EK381 Probability, Statistics, and Data Science for Engineers — Boston University
Prof Carruthers — Summer 2025

Homework 9: finish by 6/23

Reading: Notes: Chapter 7 - 9 **Videos:** 7.1 - 9.2

Problem 9.1 (**Video 7.1, 7.2, 7.3, 8.1, 8.2**, **Quick Calculations**) For each of the scenarios below, determine the requested quantities. (You should be able to do this without any long calculations or integration.)

(a) Assume that X is Uniform$(1,3)$, V is Gaussian$(1,1)$, X and V are independent, and $Y = X + V$. Determine the LLSE estimator of X given $Y = y$ and the corresponding mean-squared error.

(b) We would like to create a scalar LLSE estimator for X using either Y_1, Y_2, or Y_3. Using the correlation coefficient matrix below, determine which of Y_1, Y_2, or Y_3 will result in the smallest mean-squared error. Once you have chosen your variable, determine the corresponding scalar LLSE estimator and its mean-squared error. You may assume that $\mathbb{E}[X] = \mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \mathbb{E}[Y_3] = 0$, $\mathsf{Var}[X] = 4$, and $\mathsf{Var}[Y_1] = \mathsf{Var}[Y_2] = \mathsf{Var}[Y_3] = 1$.

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_{X,X} & \rho_{X,Y_1} & \rho_{X,Y_2} & \rho_{X,Y_3} \\ \rho_{Y_1,X} & \rho_{Y_1,Y_1} & \rho_{Y_1,Y_2} & \rho_{Y_1,Y_3} \\ \rho_{Y_2,X} & \rho_{Y_2,Y_1} & \rho_{Y_2,Y_2} & \rho_{Y_2,Y_3} \\ \rho_{Y_3,X} & \rho_{Y_3,Y_1} & \rho_{Y_3,Y_2} & \rho_{Y_3,Y_3} \end{bmatrix} = \begin{bmatrix} 1 & 0.4 & 0.7 & -0.8 \\ 0.4 & 1 & 0.5 & -0.3 \\ 0.7 & 0.5 & 1 & -0.6 \\ -0.8 & -0.3 & -0.6 & 1 \end{bmatrix}$$

(c) With the same scenario as in part (b), now we would like to create a *vector* LLSE estimator for X from $\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$. Recall that the vector LLSE formula is $\hat{x}_{\text{LLSE}}(\underline{y}) = \mathbb{E}[X] + \mathbf{\Sigma}_{X,\underline{Y}} \mathbf{\Sigma}_{\underline{Y}}^{-1} (\underline{y} - \mathbb{E}[\underline{Y}])$. Your job is to use the parameters from part (b) to specify the values of $\mathbb{E}[X]$, $\mathbb{E}[\underline{Y}]$, $\mathbf{\Sigma}_{X,\underline{Y}}$, and $\mathbf{\Sigma}_{\underline{Y}}$. You do not need to carry out the matrix inverse or the matrix-vector multiplications, just specify the requested vectors and matrices.

(d) Let $X_1, X_2, \ldots, X_{400}$ be a collection of continuous, independent, identically distributed continuous random variables, each of which is uniformly distributed over $[-5, 5]$. Let $Y = \frac{1}{400} \sum_{i=1}^{400} (X_i)^3$. Compute $\mathbb{E}[Y]$ and $\mathsf{Var}[Y]$. (This means you need to numerically evaluate at least one integral.)

(e) Let $X_1, X_2, \ldots, X_{100}$ be a collection of continuous, independent, Exponential$(\frac{1}{2})$ random variables. Let $Y = \frac{1}{100} \sum_{i=1}^{100} X_i$. Use the Central Limit Theorem to estimate $\mathbb{P}[Y \leq 1.9]$.

Problem 9.2 (Video 8.1, 8.2, 9.1, 9.2, Lecture Problem)

Let $X_1, \ldots, X_n$ be i.i.d. random variables with PMF

$$P_X(x) = \begin{cases} 1/2 & x = 2 \\ 1/2 & x = 4 \\ 0 & \text{otherwise.} \end{cases}$$

Let $S_{300} = X_1 + \cdots + X_{300}$.

(a) Determine the mean of and variance of S_{300}.

(b) Use the Central Limit Theorem approximation to estimate the probability $\mathbb{P}\big[\big|S_{300} - E[S_{300}]\big| \geq 40\big]$. You can leave your answer in terms of the standard normal CDF $\Phi(z)$.

(c) Suppose you did not know the PMF $P_X(x)$, but you knew that the standard deviation of the random variables X_k was one. You measure $X_1, \ldots, X_{300}$ and compute the sample mean $M_{300} = \frac{1}{300}\sum_{k=1}^{300} X_k$. Find a symmetric confidence interval for the true mean around the observed value M_{300} with confidence level 0.95.

Use the following assumptions: $Q(1.28) = 1 - \Phi(1.28) = 0.1$; $Q(1.645) = 1 - \Phi(1.645) = 0.05$; $Q(1.96) = 1 - \Phi(1.96) = 0.025$.

(d) Suppose you also did not know the standard deviation, but you were able to compute the sample variance of the 300 samples X_k as $V_{300} = 1.21$. Find a symmetric confidence interval for the true mean around the observed value M_{300} with confidence level 0.95.

Use the following assumptions: if W has a t-distribution with 299 degrees of freedom, its CDF satisfies: $F_W(-1.9679) = 0.025$; $F_W(-1.65) = 0.05$; $F_W(-1.2844) = 0.10$.

Problem 9.3 (Video 9.1, 9.2, Lecture Problem)

You are trying out a new cholesterol drug with a control group, consisting of 50 samples, and an experimental group, consisting of 100 samples. The variance is believed to be $\sigma_1^2 = 1$ in the control group and $\sigma_2^2 = 1.44$ in the experimental group. For the control group, you obtain sample mean $M_{50}^{(1)} = 7.9$ and for the experimental group you obtain sample mean $M_{100}^{(2)} = 7.5$. You may find one or more of the following values useful:

$$\Phi(-1.96) = Q(1.96) = 0.025; \ \Phi(-1.645) = Q(1.645) = 0.05; \Phi(-1.282) = Q(1.282) = 0.1$$

$$F_{T_{49}}(-2.0096) = 0.025; \ F_{T_{49}}(-1.677) = 0.05; F_{T_{49}}(-1.299) = 0.1;$$

$$F_{T_{99}}(-1.984) = 0.025; \ F_{T_{99}}(-1.66) = 0.05; F_{T_{99}}(-1.29) = 0.1;$$

(a) Which is larger, $\mathsf{Var}[M_{50}^{(1)}]$ or $\mathsf{Var}[M_{100}^{(2)}]$?

(b) Construct a confidence interval for the mean of the experimental group with confidence level 0.9.

(c) How many samples would we need in the experimental group so that the interval $[7.5 - 0.1645, 7.5 + 0.1645]$ has 0.9 confidence of containing the true mean?

Problem 9.4 (**Video 9.1, 9.2**)

You measure the sulfate concentration in the local water reservoir over 9 consecutive days and obtain values $X_1, \dots, X_9$, which are assumed to be i.i.d. Gaussian. (The units are mg/L and omitted below.) The sample mean is $M_9 = 6.1$ and the sample variance is $V_9 = 0.36$.

Let W have a t-distribution with 8 degrees-of-freedom. You can assume the following values are available:

- $F_W(-1.4) = \Phi(-1.3) = Q(1.3) = 0.1, \;\; F_W(1.4) = \Phi(1.3) = 0.9$
- $F_W(-1.9) = \Phi(-1.6) = Q(1.6) = 0.05, \;\; F_W(1.9) = \Phi(1.6) = 0.95$
- $F_W(-2.3) = \Phi(-2.0) = Q(2.0) = 0.025, \;\; F_W(2.3) = \Phi(2.0) = 0.975$

(a) Construct a confidence interval for the mean with confidence level 0.9.

(b) Construct a confidence interval for the mean with confidence level 0.95.

(c) Say you also go out on the 10^{th} day and collect measurement $X_{10} = 5$. What is the new sample mean M_{10}?