Second midterm practice problem solutions

Joshua Loftus

Solutions

1. The median household income in the US is about \$59,000, while the mean household income is about \$72,000. Suppose we survey households randomly, asking for household income, and use the *sample mean* $\bar{X}$ as an estimator of the median household income. What is the bias of that estimator?

$\theta = 59,000$ and $E[\hat{\theta}] = 72,000$, so $\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta = 13,000$

2. Suppose $W_1, W_2, \ldots, W_{100}$ are independent and identically distributed, with $E[W_1] = 1 + \mu/2$ and Var$(W_1) = 4$. What is the standard deviation of $\bar{W}$? Is $\bar{W}$ an unbiased estimator of μ?

$\text{sd}(\bar{W}) = \text{sd}(W_1)/\sqrt{n} = \sqrt{4}/\sqrt{100} = 2/10$

Since $\text{Bias}(\bar{W}) = E[\bar{W}] - \mu = 1 + \mu/2 - \mu = 1 - \mu/2$, this does not equal 0, so the estimator is biased.

3. A startup develops algorithms to determine if an article is "fake news." They do this by defining a parameter θ representing the trustworthiness of the given article. Their algorithms input the text of an article and output estimates $\hat{\theta}$. Engineers develop two candidate algorithms: one using advanced deep learning methods $\hat{\theta}_{\mathbf{DL}}$, and one using a simpler model called logistic regression $\hat{\theta}_{\mathbf{LR}}$. The DL estimator is unbiased, but has a variance equal to 1. The LR estimator has a bias of $-\frac{1}{2}$ and a variance of $\frac{1}{2}$. What is the MSE of the LR estimator? Which method has the lower MSE?

$\text{MSE}(\hat{\theta}_{\text{LR}}) = \text{Bias}(\hat{\theta}_{\text{LR}})^2 + \text{Var}(\hat{\theta}_{\text{LR}}) = (-1/2)^2 + 1/2 = 3/4$

And $\text{MSE}(\hat{\theta}_{\text{DL}}) = 0^2 + 1 = 1$ so the LR estimator has lower MSE.

4. Continuing problem 1 above, suppose that instead of a sample of size 100 we now continue gathering new observations of W_i, for $i = 101, 102, \ldots$. How large does the sample have to be for the standard deviation of $\bar{W}$ to be as low as 1/100? If we continue increasing the sample size indefinitely $n \to \infty$, does $\bar{W}$ converge to the true parameter μ?

Set $\text{sd}(\bar{W}) = 1/100$ and solve for n. That is, $2/\sqrt{n} = 1/100$, so $\sqrt{n} = 200$ or $n = 40,000$.

The law of large numbers says $\bar{W} \to E[W_1] = 1 + \mu/2$, so no, $\bar{W}$ does not converge to μ.

5. Suppose population household income in the US has mean $\mu = \$70,000$ and standard deviation $\sigma = \$30,000$. In this problem we know these true parameters. We survey households randomly and collect a sample of size $n = 100$, and let $\bar{X}$ denote the mean income of the sample. How would you use the normal distribution to approximate $P(\bar{X} < \$64,000)$? Why can you do this even though the distribution of incomes is not normal? (We know it is not normal because it is skewed)

$P(\bar{X} < 64000) \approx P(Z < 64000)$ for $Z \sim N(70000, 30000^2/100)$.

Since $64000 = 70000 - 2 \cdot 3000$ this is the probability of normal being less than 2 standard deviations below its mean.

We know probability of being outside 2 standard deviations from the mean is 5%, and this is split evenly between upper and lower tails. So $P(\bar{X} < 64000) \approx 2.5\%$.

We can do this because of the central limit theorem!

6. The standard deviation of a random variable is σ and the standard error of the mean of an i.i.d. sample of size n, with $n > 1$, of the same random variable is SE. Which of the following are true? Indicate with a check mark.

- $\sigma = SE$
- $\sigma > SE$ ✓
- σ decreases as n increases
- SE decreases as n increases ✓
- If the sample increases from n to $2n$, then SE decreases to $SE/2$
- If the sample increases from n to $2n$, then σ decreases to $\sigma/2$
- Neither σ nor SE decrease as n increases

(Remember: $SE = \sigma/\sqrt{n}$)

7. Suppose $U_1, U_2, \ldots, U_n$ are i.i.d., $E[U_1] = \mu$, Var$(U_1) = \sigma^2$, and the overall distribution of U_1 is right-skewed. Is the normal distribution $N(\mu, \sigma^2)$ a good approximation for the distribution of U_1? Why or why not? What about for $\bar{U}$? Why or why not?

No, because U_1 is skewed (so for example $P(U_1 > \mu + \sigma)$ will be much larger than $P(Z > \mu + \sigma)$).

And no, because the right variance for $\bar{U}$ is σ^2/n, not σ^2.

8. Continuing problem 7, let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \bar{U})^2$. What is the T statistic for this sample? Suppose $n = 100$, $\bar{U} = 1.8$, $S = 10$, and $P(T > 2.62) = 0.005$ (so $P(T < -2.62) = 0.005$). What is the 99% confidence interval for μ based on this sample? (You don't need to simplify expressions with numbers)

$$T = \frac{\bar{U} - \mu}{SE} = \frac{\bar{U} - \mu}{S/\sqrt{n}} = \frac{1.8 - \mu}{10/\sqrt{100}}$$

99% confidence interval: $\bar{U} \pm 2.62SE = \bar{U} \pm 2.62$

9. (Continuing from problem 8 above) True or false, and explain: the probability that the interval computed above contains μ is 99%.

False. Once we observe specific numbers for the random variables, the interval either containts μ or does not. This is enough of an answer, but I'll give more explanation now because this is a bit of a subtle point.

Probability is a model of uncertainty for before the outcome is observed. It doesn't make sense to ask what is the probability a dice lands on 6 when the dice has already been rolled–either it has already landed on 6 or it hasn't.

Like in Problem 5, if we collect data and observe a specific value for the sample mean $\bar{x}$ (notice the lower case instead of upper case) there is no longer any "randomness" to calculate $P(\bar{x} < 64,000)$. Either $\bar{x}$ is less than 64,000 or it isn't.

99% confidence means that if we repeated the experiment by collecting another sample, calculating the interval for that sample, and repeating many times, then about 99% of those intervals would contain the true μ.