Study guide for second midterm exam

Joshua Loftus

Suggested studying practices:

- Read the lecture notes posted on the course page. You can mostly skip over `R` code, since **there will be no coding on the exam**. You might want to look at the output from the code, like for some of the plots, since that may help understand the given example. These notes should be sufficient to answer any exam problems.
- Review homework solutions.
- Any topics covered in lectures may appear on the exam. This includes, but is not limited to: sampling distribution of the mean, parameter estimation, unbiasedness, mean-squared error and bias-variance tradeoff, the law of large numbers, the central limit theorem, and confidence intervals.

Practice questions

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?

2. Suppose $W_1, W_2, \dots, W_{100}$ 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?

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?

4. Continuing problem 1 above, suppose that instead of a sample of size 100 we now continue gathering new observations of W_i, for $i = 101, 102, \dots$. 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?

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)

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

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?

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 $P(T > 2.62) = 0.995$. What is the 99% confidence interval for μ based on this sample? (You don't need to simplify expressions with numbers)

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