18.650. Fundamentals of Statistics
Fall 2023. Problem Set 2

Due Monday, October 16

Problem 1

Let $X \sim N(1, 2.25)$. Compute the following probabilities

1. $\mathbb{P}(X > 1)$
2. $\mathbb{P}(|X - 2| \leq 1)$
3. $\mathbb{P}(|X| < 1)$
4. $\mathbb{P}(X^2 - 2X - 1 > 0)$

Problem 2 Let

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

Compute the following quantities

1. $\mathbb{V}[X]$
2. $\mathbb{E}[Y^2 + X]$
3. $\mathbb{E}[(X - Y)^2]$
4. $\mathbb{V}[X + 2Y]$
5. Find $\alpha > 0$ such that $\alpha X = Y$ with probability 1 or prove that no such α exists.

Problem 3 Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Ber}(.5)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathsf{Exp}(1)$. Assume further that all the random variables are mutually independent. Write a central limit theorem (compute the variance/covariance matrix) for each of the following quantities

1. $(\bar{X}_n, \bar{Y}_n)$
2. $\bar{X}_n - \bar{Y}_n$
3. $$X\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \,.$$

4.

$$\frac{(\bar{X}_n)^2}{\bar{Y}_n}$$

Problem 4 Estimation of a Bernoulli parameter

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables, with unknown parameter $p \in (0, 1)$.

1. Suppose we observe these $n \geq 4$ random variables. Write down a valid statistical model for the resulting data.
2. Define the estimator of p, $\hat{p}_3 = \dfrac{X_1 + X_2 + X_3}{3}$. What is the bias of $\hat{p}_3$?
3. What is the variance of $\hat{p}_3$?
4. What is the MSE of the estimator $\hat{p}_3$?
5. What is the MSE of the estimator $\bar{X}_n$.
6. Which estimator is better? Justify your answer.

Problem 5

Let $X_1, \ldots, X_n$ be i.i.d. $\mathsf{Exp}(\lambda)$ random variable, where λ is unknown. Thus, each X_i has density $e^{-x/\lambda}/\lambda, x \geq 0$[1].

1. What is the distribution of $\min_i X_i$ (compute the CDF and take its derivative)?
2. Use the previous question to give an unbiased estimator for λ.
3. What is MSE of the above estimator?
4. What is MSE of the plugin estimator?

Problem 6 Let $X_n \sim \mathsf{Unif}\big(-\frac{1}{n}, \frac{1}{n}\big)$ and let X be a random variable such that $\mathbb{P}(X = 0) = 1$.

1. Compute and draw the CDF $F_n(x)$ and $F(x)$ of X_n and X respectively.
2. Does $X_n \xrightarrow{\mathbb{P}} X$? (prove or disprove)
3. Does $X_n \rightsquigarrow X$? (prove or disprove)

[1]Some people use the alternative definition with density $\lambda e^{-\lambda x}$, but we stick with Wasserman's convention (see Section 2.4 in the textbook).

Problem 7

In a simple model of inheritance, a given person has genotypes AA, aa or Aa. Each person has two alleles (each of which is either a or A), and the first and the second allele are inherited independently and are identically distributed. Note that Aa is the same as aA: the order of alleles does not matter. Therefore $\mathbb{P}(\mathsf{AA}) = \theta^2$, $\mathbb{P}(\mathsf{aa}) = (1-\theta)^2$ where $\theta \in (0,1)$ is an unknown parameter. Our goal here is to estimate this parameter.

1. Compute $\mathbb{P}(\mathsf{Aa})$ in terms of θ.
2. Define the random vector $X = \{0,1\}^3$ associated to a random person with genotype g by
$$X = \begin{cases} (1,0,0) & \text{if } g = \mathsf{AA} \\ (0,1,0) & \text{if } g = \mathsf{aa} \\ (0,0,1) & \text{if } g = \mathsf{Aa} \end{cases}$$
What is the pmf $p(x)$ of X for $x = (x_1, x_2, x_3) \in \{(1,0,0),(0,1,0),(0,0,1)\}$?
   - $p(x) = \theta^x(1-\theta)^{1-x}$
   - $p(x) = (\theta^2)^{x_1}(1-\theta^2)^{x_2}(1-2\theta)^{x_3}$
   - $p(x) = (1-\theta)^{2x_1}(\theta)^{2x_2}([2\theta(1-\theta)]^{x_3}$
   - $p(x) = (\theta)^{2x_1}(1-\theta)^{2x_2}[2\theta(1-\theta)]^{x_3}$
   - $p(x) = (\theta)^{2x_3}(1-\theta)^{2x_1}[2\theta(1-\theta)]^{x_2}$
3. A sample of 942 males from Zimbabwe was collected, and the following genotypes were observed: 501 of type AA, 83 of type aa and 358 of type Aa. Compute the maximum likelihood estimator for θ.

Problem 8 Let $X_1, \ldots, X_n$ be independent copies of the random variable X where X is a mixture of two uniform random variables and has pdf:

$$f(x) = \frac{1}{4\theta}\mathbb{1}(x \in [0, 2\theta]) + \frac{1}{4\theta}\mathbb{1}(x \in [\theta, 3\theta])$$

for some unknown $\theta > 0$. For this problem, we call $\mathsf{Unif}[0, 2\theta)$, the first component.

1. Compute the proportion π of the first component.
2. Compute $\mathbb{E}[X]$ and $\mathbb{V}[X]$.
3. Assume that we start the k-th E-step of the EM algorithm with a candidate θ_k from the previous M step. Let $w_1, \ldots, w_n$ be the weights obtained in the E-step of the EM algorithm. Compute these weights and show that they can take only three values depending on X_i and θ_k.

4. Assume that $n = 8$ and the observations are (in order)

$$1.01 \quad 1.02 \quad 1.19 \quad 1.19 \quad 1.28 \quad 2.39 \quad 2.56 \quad 2.58$$

and that the the EM algorithm is initialized at $\theta_0 = 3$, what are the values of the iterates: θ_1, θ_2 and θ_3?

Problem 9

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, I)$, where $\mu \in \mathbb{R}^p$ and I is the $p \times p$ identity matrix (the X_i are random vectors).

1. Let A be a $p \times p$ matrix such that $A\mu = 0$ and $\mathsf{Tr}(A) = 1$ (here $\mathsf{Tr}(A)$ denotes the trace of A). Compute $\mathbb{E}[X^\top A X]$.
2. What is the likelihood function for μ?
3. Compute the maximum likelihood estimator $\hat{\mu}_{MLE}$ for μ. Prove that it actually maximizes the likelihood function.
4. What is the distribution of $\hat{\mu}_{MLE}$?
5. Let B be a fixed $m \times p$ matrix. What is the variance of $B\hat{\mu}_{MLE}$?
6. Define the function $g(X) = \|X\|^2$. What is the asymptotic variance of $g(\hat{\mu}_{MLE})$?

Problem 10

The lifetime (in months) of a cell phone battery is modeled by a random variable X that has pdf

$$f_\theta(x) = K\theta^x \mathbf{1}(x > 0)$$

for an unknown $\theta \in (0, 1)$. Assume that we have n independent observations $X_1, \ldots, X_n$ of the lifetime of n cell phones. We want to use them to estimate $\theta \in (0, 1)$.

1. Show that $K = \log(1/\theta)$.
2. Compute the expected value and the variance of X.
3. Compute the maximum likelihood estimator $\hat{\theta}$ of θ.
4. Using the Delta method, show that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$, where σ^2 is to be made explicit.
5. Compute the Fisher information $I(\theta)$.
6. We observe $\hat{\theta} = 0.62$ for $n = 100$. Show that $(0, 0.67]$ is a valid confidence interval at asymptotic level 95% for θ.

Problem 11 Predicting heat waves

The Boston Health Department is monitoring the average daily temperatures of a city over the summer months to understand patterns and prepare for potential heatwaves. They've collected daily temperature data for the past two months. The department believes that the daily temperatures follow a normal distribution. As an 18.650 student, the city council decided to hire you to help them with some very important tasks.

1. Model the daily temperatures as a random variable and specify its potential distribution. That is, formalize the conjecture made by the health department. Justify why this might be a suitable distribution.
2. Given your model, describe how you would estimate the parameters of the distribution using the provided data. Discuss any assumptions you're making.
3. The department is particularly concerned about extremely high temperatures. Using your model, explain how you would compute the probability that the temperature on a given day exceeds 100F. Also, discuss the implications of your model's assumptions on this probability estimate.

Hedge fund interview question

In every PSet, we have an additional question taken from a hedge fund interview. This question is not mandatory and does not hold any point but you are welcome to give it a shot.

Problem 12 (Source: Quantlab)

Given i.i.d. random variables $X_1, \dots, X_n$, what is $\mathbb{E}[\max_i X_i]$ if:

1. $X_1, \dots, X_n \sim \mathsf{Unif}\big(0, 1\big)$?
2. $X_1, \dots, X_n \sim \mathsf{Exp}\big(1\big)$?

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

The table lists $P(Z \leq z)$ where $Z \sim N(0,1)$ for positive values of z.