BEE552 Biometry Week 6

Maria Feiler

03/02/2022

My Learning Journey

Over the last week, I participated in Biometry in the following ways:

- I asked / answered **2** questions posed in class.
- I asked **3** questions in Slack.
- I answered **0** questions posed by other students on Slack.
- I came to Heather's office hours: **Yes**
- I came to Jose's office hours: **No**
- I met with Heather or Jose separately from office hours: **Yes**

Anything not falling into one of the above categories?

> **No**

On a scale of 1 (no knowledge) to 10 (complete expert), how would I rate my comfort with R programming after this week?

> **7**

Any topics from last week that you are still confused about?

> **. . . everything? The quiz psyched me out a bit.**

Problem Set

Part I

Listen to Planet Money's podcast Episode 677: The Experiment Experiment. http://www.npr.org/sections/money/2016/01/15/463237871/episode-677-the-experiment-experiment. List three take-home messages from the podcast.

- The "file drawer" effect can explain why so many positive results are published, and why so many of those experiments are not reproducible. By relegating negative results to the file drawer, you are inadvertently removing relevant data from the overall body of scientific knowledge.
- Furthermore, if you are relegating negative results to the file drawer, students in the future will never know that those experiments were conducted. Then they'll go and try something that seems interesting without ever knowing that it had been done and was considered not valuable in the past.
- Even with the best intentions, scientists can introduce bias into their own work by running an experiment a few extra times "just to make sure it's interesting." By changing the rules in the middle of the experiment, you are not increasing sample size to increase scientific rigor. Instead you are increasing the chances that your experiment will find a positive result by chance.

Part II

Suppose an experimenter plans to collect data on a coin-flipping experiment based on a two-tier stopping criterion (assume the coin is a fair coin). The experimenter will collect an initial batch of data with N=30 and then do a null hypothesis significance test. If the result is not significant, then an additional 15 subjects' data will be collected, for a total of 45. Suppose the researcher intends to use the standard critical values for determining significance at both the N=30 and N=45 stages. Our goal is to determine the actual false alarm rate (the Type I error rate α) for this two-stage procedure, and to ponder what the mere intention of doing a second phase implies for interpreting the first stage, even if data collection stops with the first stage.

A *For N=30, what are the lower (z_{low}) and upper (z_{high}) limits of the 95th percentile confidence interval for z (z=number of heads)*

$$p(z \leq z_{low}|N = 30, \theta = 0.5) < 0.025$$
$$p(z \geq z_{high}|N = 30, \theta = 0.5) < 0.025$$

assuming a two-tailed Type I error rate of 0.05 or less?

```
z30Low <- qbinom(p = 0.025,
                 size = 30,
                 prob = 0.5
                 )-1
# Subtract one because you want less but not equal to the critical value and these are
# discrete observations.

z30High <- qbinom(p = 0.975,
                  size = 30,
                  prob = 0.5
                  )+1
# Add one for the same reasoning as subtracting one from zLow
```

The 95 percent confidence interval is (9, 21).

B *For N=45, what are the lower (z_{low}) and upper (z_{high}) limits of the 95th percentile confidence interval for z (z=number of heads)*

$$p(z \leq z_{low}|N = 45, \theta = 0.5) < 0.025$$
$$p(z \geq z_{high}|N = 45, \theta = 0.5) < 0.025$$

assuming a two-tailed Type I error rate of 0.05 or less?

```
z45Low <- qbinom(p = 0.025,
                 size = 45,
                 prob = 0.5
                 )-1

z45High <- qbinom(p = 0.975,
                  size = 45,
                  prob = 0.5
                  )+1
```

The 95 percent confidence interval is (15, 30).

For the next part of the exercise, consider the table provided.

	First 30 flips										
Second 15 flips	0	...	9	10	...	15	16	...	20	21	...	30
0	✠★		✠★	★		★	–		–	✠		✠★
1	✠★		✠★	★		–	–		–	✠		✠★
2	✠★		✠★	★		–	–		–	✠		✠★
3	✠★		✠★	★		–	–		–	✠		✠★
4	✠★		✠★	★		–	–		–	✠		✠★
5	✠★		✠★	★		–	–		–	✠		✠★
6	✠★		✠★	–		–	–		–	✠		✠★
7	✠★		✠	–		–	–		–	✠		✠★
8	✠★		✠	–		–	–		–	✠		✠★
9	✠★		✠	–		–	–		–	✠★		✠★
10	✠★		✠	–		–	–		★	✠★		✠★
11	✠★		✠	–		–	–		★	✠★		✠★
12	✠★		✠	–		–	–		★	✠★		✠★
13	✠★		✠	–		–	–		★	✠★		✠★
14	✠★		✠	–		–	★		★	✠★		✠★
15	✠★		✠	–		★	★		★	✠★		✠★

Each cell of the table corresponds to a certain outcome from the first 30 flips of a fair coin and a certain outcome from the second 15 flips of the same fair coin. A cell is marked by a dagger, ✠, if it has a result for the first 30 flips that would reject the null hypothesis. A cell is marked by a star, ★, if it has a result for the total of 45 flips that would reject the null hypothesis. For example, the cell with 10 heads from the first 30 flips and 1 head from the second 15 flips is marked with a ★ because the total number of heads for that cell, 10+1=11, is less than 15 (which is z_low for N=45 [a hint for part B!]). That cell has no dagger, ✠, because getting 10 heads in the first 30 flips is not extreme enough to reject the null. If neither the first 30 coin flips, nor the second 15 coin flips, would reject the null hypothesis of a fair coin, than the cell is marked with a dash -.

C *Denote the number of heads in the first 30 flips as z_1, and the number of heads in the second 15 flips as z_2. Explain why it is true that the z_1,z_2 cell of the table has a joint probability equal to* `dbinom(z1,30,0.5)*dbinom(z2,15,0.5)`

The probability of obtaining z_1 heads in the first thirty flips is independent of obtaining z_2 heads in the following fifteen. Therefore, the probability of obtaining z_1 and z_2 heads in forty-five flips is the joint probability of obtaining z_1 and z_2 heads in the first thirty and second fifteen, respectively. $P(z_1 \cap z_2) = P(z_1) \times P(z_2)$

`dbinom(z1,30,0.5)` produces the sum at and before z_1 under the curve of the PDF of a discrete probability distribution that describes thirty independent trials, or the equivalent of the probability of obtaining z_1 heads, $P(z_1)$. Likewise, `dbinom(z2,15,0.5)` produces the sum at and before z_2 under the curve of the PDF of a discrete probability distribution that describes fifteen independent trials, or the equivalent of the probability of obtaining z_2 heads, $P(z_2)$. Therefore, the product of `dbinom(z1,30,0.5)*dbinom(z2,15,0.5)` is equivalent to $P(z_1 \cap z_2) = P(z_1) \times P(z_2)$.

D *What is the sum of the probabilities of all the cells that contain a ✠ (whether or not it contains a ★)? Explain how you got your answer!*

```
probDagger <- 2*sum(dbinom(x = 0:z30Low,
                           size = 30,
                           prob = 0.5
                           )
                    )
```

`dbinom()` provides the probabilities of z_1 heads in thirty independent trials. The cells that contain a ✠ on the lower end of thirty trials is defined as 0 through 9, as calculated as `z30Low` in Part A and shown in the table. Passing a vector to `dbinom()` will produce a vector of probabilities, so `sum(dinom())` will add the discrete probabilities from 0 to 9. Since that is the lower tail, you then multiply by two to get the two-tailed probability, or to include the equivalent of `dbinom(x = z30High:30, size = 30, prob = 0.05))`.

Therefore, the sum of the probabilities of all cells that contain ✠ is 0.043.

E *What is the sum of the probabilities of all the cells that contain a ★ (whether or not it contains a ✠)? Explain how you got your answer!*

```
probStar <- 2*sum(dbinom(x = 0:z45Low,
                         size = 45,
                         prob = 0.5
                         )
                  )
```

This was calculated very similarly to Part D. `dbinom()` provides the probabilities of z_2 heads in forty-five independent trials. The cells that contain a ★ on the lower end of forty-five trials is defined as 0 through 15, stored as `z45Low` in Part B and shown in the table. Passing a vector to `dbinom()` will produce a vector of probabilities, so `sum(dinom())` will add the discrete probabilities from 0 to 15. Since that is the lower tail, you then multiply by two to get the two-tailed probability, or to include the equivalent of `dbinom(x = z45High:45, size = 45, prob = 0.05))`.

Therefore, the sum of the probabilities of all cells that contain ★ is 0.036.

F *What is the sum of the probabilities of all the cells that contain either a ✠ or a ★? (Note: This is the Type I error rate for the two-stage design, because these are all the ways you would decide to reject the null even when it is true.) Explain how you got your answer!*

See code notes for explanation.

```
# Get the probabilities of every potential outcome of the table
# Produces a 15 by 30 matrix with probabilities of the results of the first 30 trials
# followed by the next 15 trials. Probabilities correspond with the table of daggers
# and stars provided
probs <- matrix(data = outer(dbinom(0:15, 15, 0.5), # Prob of j heads in 15 trials
                             dbinom(0:30, 30, 0.5)  # Prob of i heads in 30 trials
                             ),
                nrow = 16,                          # Since 0 is not a valid matrix
                ncol = 31,                          # dimension call
                dimnames = list(seq(0, 15),         # Names for my sanity
                                seq(0, 30)
                                )
```

```
                )

# Create vector to catch the relevant p-values
probDashes <- c()

# Produce a vector of the probabilities corresponding with any experimental iteration
# that produced an insignificant number of heads in 45 trials

for (i in 10:20){           # For the insignificant # of heads in the first 30 trials
        for (j in 0:15){    # For the # of heads in the following 15 trials
                # Determines if the number of heads is significant for 45 trials
                # aka between and including 16 to 29
                if (16 <= (i + j) & (i + j) <= 29){
                        # Collect the probabilities of interest
                        probDashes <- append(probDashes,
                                             probs[j+1, i+1]
                                             )
                        # +1 to the prob dimension calls since a matrix cannot have a
                        # 0th dimension, 0 heads in is column 1 or row 1
                }
        }
}

# The complement of the sum of the probabilities of an insignificant number of heads is
# the sum of the probabilities of all the cells that contain either a dagger or a star
probStarDagger <- 1 - sum(probDashes)
```

The sum of the probabilities of all the cells that contain either a ✠ or a ★ is 0.061.

G *Suppose that the researcher intends to run an experiment using this two-stage stopping criterion. She collects the first 30 flips and finds 8 heads. She therefore rejects the null hypothesis and reports that $p < 0.05$. Is that correct? Explain.*

Since the increased Type I Error rate of a two stage experiment is 0.061, as calculated in Part F, then she cannot claim that $p < 0.05$. Her new Type I Error rate is what defines her critical value, α_c, which means $p < 0.061$ at the least.

H *Whenever we run an experiment and get a result that trends away from the null experiment, but isn't quite significant, it's natural to consider collecting more data. We saw in the previous part that even intending to collect more data, but not actually doing it, inflates the Type I error rate. Doesn't the fact that we always consider collecting more data mean that we always have a much higher Type I error rate than we pretend we do? Doesn't the actual Type I error rate of an experiment depend on the maximal number of data points we'd be willing to collect over the course of our lifetimes? In 1-2 paragraphs, discuss this conundrum and decide whether or not you think this poses a fundamental problem with null hypothesis testing.*

When collecting extra data after the experiment is planned, the experimenter inflates the Type I Error rate. The issue we then face is the inflation of Type I Error after every single follow up study or reassessment of the experiment. However, the desire to increase sample size, which in theory increases the power of our statistical analyses then compete with our desire to keep the Type I Error rate at our predetermined critical value, α_c. This definitely highlights an issue with null hypothesis testing that is rarely acknowledged by the academic community at large and results in the distribution of non-reproducible experiments as shown in the podcast in Part I. Most researchers, if they haven't considered this factor of their results, need to contend with the fact that they have definitely rejected null hypotheses that were true.

However, I would argue that in the event that you are designing an experiment, null hypothesis testing is not a completely undesirable option. As we've discussed in class, though Bayesian statistical methods would mitigate inflating Type I Error rates due to multiple comparisons, they are computationally more difficult to perform and to understand. In order to mitigate the effects of multiple end points or intermittent analyses, one should set α_c ahead of time. Then, in the event that the null is not rejected, a power analysis can be performed to determine if the experiment bears repeating. This is of course assuming that you did not calculate the sample size you need for a sufficient statistical power ahead of time or if there are no other constraints on sample size (time, money, etc.). This methodology would use the "Gold Standard" of $\alpha_c = 0.05$ correctly. Clearly this does not account for other ways that multiple comparisons can sneak into your experimental design, but it does directly counter on of the most common and understandable mistakes made by scientists.