Homework-7

waheeb Algabri

Question 1

Let X_1, X_2, . . . , X_n be n mutually independent random variables, each of which is uniformly distributed on the integers from 1 to k. Let Y denote the minimum of the X_i's. Find the distribution of Y.

The Solution

Our objective is to ascertain a function that characterizes the distribution of Ys, namely the probability mass function $P(Y = y)$. To begin, let's establish the cumulative distribution of Y, denoted as $F(Y) = P(Yy)$. Since Y represents the minimum of each X_i, the likelihood that Y will be less than or equal to a certain value y is equivalent to the probability that at least one of the random variables X is less than or equal to y. In simpler terms, it's the complement of the joint probability that all variables X are greater than y.

$$P(Y \leq y) = 1 - P(X_1 > y) \times P(X_2 > y) \times \ ... \ \times P(X_n > y)$$

Each individual variable X is uniformly distributed, so $P(X_i > y) = (k-y)/k$, and the joint probability can be expressed as $P(Y > y) = [(k-y)/k]^n$, and our CDF can be expressed as $P(Y \leq y) = 1 - [(k-y)/k]^n$.

Because all variables X are discrete, we can define the probability that Y is a specific value y as the difference between the CDFs given y and $y - 1$. In other words:

$$P(Y = y) = P(Y \leq y) - P(Y \leq y - 1)$$

If we substitute our CDF definitions above, we get the following.

$$P(Y = y) = 1 - (\frac{k-y}{k})^n - [1 - (\frac{k-(y-1)}{k})^n]$$

We can rearrange this and simplify as follows.

$$P(Y = y) = (\frac{k-y+1}{k})^n - (\frac{k-y}{k})^n$$

We can put together a function to test this distribution for certain values of n and k, given a certain value y. We generate plots for different combinations of n and k to observe the varied distributions.

```
# Function to calculate probability mass function (PMF) for Y
Y_probability_mass_function <- function(y, n, k) {
  # Calculate probability based on the formula
  probability <- ((k - y + 1) / k)^n - ((k - y) / k)^n
  return(probability)
}
```

```r
# Values for n and k
n_values <- c(5, 10, 5, 15)
k_values <- c(20, 20, 15, 15)

# Initialize empty data frame to store results
results <- data.frame()

# Iterate over each pair of n and k values
for (i in 1:4) {
  n <- n_values[i]  # Extract current n value
  k <- k_values[i]  # Extract current k value

  # Iterate over possible values of y
  for (y in 1:k) {
    # Calculate probability for the current y, n, and k
    probability <- Y_probability_mass_function(y, n, k)

    # Store the result in a list
    result <- list(y = y, n = n, k = k, probability = probability)

    # Append the result to the results data frame
    results <- rbind(results, result)
  }
}

# Plot the results
results %>%
  ggplot(aes(y, probability, color = interaction(n, k))) +
  geom_line()
```

Homework-7_files/figure-latex/unnamed-chunk-1-1.pdf

```r
# Plot the second set of results with a different visualization
ggplot(results, aes(y, probability, fill = factor(k))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Probability Mass Function (PMF) for Y",
       x = "Y",
       y = "Probability",
       fill = "k",
       caption = "Source: Your Data") +
  theme_minimal()
```

Homework-7_files/figure-latex/unnamed-chunk-2-1.pdf

Question 2

Your organization owns a copier (future lawyers, etc.) or MRI (future doctors). This machine has a manufacturer's expected lifetime of 10 years. This means that we expect one failure every ten years. (Include the probability statements and R Code for each part.)

Part A

What is the probability that the machine will fail after 8 years? Provide also the expected value and standard deviation. Model as a geometric. *Hint: the probability is equivalent to not failing during the first 8 years.*

The Solution we can define the machine failing as a "success," and each year as a separate "trial." In this scenario, our probability of success per trial, the likelihood of failure each year) is 1/10. If we're seeking the probability that the machine fails after 8 years, we're essentially interested in the cumulative probability of failure over those 8 years, then we are looking for $P(X > 8)$.

Formula for geometric:_ $P(X = x) = (1 - p)^{x-1}p$

We can use the CDF of the geometric distribution to find $P(X \leq 8)$, then take the complement (i.e. $1-P(X \leq 8)$) to find $P(X > 8)$. The CDF is defined as $F(x) = P(X \leq x) = 1 - (1 - p)^{x+1}$, so we simply need to plug in our values for p and x (i.e., 1/10 and 8, respectively).

The expected value and standard deviation of the geometric distribution are defined as $1/p$ and $\sqrt{(1-p)/p^2}$.

```
# Parameters
p <- 1/10
x <- 8

# Calculations
probability <- 1 - (1 - (1 - p)^(x+1) )
expected_value <- 1/p
standard_deviation <- sqrt( (1-p) / p^2 )

# Results
cat("Probability of failure after 8 years:", probability, "\n")
```

```
## Probability of failure after 8 years: 0.3874205
```

```
cat("Expected value:", expected_value, "\n")
```

```
## Expected value: 10
```

```
cat("Standard deviation:", standard_deviation, "\n")
```

```
## Standard deviation: 9.486833
```

We'll check our results with the pgeom function.

```
probability <- 1 - pgeom(x, prob = p)
cat("Probability of failure after 8 years:", probability, "\n")
```

```
## Probability of failure after 8 years: 0.3874205
```

Our results match! However, this results appears somewhat lower than I would expect, given the expected value is greater than our x. As a sanity check, I'll run a quick simulation, in which we randomly sample a 1 or 0 (representing success or failure) using $p = 0.1$, then extract the index of the first appearance of a 1. This gives us an empirical measure of how many trials are required before we realize a success.How many years before a machine failure occurs). We'll iterate this simulation 100,000 times and aggregate results.

```r
# Set parameters
num_simulations <- 10000
sample_size <- 50
probability_of_success <- 1/10
results <- data.frame()

# Simulate multiple scenarios
for (sim in 1:num_simulations) {
  # Generate samples with replacement
  samples <- sample(0:1, size = sample_size, replace = TRUE, prob = c(1 - probability_of_success, probal

  # Find the index of the first success
  first_success_index <- min(which(samples == 1))

  # Store the result in a list
  result <- list(simulation = sim, first_success_index = first_success_index)

  # Append the result to the results data frame
  results <- rbind(results, result)
}

# Plot the histogram of first success indices
ggplot(results, aes(first_success_index, ..density..)) +
  geom_histogram(bins = 50, fill = "lightblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = 9), linetype = "dashed", color = "red") +
  labs(title = "Distribution of First Success Index",
       x = "Index of First Success",
       y = "Density",
       caption = "Source: Simulation Results") +
  theme_minimal()
```

Homework-7_files/figure-latex/unnamed-chunk-5-1.pdf

To my surprise, the plot aligns with our previous findings

Part B

What is the probability that the machine will fail after 8 years?. Provide also the expected value and standard deviation. Model as an exponential.

The Solution We can define the machine failing as a "success". We don't have discrete trials when modeling as an exponential process, but instead use our probability of success as a rate of success. This rate (i.e. λ) is 1/10. If we are looking for the probability that the machine fails after 8 years, then we are looking for $P(X > 8)$.

Formula for exponential:_ $P(X = x) = \frac{1}{\theta} e^{\frac{-x}{\theta}}$.

The CDF of the exponential distribution is defined as $F(x) = P(X \leq x) = 1 - e^{-\lambda x}$, so we simply need to plug in our values for λ and x (i.e., 1/10 and 8, respectively). Again, we'll take the compliment (i.e. $1 - P(X \leq 8)$) to find $P(X > 8)$.

The expected value and standard deviation of the geometric distribution are both defined as $1/\lambda$.

```r
# Parameters
lambda <- 1/10
x <- 8

# Calculations
probability <- 1 - (1 - exp(-lambda*x))
expected_value <- 1/lambda
standard_deviation <- sqrt( 1/lambda^2 )

# Results
cat("Probability of failure after 8 years:", probability, "\n")
```

```
## Probability of failure after 8 years: 0.449329
```

```r
cat("Expected value:", expected_value, "\n")
```

```
## Expected value: 10
```

```r
cat("Standard deviation:", standard_deviation, "\n")
```

```
## Standard deviation: 10
```

We'll check our results with the `pexp` function.

```r
probability <- 1 - pexp(x, rate = lambda)
cat("Probability of failure after 8 years:", probability, "\n")
```

```
## Probability of failure after 8 years: 0.449329
```

Our results match!

Part C

What is the probability that the machine will fail after 8 years? Provide also the expected value and standard deviation. Model as a binomial. *Hint: 0 success in 8 years.*

The Solution As with the first problem, we can define the machine failing as a "success", and each year as a separate "trial". In this case, our probability of success per trial (i.e. the likelihood of failure each year) is 1/10. If we are looking for the probability that the machine fails after 8 years, then we define our random variable as $X \sim B(n, p)$ with $n = 8$ and $p = 0.1$. Finally, we are looking for the probability that X remains zero across all 8 trials, i.e. $P(X = 0)$.

Formula for binomial:_ $P(X = x) = \binom{n}{x} p^x q^{n-x}$

We can use the PMF of the binomial distribution to find $P(X = 0)$, defined as $\binom{n}{k} p^k (1-p)^{n-k}$. The expected value and standard deviation of are defined as $n * p$ and $\sqrt{n * p * (1 - p)}$.

```r
# Parameters
p <- 1/10
n <- 8
k <- 0

# Calculations
probability <- choose(n,k) * p^k * (1-p)^(n-k)
expected_value <- n*p
standard_deviation <- sqrt( n*p*(1-p) )

# Results
cat("Probability of failure after 8 years:", probability, "\n")
```

```
## Probability of failure after 8 years: 0.4304672
```

```r
cat("Expected value:", expected_value, "\n")
```

```
## Expected value: 0.8
```

```r
cat("Standard deviation:", standard_deviation, "\n")
```

```
## Standard deviation: 0.8485281
```

We'll check our results with the `dbinom` function.

```r
probability <- dbinom(0, n, p)
cat("Probability of failure after 8 years:", probability, "\n")
```

```
## Probability of failure after 8 years: 0.4304672
```

Our results match!

Part D

What is the probability that the machine will fail after 8 years? Provide also the expected value and standard deviation. Model as a Poisson.

The Solution Typically, the Poisson distribution is used to estimate the probability of a certain number of occurrences happening over a fixed period. Here, we are looking specifically if the number of occurrences is zero over an 8-year period. If start with the same annual lambda as we had in Part B (1/10), we can convert this to an 8-year lambda, which we'll dub λ_8. We can then calculate the probability that X remains zero across the 8-year period, i.e. $P(X = 0)$ given λ_8.

_**Formula for binomial:** $P(X = x) = \frac{\lambda^x e^- \lambda}{x!}$

The PMF of the Poisson distribution is defined as $P(X = x) = (\lambda^x e^{-\lambda})/x!$. The expected value and standard deviation are defined as $lambda$ and $\sqrt{\lambda}$.

```r
# Parameters
lambda <- 1/10
lambda_8 <- lambda * 8
x <- 0

# Calculations
probability <- lambda_8^x * exp(-lambda_8) / factorial(x)
expected_value <- lambda
standard_deviation <- sqrt(lambda)
expected_value_8 <- lambda_8
standard_deviation_8 <- sqrt(lambda_8)

# Results
cat("Probability of failure after 8 years:", probability, "\n\n")
```

```
## Probability of failure after 8 years: 0.449329
```

```r
cat("Expected number of failures for 1-year period:", expected_value, "\n")
```

```
## Expected number of failures for 1-year period: 0.1
```

```r
cat("Standard deviation for 1-year period:", standard_deviation, "\n\n")
```

```
## Standard deviation for 1-year period: 0.3162278
```

```r
cat("Expected number of failures for 8-year period:", expected_value_8, "\n")
```

```
## Expected number of failures for 8-year period: 0.8
```

```r
cat("Standard deviation for 8-year period:", standard_deviation_8, "\n")
```

```
## Standard deviation for 8-year period: 0.8944272
```

Finally, we'll check out results with `dpois`.

```r
probability <- dpois(x, lambda_8)
cat("Probability of failure after 8 years:", probability)
```

```
## Probability of failure after 8 years: 0.449329
```

Another way to approach the same result!

```r
# Parameters
lambda <- 1/10  # Rate parameter (failures per year)
t <- 8  # Time in years

# Probability of machine failure after 8 years
probability <- exp(-lambda * t)
cat("Probability of failure after 8 years:", probability, "\n")
```

```
## Probability of failure after 8 years: 0.449329
```

```r
# Expected value and standard deviation for exponential distribution
expected_value <- 1 / lambda
standard_deviation <- 1 / lambda
cat("Expected value (mean):", expected_value, "\n")
```

```
## Expected value (mean): 10
```

```r
cat("Standard deviation:", standard_deviation, "\n")
```

```
## Standard deviation: 10
```

```r
# Parameters
lambda <- 1/10  # Rate parameter (failures per year)
t <- 8  # Time in years

# Generate random samples from exponential distribution
samples <- rexp(10000, rate = lambda)

# Create a histogram
ggplot() +
  geom_histogram(aes(x = samples), bins = 30, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Machine Failure Times",
       x = "Time (years)",
       y = "Frequency") +
  theme_minimal()
```

Homework-7_files/figure-latex/unnamed-chunk-13-1.pdf