

**V1**

Name: _____

PUID _____

Instructor (circle one): Heekyung Ahn Evidence Matangi Timothy Reese Halin Shin

Class Start Time: ☐ 11:30 AM ☐ 12:30 PM ☐ 1:30 PM ☐ 2:30 PM ☐ 3:30 PM ☐ 4:30 PM ☐ Online

As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do.

Accountable together - we are Purdue.

Instructions:

1. **IMPORTANT** Please write your **name** and **PUID** clearly on every **odd page**.
2. **Write your work in the box. Do not run over into the next question space.**
3. The only materials that you are allowed during the exam are your **scientific calculator, writing utensils, erasers, your crib sheet, and your picture ID**. If you bring any other papers into the exam, you will get a **zero** on the exam. Colored scratch paper will be provided if you need more room for your answers. Please write your name at the top of that paper also.
4. The crib sheet can be a handwritten or type double-sided 8.5in x 11in sheet.
5. Keep your bag closed and cellphone stored away securely at all times during the exam.
6. If you share your calculator or have a cell phone at your desk, you will get a **zero** on the exam.
7. The exam is only 60 minutes long so there will be no breaks (including bathroom breaks) during the exam. If you leave the exam room, you must turn in your exam, and you will not be allowed to come back.
8. **For free response questions you must show ALL your work to obtain full credit.** An answer without showing any work may result in **zero** credit. If your work is not readable, it will be marked wrong. Remember that work has to be shown for all numbers that are not provided in the problem or no credit will be given for them. All explanations must be in complete English sentences to receive full credit.
9. All numeric answers should have **four decimal places** unless stated otherwise.
10. After you complete the exam, please turn in your exam as well as your table and any scrap paper that you used. Please be prepared to **show your Purdue picture ID**. You will need to **sign a sheet** indicating that you have turned in your exam.
11. You are expected to uphold the honor code of Purdue University. It is your responsibility to keep your work covered at all times. Anyone caught cheating on the exam will automatically fail the course and will be reported to the Office of the Dean of Students.
12. It is strictly prohibited to smuggle this exam outside. Your exam will be returned to you on Gradescope after it is graded.

Your exam is not valid without your signature below. This means that it won't be graded.

I attest here that I have read and followed the instructions above honestly while taking this exam and that the work submitted is my own, produced without assistance from books, other people (including other students in this class), notes other than my own crib sheet(s), or other aids. In addition, I agree that if I tell any other student in this class anything about the exam BEFORE they take it, I (and the student that I communicate the information to) will fail the course and be reported to the Office of the Dean of Students for Academic Dishonesty.

Signature of Student: _____

**You may use this page as scratch paper.
The following is for your benefit only.**

Question Number	Total Possible	Your points
Problem 1 (True/False) (2 points each)	12	
Problem 2 (Multiple Choice) (3 points each)	15	
Problem 3	20	
Problem 4	26	
Problem 5	32	
Total	105	

The rest of this page can be used for scratch work

1. (12 points, 2 points each) **True/False Questions.** Indicate the correct answer by completely filling in the appropriate circle. If you indicate your answer by any other way, you may be marked incorrect.

1.1. Employees in a certain UPS branch collected the types of mail customers brought for a month. They plan to present the data appropriately to the manager and discuss how to utilize empty space efficiently.

☐ T or ☒ F A histogram is appropriate to use because the variable is categorical.

1.2. A hardware manufacturer is about to ship 20,000 of its products to a client. To estimate the defect rate of this shipment, they randomly selected 100 products for a last-minute inspection. For each product, they assign a value of 0 if the product is good and 1 if it is defective. The defect rate is then calculated as the average of these 0's and 1's.

☐ T or ☒ F If the company had the resources to inspect all 20,000 products, the defect rate calculated using all 20,000 products would represent a sample statistic.

1.3. Suppose the number of visitors to a mall follows a **Poisson distribution** with an **average rate of 45 visitors per 30 minutes**.

☐ T or ☒ F In this mall, the **variance** in the number of visitors arriving between **2:00 PM and 3:00 PM** is **equal** to the **variance** in the number of visitors arriving between **3:00 PM and 5:00 PM**.

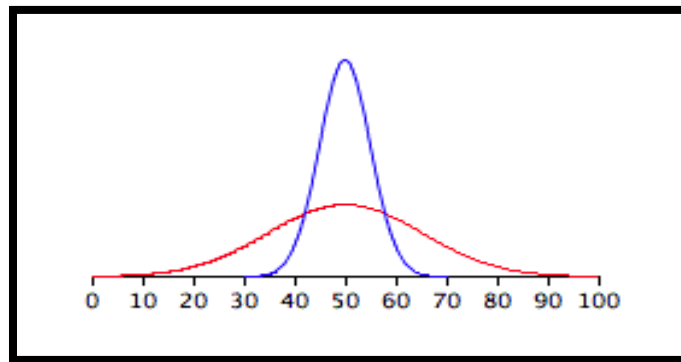
1.4. Let V be a random variable with a **probability density function** $f_V(v)$ that is **non-zero only** on the **interval** $[-5, -2)$. Let $F_V(\cdot)$ denote the **cumulative distribution function (CDF)** of V .

☒ T or ☐ F Then, $F_V(c) = 1$ holds for any $c > 0$.

1.5. A student scored 85 on two different math exams. For Exam 1, the mean score is 75 with a standard deviation of 5, and for Exam 2, the mean score is 70 with a standard deviation of 10.

☒ T or ☐ F The student performed better on Exam 1 compared to Exam 2.

1.6. For the figure below,



☐ or ☒ the blue normal distribution has more area underneath its curve than the red normal distribution does.

2. (15 points, 3 points each) **Multiple Choice Questions.** Indicate the correct answer by completely filling in the appropriate circle. If you indicate your answer by any other way, you may be marked incorrect. **For each question, there is only one correct option letter choice.**

2.1. The number of customers arriving at a UPS branch during working hours follows a **Poisson distribution** with an **average rate of 4 customers per hour**. Let X denote the number of customers arriving between **9:00 AM** and **10:00 AM** and let Y denote the number of customers arriving between **10:30 AM** and **12:00 PM**.

What is the **conditional probability** that exactly **3 customers** arrive between **10:30 AM** and **12:00 PM**, given that **6 customers** arrived between **9:00 AM** and **10:00 AM**?

- ☐ $P(Y = 3|X = 6) = 0$
- ☐ $P(Y = 3|X = 6) = 0.0093$
- ☒ $P(Y = 3|X = 6) = 0.0892$
- ☐ $P(Y = 3|X = 6) = 0.1954$
- ☐ $P(Y = 3|X = 6) = 0.8564$

2.2. The time between customer arrivals at the same UPS facility follows an exponential distribution with an **average of 15 minutes** between **customer arrivals**. Let T denote the time between customer arrivals. If no customer has arrived in the **last 20 minutes**, what is the probability that the next customer arrives after waiting more than **15 additional minutes**.

- Ⓐ $P(T > 35 | T > 20) = 0$
- Ⓑ $P(T > 35 | T > 20) = 0.097$
- Ⓒ $P(T > 35 | T > 20) = 0.2636$
- Ⓓ $P(T > 35 | T > 20) = 0.3679$
- Ⓔ $P(T > 35 | T > 20) = 0.6321$

2.3. Suppose $X \sim \text{Binomial}(n = 10, p = 0.1)$ and $Y \sim \text{Binomial}(n = 10, p = 0.9)$.

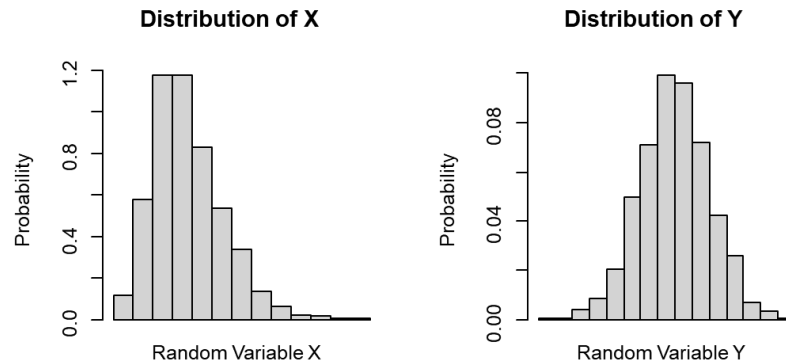
Which statement is **not always true** about X and Y ?

- Ⓐ The **mode** of X is less than the **mode** of Y .
- Ⓑ $SD(X) - \left| \sqrt{Var(Y)} \right| = 0$
- Ⓒ $P(X = 1 \cap Y = 8) = 0.1943$
- Ⓓ $E[X^2] = (10)(0.1)(0.9) + [(10)(0.1)]^2$
- Ⓔ $P(X = 1) = P(Y = 9)$

2.4. Suppose X is a random variable with $E[e^X] = 2$ and $Var(e^X) = 5$, and Y is a random variable **independent** of X , satisfying $E(Y) = -10$, $Var(Y) = 3$. What is $E[(e^X - 3Y)^2]$?

- Ⓐ 1056
- Ⓑ 240
- Ⓒ 1024
- Ⓓ -752
- Ⓔ None of the above

2.5. The figure below shows the shape of the distribution for two continuous random variables X and Y .



Which of the following statements is TRUE about the random variable X ?

- Ⓐ The **mean** is a better measure of central tendency than median.
- Ⓑ The distance between Q_3 and the **median** is narrower than the distance between Q_1 and the **median**.
- Ⓒ **IQR** is a robust (resistant) measure of the **spread**.
- Ⓓ The distribution is negatively skewed with one peak.
- Ⓔ The **mode** will have the largest value among all the measures of central tendency.

Free Response Questions 3-5. Show all work, clearly label your answers, and use **four decimal places**.

3. (20 points) The stated speed limit on I-65 is 65 mph. The speeds of vehicles along a certain stretch of I-65 follow an approximately **normal distribution** with a **mean** of **71 mph** and a **standard deviation** of **8 mph**.

- a) (2 points) What is the probability that the speed of a vehicle on this stretch of I-65 is below $\mu + 3\sigma$?

Let V denote the speed of a random Vehicle on I-65.

Simply using the **Empirical Rule**:

$$P(V < \mu + 3\sigma) \approx 0.9985$$

- b) (2 points) Calculate the **z-score** for the stated speed limit of 65 mph.

$$z = \frac{x - \mu}{\sigma} = \frac{65 - 71}{8} = -0.75$$

- c) (8 points) What is the probability that a vehicle's speed is between 61 mph and 71 mph on this stretch of I-65?

$$\begin{aligned} P(61 < V < 71) &= P\left(\frac{61 - 71}{8} < \frac{X - \mu}{\sigma} < \frac{71 - 71}{8}\right) \\ &= P(-1.25 < Z < 0) = 0.5 - \Phi(-1.25) \\ &= 0.5 - 0.1056 = 0.3944 \end{aligned}$$

d) (8 points) State patrol officers will issue radar tickets to vehicles whose speeds are in the **top 4%** of this distribution. What is the speed cutoff for issuing tickets?

The top 4% corresponds to the 96th percentile.

$$z = 1.75$$

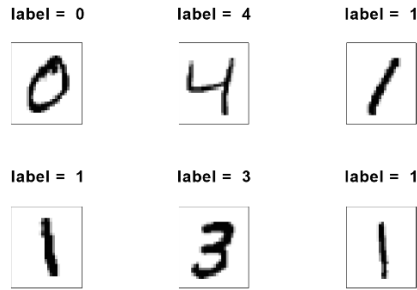
Transform to the distribution of car speeds on I-65:

$$v_{0.96} = \mu + z \times \sigma = 71 + 1.75 * 8 = 85$$

The cutoff for the top 4% of vehicle speeds on I-65 is 85 miles per hour.

The rest of this page can be used for scratch work

4. (26 points) Kristin, a data science major, is working on a term project to build a predictive model that can classify images of handwritten digits (0-4).



She has a dataset containing 1600 images, each displaying a single digit. Kristin divided the dataset into a **training set of 1000 images** and a **test set of 600 images**. The training set is used to teach the model, while the test set is used to evaluate its performance.

After training, Kristin used the test set to create a **confusion matrix**, which shows the number of **correctly** and **incorrectly classified images**. In the matrix below, **rows** indicate the **actual labels (ground truth)**, and **columns represent the predicted labels** made by the **model**:

		Predicted Label					
True Label	Digits	0	1	2	3	4	Total
	0	107	0	0	1	8	116
	1	0	117	1	0	4	122
	2	0	4	92	11	1	108
	3	3	1	15	112	1	132
	4	4	0	0	4	114	122
	Total	114	122	108	128	128	600

Reading the Table: The **highlighted cell** with the value **117** indicates that the model correctly predicted the digit '1' for **117 images** that had **True Label** as '1'. This number represents the model's **accurate classifications** for the digit '1' in the **test set**.

All questions below refer to the data presented in the confusion matrix (table).

- a) (3 points) Define the events:

- $E_1 = \{\text{true label is 4}\}$
- $E_2 = \{\text{true label is 1 or 2}\}$
- $E_3 = \{\text{predicted label is 0}\}$

Which of the following statements is TRUE?

Ⓐ Two events E_1 and E_3 are mutually exclusive.

Ⓑ $P(E_1 \cap E_3) = P(E_1)P(E_3)$.

Ⓒ Two events E_1 and E_2 are disjoint.

Ⓓ $P(E_2 \cup E_3) > P(E_2) + P(E_3)$.

[Questions b)-e)] Kristin wants to know if the model performs better than random guessing at classifying images of the digit three. Define the events:

- $T_3 = \{\text{true label is 3}\}$
- $P_3 = \{\text{predicted label is 3}\}$

b) (5 points) What is the probability that a randomly selected image has the true label three?

$$P(T_3) = \frac{132}{600} = 0.22$$

c) (5 points) What is the probability that a randomly selected image is predicted to be three?

$$P(P_3) = \frac{128}{600} = 0.2133$$

d) (8 points) What is the probability that an image of digit three is correctly predicted to be three?

$$P(P_3|T_3) = \frac{112}{132} = 0.8485$$

- e) **(5 points)** Are the events T_3 and P_3 independent? State your answer and provide a mathematical justification.

No, they are not independent as the conditional probability does not equal the unconditional probability.

$$P(P_3|T_3) = \frac{112}{132} = 0.8485 \neq 0.2133 = \frac{128}{600} = P(P_3)$$

The rest of this page can be used for scratch work

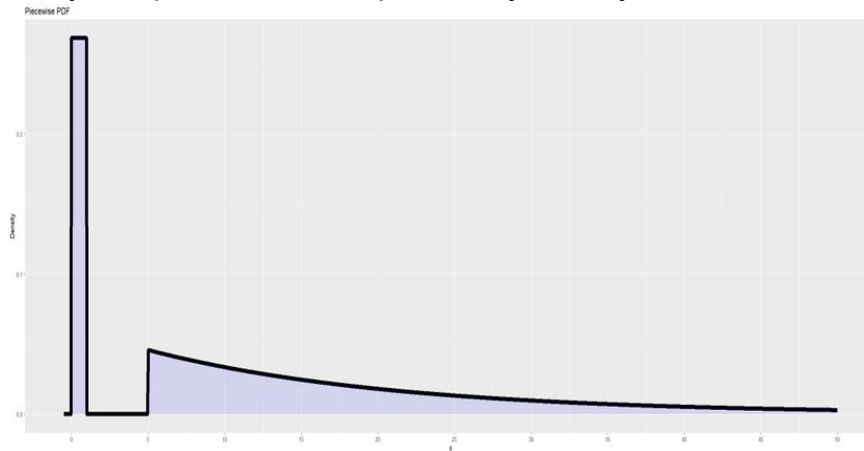
5. (32 points) Robust-ish Devices Inc. manufactures devices whose lifetimes are divided into three distinct phases: early failure, stable operation, and wear-out.

- **Phase 1 (Early Failure):** During the first year ($0 \leq x \leq 1$), the device has a constant likelihood of failing due to manufacturing defects, meaning the probability density function (pdf) for the device's **lifetime** is **constant** in this interval.
- **Phase 2 (Stable Operation):** After surviving the early failure phase, the device operates reliably with virtually no chance of failure for the next 4 years ($1 < x \leq 5$), meaning the pdf is zero during this phase, as the device is highly reliable.
- **Phase 3 (Wear Out):** Beyond 5 years ($x > 5$), the device enters a wear-out phase where the likelihood of failure increases over time. The **lifetime** is modeled by an exponentially decaying function, meaning the chance of the device surviving much longer decreases, and the risk of failure increases as the device ages.

The probability density function for X (**the lifetime of the device**) is given by the following piecewise function.

$$f_X(x) = \begin{cases} 1 - e^{-\frac{5}{16}} & 0 \leq x \leq 1 \\ \frac{1}{16} e^{-\frac{x}{16}} & x \geq 5 \\ 0 & \text{otherwise} \end{cases}$$

a) (10 points) Verify that $f_X(x)$ is a valid probability density function.



Axiom 1: $f_X(x) \geq 0$ clearly by the graph of the pdf or because it is a positive constant over $0 \leq x \leq 1$, an exponentially decaying function over $x \geq 5$ and 0 everywhere else.

Axiom 2:

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x) dx &= \int_0^1 \left(1 - e^{-\frac{5}{16}}\right) dx + \int_5^{\infty} \frac{1}{16} e^{-\frac{x}{16}} dx \\ &= \left(1 - e^{-\frac{5}{16}}\right) - e^{-\frac{x}{16}} \Big|_5^{\infty} = \left(1 - e^{-\frac{5}{16}}\right) + e^{-\frac{5}{16}} = 1 \end{aligned}$$

- b) (10 points) Determine the missing value of the cumulative distribution function (CDF) $F_X(x)$, which is partially given below.

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \left(1 - e^{-\frac{5}{16}}\right)x & 0 \leq x \leq 1 \\ 1 - e^{-\frac{5}{16}} & 1 \leq x \leq 5 \\ [\text{Unknown}] & x \geq 5 \end{cases}$$

[Unknown]

$$\begin{aligned} &= \left(1 - e^{-\frac{5}{16}}\right) + \int_5^x \frac{1}{16} e^{-\frac{t}{16}} dt \\ &= \left(1 - e^{-\frac{5}{16}}\right) - e^{-\frac{t}{16}} \Big|_5^x \\ &= 1 - e^{-\frac{x}{16}} \end{aligned}$$

- c) (4 points) Determine the probability that the device lasts longer than 1 year.

$$P(X > 1) = 1 - F_X(1) = 1 - \left(1 - e^{-\frac{5}{16}}\right) = e^{-\frac{5}{16}} = 0.7316$$

- a) (8 points) Find the **25th percentile** for the **lifetime** of devices manufactured by Robust-ish Devices Inc..

The 25th percentile happens in the region between 0 and 1 because the CDF is $1 - e^{-\frac{5}{16}} \approx 0.2684$ between 1 and 5 and will be larger than this after 5.

Solve $F_X(x) = 0.25$ for the region $[0, 1)$

$$\left(1 - e^{-\frac{5}{16}}\right)x^* = 0.25$$

$$x^* = \frac{0.25}{\left(1 - e^{-5/16}\right)} \approx 0.9315$$

The 25th percentile of lifetime is **0.9315**.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

[illegible]