Brown University
DATA 1010
Fall 2019: Practice Midterm II
Samuel S. Watson

Name:

You will have three hours to complete this exam. The exam consists of 24 written questions. No calculators or other materials are allowed, except the Julia-Python-R reference sheet and your record of medals from the first exam.

You are responsible for explaining your answer to **every** *question. Your explanations do not have to be any longer than necessary to convince the reader that your answer is correct.*

I verify that I have read the instructions and will abide by the rules of the exam: _____

Problem 1 [BAYES]

(a) Suppose that the conditional probability of an email (chosen uniformly at random from a large collection of emails) containing the phrase "additional income", given that the email is spam, is 14%. Suppose that the conditional probability of an email being spam, given that it contains the phrase "additional income", is 88%. Find the ratio of the probability that an email is spam to the probability that an email contains the phrase "additional income".

(b) We flip a weighted coin that has probability $\frac{3}{4}$ of turning up heads. If we get heads, we roll a six-sided die, and otherwise we roll an eight-sided die. Given that the die turns up 4, what is the conditional probability that the coin turned up heads?

Solution

(a) By the definition of conditional probability, we have

$$\frac{\mathbb{P}(A \mid B)}{\mathbb{P}(B \mid A)} = \frac{P(A \cap B)/\mathbb{P}(B)}{\mathbb{P}(B \cap A)/\mathbb{P}(A)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Therefore, the desired ratio is $88\%/14\% = 44/7$.

(b) We have

$$\mathbb{P}(\text{roll} = 4) = \frac{3}{4} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{1}{8} = \frac{5}{32},$$

and the proportion of that probability mass which comes from the 'heads' branch of the tree is

$$\frac{\frac{3}{4} \cdot \frac{1}{6}}{\frac{5}{32}} = \frac{4}{5}.$$

Final answer:

$\frac{44}{7}, \frac{4}{5}$

Problem 2 [IND]

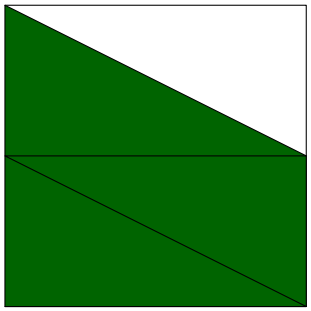
(a) Suppose that $X_1, \ldots, X_{10}$ are independent Bernoulli(p) random variables defined on a probability space Ω. What is the smallest possible cardinality of Ω?

(b) Suppose that U and V are independent random variables, each selected uniformly at random from $[0, 1]$. Find the probability of the event $\{\frac{1}{2}U + V \leq 1\}$.

Solution

(a) The range of the random vector $[X_1, \ldots, X_{10}]$ (thought of as a map from Ω to $\mathbb{R}^{10}$) has $2^{10} = 1024$ points in it. Therefore, Ω must have at least 1024 points. It can have exactly 1024 points; for example, we could take $\Omega = \{0, 1\} \times \{0, 1\} \times \cdots \times \{0, 1\}$ and $X(\omega) = \omega$.

(b) The joint distribution of (U, V) is the area measure on the unit square. The given event corresponds to the set of points shown shaded in the figure, which we can see from symmetry (using the extra lines drawn there) is $\frac{3}{4}$ of the square.



Problem 3 **[EXP]**

(a) Find the expected value of the sum of the sum and product of two independent die rolls.

(b) You roll a die, and if the result is prime you roll two more dice, and if it isn't prime you roll *three* more dice. Find the expected number of pips showing on the top faces of all of the dice rolled (so, either three dice or four dice).

Solution

(a) Let X and Y be the two die rolls. Then

$$\mathbb{E}[X + Y + XY] = \mathbb{E}[X] + \mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] = 3.5 + 3.5 + 3.5^2 = 19.25.$$

(b) Let us adjust the experiment by rolling the fourth die anyway, If the first die roll isn't prime, we won't count the last one. Then the desired sum is

$$X_1 + X_2 + X_3 + YX_4,$$

where Y is the indicator of the event that X_1 is prime. Then by linearity of expectation,

$$\mathbb{E}[X_1 + X_2 + X_3 + YX_4] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3] + \mathbb{E}[YX_4].$$

Since Y and X_4 are independent, this expression simplifies to

$$\mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3] + \mathbb{E}[Y]\mathbb{E}[X_4] = \frac{7}{2} + \frac{7}{2} + \frac{7}{2} + \left(\frac{1}{2}\right)\left(\frac{7}{2}\right) = \frac{49}{4}.$$

Final answer:

$$\frac{49}{4}$$

Problem 4 **[COV]**

Suppose that X_1 and X_2 are independent and identically distributed.

(a) Find the covariance of $X_1 + X_2$ and $X_1 - X_2$.

(b) Show that if X_1 and X_2 are normal random variables, then $X_1 + X_2$ and $X_1 - X_2$ are independent. Hint: use your knowledge of the multivariate normal distribution density.

Solution

(a) We have $\mathbb{E}[(X_1 + X_2)(X_1 - X_2)] = \mathbb{E}[X_1^2] - \mathbb{E}[X_2^2]$, and we have $\mathbb{E}[X_1 + X_2]\mathbb{E}[X_1 - X_2] = \mathbb{E}[X_1]^2 - \mathbb{E}[X_2]^2$. Subtracting, we find that the covariance of $X_1 + X_2$ and $X_1 - X_2$ is $\mathrm{Var}(X_1) - \mathrm{Var}(X_2)$, which is zero since X_1 and X_2 have the same distribution and hence also the same variance. We didn't even need the independence hypothesis!

(b) If X_1 and X_2 are independent normal random variables with the same mean $\boldsymbol{\mu}$ and variance σ^2, then the distribution of

$$\begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = A \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

is multivariate Gaussian since it's an affine transformation of a vector of independent standard normals. The covariance is $2\sigma^2 I$, and the mean is $\boldsymbol{\mu} = [2\mu, 0]$, so the density is

$$\mathbf{y} \mapsto \frac{1}{\sqrt{(2\pi)^2(2\sigma^2)^2}} e^{-\frac{1}{2(2\sigma^2)}(\mathbf{y}-\boldsymbol{\mu})'(\mathbf{y}-\boldsymbol{\mu})}.$$

Therefore, the density can be written as

$$\frac{1}{\sqrt{2\pi}(\sqrt{2\sigma^2})} e^{-(y_1-2\mu)^2/(2\cdot 2\sigma^2)} \frac{1}{\sqrt{2\pi}(\sqrt{2\sigma^2})} e^{-y_2^2/(2\cdot 2\sigma^2)},$$

which is the product of the density of Y_1 and the density of Y_2. Therefore, the two random variables are independent.

Problem 5 **[CONDEXP]**

(a) Suppose that, for all $x \in \mathbb{R}$, the conditional distribution of Y given $X = x$ is exponential with parameter $\lambda = 2|x| + 1$. Find $\mathbb{E}[Y \mid X]$.

(b) What is the strongest conclusion that can be drawn about the distribution of X, based on the information in (a)?

Solution

(a) The conditional expectation is the expectation calculated with respect to the conditional measure. Therefore, the conditional expectation given $X = x$ is the mean of the exponential distribution with parameter $2|x| + 1$, which is $\frac{1}{2|x|+1}$. Upper-casing x gives $\frac{1}{2|X|+1}$.

(b) The only conclusion that can be drawn is that X has either probability mass or probability density at every point on the number line (since otherwise we couldn't make sense of the conditional distribution of Y there). Besides that, it can have any distribution whatsoever, since we can generate X from any distribute we like and then generate Y from the exponential distribution with parameter $2|X| + 1$. The resulting pair (X, Y) will satisfy the conditions of the problem and have the chosen marginal distribution on X.

Final answer:

$$\frac{1}{2|X|+1}$$

Problem 6 **[COMDISTD]**

Suppose that $S = X_1 + \ldots + X_n$, where the X_i's are independent $\text{Ber}(p)$ random variables.

(a) The distribution of S is a named probability measure. Which one is it, and what are the parameters?

(b) Find the probability mass function for the conditional distribution of S given $\{X_1 = 1\}$.

(c) You collect some data over a few years, and you find that the number of near-doorings you experience per month on your bicycle commute is approximately Poisson distributed. Give an explanation for why the Poisson distribution might be expected to emerge in this context.

Solution

(a) The distribution of S is a Binomial distribution with parameters n and p.

(b) Given that $X_1 = 1$, the sum of the remaining random variables is a Binomial with parameters $n - 1$ and p. Therefore, the conditional distribution of S_n given X_1 is one plus a $\text{Bin}(n - 1, p)$:

$$m(k) = \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}.$$

(c) Your probability of getting doored on a particular block is low, but you traverse many blocks on your commute. Therefore, the number of doorings is a binomial random variable with large n and small p (small enough that np is still modest; otherwise you'd have stopped commuting by bike). The Poisson approximation says that such a distribution is approximately Poisson with parameter $\lambda = np$.

Problem 7 **[COMDISTC]**

(a) Find the probability density function of the distribution of $\sqrt{X}$, where X is an exponential random variable with parameter λ.

(b) Find $\mathbb{P}(Z = 0.5)$, where Z is a standard normal random variable.

Solution

(a) We calculate $\mathbb{P}(\sqrt{X} > t) = \mathbb{P}(X > t^2) = e^{-\lambda t^2}$, which implies that the density function of $\sqrt{X}$ is

$$\frac{d}{dt}\mathbb{P}(X^2 \leq t) = -\frac{d}{dt}\mathbb{P}(X^2 > t) = 2\lambda t e^{-\lambda t^2}.$$

(b) The probability that a normal random variable equals any particular value is 0.

Problem 8 **[CLT]**

The **chi-squared distribution** with parameter n is the distribution of the sum of the squares of n independent standard normal random variables.

Let S_k be the sum of k independent chi-squared random variables with parameter 8. Find the limit as $k \to \infty$ of

$$\mathbb{P}(8k \leq S_k \leq 8.01k).$$

Solution

The mean of the chi-squared distribution is

$$\mathbb{E}[Z_1^2 + \cdots + Z_8^2],$$

where Z_i's are independent standard normals. Applying linearity and using the fact that $\mathbb{E}[Z_i^2] = \operatorname{Var} Z_i = 1$, we find that the mean of the chi-squared distribution is 8. The variance of the chi-squared distribution is not as straightforward to calculate explicitly; let's call it σ^2.

The sum S_k has mean $8k$ and variance $k\sigma^2$. Therefore, its typical values are close to $8k$, with fluctuations on the order of $\sigma\sqrt{k}$. Since $0.01k$ is much larger than $\sigma\sqrt{k}$ when k is large (and since the normal distribution is symmetric), approximately $\boxed{\frac{1}{2}}$ of the mass is between $8k$ and $8k + 0.01k$.

Final answer:

$$\frac{1}{2}$$

Problem 9 **[POINTEST]**

(a) Consider the statistical functional $T(\nu)$ which returns the second moment of ν (in other words, $T(\nu) = \mathbb{E}[X^2]$ where X is ν-distributed), and let $\theta = T(\nu)$. Is the plug-in estimator of θ biased? Is it consistent?

(b) Now consider the estimator $\widehat{\theta}$ of θ which is defined to be the sum of (i) the square of the plug-in estimator of the mean of ν and (ii) the plug-in estimator of the variance of ν. Is $\widehat{\theta}$ biased? Is it consistent?

Solution

(a) The plug-in estimator of θ is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$, which is unbiased by linearity of expectation and consistent by the law of large numbers.

(b) We have

$$\begin{aligned}\widehat{\theta} &= \overline{X}^2 + \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \overline{X}^2 + \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n} X_i + \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2.\end{aligned}$$

Therefore, this estimator is actually the same as the estimator in (a), and it is therefore also unbiased and consistent.

Problem 10 **[BOOT]**

One thousand voters are polled about their position on a given ballot initiative, and 637 of them respond that they are in favor of the initiative.

(a) Find the value of the plug-in estimator $\widehat{p}$ of the proportion p of voters who are in favor of the initiative.

(b) Write an expression which approximates the standard error of $\widehat{p}$.

(c) Describe how the bootstrap methodology would be used to produce an estimate of the standard error of $\widehat{p}$. Which approach do you find preferable in this case?

Solution

(a) The value of the plug-in estimator for the observed samples is $\widehat{p} = 637/100 = 63.7\%$.

(b) We can approximate the standard error of $\widehat{p}$ in terms of the value of $\widehat{p}$ using the fact that the variance of a binomial random variable is $np(1-p)$, and therefore the variance of a binomial random variable divided by n is $p(1-p)/n$. So we estimate the standard error as

$$\sqrt{\frac{(0.637)(1-0.637)}{1000}}.$$

(c) We could use the bootstrap to accomplish the same objective by drawing from the 1000 responses 1000 times with replacement, calculating the value of the estimator $\widehat{p}$ for each of them, and computing the sample variance of the resulting list.

In this case, using the formula is preferable, since it provides the exact limiting value of the bootstrap procedure with far less computational expense. The reason that the limiting bootstrap value is equal to the value returned by the formula is that drawing with replacement from 1000 samples, 637 of which are 1's, is exactly the same as sampling independent Bernoulli random variables with $p = 63.7\%$.

Problem 11 [HYPTEST]

One Bayesian criticism of the hypothesis test framework is that it doesn't account for the *a priori* plausibility of the alternative hypothesis.

(a) You have a magician's coin, and you don't know whether it's a regular coin or a two-headed or two-tailed coin. Consider the null hypothesis that the coin is fair, with the alternative hypothesis that the coin favors one of the two sides. You flip the coin 10 times, and it comes up heads all 10 times. The null hypothesis is rejected with what p-value? What do you actually believe about the coin?

(b) Now suppose you have a coin that you just got from the cashier at Trader Joe's. You carefully inspect it and determine that it appears to be an entirely ordinary U.S. quarter. Once again, consider the null hypothesis that the coin is fair, with the alternative hypothesis that the coin favors one of the two sides. Once again, you flip the coin 10 times, and it comes up heads all 10 times. What do you actually believe about this coin?

Solution

(a) Under the null hypothesis, the probability of getting all 10 heads or all 10 tails is $2/1024 = 1/512$. Therefore, the p-value is $1/512$. I would believe that this coin is likely two-headed.

(b) The p-value is the same as in (a), but I would believe that the coin is fair and it just happened to come up heads 10 times in a row.

Problem 12 [MLE]

(a) Find the maximum likelihood estimator for the family of geometric distributions with parameter $0 < p < 1$. (You don't need to prove that the value you find is actually a maximum; just differentiate the log-likelihood and solve for the zero).

(b) I simulated 10 independent samples from a geometric distribution with parameter p and got

$$0, 4, 1, 3, 4, 3, 1, 14, 0, 13$$

Use the maximum likelihood estimator to estimate the value of p that I used.

Solution

(a) Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with common distribution Geometric(p). The likelihood function is

$$\mathcal{L}(p) = p(1-p)^{X_1-1} p(1-p)^{X_2-1} \cdots p(1-p)^{X_n-1},$$

so the log likelihood is

$$n \log p + \left(\sum_{i=1}^{n} X_i - n \right) \log(1-p).$$

Setting the derivative of the log likelihood equal to zero yields

$$\frac{n}{p} - \frac{\sum_{i=1}^{n} X_i - n}{1-p} = 0,$$

and solving for p gives $p = n / \sum_{i=1}^{n} X_i = 1/\overline{X}_n$.

(b) The average of the 10 provided values is 4.3, so the maximum likelihood estimator of p is $1/4.3$.