DATA605 - Assignment 7

Jai Jeffryes

4/13/2020

Note on written style

In no way is my written style here academic, this I know. I include my extra internal dialog with its detours and mistakes. I want you to see that I'm doing my own work. I don't just want to fish for answers on the internet to get credit. I want these to be teachable moments for me. Moments? Try hours, days, and weeks. Dark nights of the soul complete with hair pulling. I want to get this!

Problem 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.

Thinking it through

It's a little late in the game to read a variable's case incorrectly. It said "random variables," it didn't say outcomes. I looked at it and saw a vector and thought, the minimum is just one value. Nope. There are n random variables, and each one X_i is made of outcomes x_i whose range is [1, k].

At least I think I know what the question is asking. I want to see what these outcomes would look like. I'll generate n X's, each sampled from the range [1, k], and assign the minimum of each X to Y. Then I can view the frequency of the minimum values that appear across all the X outcomes.

Note: I think there must be a better way to generate integers than taking the floor of `runif()`. The interval [19, 20) generates the value 19, but the maximum, 20, is only a single value. The maximum value is under represented in frequency. Nevertheless, I see what I want from the simulation because it's so rare that all outcomes in a single X would equal the maximum sampling value.

```
simulate <- function(n, k) {
    # Allocate some space for holding Y.
    Y <- rep(as.integer(NULL), n)

    # For each X, generate a random uniform distribution.
    # Truncate the decimal part to return integers.
    for (i in 1:n){
        X <- floor(runif(k, min = 1, max = k))

        # If n isn't too big, print the data.
        if (n <= 5){
            print(paste0("X[", i, "]"))
            print(X)
        }

        # Pick off the minimum outcome from X.
        Y[i] <- min(X)
    }
```

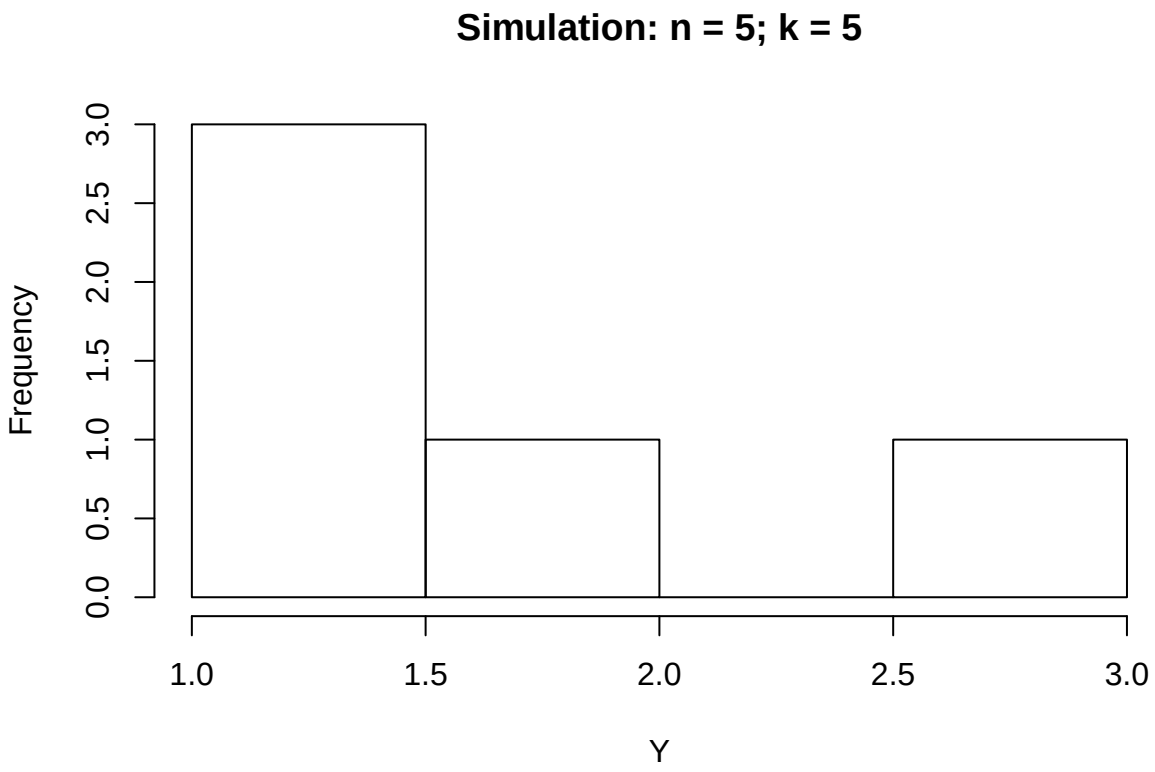
```
    if (n <= 5) {
        print("Y")
            print(Y)
    }
    return(Y)
}
```

First, I'll run this one time with a small enough sample to print them out. I want to see the data.

```
n <- 5
k <- 5
set.seed(59)
Y <- simulate(n, k)
```

```
## [1] "X[1]"
## [1] 1 3 4 3 3
## [1] "X[2]"
## [1] 4 4 2 4 1
## [1] "X[3]"
## [1] 2 1 2 1 1
## [1] "X[4]"
## [1] 3 4 3 3 4
## [1] "X[5]"
## [1] 4 3 2 4 3
## [1] "Y"
## [1] 1 1 1 3 2
```

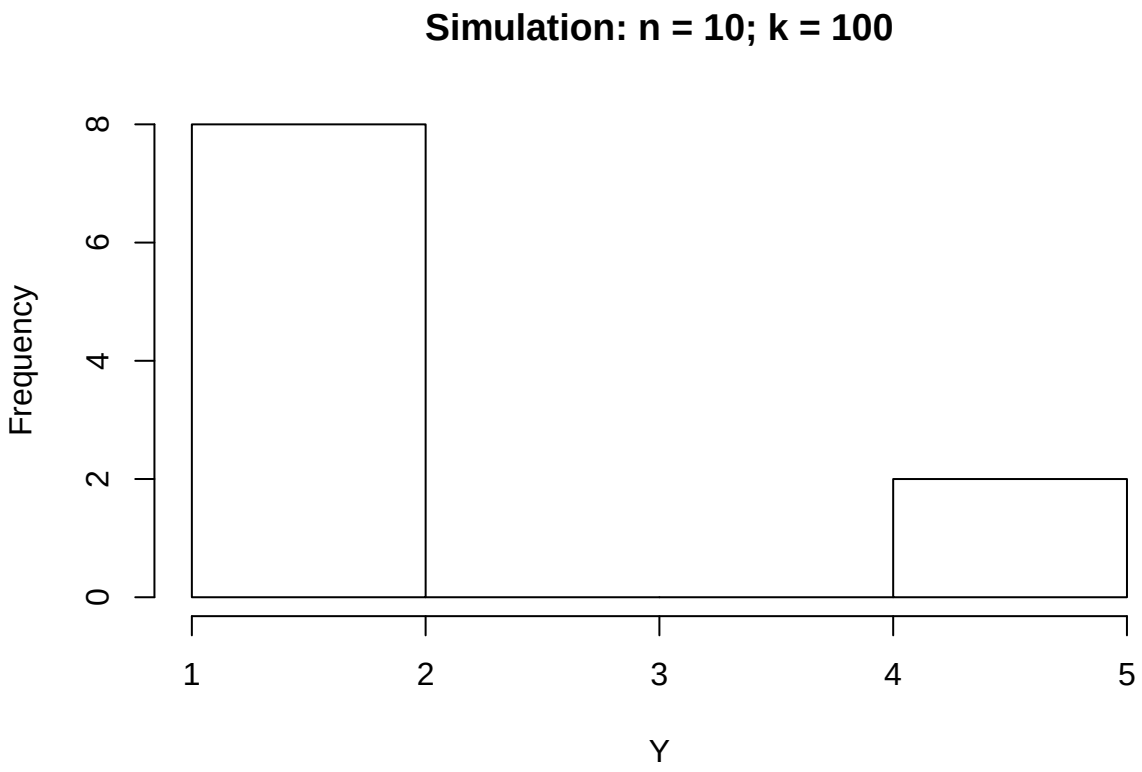
```
hist(Y, main = paste0("Simulation: n = ", n, "; k = ", k))
```



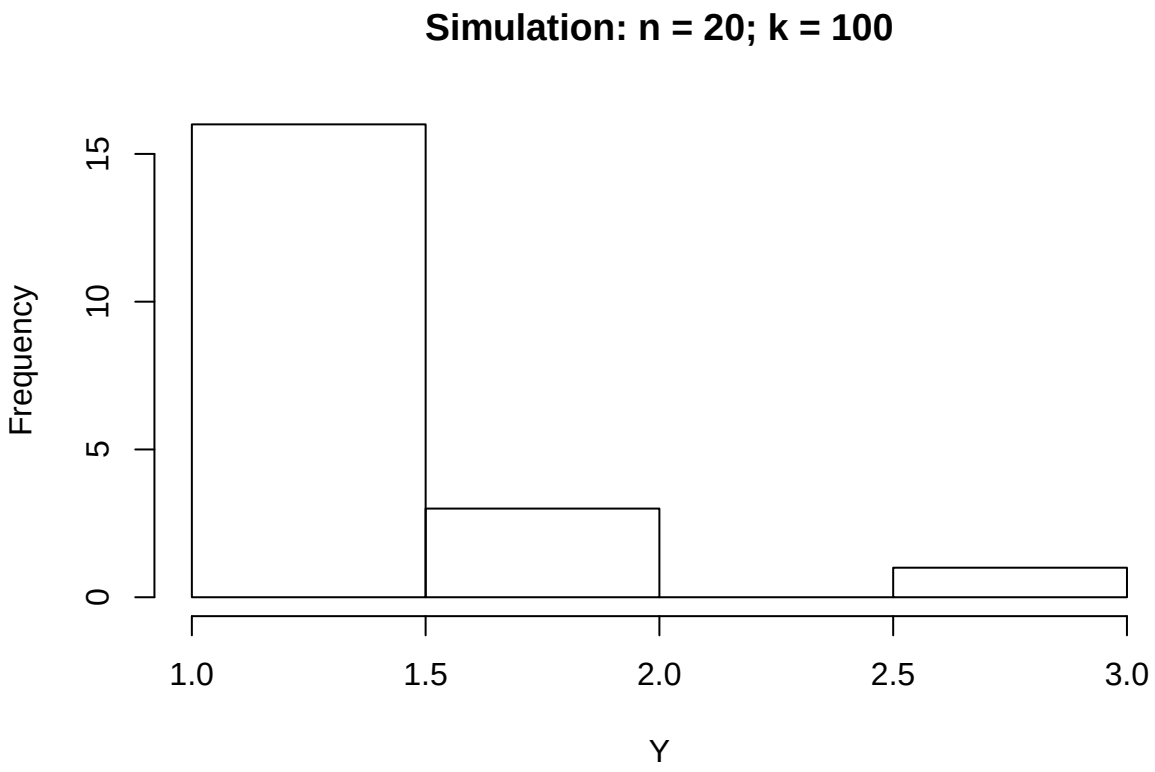
I see in the data that it isn't a simple sequence. I do a random sample with a uniform probability of getting back any value from a range of values. There are some duplicate outcomes within a single random variable, and the minimum among them is not necessarily 1.

Let's run a few larger simulations and look at the plots.

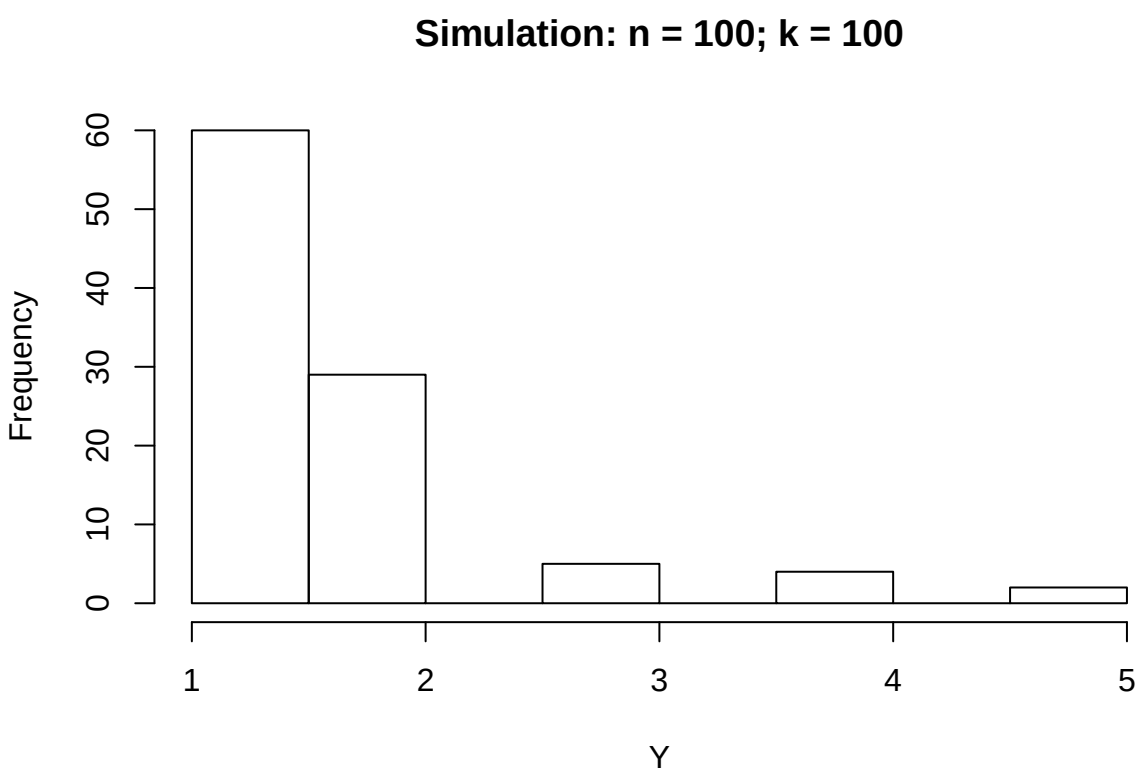
```
n <- 10
k <- 100
Y <- simulate(n, k)
hist(Y, main = paste0("Simulation: n = ", n, "; k = ", k))
```




```r
n <- 20
k <- 100
Y <- simulate(n, k)
hist(Y, main = paste0("Simulation: n = ", n, "; k = ", k))
```




```r
n <- 100
k <- 100
Y <- simulate(n, k)
hist(Y, main = paste0("Simulation: n = ", n, "; k = ", k))
```



That looks about like what I expected. The minimum value of X is most frequently 1, but not always.

So let's think about this. I'm asked for a probability distribution.

> A probability distribution is a table of all disjoint outcomes and their associated probabilities (or it can be expressed by a function).

What is my outcome of interest? For each random variable X I'm getting back the minimum outcome. That can be any value in the range [1, k]. For one random variable X_1 there are k possible values for a minimum. Nevermind the low probability of getting back the value k itself as the minimum. I just want to ennumerate the possibilities. For X_2, it's another k possible values for a minimum. The combination of possible outcomes for X_1 and X_2 together is k^2. Indeed, the possible outcomes for all n random variables X_i is k^n.

This messed me up for days. I fretted over the combinations *within* a single random variable X_i. I was stuck until I realized I don't care about all the x_i (small x) values. I only care about one value, the minimum, and I don't care about its position.

A probability is a proportion. I figured out the denominator, k^n. Now for the numerator.

The probabilities for whose disjoint outcomes I'm looking for:

- $P(Y = 1)$
- $P(Y = 2)$
- ...
- $P(Y = k)$

Since probability is a proportion, I need to count outcomes and divide them by the number of all possible outcomes, which is k^n.

What is the probability $P(Y = 1)$? The difference between 1 and the quotient of outcomes which do not include 1 divided by the number of all possible outcomes.

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

Continuing for the other disjoint outcomes.

$P(Y = 2) = \frac{k^n-(k-2)^n-(k^n-(k-1)^n)}{k^n}$ This simplifies to:

$P(Y = 2) = \frac{k-1)^n-(k-2)^n}{k^n}$

For a value a such that $Y = a$, the probability is.

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

Let's express that as a distribution function.

Answer

For $1 \leq a \leq k$, the distribution function $m(a) = \frac{k-a+1)^n-(k-a)^n}{k^n}$

Finally!

Problem 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..)

We're given that the expected lifetime is 10 years. Therefore, the product is expected to fail in the 11th year. Keep the negatives straight. Success in the mathematical model is defined as product failure. A year that the product works is a failure to fail. I'm going to use variables that describe the events, not calling it distribution "success" and "failure."

The expected value, or mean, is computed as: $\mu = \frac{1}{p}$

I waffled on the expected value. We're given that the expected lifetime is 10 years. My first line of reasoning was this. Remembering that machine life is 10 years, then machine failure is expected in year 11, so the expected value is $11 = \frac{1}{p}$ and $p = \frac{1}{11}$.

After viewing the worked solution in the video provided for the exponential distribution, I suspect I'm thinking about that wrong. The expected value is $10 = \frac{1}{p}$ and $p = \frac{1}{10}$.

Now that we know the probability for machine failure, we can compute the probability given a different number of trials, each year being a trial. We want the probability of machine failure *after* 8 years. That's any year, 9 plus. So let's calculate the probability of failure within those first 8 years and subtract it from 1.

This is the idea, the probability of machine failure within the first 8 years:

$(0.9)^0(0.1) + (0.9)^1(0.1) + \cdots + P(0.9)^8(0.1)$

Subtract that from 1, and I'll get the probability of machine failure in year 9 or any time after.

I'm going to code this two ways in order to check my work, first using the formula I illustrated, and then a cumulative calculation `pgeom()`.

```
p.machine_failure <- 1 / 10

# Calculate a machine failure within the 1st 8 years.
machine_failure_1st_year <- 9

p.machine_failure_8yrs <- 0

for (i in 1:machine_failure_1st_year) {
    p.machine_failure_8yrs <- p.machine_failure_8yrs +
        (1 - p.machine_failure)^(machine_failure_1st_year - i) * p.machine_failure
}

p.machine_failure_8yrs_pgeom <- pgeom(8, p.machine_failure)
```

Probability of machine failure within 8 years:

- Approach with sums: 0.6126
- Approach with `pgeom()`: 0.6126

Bang on. I'm using `pgeom()` correctly. Go ahead and use that term in the calculation of the answer.

Answer

- $P(failure\ after\ 8\ years) = 1 - P(failure\ within\ 8\ years) = 0.3874$.
- Expected value was given: 10.
- Standard deviation $= \sigma = \sqrt{\frac{1-p}{p^2}} = 9.4868$

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 exponential probability distribution is: $P(X \leq k) =$

- $1 - e^{\frac{-k}{\mu}}$, $k \geq 0$
- 0, $k < 0$

This would give us the probability of machine failure *within* 8 years. We want the complement:

- $P(x \geq k) = e^{\frac{-k}{\mu}}$
- $e^{\frac{-8}{10}}$

```
k <- 8
mu <- 10
p <- exp(-k / mu)
```

The standard deviation of the exponential probability distribution equals the mean, which was given as 10.

Answer

- Probability of machine failure after 8 years: 0.4493
- Mean: 10
- Standard deviation: 10

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 successes in 8 years.)

These are good exercises for me. I never remember that the `p` version of `R`'s distribution functions are the cumulative ones. I could just use `pbinom()` to calculate this in one line, but I want the review of the binomial function so I'm going to hand roll the calculation first and compare.

$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$

$\binom{8}{0} 0.1^0 (1-.1)^{8-0} = \frac{8!}{0!(8-0)!} .1^8 (1-.1)^{8-0}$

```
p.by_hand <- choose(8, 0) * 0.1^0 * (1 - 0.1)^(8 - 0)
p.pbinom <- pbinom(0, 8, 0.1)
```

Answer

- Probability (by hand): 0.4305
- Probability (`pbinom()`): 0.4305
- Mean ($\mu = np = 8 \cdot 0.1$): 0.8
- Standard deviation ($\sigma = \sqrt{np(1-p)}$): 0.8485

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.

I think this says we have an event whose rate is 1 per 10 years. But it asks for the probability of observing the event after 8 years. We need to put the rate in a common unit of year. The given rate is 0.1 per year and we're looking at 8 years or trials. We want to observe no failures in those 8 trials.

The Poisson distribution looks like this:

$P(observe\ k\ events) = \frac{\lambda^k e^{-\lambda}}{k!}$

We want to observe no event for 8 periods. We need to take the probability of no observiation in one trial and raise it to the 8th power.

$(\frac{0.1^0 e^{-0.1}}{0!})^8$

```
p.by_hand <- ((0.1^0 * exp(-0.1)) / factorial(0))^8
p.ppois <- ppois(0, 0.1)^8
```

Answer

- Probability (by hand): 0.4493
- Probability (`ppois()^8`): 0.4493
- Mean ($\mu = \lambda$): 0.1
- Standard deviation ($\sigma = \sqrt{\lambda}$): 0.3162

Victory!

I liked this assignment a lot. I really got a chance to see all of these distributions. There's still a lot for me to think about. All of those different values for standard deviation. I must ponder that.

Thank you, Dr. Fulton, for allowing me to turn this in. I learned a lot from sticking it out with this assignment.