Section: ___________ Name: ____________________________________

Read the following directions carefully. DO NOT turn to the next page until the exam has started.

Write your name and section number at the top right of this page:

Class	Section Number
Miranda 8:50	001
Sahifa 12:05	002
Miranda 9:55	004

As you complete the exam, write your initials at the top right of each other page.

When the exam start time is called, you may turn the page and begin your exam.
If you need more room, there is a blank page at the end of the exam, or we can give you some scratch paper.

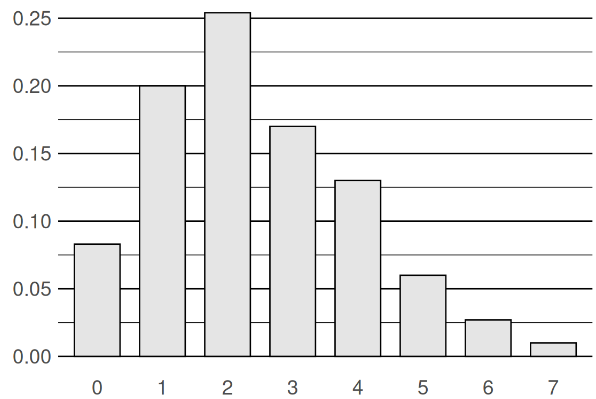
Some multiple choice questions are "Select ONE" while others are "Select ALL that apply". Pay attention to the question type and only mark one option if it says "Select ONE". Fill in the circles completely.

No calculators are allowed for this exam. For questions that require a numeric solution, you can leave your answer as an algebraic expression without needing to simplify it.

If you finish early, you can hand your exam to your instructor or TA and leave early.

Otherwise, stop writing and hand your exam to your instructor or TA when the exam stop time is called.

1. The number of goals in a game in the FIFA World Cup is given by the random variable X. The distribution of X is given in the plot below.



(a) (3 points) Find the 25th, 50th and the 75th percentile of X.

The percentiles are based on the cumulative probabilities. $P(X \leq 0)$ is about 0.08, $P(X \leq 1)$ is about $0.08 + 0.2 = 0.28$, $P(X \leq 2)$ is about $0.28 + 0.25 = 0.53$, $P(X \leq 3)$ is about $0.53 + 0.17 = 0.7$, and $P(X \leq 4)$ is about $0.7 + 0.13 = 0.83$.

So, the 25th percentile is 1, the 50th percentile is 2, and the 75th percentile is 4. These are the values of X that capture at least this much cumulative probability.

(b) (3 points) Is $E(X)$ equal to, smaller than, or larger than the 50th percentile of X? Explain how you can tell without knowing $E(X)$.

$E(X)$ must be larger than the 50th percentile (the median) because of the skew in the distribution of X. Most of the values are on the left side of the graph, around 1 or 2, but there is a tail of bigger values going to the right. The mean will be pulled higher because of this right tail.

(c) (4 points) Which option is closest to the probability of scoring 5 or more goals in a FIFA World Cup game? **Select ONE**.

- [x] `0.11`
- [ ] `0.02`
- [ ] `0.5`
- [ ] `0.96`

2. Let $M \sim N(10, 1)$ denote the shoe sizes for males in the U.S. The area under the curve between shoe size 8 and 12 is 95%, i.e. $P(8 \leq M \leq 12) = 0.95$.

   Let $F \sim N(8, 0.8)$ denote the shoe sizes for females in the U.S.

   (a) (3 points) Which is taller - the height of the peak of the M bell-curve or the height of the peak of the F bell-curve? Briefly explain your answer.

   The height of F must be taller than the height of M because F has a smaller standard deviation, and therefore the bell curve for F is more concentrated around 8 than the bell curve for M is concentrated around 10.

   (b) (4 points) Find the points lo and hi such that the area under the curve for F is 95%. In other words, $P(lo \leq F \leq hi) = 0.95$. You do not need to simplify your answers and can leave them as an expression.

   We are given the percentiles that cover 95% of the area under the curve for M. If we convert these into z-scores, we get

$$z_{lo} = \frac{8-10}{1} = -2, \qquad z_{hi} = \frac{12-10}{2} = 2$$

   (two standard deviations below and two standard deviations above the mean). Then we can convert these into values on F.

$$lo = 0.8(-2) + 8 = 6.4, \qquad hi = 0.8(2) + 8 = 9.6$$

   So $P(6.4 \leq F \leq 9.6) = 0.95$.

   (c) (3 points) We want to confirm that the values `lo` and `hi` from part (a) actually do have 95% of the area between them. Which lines of R code return the area between `lo` and `hi`? **Select ALL that apply**.

   ○ `2*pnorm(lo, 8, 0.8)`
   ○ `2*pnorm(hi, 8, 0.8)`
   ● `1 - 2*pnorm(lo, 8, 0.8)`
   ○ `1 - 2*pnorm(hi, 8, 0.8)`
   ○ `pnorm(lo, 8, 0.8) - pnorm(hi, 8, 0.8)`
   ● `pnorm(hi, 8, 0.8) - pnorm(lo, 8, 0.8)`

3. Based on historic data, a carnival believes that a player has a 7% chance of winning at ring toss. They model the number of winners (out of 350 total players) as $\text{Binom}(350, 0.07)$.

   (a) (3 points) Identify the four assumptions that are being made for this binomial model, and briefly discuss how each one applies to this situation. Name one assumption that might not be well met, and explain.

   - B - the outcomes are binary, which is true for a game (each player either wins or loses)
   - I - the players are independent of each other, which is probably true
   - N - there is a fixed number of players, which is true ($n = 350$)
   - S - each player is equally likely (7% chance) to win.

   Out of all of the assumptions, the same probability assumption may not be well met. Some players might be more or less skilled at the game and therefore have a different probability of winning.

   (b) (3 points) Assuming the binomial model is accurate, fill in the three blanks of `pbinom()` below to calculate the probability that 30 or more players out of 350 win ring toss.

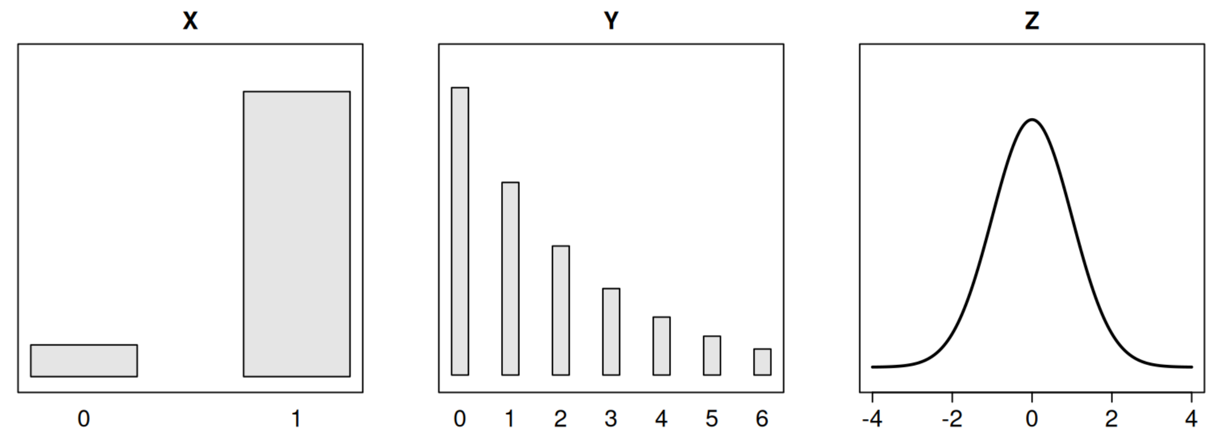
   ```
   1 - pbinom( 29 , 350 , 0.07 )
   ```

   (c) (4 points) Consider performing the calculation in (b) with a normal approximation. Fill in the four blanks of `pnorm()` below to approximate the probability that 30 or more players out of 350 win ring toss.

   You don't need to simplify your answers for the blanks.

   `pnorm(` 30 `,` 350*0.07 `,` $\sqrt{350 * 0.07 * 0.93}$ `,` `lower.tail = F` `)`

4. Consider the following random variables: X is discrete and can take on values 0 or 1, Y is a different skewed discrete variable, and Z is normal.



(a) (3 points) Which of the following will have a normal or approximately normal distribution? **Select ALL that apply**.

- ● The average of 200 samples from X.
- ○ The interquartile range (the 75th percentile minus 25th percentile) of 20 samples of Z.
- ● The average of 100 samples from Y.
- ○ The median of 100 samples from Y.
- ● The average of 1 sample from Z.
- ● The sum of 200 samples from X.

A researcher who does not know the true distribution of X wants to estimate $p = P(X = 1)$. They report a 95% confidence interval for p as $(0.845, 0.921)$.

(b) (4 points) Which line of R code correctly calculates the critical value / quantile score for the 95% CI? **Select ONE**.

- ○ `qnorm(0.9)`
- ○ `qnorm(0.95)`
- ● `qnorm(0.975)`
- ○ `qnorm(0.99)`

(c) (3 points) Suppose the researcher instead set $\alpha = 0.02$ to determine the confidence level. How would this change the interval compared to the 95% CI above? Assume the data does not change. **Select ALL that apply**.

- ○ The new CI would have a smaller margin of error than the 95% CI.
- ○ The new CI would have the same margin of error as the 95% CI.
- ● The new CI would have a larger margin of error than the 95% CI.
- ○ The new CI would have a smaller point estimate than the 95% CI.
- ● The new CI would have the same point estimate as the 95% CI.
- ○ The new CI would have a larger point estimate than the 95% CI.

5. Carabiners made at a factory must have a breaking strength of at least 4500 lbs ($H_0 : \mu \geq 4500$). A manager thinks that the mean breaking strength may be less than 4500 lbs ($H_A : \mu < 4500$) so he takes a sample of carabiners to test his hypothesis with $\alpha = 0.1$. If he has evidence that $\mu < 4500$, he will halt carabiner production.

   (a) (5 points) Classify each scenario as a type I error, type II error, or correct decision. **Select ONE each.**

   The true mean breaking strength is 4500 lbs. The manager allows production to continue.
   ○ Type I error ○ Type II error ● Correct decision

   The true mean breaking strength is 4500 lbs. The manager halts production.
   ● Type I error ○ Type II error ○ Correct decision

   The true mean breaking strength is 4300 lbs. The manager allows production to continue.
   ○ Type I error ● Type II error ○ Correct decision

   The true mean breaking strength is 4300 lbs. The manager halts production.
   ○ Type I error ○ Type II error ● Correct decision

   The true mean breaking strength is 4800 lbs. The manager allows production to continue.
   ○ Type I error ○ Type II error ● Correct decision

   (b) (3 points) Explain whether the manager should make α larger or smaller to avoid mistakenly selling carabiners that are not strong enough.

   The manager should make α larger in order to avoid making a type II error. A type II, or false negative error, would mean that the alternative is true (the breaking strength is too low) but the manager's test failed to detect this. By making α larger, he makes it easier to reject the null in favor of the alternative.