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

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

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

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

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

Problem 6 **[COMDISTD]**

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

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

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

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

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

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

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