1 Codebreaking

Continued from previous section.

1.1 Interlude: G-Test

Suppose that every registered voters in an imaginary county in the United States is classified into the mutually exclusive and exhaustive racial groups "White," "Black," "Hispanic," and "Other." Suppose that, by inspecting the voter rolls, we find that the racial distribution of this county is

	White	Black	Hispanic	Other	Total
Distribution	72%	7%	12%	9%	100%

Since jurors are supposed to be drawn from the list of registered voters, we might hope that a random sample of jurors would follow this same racial distribution. Suppose we sample 275 jurors and observe the racial distribution displayed in the second row:

	White	Black	Hispanic	Other	Total
Distribution	72%	7%	12%	9%	100%
Observed	210	10	20	35	275
Expected	198	19.25	33	24.75	275

Now, if our random sample of jurors followed the overall racial distribution of registered voters, we would expect that 72% of them would be White, which would be $0.72 \cdot 275 = 198$ people. We can calculate the expected numbers of jurors in the other groups similarly to fill in the third row above.

Note that we can, and should, expect *some* deviation from the expected counts. Remembering that our categories are mutually exclusive, so we could not possibly observe a sample of 19.25 Black jurors. However, if we had expected something like 198 White jurors, 19 Black jurors, 33 Hispanic jurors, and 25 Other jurors – or something close to that – we probably would not be surprised with our results.

Stated differently, *the data we collected would feel consistent with the hypothesis that the racial distribution of jurors matches the racial distribution of the electorate.* However, what we observed was pretty far from the expected counts. **How do we quantify and make sense of this observation?**

1.1.1 The G-Test

The idea is to introduce a number that measures the difference between the observed and expected rows. There are a variety of numbers that can be used, but let us consider one that is often denoted G. It is defined as follows:

Definition 1.1

Suppose X is a random variable with finitely many values $a_1, \dots, a_n$ and let $p_i = \mathbb{P}[X = a_i]$. Suppose we make N observations of the values $a_1, \dots, a_n$ and that O_i is the number of observations of a_i that we made. Let $E_i = Np_i$ and then define

$$G = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right).$$

If $O_i = 0$ for some i, we set the corresponding summand $O_i\left(\frac{O_i}{E_i}\right) = 0$. If there exists an i such that $E_i = 0$ but $O_i \neq 0$, set $G = \infty$.

(Example.) Consider the motivating example with the voters. Define

- The random variable X represents observing the race of a randomly drawn voter from our county. It has 4 possible values (White, Black, Hispanic, Other), so $n = 4$.
- The values p_i are the percentages of the electorate in each racial group.
- The values O_i are the observed counts.
- The values E_i are the expected counts.
- For fun, $N = 275$ (we have 275 total observations across a_1, a_2, a_3, a_4).

Then,

$$G = 2\left(210\ln\left(\frac{210}{198}\right) + 10\ln\left(\frac{10}{19.25}\right) + 20\ln\left(\frac{20}{33}\right) + 35\ln\left(\frac{35}{24.75}\right)\right) \approx 15.84.$$

Remark: If you're inclined to see why $E_i = Np_i$, note that $N = 275$ (that's the number of observations of all the values) and p_i is the percent of the electorate in the racial group i. So, for Black, $E = 275 \cdot 0.07 = 19.25$.

Theorem 1.1: Gibbs' Inequality

We always have $G \geq 0$. Moreover, $G = 0$ if and only if $O_i = E_i$ for all $i = 1, \ldots, n$.

The question is simply, how big is "big?" In particular, in our example, can we say 15.84 is a "*big*" value of G? The answer to this question is provided by the following theorem, which we'll state slightly imprecisely and explain in a bit more detail later.

Theorem 1.2: Wilks' Theorem

Suppose the N observations of the values $a_1, \ldots, a_n$ that we make are in fact independent observations of the random variable X. For large values of N, the values of G are well-approximated by a chi-square distribution with $n - 1$ degrees of freedom.

There are several points of explanation to make.

- A "chi-square distribution with k degrees of freedom" is a certain function f_k defined on $[0, \infty)$ and taking non-negative values everywhere with total integral equal to 1. In other words, we have $f_k(x) \geq 0$ for all $x \geq 0$ and
$$\int_0^\infty f_k(x)dx = 1.$$
The formula for $f_k(x)$ is complicated and also unimportant for our purposes.
- To say that "the values of G are well-approximated by a chi-square distribution with $n - 1$ degrees of freedom" is to say that, for any (not necessarily finite) interval (a, b), the probability that G lands inside the interval (a, b) is approximately
$$\int_a^b f_k(x)dx.$$

(Example.) Notice that

$$\int_0^{15.84} f_3(x)dx \approx 0.999.$$

It follows that

$$\int_{15.84}^{\infty} f_3(x)dx = 1 - \int_0^{15.84} f_3(x) \approx 1 - 0.999 = 0.001.$$

The number 0.001 is our p-value, and it means that the probability of observing a value of G that is bigger than 15.84 is only about 0.1%. That is quite a small probability, so our calculation suggests that the value of G that we saw is in fact quite large.

Stated differently, with a p-value of 0.001, this indicates that if jurors in this county were truly representative of the county's electorate, there would only be roughly a 0.1% chance of seeing a sample that deviated at least as much from the expected counts as the data that we saw. Because that's such a small probability, this suggests that it's very unlikely that our sample of jurors is actually representative of the county's electorate. We have quantified the observation we made informally above.

- Another thing to look at is "for large values of N." In particular, that this theorem would only work for large values of N. Was $N = 275$ large enough to justify what we did? The answer *depends* on how well you want the values of G to be approximated by a chi-square distribution. The better an approximation you want, the higher a value of N you need. That being said, the following heuristic generally works well.

Theorem 1.3: Heuristic Addendum to Wilks' Theorem

The approximation of H by a chi-square distribution with $n-1$ degrees of freedom is "good enough" as long as the vast majority of the expected counts $E_1, \dots, E_n$ are all at least 5.

Because all of our expected counts are well above 5, we do not need to worry.

The process (computing expected counts, finding an observed value of G, using a chi-square approximation to find a p-value, i.e., the probability of observing a larger value of G than what we observed if the observations do in fact come from the theoretical distribution) is called a **G-test**. It's a useful technique for a lot of problems in statistics and can be used in codebreaking.

(Exercise.) A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester she asks her students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online. How well does this data fit the professor's predictions? Run a G-test to find out!

Similar to the introduction of this section, we can create a table.

	Hard Copy	Web	Online	Total
Distribution	60%	25%	15%	
Observed	71	30	25	126
Expected	75.6	31.5	18.9	126

Note that

- X represents observing whether a person reads from a hard copy, web, or online. Therefore, $n = 3$.
- The values p_i represents the percentages that a person chooses to either purchase a hard copy, or print it out, or read it online.
- The values O_i are the observed counts.
- The values E_i are the expected counts.

Calculating the G value, we have

$$G = 2\left(71\ln\left(\frac{71}{75.6}\right) + 30\ln\left(\frac{30}{31.5}\right) + 25\ln\left(\frac{25}{18.9}\right)\right) \approx 2.14403.$$

We now want to see what the probability is of observing a value of G that is bigger than 2.14403. To do this, note that

$$\int_{2.14403}^{\infty} f_2(x) = 1 - \int_0^{2.14403} f_2(x) \approx 1 - 0.65768194886549 = 0.342318.$$

So, 0.342318 is our p-value, and it follows that the probability that we find a higher G value is about 34.23%. In other words, there would be a 34.23% chance of seeing a sample that deviated as least as much from the expected counts as the data we just saw.