ORIE 4580/5580 F25 Midterm Exam
ORIE 5581 F25 Final Exam

Honor Code: I have neither given nor received unauthorized aid on this exam.

Printed Name: _____

Signature: _____

NetID (e.g. sb2373): _____

.

Instructions:

The booklet consists of this cover page and 11 pages of questions with space for answers. We will take extra paper if needed. If you take extra pages to write an answer, then indicate clearly which question by writing "**Extra work for Question** _____" on top.

- Do not communicate with anyone or access the internet via any electronic devices during the test.
- Time allowed: 90 minutes
- For numeric answers, you can leave answers in closed form (without calculating).
- For 95% confidence interval calculations, **you should use** $z_{\alpha/2} = 2$ (instead of 1.96).

Question	Points	Out of
1		36
2		20
3		15
4		15
5		14
Total		100

List of useful formulae

- Given n samples $X_1, X_2, \ldots, X_n$, their empirical mean is $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and their empirical variance is given by $S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)$
- For computing 95% confidence intervals (i.e., $\alpha = 0.05$), **you should use** $z_{\alpha/2} = 2$ (instead of 1.96).
- Random variable $U \sim$Uniform$[0,1]$ has mean $\mathbb{E}[U] = 1/2$ and variance $Var(U) = 1/12$
- Random variable $X \sim$ Bernoulli(p) has mean $\mathbb{E}[X] = p$ and variance $Var(X) = p(1-p)$
- Random variable $Y \sim$Exponential(λ) has PDF $f(x) = \lambda e^{-\lambda x}$, and also CDF $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$; moreover, its mean is $1/\lambda$.
- For standard Normal rv $Z \sim \mathcal{N}(0,1)$, its PDF is denoted as $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and its CDF is denoted as $\Phi(x) = \mathbb{P}[Z \leq x] = \int_{-\infty}^{x} \phi(x)$

Space for scratch work

1. (36 pts) Multiple choice questions (3pts each). **Circle** the best answer; no need to show work.

(a) You are testing an LLM to see how it performs on your Simulation test. The test has 20 multiple-choice questions, each with 5 options.

For a 'control' benchmark, you first simulate uniform random guessing: what distribution models the number of correct responses N for this?

A Sum of 4 Binomial$(5, 0.75)$ rvs **B** Binomial$(20, 0.2)$ rv
C Poisson(4) rv **D** Both A and B are fine **E** A, B and C are all fine

(b) The test covers 4 topics, with 5 questions on each topic. To simulate a 'random student', you assume the student either studies a topic independently with probability 0.5 (in which case she correctly answers all related questions), or else she randomly guesses. What distribution models the number of correct responses?

A Sum of 4 iid Binomial$(5, 0.6)$ rvs **B** Sum of 4 iid Binomial$(5, 0.55)$ rv
C Product of an independent Binomial$(4, 0.5)$ and a $(5 +$ Binomial$(5, 0.5))$ rv
D Binomial$(20, 0.6)$ **E** None of the above

(c) Suppose the LLM's has seen each question in its training data independently with probability 0.5 (in which case it answers correctly); otherwise it makes a uniform random guess. What distribution models the number of correct responses?

A Sum of 10 Bernoulli(0.5) rvs and 10 Bernoulli(0.2) rvs
B Sum of 20 Bernoulli(0.55) rvs **C** Binomial$(20, 0.6)$ rv
D Poisson(0.7) rv **E** None of the above

(d) If we simulate the random benchmark and student, and run the LLM, which will do the best on average? (*Hint: we only want to compare expectations*)

A The LLM **B** The random student **C** The benchmark
D The LLM and student **E** All are the same

(e) Which of the following properties are *necessary* for a function $g(x), x \in [0,1]$ to be the CDF of a distribution?

A *strictly* increasing (i.e., $g(x) > g(y)$ for any $x > y$) **B** continuous
C $g(0) = 0, g(1) = 1$ **D** convex **E** A, B and C

(f) In any simulation, which of the following require you to run more replications?

A reduce halfwidth **B** increase confidence **C** test multiple hypotheses
D unexpectedly high variance of pilot samples **E** all of these

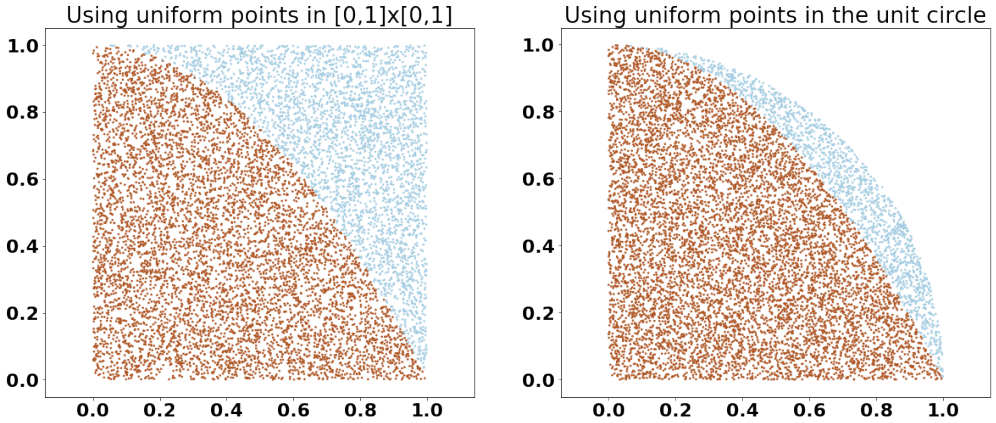
(g) We are given 12 Uniform$[0,1]$ samples $U_1, U_2, \ldots, U_{12}$. Which of the following has a distribution closest to a $\mathcal{N}(0,1)$ rv? (Note: $\mathbb{E}[U_1] = 0.5, Var(U_1) = 1/12$)

A $\frac{1}{\sqrt{12}}\left(\sum_{i=1}^{12} U_i - 6\right)$ **B** $\sum_{i=1}^{12} U_i - 6$ **C** $\sum_{i=1}^{12} U_i - 12$
D $\frac{1}{12}\left(\sum_{i=1}^{12} U_i - 6\right)$ **E** All are as good

(h) We want $\mathbb{E}[Y]$ for the rv $Y = f(Z)$, where $Z \sim \mathcal{N}(\mu, 1)$ is a Normal rv with mean μ and variance 1. Which of the following is an antithetic estimator for Y?
Hint: Note that for any function f, your estimator should have the correct expected value. Now consider some simple f...

A $0.5\big(f(Z) + f(1-Z)\big)$ **B** $0.5\big(f(Z) + f(-Z)\big)$ **C** $0.5\big(f(Z) - f(-Z)\big)$
D $0.5\big(f(Z) + f(\mu - Z)\big)$ **E** $0.5\big(f(Z) + f(2\mu - Z)\big)$

(Parts – (i) to (l)) are based on the following figure) We want to estimate the **integral** $q = \int_0^1 (1-x^2)dx$; below we show two ways using acceptance-rejection.



(i) In the figure on the left, we generate $10,000$ uniform random points in $[0,1]^2$, count the number of points Q that lie under the curve $y = 1-x^2$, and compute estimate $\widehat{M} = Q/10000$. What is the **standard deviation** of $\widehat{M}$?
A $\sqrt{q(1-q)}$ **B** $0.01q(1-q)$ **C** $0.01\sqrt{q(1-q)}$ **D** $0.0001q(1-q)$
E None of the above

(j) Suppose we use $\widetilde{M} = 1 - \widehat{R}$, where $\widehat{R}$ is the fraction of points in $[0,1]^2$ that lie **above** $y = 1-x^2$. How does the variance of $\widetilde{M}$ compare to $\widehat{M}$ in part (j)?
A it is lower **B** it is higher **C** remains the same
D depends on the random samples **E** can not be determined

(k) In the figure on the right, we sample $10,000$ points u.a.r. in the positive quarter-circle of radius 1 centered at the origin (i.e., in $\{(x,y)|x \geq 0, y \geq 0, x^2+y^2 \leq 1\}$). Suppose we find 8400 points lie under the curve $y = 1-x^2$. Using this data, what estimate can you give for q?

A 0.84 **B** $0.84/\pi$ **C** 0.42π **D** $0.21/\pi$ **E** 0.21π

(l) For the estimator for q in the previous part (using u.a.r. points in the quarter circle), what is the variance of the estimate you get from a **single** sample point?

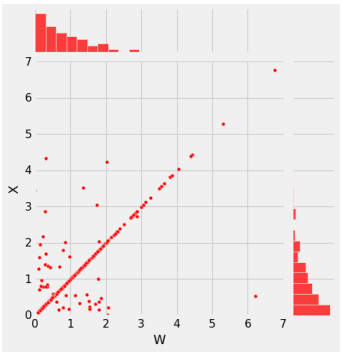
A $\frac{\pi q}{2}\left(1-\frac{2q}{\pi}\right)$ **B** $q\left(1-\frac{q}{\pi}\right)$ **C** $\frac{q}{\pi}\left(1-\frac{q}{\pi}\right)$ **D** $\frac{\pi q}{4}\left(1-\frac{4q}{\pi}\right)$ **E** $\frac{4q}{\pi}\left(1-\frac{4q}{\pi}\right)$

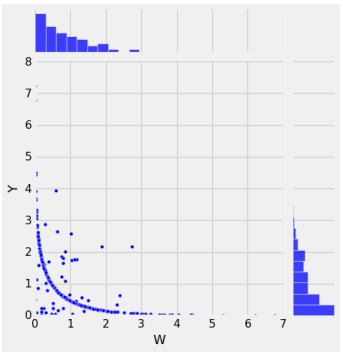
2. (20 pts) Short-answer questions. **Please provide justifications for your answers**.

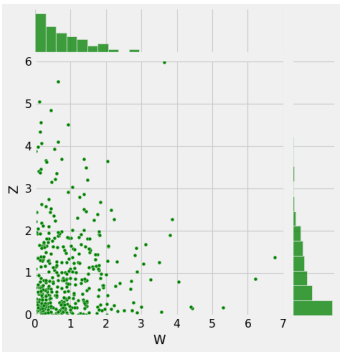
(a) (5 pts) A data scientist builds a simulation model of air-traffic control at Newark airport, and uses it to report 90% **confidence intervals** for average delays in departures over *each* of the 24 hours in a typical day. If departures in any two different hours are modeled as being **independent** of each other, what is the probability that expected delays (under the model) for *all* 24 *hours* are simultaneously captured in their CIs? (Answer as a closed-form expression without simplifying.)

(b) (5 pts) In reality, different hours are **correlated** in complicated ways due to spillover effects. To account for this and yet ensure that expected delays (under the model) in *all 24 hours to simultaneously lie in their CI* with 90% confidence, what confidence level α should we use for reporting each individual delay?

(c) (5 pts) A financial analyst collects datasets $W = (W_1, W_2, \ldots), X = (X_1, X_2, \ldots), Y = (Y_1, Y_2, \ldots), Z = (Z_1, Z_2, \ldots)$ of closing prices of 4 different stocks over a year, and then plots them via the following joint scatter plots (with histograms of the corresponding marginals on the x and y axis).







Based on these plots, how do you think the vectors W, X, Y and Z are correlated (i.e., what are the signs of $Cov(W, X), Cov(W, Y)$ and $Cov(W, Z)$, the pairs of vectors whose scatter plots are provided). Briefly explain your reasoning.

(d) (5 pts) What can you infer about $Cov(X, Y), Cov(X, Z), Cov(Y, Z)$ (i.e., the correlation between the pairs of vectors whose scatter plots are not provided)?

3. (15 pts) A powerful feature of pseudorandom numbers is the ability to 'replay' a simulation by controlling the PRNG seed. This however needs to be done carefully...

(a) (5 pts) Consider the following function (recall Numpy's `rand()` generates a $U[0,1]$ rv)

```
from numpy.random import rand
def mystery_rng(p):
  N = 1
  while 1:
    U = rand()
    if U < p: return N
    else: N = N+1
```

What is the probability mass function of the output of the algorithm?

(b) (5 pts) You modify the above to fix PRNG seed to 0 (using `np.random.seed()`) as follows.

```
from numpy.random import rand,seed
def mystery_rng(p):
  seed(1)
  N = 1
  while 1:
    U = rand()
    if U < p: return N
    else: N = N+1
```

You call this function once with $p = 0.7$, and get an output 5. Next you call the function 5 more times with $p = 0.7$; what output will you get each time?

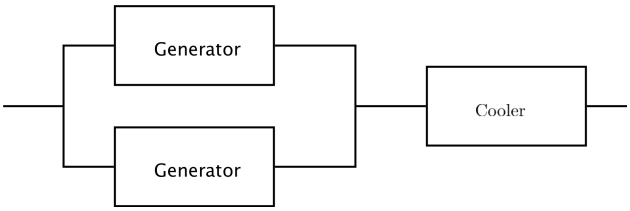
.

(c) (5 pts) ChatGPT suggests you should remove `np.random.seed(0)` from inside the function, and instead, use a loop to call the function with seeds $1, 2, \ldots$, as follows:

```
for i in range(5):
    seed(i+1)
    print(mystery_rng(0.7))
```

Suppose `np.random.rand()` is using the LCG $x_{n+1} = (ax_n + c) \mod m$ with parameters $a = 5$, $c = 44$ and $m = 49$, and outputting $u_n = \frac{x_n+1}{m+1}$. What is the output of the above code?

4. (15 pts) A quantum computer needs two parallel generators and one freezer for its operations; it must be run continuously, but halts if either both generators fail OR the freezer fails.



The lifetime of each generator is exponentially distributed with mean 2 years. The cdf for the lifetime of the freezer is given by (where x is in units of years)

$$F(x) = 1 - \frac{1}{(1+x)^2} \qquad , \qquad \text{for } x \geq 0$$

Note: Recall an Exponential(λ) rv has mean $1/\lambda$ and pdf $f(x) = \lambda e^{-\lambda x}$

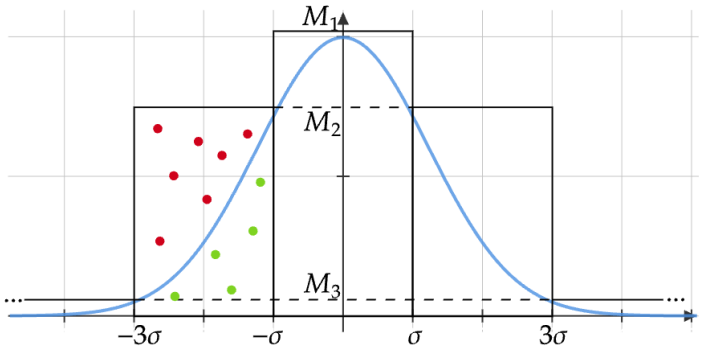
(a) (9 pts) Let T be the first time the computer halts. Given i.i.d. samples U_1, U_2, U_3 from Uniform$[0, 1]$ distribution, how can you use these to generate a sample of T?

(b) (6 pts) Using your Monte Carlo simulation model, you perform a pilot run of 11 trials to obtain the following 11 samples of the lifetime of the system in years

$$4, 3, 5, 4, 3, 4, 5, 3, 6, 3, 4.$$

(Note: these 11 numbers sum to 44) Using these samples, compute the number of replications needed to estimate the expected lifetime with ± 0.2 accuracy and 95% confidence. (*Remember to use* $z_{\alpha/2} = 2$ *for* $\alpha = 0.05$)

5. (14 pts) We will next look at another variance reduction technique called *stratified sampling*. As an example, in class we said that basic AR (i.e., using a rectangle) does not work for generating Gaussian random variables as the range is $[-\infty, \infty]$; however, we *can* use basic AR to sample a *truncated Gaussian* (i.e., restricted to a finite range). In this question, we discuss how we can use this to generate a Gaussian rv. $X \sim \mathcal{N}(0, \sigma^2)$, as suggested in the figure below. Recall $\Phi(x) = \int_{-\infty}^{x} \phi(x)dx$



(a) (3 pts) Suppose we break up the interval $[-\infty, \infty]$ into five finite intervals $I_0 = [-\sigma, \sigma]$, $I_1 = [\sigma, 3\sigma]$, $I_{-1} = [-3\sigma, -\sigma]$, $I_2 = [3\sigma, 5\sigma]$, $I_{-2} = [-5\sigma, -3\sigma]$, and two infinite intervals $I_3 = [5\sigma, \infty]$ and $I_{-3} = [-\infty, -5\sigma]$. Define a discrete random variable $B = k$ if $X \in I_k$ for $k \in \{-3, -2, -1, 0, 1, 2, 3\}$. What is the PMF of B (i.e., what is $p(k) = \mathbb{P}[B = k]$)? Express your answers in terms of the $\Phi(\cdot)$ function.

(b) (6 pts) Suppose we are given a function `sample_B()` that returns the random number in $\{-3, -2, -1, 0, 1, 2, 3\}$ according to the pmf in $p(i)$ part (a). Provide pseudocode showing how to use this to return a sample of $X \sim \mathcal{N}(0, \sigma^2)$ using basic AR (i.e., by sampling from an appropriate rectangle for each interval). If your interval is I_{-3} or I_3 (i.e., $[-\infty, -5\sigma]$ or $[5\sigma, \infty]$), you can return "Out of Bounds".
(Truncating like this adds bias, but you can remove it by using generalized AR in sets I_3, I_{-3}, or reduce it by using more intervals...)

(c) (5 pts) The advantage of this technique is that you can now choose any number of points in each set! Given $n(k)$ i.i.d. samples $(X_1^k, X_2^k, \ldots, X_{n(k)}^k)$ in each interval $I_k, k \in \{-3, -2, -1, 0, 1, 2, 3\}$, and some given function $G(\cdot)$, provide an unbiased estimator for $\mathbb{E}[G(X)]$.

(d) (OPTIONAL) Finally, you can choose n_k to reduce variance! Suppose you do a pilot run to get variance estimates $S(k)^2$ for the estimates in each interval I_k. Argue that the overall variance of your estimate in part (c) is (approximately) $\sum_{k=-3}^{3} S(k)^2/n(k)$. Now given a budget of M samples, how can you divide them between the intervals to minimize variance?

Note: Try to make this an unconstrained optimization via Lagrange multipliers – now what do you need for the solution to be optimal. . . ?

Extra work for Question _____ only: