Lab 4

2022-10-21

The dangers of hard coding

One of the most common problems from the last problem set was in question 1e, which asked you to write a function called mymean(). Many people had the correct solution, or something which worked similarly:

```
mymean <- function(vec){
  sum(vec)/length(vec)
}
```

x	$P(X = x)$
0	1/8
1	3/8
2	3/8
3	1/8

We can code this distribution like this:

```
X <- c(0, 1, 2, 3)
probs <- c(1/8, 3/8, 3/8, 1/8)
```

Then we can sample from it like this:

```
set.seed(2418)

sample_1 <- sample(x = X,
                   size = 1,
                   prob = probs)
sample_1
```

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

```
sample_10000 <- sample(x = X,
                       size = 10000,
                       prob = probs,
                       replace = TRUE)
```

Now, back to our function. Let's check to make sure it works:

```
mean(sample_10000)
```

```
## [1] 1.4965
```

```
mymean(sample_10000)
```

```
## [1] 1.4965
```

Many assignments that something was "hard-coded," and that even though the function may have produced the correct answer in this case, it would not work as intended in general.

First, consider this version of mymean:

```
n <- 10000

mymean_2 <- function(vec){
  sum(vec)/n
}

mymean_2(sample_10000)
```

```
## [1] 1.4965
```

It looks like it works, but notice that the denominator is n instead of length(vec). In this case, the denominator has been "hard-coded" into the function. It does not change when we pass in a different vector. The downside of the hard-coding is that the function only works correctly if you pass in a vector with 10000 entries.

```
sample_5000 <- sample(x = X,
                      size = 5000,
                      prob = probs,
                      replace = TRUE)

mymean_2(sample_5000)
```

```
## [1] 0.742
```

This value is roughly half of what we would expect. Why?

To be clear, there isn't anything wrong with mymean_2. It correctly calculates the mean under certain conditions (specifically, when you have exactly 10000 data points). But, it's usefulness is farily limited. This is almost always true when you hard-code values, parameters, etc. in your functions, so it is a practice to avoid if possible (and it usually is).

Conditional Probability of two Random Variables

When writing the PMF of X conditional on Y, X is the random variable we defined above and Y is a random variable which takes on the value 1 if all three flips are heads and 0 otherwise.

Conditional probability is a concept that will come up again and again when we talk about inference and regression.

x	y	$P(X = x\|Y = y)$
0	0	1/7
1	0	3/7
2	0	3/7
3	1	1

One thing that always helps me conceptualize conditional probabilities is to read | as "assuming that. . ." So, if we want to know P(X = x|Y = y), we want the "probability that X = x assuming that Y = y.

First, assume that Y = 1. There is only one possible event that maps to this value of Y : {HHH}. Therefore, the probability that X = 3 given that Y = 1 is one, and the probability that X takes any other value is 0. Now, assume that Y = 0. How many events map to this value of Y ? Seven: {T T T, T TH, T HH, HT T, HT H, T HT, HHT }.

Now, calculating the conditional probabilities in the table above is as simple as counting how many of the seven map into particular values of X. There is one for which X = 0, 3 for which X = 1, and 3 for which X = 2.

Final Projects!

Hopefully everyone has had a chance to locate an interesting data set you want to work with. We are nearly halfway through the quarter and I want to make sure that everyone is making progress. Hopefully this week's material has prompted deeper thinking about what your data can tell you.

Pull up your data and briefly describe it to the person sitting next to you. Where does it come from? What is the unit of observation?

If you haven't chosen a dataset, what are you considering? What topics interest you? See if anyone around you has suggestions.

Have you run into any problems wrangling your data? Discuss ongoing challenges with the people around you.

What are some questions your data can help you answer? Are they causal or descriptive questions?