STA 130A – Homework 7

Submission due: Thu, December 3 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a single PDF file to Gradescope via Canvas ("Homework 7" under "Assignments"). Name the file using the prefix of your UC Davis email ID and the homework number (e.g., `dgsong_hw7.pdf`). Include "STA 130A," your name, and the last four digits of your student ID on the front page. No late submissions will be accepted; any submission received after the deadline will receive 0 points. For full information about submission requirements and the late submission policy, see the syllabus. Show your work and briefly justify your conclusions.

Problem 1 (20 points in total). Sampling distributions and estimation error

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where μ and $\sigma^2 > 0$ are unknown and $n \geq 2$. Define

$$\bar{X}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \text{and} \qquad S_n = \sqrt{S_n^2} \quad \text{where} \quad S_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

(a) (5 points) Which of $\bar{X}_n$, S_n^2, and $\bar{X}_n - \mu$ are statistics? Explain briefly. For $n = 16$, $\bar{x} = 12.4$, and $s = 2$, report the value and estimated standard error of the estimator $\widehat{\mu} = \bar{X}_n$.

(b) (5 points) State the exact distributions of

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}, \qquad U = \frac{(n-1)S_n^2}{\sigma^2}.$$

Are Z and U independent? State the normal-sample result you use.

(c) (5 points) Use the mean and variance of a chi-square random variable to find $\mathbb{E}[S_n^2]$ and $\mathrm{Var}(S_n^2)$.

(d) (5 points) Consider $V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ as an estimator of σ^2. Find its bias and MSE. Does the unbiased estimator S_n^2 have smaller MSE here?

Problem 2 (20 points in total). Confidence about a population mean

The following nine fill volumes (in mL) form a random sample from a normal population:

$$43, \quad 44, \quad 48, \quad 51, \quad 53, \quad 55, \quad 56, \quad 58, \quad 60.$$

Hint: You may use the lower-tail critical values $z_{0.975} \approx 1.960$ and $t_{8,0.975} \approx 2.306$.

(a) (5 points) Compute the sample mean. Suppose the population standard deviation is known to be $\sigma = 6$ mL. Construct a 95% confidence interval for the population mean μ. Report the standard error and margin of error.

(b) (5 points) Now suppose σ is unknown. Compute the sample standard deviation s and construct a 95% confidence interval for μ. State the exact distribution of $(\bar{X} - \mu)/(S/\sqrt{n})$.

(c) (5 points) Interpret the 95% confidence level in terms of repeated sampling. Does it mean that 95% of individual fill volumes lie in the interval from part **(b)**?

(d) (5 points) With $\sigma = 6$ mL known, consider the interval $[\bar{X}_n - 1.5, \bar{X}_n + 1.5]$. Find the smallest n giving at least 95% coverage under normality. Without assuming normality, find the smallest n for which Chebyshev's inequality guarantees this coverage.

Problem 3 (20 points in total). Method of moments and maximum likelihood

Let $X_1, \ldots, X_n$, with $n \geq 2$, be i.i.d. completion times, each modeled as a start time $\delta \in \mathbb{R}$ plus an exponential waiting time with mean $\theta > 0$. Thus

$$f_{\delta,\theta}(x) = \begin{cases} \dfrac{1}{\theta} e^{-(x-\delta)/\theta}, & x \geq \delta, \\ 0, & x < \delta, \end{cases} \qquad \mathbb{E}[X_i] = \delta + \theta, \qquad \mathrm{Var}(X_i) = \theta^2.$$

For parts **(a)**–**(c)**, suppose $\delta = 2$ is known. Use the observations $3, 4, 4, 7$ for numerical estimates.

(a) (5 points) Derive the method-of-moments estimator of θ using the first moment and compute its estimate for the given data.

(b) (5 points) For observations $x_1, \ldots, x_n > 2$, write the likelihood and log likelihood for θ. Derive the maximum likelihood estimator and justify the maximum. Compare the estimate with part **(a)**.

(c) (5 points) For the estimator from part **(a)** and $T_n = X_1 - 2$, find the bias and MSE. Determine which are consistent for θ and justify your conclusions.

(d) (5 points) Now suppose both δ and θ are unknown. Match the first two population moments to the first two sample moments to derive method-of-moments estimators of both parameters. Compute their estimates for the same data.

Problem 4 (20 points in total). Estimating an unknown upper limit

Let $X_1, \ldots, X_n$ be i.i.d. Uniform$[0, \theta]$, where $\theta > 0$ is unknown and $n \geq 2$. The density is $1/\theta$ on $[0, \theta]$ and zero elsewhere. Write $M_n = \max\{X_1, \ldots, X_n\}$.

(a) (5 points) For the four observations $1, 2, 2, 7$, find the method-of-moments and maximum likelihood estimates of θ. Show the moment equation and likelihood used. Does each fitted interval $[0, \widehat{\theta}]$ contain all four observations?

(b) (5 points) Find the CDF of M_n for all real arguments and its PDF. Indicate where independence is used.

(c) (5 points) Find an unbiased estimator of θ of the form $a_n M_n$. Compute its MSE and compare it with $\theta^2/(3n)$, the MSE of the method-of-moments estimator $2\bar{X}_n$.

Hint: You may use

$$\mathbb{E}[M_n] = \frac{n\theta}{n+1}, \qquad \mathbb{E}[M_n^2] = \frac{n\theta^2}{n+2}.$$

(d) (5 points) For $0 < \alpha < 1$, find a constant $c > 1$ depending only on n and α such that $[M_n, cM_n]$ is an exact $100(1-\alpha)\%$ confidence interval for θ. Verify its coverage using the CDF from part **(b)**.

Hint: Rewrite $\{\theta \in [M_n, cM_n]\}$ as an event involving M_n.

Problem 5 (20 points in total). Estimating a defect rate from imperfect screening

Items are sampled and screened independently of one another. Each item is defective with probability p, where $0 \le p \le 1$ is unknown. A screening device reports a positive result with probability 0.9 for a defective item and 0.1 for a nondefective item. Let $Y_i = 1$ if item i tests positive and $Y_i = 0$ otherwise, and let $K = \sum_{i=1}^{n} Y_i$. Only the screening outcomes are observed; the true defect statuses are unknown.

(a) (5 points) Find $q = P(Y_i = 1)$ in terms of p. Find the probability that an item is defective given a positive result.

(b) (5 points) Identify the distribution of K and write the likelihood for p when $K = k$. Find the maximum likelihood estimate of p for every possible k, respecting $0 \le p \le 1$.

(c) (5 points) Consider $A_n = (K/n - 0.1)/0.8$. Find its expectation and variance. Can the MLE from part **(b)** have a larger MSE than A_n for any $p \in [0, 1]$? Justify your answer by comparing their squared errors for each possible value of A_n.

(d) (5 points) Suppose 30 of 100 items test positive. Use A_n to estimate p and construct an approximate 95% confidence interval using a normal approximation without continuity correction. Estimate q by K/n in the standard error and use $z_{0.975} \approx 1.96$.

For planning, suppose $p = 1/4$. Find the number of screened items needed for A_n to have the same variance as the sample proportion from 100 perfectly classified items.

Problem 6* (up to 10 bonus points). Further questions about inference

Each part is optional and worth up to 5 additional points.

(a) (5 bonus points) Predicting a future observation. Under the normal sampling model in Problem 1, let X_{n+1} be an independent future observation from the same population. Find a constant $a_n > 0$ such that

$$P\big(|X_{n+1} - \bar{X}_n| \le t_{n-1,0.975}\, a_n S_n\big) = 0.95.$$

Justify your choice using Student's t distribution and compare the resulting interval's width with that of a 95% confidence interval for μ based on the same sample.

Hint: $X_{n+1} - \bar{X}_n$ is normal and independent of S_n^2.

(b) (5 bonus points) When more data are not enough. In Problem 5, keep the 0.9 detection probability but let the false-positive probability $a \in [0, 0.9)$ also be unknown. Find two pairs (p, a) with different p that make $Y_1, \ldots, Y_n$ i.i.d. Bernoulli(0.3) for every n. Explain whether increasing the sample size can distinguish these two defect rates.