DATA 1010 PRACTICE PROBLEMS
MIDTERM I
SAMUEL S. WATSON

Note: I have no inside knowledge of what's going to be on the exam. Furthermore, these problems do not adhere to the multiple choice format, and their level of difficulty is not representative. The intention is to provide some basic practice with the topics covered in the lecture notes.

Problem 1

Suppose that X is a random variable which is generated as follows: a fair coin is flipped, and if the coin comes up heads then we return $X = \frac{1}{2}$. If the coin comes up tails, then we generate a uniform random number U between 0 and 1 (which is independent of the coin flip) and return $X = 2U$.

(a) Sketch the cdf of X.

(b) Does X have a pdf? Does X have a pmf?

(c) Find $\mathbb{E}[X]$

Problem 2

Suppose that 1% of the population has a particular medical condition. A diagnostic test for the condition correctly detects the condition 95% of the time (in the cases where the person does indeed have it). The test for the condition also correctly identifies a person as not having it 90% of the time (again, in the cases where the person indeed does not have the condition). Suppose person X, randomly selected from the population, is identified by the test as having the condition. What is the probability that person X actually has the condition?

Problem 3

Suppose that the random vector (X, Y, Z) is chosen uniformly from the set

$$\{(0,0,0),(1,1,0),(1,0,1),(0,1,1)\}.$$

Show that X and Y are independent, and that Y and Z are independent.

Problem 4

Suppose that the probabilities of rain for the next three days are 20%, 30%, and 25%. Let D be the number of days of rain in the next three days. (So D is a random variable which takes values in $\{0,1,2,3\}$.)

(a) Does it make sense (from a meteorology point of view) that the event that it rains tomorrow and the event that it rains the next day would be independent?

(b) Can you determine $\mathbb{P}(D = 2)$ on the basis of the given information?

(c) Can you determine $\mathbb{E}[D]$ on the basis of the given information?

Problem 5

Suppose $F(x)$ is a right-continuous, increasing function which goes from 0 to 1 as x goes from $-\infty$ to ∞. Let $X_1, \ldots, X_{10}$ be independent random variables each of which has cdf F. Suppose that the variance of X_1 is 3.

(a) Find the variance of the random variable

$$\tfrac{1}{10}\left(X_1 + \cdots + X_{10}\right).$$

(b) Suppose that you sample $X_1, \ldots, X_{10}$ and work out that $\frac{1}{10}\left(X_1 + \cdots + X_{10}\right)$ is equal to 3.6 for that run. Give an interval around 3.6 which you are 95% confident contains the mean of the distribution F.

Problem 6

Shuffle a standard 52-card deck (which contains half red cards and half black cards). Let X be the number of consecutive pairs of cards in the shuffled deck which are both red. Find $\mathbb{E}[X]$.

Problem 7

A box contains four balls with distinct colors. We sample four balls from this box independently and with replacement. What is the expected number of distinct colors in our sample?

Hint: consider using the four random variables which indicate whether each color is absent from our sample.

Problem 8

Which of the following could possibly be the covariance matrix of a vector of independent random variables?

$$\begin{bmatrix} 0.3024 & 0.8967 & -0.0802 & -1.3614 \\ -0.0364 & -0.5135 & -1.0912 & -0.1145 \\ 0.142 & -0.7648 & -0.5805 & 0.1658 \\ 0.5213 & -1.5414 & -0.3154 & -0.4084 \end{bmatrix} \quad \begin{bmatrix} 0.405 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.4995 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.6588 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.5156 \end{bmatrix}$$

$$\begin{bmatrix} 0.0 & 0.8644 & 0.5023 & 0.1281 \\ -0.9017 & 0.0 & -0.517 & 1.8528 \\ -0.4945 & 0.5328 & 0.0 & -0.8278 \\ -0.9029 & -0.2717 & -0.0193 & 0.0 \end{bmatrix} \quad \begin{bmatrix} -0.2512 & 1.5642 & -3.2114 & 0.0 \\ 0.3697 & -1.3967 & 0.0 & -0.6027 \\ 0.0721 & 0.0 & 0.151 & -1.2797 \\ 0.0 & -1.1067 & 0.7693 & 0.9973 \end{bmatrix}$$

Problem 9

The kh moment of a random variable X is defined to be $\mathbb{E}[X^k]$. Find the second moment of a unit mean exponential random variable. Note: a unit mean exponential random variable has density e^{-x} on $[0, \infty)$.

Problem 10

Let M be a large positive integer. Suppose we propose a pseudo-random number generator which begins with a *seed* s (which is any integer between 0 and M) and returns the sequence of numbers:

$$\tfrac{s}{M}, \tfrac{s+1}{M}, \cdots, \tfrac{s+2}{M}, \ldots, \tfrac{M}{M}, \tfrac{0}{M}, \tfrac{1}{M}, \tfrac{2}{M}, \ldots \tfrac{s-2}{M}, \tfrac{s-1}{M},$$

repeating from there.

(a) Is this a sensible PRNG?

(b) Give an example of a PRNG test that this PRNG fails.

(c) Give an example of a PRNG test that this one *passes*.

Problem 11

Suppose some genes are studied for impact on a particular trait, with the null hypothesis for each gene being that it has no influence on the trait. The alternative hypothesis for each gene is that it does have an effect on the trait. We run 7 hypothesis tests, one for each gene, and get p-values of

$$\begin{bmatrix} 0.0401 & 0.2527 & 0.0734 & 0.0647 & 0.1482 & 0.5734 & 0.3619 \end{bmatrix}$$

(a) Is there enough evidence to reject the null hypothesis at the 95% confidence level for any of the genes?

(b) Is there evidence to conclude that the alternative hypothesis is incorrect for all the genes?

Problem 12

Suppose $h : [0,1]^{40} \to [0,1]$ is a continuous function. How many Monte Carlo samples do we need if we want to identify an interval of width 0.02 which contains $\int_0^1 \cdots \int_0^1 h(x_1, \cdots, x_{40}) \, \mathrm{d}x_1 \cdots \mathrm{d}x_{40}$ with a confidence of at least 95%?

Problem 13

The code

```
n = 1000
sum(norm(rand(5) - 1/2) < 0.5 for trial=1:n)
```

generates a point selected uniformly at random from the cube $[0,1]^5$, computes its distance to the center of the cube, and determines whether the randomly selected point lies in the 5-dimensional sphere inscribed in the cube. It then repeats this $n-1$ more times and returns the number of times that the point did lie in the sphere.

(a) Suppose that the code is run once and returns 167. Give an estimate of the volume of the sphere, together with a 95% confidence interval around this value.

Note: The code `[pi^(d/2)/(2^d*gamma(d/2+1)) for d=2:10]` gives the percentage of the volume of a d-dimensional cube which is occupied by the inscribed d-dimensional sphere, as d ranges from 2 to 10. It returns

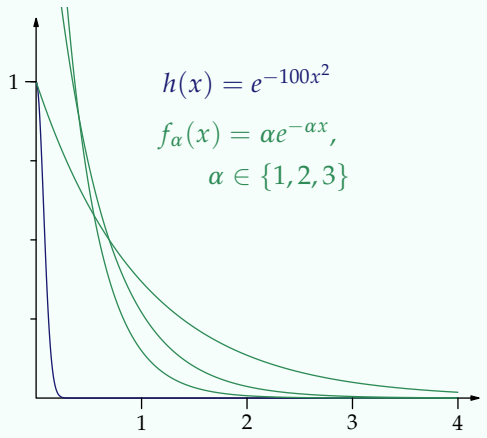
$$\begin{bmatrix} 0.7854 & 0.5236 & 0.3084 & 0.1645 & 0.0807 & 0.0369 & 0.0159 & 0.0064 & 0.0025 \end{bmatrix}$$

You can use this information to help check your answer.

(b) One way to measure the "size" of a bounded region D in $\mathbb{R}^n$ is to calculate the average distance between two points selected uniformly from D. Explain how to set up a Monte Carlo experiment that would estimate this value for the 100-dimensional unit cube $[0,1]^{100}$.

Problem 14

Suppose we want to estimate the integral of the function $h(x) = e^{-100x^2}$ on $[0,4]$. Explain how to do this using importance sampling with an i.i.d. sequence of exponential random variables $X_1, X_2, \ldots$, and explain how you might choose the parameter α for the density $\alpha e^{-\alpha x}$ of the exponential random variables.



Problem 15

The *Cauchy distribution* has the property that the average of any finite number of independent Cauchy-distributed random variables is also Cauchy distributed. In other words, if $X_1, X_2, \ldots$ is an i.i.d. sequence of Cauchy-distributed random variables, then

$$X_1 \sim \tfrac{1}{2}(X_1 + X_2) \sim \tfrac{1}{3}(X_1 + X_2 + X_3) \sim \cdots$$

Without having any more information about the Cauchy distribution, we can conclude from the law of large numbers and/or the central limit theorem that (select all that apply):

(a) $\frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$ converges to the mean of the Cauchy distribution

(b) $\frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$ has approximately the same distribution as a standard normal random variable

(c) $\frac{1}{\sqrt{n}}(X_1 + X_2 + X_3 + \cdots + X_n)$ has approximately the same distribution as a standard normal random variable

(d) The Cauchy distribution does not have finite variance.

Problem 16

Is the sum of two independent exponential random variables again an exponential random variable? Convolve the density $e^{-x}\mathbf{1}_{[0,\infty)}(x)$ with itself to determine the answer to this question.

Problem 17

Consider the density function $f(x) = \frac{1}{2\sqrt{2\pi}} e^{-(x-2)^2/8}$. Find $\overbrace{f * f * \cdots * f}^{\text{100 total}}$.

Problem 18

Consider the sum $X_1 + \cdots + X_{10}$, where X_k has mean 0 and variance 5 for all $1 \leq k \leq 10$.

(a) Find $\mathbb{E}[(X_1 + \cdots + X_{10})^2]$ under the assumption that the X_k's are independent

(b) Find $\mathbb{E}[(X_1 + \cdots + X_{10})^2]$ under the assumption that the X_k's are perfectly correlated (that is, $X_1(\omega) = X_2(\omega) = \cdots = X_{10}(\omega)$ for all ω in the underlying sample space Ω).

(c) Find $\mathbb{E}[(X_1 + \cdots + X_{10})^2]$ under the assumption that the X_k's have pairwise correlation $0 < \alpha < 1$. This means that $\alpha = \frac{\mathbb{E}[X_i X_j]}{\mathbb{E}[X_i^2]}$ for all $i \neq j$.

Problem 19

Your friend is thinking of a distribution, and it happens to be a *Poisson distribution*, which has pmf

$$f(k) = \frac{2^k e^{-2}}{k!}, \quad \text{for all } k = 0, 1, 2, \ldots.$$

Let X_1 and X_2 be independent random variables drawn from this distribution, and define the random variable

$$V = \frac{1}{2}\left(\left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2\right).$$

(a) You are interested in knowing the variance of the random variable V, but your friend doesn't reveal the distribution they're thinking of. They only reveal the result of two independent draws from the distribution, which happen to be 2 and 1. What is the *exact limiting value* of the bootstrap estimator of Var V?

(b) The actual variance of V can be shown to be exactly $\frac{9}{4}$. You should find that the answer to (a) is a really terrible estimate of this variance, even though you were able to perform an exact calculation in (a). Why?