Discrete Random Variables Application Solutions

Lt Col Ken Horton   Professor Bradley Warner

16 June, 2020

Exercises

1. Suppose we are flipping a fair coin, and the result of a single coin flip is either heads or tails. Let X be a random variable representing the number of flips until the first heads.

a) Is X discrete or continuous? What is the domain/support of X?

X is discrete since number of flips is a discrete process (I can't perform a fraction of a flip). The wording is specific in that it is the number of flips until the first heads, so we must flip at least once. The domain of X is $S_X = \{1, 2, ...\}$.

b) What values do you *expect* X to take? What do you think is the average of X? Don't actually do any formal math, just think about if you were flipping a regular coin, how long it would take you to get the first heads.

I would *expect* X to be 0 or 1 fairly often, since the coin is fair and has an even chance of landing on heads or tails. I would expect large values of X to be rare. For these reasons, I think the average of X should be around 2 flips or a little less than 2.

c) Advanced: In `R`, generate 10,000 observations from X. What is the average value of X based on this simulation?

Note: There are many ways to do this. Below is a description of one approach.

```r
set.seed(68)
which(sample(c("H","T"),1000,replace=TRUE)=="H")[1]
```

```
## [1] 2
```

Now repeat using `replicate()` or `do()`. We will repeat 10000 times.

```r
results <- do(10000)*which(sample(c("H","T"),1000,replace=TRUE)=="H")[1]
```

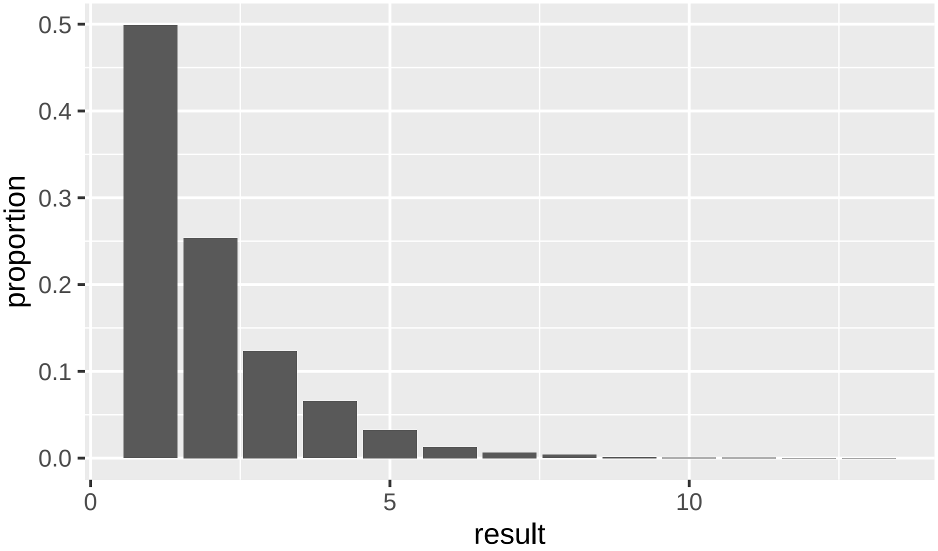
```r
mean(~result,data=results)
```

```
## [1] 1.9849
```

```
tally(~result,data=results,format="percent")
```

```
## result
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 49.89 25.35 12.33  6.56  3.25  1.27  0.66  0.39  0.12  0.07  0.06  0.03  0.02
```

```
results %>%
  gf_props(~result)
```



As predicted, the mean is close to 2, and the most common values of X are 1 and 2. The most common is 1 occurring 50% of the time, this is what we would think since the coin comes up Heads 50% of the time.

d) We know that $\mathrm{P}(X=1)=\frac{1}{2}$ and $\mathrm{P}(X=2)=\frac{1}{2^2}$ so in general $\mathrm{P}(X=x)=\frac{1}{2^x}$. To show that the sum of the infinite sequence of values is 1 requires some Calculus knowledge. Let's start with a partial sum:

$$S_n = \frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^n}$$

Now multiply both sides by $\frac{1}{2}$.

$$\frac{1}{2}S_n = \frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^{n+1}}$$

The difference between these two sums is

$$S_n - \frac{1}{2}S_n = \frac{1}{2}S_n = \frac{1}{2}-\frac{1}{2^{n+1}}$$

Now as

$$\lim_{n\to+\infty}\frac{1}{2^{n+1}}=0$$

So

$$\lim_{n\to+\infty}\left[\frac{1}{2}S_n=\frac{1}{2}-\frac{1}{2^{n+1}}\right]$$

This implies that $S=1$.

2. Repeat Problem 1, but with a different random variable. Consider Y: the number of coin flips until the *fifth* heads.

a) Y is discrete for the same reasons as X. The domain of Y is $S_Y = \{5, 6, ...\}$.

b) In order to land on heads five times, it would be reasonable to expect around 9 to 13 flips. Thus, I would expect Y to take values 8, 9, 10, 11, and 12 fairly often, and values outside of that range less often. I think the average of Y should be around 10 or so.

c)

```
set.seed(102)
results <- do(10000)*which(sample(c("H","T"),1000,replace=TRUE)=="H")[5]
```

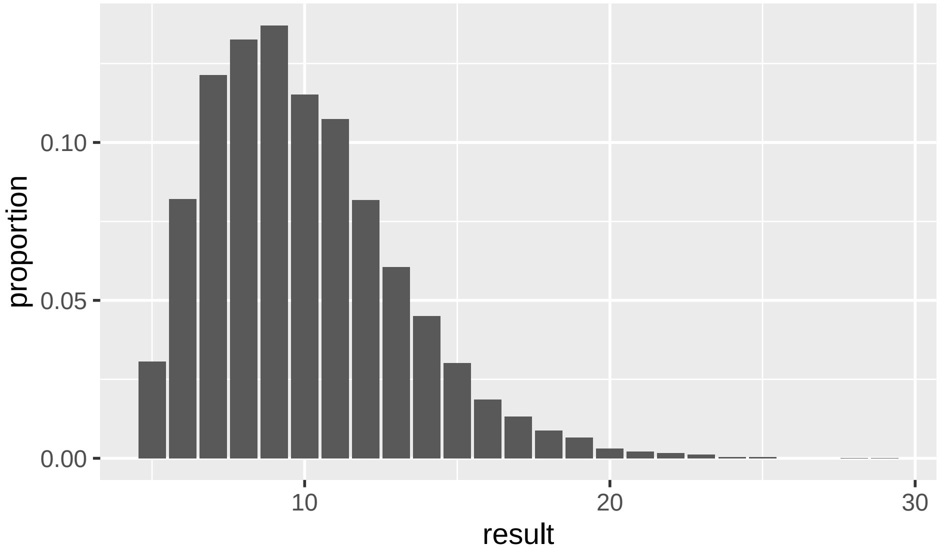
```
mean(~result,data=results)
```

```
## [1] 9.9728
```

```
tally(~result,data=results,format="percent")
```

```
## result
##     5     6     7     8     9    10    11    12    13    14    15    16    17
##  3.06  8.21 12.14 13.26 13.71 11.52 10.74  8.17  6.06  4.50  3.02  1.86  1.32
##    18    19    20    21    22    23    24    25    28    29
##  0.88  0.65  0.31  0.21  0.16  0.12  0.04  0.04  0.01  0.01
```

```
results %>%
  gf_props(~result)
```



The most common values of Y are between 6 and 11. The average of Y in this simulation is 9.97, close to what we predicted.

3. Suppose you are a data analyst for a large international airport. Your boss, the head of the airport, is dismayed that this airport has received negative attention in the press for inefficiencies and sluggishness. In a staff meeting, your boss gives you a week to build a report addressing the "timeliness" at the airport. Your boss is in a big hurry and gives you no further information or guidance on this task.

Prior to building the report, you will need to conduct some analysis. To aid you in this, create a list of at least three random variables that will help you address timeliness at the airport. For each of your random variables,

a) Determine whether it is discrete or continuous.

b) Report its domain.

c) What is the experimental unit?

d) Explain how this random variable will be useful in addressing timeliness at the airport.

I will provide one example:

Let D be the difference between a flight's actual departure and its scheduled departure. This is a continuous random variable, since time can be measured in fractions of minutes. A flight can be early or late, so domain is any real number. The experimental unit is each individual (non-canceled) flight. This is a useful random variable because the average value of D will describe whether flights take off on time. We could also find out how often D exceeds 0 (implying late departure) or how often D exceeds 30 minutes, which could indicate a "very late" departure.

There are many correct answers.

X: Time it takes for a passenger to go through security (defined as time from entering security line to departing security with all belongings). Continuous. Experimental unit is individual passenger. This variable would help identify whether security line is too long. We could also explore how X changes based on day or time of day.

Y: Status of each scheduled departure (on time, somewhat late, very late, canceled). Discrete. Experimental unit is each scheduled departure. This variable will help describe how often flights are canceled or late. We could also explore Y by airline, destination, time of day, etc.

Z: Number of time-related complaints at customer service desk in a given day. Discrete. Experimental unit is day. This variable will describe attitudes/perceptions of customers. It is probably a bad sign if customers feel like the airport is not working efficiently. We can explore how Z changes over time.

4. Consider the experiment of rolling two fair six-sided dice. Let the random variable Y be the absolute difference between the two numbers that appear upon rolling the dice.

a) What is the domain/support of Y?

$S_Y = \{0, 1, 2, 3, 4, 5\}$.

b) What values do you *expect* Y to take? What do you think is the average of Y? Don't actually do any formal math, just think about the experiment.

I'd say that Y should take values 0,1 and 2 fairly often. I'd guess that the average should be around 1.5.

c) Find the probability mass function and cumulative distribution function of Y.

Using counting methods, we know there are 36 possible values. We can just count them. The number 0 will occur when both numbers are the same, which happens six times. The number 1 happens when the first die is one larger than the second, 5 times, or vice versa. Thus 1 happens 10 times. Continue this process. Thus, the pmf of Y becomes:

$$f_Y(y) = \begin{cases} \frac{6}{36}, & y = 0 \\ \frac{10}{36}, & y = 1 \\ \frac{8}{36}, & y = 2 \\ \frac{6}{36}, & y = 3 \\ \frac{4}{36}, & y = 4 \\ \frac{2}{36}, & y = 5 \\ 0, & \text{otherwise} \end{cases}$$

We could also create a table and count the entries.

		Die 2					
		1	2	3	4	5	6
Die 1	1	0	1	2	3	4	5
	2	1	0	1	2	3	4
	3	2	1	0	1	2	3
	4	3	2	1	0	1	2
	5	4	3	2	1	0	1
	6	5	4	3	2	1	0

The cdf of Y is thus,

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \frac{6}{36}, & 0 \leq y < 1 \\ \frac{16}{36}, & 1 \leq y < 2 \\ \frac{24}{36}, & 2 \leq y < 3 \\ \frac{30}{36}, & 3 \leq y < 4 \\ \frac{34}{36}, & 4 \leq y < 5 \\ \frac{36}{36}, & y \geq 5 \end{cases}$$

d) Find the expected value and variance of Y.

$$\mathrm{E}(Y) = \sum_{y=0}^{5} y\mathrm{P}(Y = y) = 0 \times \frac{6}{36} + 1 \times \frac{10}{36} + 2 \times \frac{8}{36} + 3 \times \frac{6}{36} + 4 \times \frac{4}{36} + 5 \times \frac{2}{36} = \frac{70}{36} = 1.944$$

```
y<-c(0,1,2,3,4,5)
sum(y*c(6,10,8,6,4,2)/36)
```

```
## [1] 1.944444
```

e) Advanced: In R, obtain 10,000 realizations of Y. In other words, simulate the roll of two fair dice, record the absolute difference and repeat this 10,000 times. Construct a frequency table of your results (what percentage of time did you get a difference of 0? difference of 1? etc.) Find the mean and variance of your simulated sample of Y. Were they close to your answers in part d?

```r
set.seed(9)
sim_diffs<-do(10000)*abs(diff(sample(1:6,2,replace=T)))
```

```r
tally(~abs,data=sim_diffs,format="proportion")
```

```
## abs
##      0      1      2      3      4      5
## 0.1643 0.2752 0.2273 0.1618 0.1116 0.0598
```

```r
mean(~abs,data=sim_diffs)
```

```
## [1] 1.9606
```

```r
var(sim_diffs)*9999/10000
```

```
##          abs
## abs 2.077248
```

```r
true_mean<-sum(c(6,10,8,6,4,2)/36*c(0,1,2,3,4,5))
true_mean
```

```
## [1] 1.944444
```

```r
sum(c(6,10,8,6,4,2)/36*(c(0,1,2,3,4,5)-true_mean)^2)
```

```
## [1] 2.052469
```

We got similar mean and variance to the theoretical values.

5. Prove the Lemma from the Notes: Let X be a discrete random variable, and let a and b be constants. Then $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

$$\mathrm{E}(aX+b) = \sum_x (ax+b) f_X(x) = \sum_x ax f_X(x) + \sum_x b f_X(x) + a \sum_x x f_X(x) + b \sum_x f_X(x)$$

Since $\sum_x x f_X(x) = \mathrm{E}(X)$ and $\sum_x f_X(x) = 1$, this reduces to $a\mathrm{E}(X) + b$.

$$\begin{aligned}\mathrm{Var}(aX+b) &= \mathrm{E}\left[(aX+b-\mathrm{E}(aX+b))^2\right] = \mathrm{E}\left[(aX+b-a\mathrm{E}(X)-b)^2\right] = \mathrm{E}\left[(aX-a\mathrm{E}(X)^2\right] \\ &= \mathrm{E}\left[a^2(X-\mathrm{E}(X))^2\right] = a^2\mathrm{E}\left[(X-\mathrm{E}(X))^2\right] = a^2\mathrm{Var}(X)\end{aligned}$$

6. In the Notes, we saw that $\mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)^2]$. Show that $\mathrm{Var}(X)$ is also equal to $\mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2] = \mathrm{E}[X^2 - 2\mu_X X + \mu_X^2] = \mathrm{E}(X^2) - \mathrm{E}(2\mu_X X) + \mathrm{E}(\mu_X^2)$$

The quantity μ_X is a constant with respect to X, so

$$= \mathrm{E}(X^2) - 2\mu_X \mathrm{E}(X) + \mu_X^2 = \mathrm{E}(X^2) - 2\mu_X^2 + \mu_X^2 = \mathrm{E}(X^2) - \mu_X^2$$