2 and s_S^2 be the sample variances. Then the CLT implies that $\hat{\mu}_F - \hat{\mu}_S$ is approximately Gaussian, with mean $\mu_F - \mu_S$. We claim that the variance is $\frac{\sigma_F^2}{n_F} + \frac{\sigma_S^2}{n_S}$, where n_F and n_S are the number of freshman and seniors samples, respectively. (To see this, we do a similar calculation to the one you did in Problem 2). Thus, we can estimate the variance by $\frac{s_F^2}{n_F} + \frac{s_S^2}{n_S}$. So we can construct a 95% confidence interval like we did in a previous problem:

$$(\hat{\mu}_F - \hat{\mu}_S) \pm 1.96 \cdot \sqrt{\frac{s_F^2}{n_F} + \frac{s_S^2}{n_S}}.$$

If 0 does not lie in this confidence interval, we reject.

(b) Given that the groups are small, and the central limit theorem may not apply, what can you do?

Hint: Imagine the two data sets were from the same population, as in the hypothesis. How likely would it be that these heights were split up between the two groups in this way?

Solution

We could use resampling, as we did in the previous problem, although we don't recommend that for samples so small. However, we can reason about this directly, following the hint. Suppose that all these numbers did come from the same distribution. Notice that *all* of the Seniors numbers are bigger than *all* of the Freshman's numbers. What are the odds of that happening, if they all came from the same distribution? It's at most $\frac{1}{\binom{10}{5}}$. That's because there are $\binom{10}{5}$ ways to split 10 numbers into two groups, and only one way to do so that puts all of the numbers ≥ 2 in one group and all of the numbers ≤ 1 in the other group. Since $\frac{1}{\binom{10}{5}} \approx 0.004$ is very small, we can confidently reject the null hypothesis that all these numbers came from the same distribution. (In case you are interested, this counting method illustrates the logic behind the *Mann-Whitney U test*, a nonparametric statistical test that performs a similar analysis on the ranks of the combined data.)