Class size paradox

```
library(tidyverse)
county_complete <- read_rds("county_complete.rds")
```

The following example assumes you have already read the standard reading on *Probability mass functions*.

Let's consider another PMF computation that illustrates something that we may call the "class size paradox".

At many American colleges and universities, the student-to-faculty ratio is about 10:1. But students are often surprised to discover that their average class size is bigger than 10. There are two reasons for the discrepancy:

- Students typically take 4–5 classes per semester, but professors often teach 1 or 2.
- The number of students who enjoy a small class is small, but the number of students in a large class is (unsurprisingly) large.

The first effect is obvious, at least once it is pointed out; the second is more subtle. Let's look at an example. Suppose that a college offers 65 classes in a given semester, with the following distribution of sizes:

size	count
5–9	8
10–14	8
15–19	14
20–24	4
25–29	6
30–34	12
35–39	8
40–44	3
45–49	2

If you ask the Dean for the average class size, he or she would construct a PMF, compute the mean, and report that the average class size is 23.7. Here's the code:

```
class_sizes <- tribble(
  ~`class size`, ~count,
  7, 8,
  12, 8,
  17, 14,
  22, 4,
  27, 6,
  32, 12,
  37, 8,
  42, 3,
  47, 2)

class_sizes %>%
  summarize(`Average class size` = round(weighted.mean(`class size`, count), 1))
```

Average class size
23.7

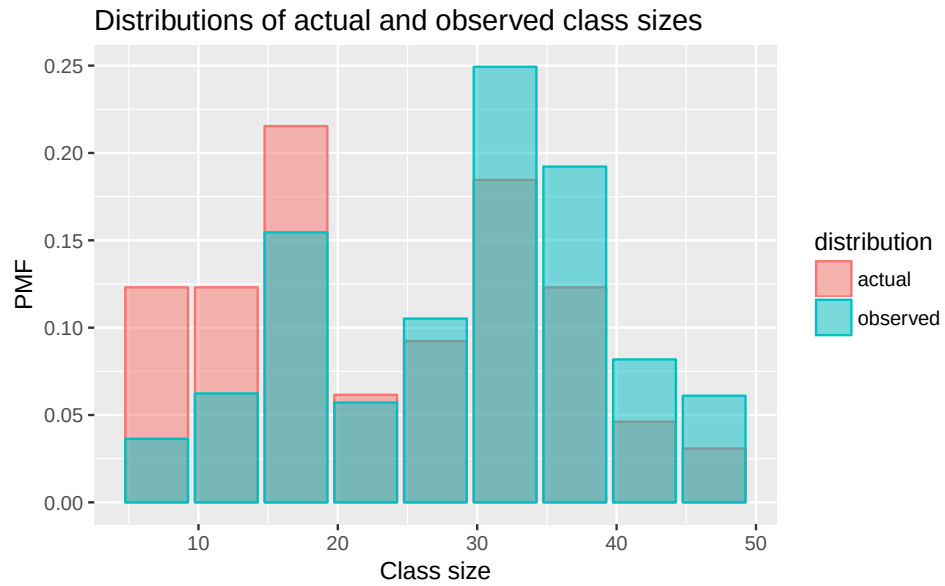
But if you survey a group of students, ask them how many students are in their classes, and compute the mean, you would think the average class was bigger. Let's see how much bigger.

First, let's compute the distribution as observed by students, where the probability associated with each class size is "biased" by the number of students in the class. For each class size we multiply the probability by the number of students who observe that class size. The result is a new PMF that represents the biased distribution:

```
class_sizes2 <- class_sizes %>%
  mutate(actual = count / sum(count)) %>%
  mutate(observed = (actual * `class size`) / sum(actual * `class size`)) %>%
  gather(key = distribution, value = PMF, actual:observed)
```

Now we can plot the actual and observed distributions together:

```
class_sizes2 %>%
  ggplot(
    mapping = aes(x = `class size`, y = PMF, fill = distribution,
                  color = distribution)) +
  geom_col(position = "identity", alpha = 0.5)
```



As we see in the above figure, the biased distribution corresponds to fewer small classes and more large ones. The mean of the biased distribution is,

```
class_sizes2 %>%
  group_by(distribution) %>%
  summarize(`Average class size` = round(weighted.mean(`class size`, `PMF`), 1))
```

distribution	Average class size
actual	23.7
observed	29.1

which is almost 25% higher than the actual mean.

It is also possible to invert this operation. Suppose you want to find the distribution of class sizes at a college, but you can't get reliable data from the Dean. An alternative is to choose a random sample of students and ask how many students are in their classes. The result would be biased for the reasons we've just seen, but you can use it to estimate the actual distribution. Here's the code that can be used to unbias a PMF:

```
class_sizes2 %>%
  spread(key = distribution, value = PMF) %>%
  mutate(
    unbiased = (observed * 1 / `class size`) /
      sum(observed * 1 / `class size`)) %>%
  select(`class size`, count, observed, unbiased, actual)
```

class size	count	observed	unbiased	actual
7	8	0.0363636	0.1230769	0.1230769
12	8	0.0623377	0.1230769	0.1230769
17	14	0.1545455	0.2153846	0.2153846
22	4	0.0571429	0.0615385	0.0615385
27	6	0.1051948	0.0923077	0.0923077
32	12	0.2493506	0.1846154	0.1846154
37	8	0.1922078	0.1230769	0.1230769
42	3	0.0818182	0.0461538	0.0461538
47	2	0.0610390	0.0307692	0.0307692

It's similar to before; the only difference is that it divides each probability by the class size instead of multiplying.

Credits

This work, *Class size paradox*, is a derivative of Allen B. Downey, "Chapter 3 Probability mass functions" in *Think Stats: Exploratory Data Analysis*, 2nd ed. (O'Reilly Media, Sebastopol, CA, 2014), used under CC BY-NC-SA 4.0. *Class Size Paradox* is licensed under CC BY-NC-SA 4.0 by James Glasbrenner.