Hypothesis testing and chi-squared tests

Kenneth A. Ribet

UC Berkeley

Math 10B
February 32, 2016

Meals

We had a dinner last night at the Chengdu Style Restaurant

. . . and a breakfast this morning at the Faculty Club.

There are still a few places available for the March 18 breakfast at 8AM. Sign-ups by email.

Office hours

Monday 2:10–3:10 and Thursday 10:30–11:30 in Evans

Tuesday 10:30–noon at the SLC

Regrade requests

TOMORROW IS THE LAST DAY TO REQUEST A REGRADE ON MT #1

Chi-squared for idiots like me

As you'll see very quickly, I don't know this subject at all well. I learned a little bit about χ^2 over the last 48 hours. A good source is `https://en.wikipedia.org/wiki/Chi-squared_distribution`.

Suppose that we have a coin and suspect that it comes up `H` with probability p ($0 < p < 1$).

The null hypothesis is that the coin is "biased" with probabilities p (for heads) and $q = 1 - p$ (for tails). (If $p = q = 1/2$, the coin is actually fair.

How would we test the null hypothesis?

The only thing to do is to flip the coin lots of times, say N times, and to record the results. We expect to get around pN heads, and suppose that we actually get m heads. If m is *too far* from pN, then we will end up rejecting the null hypothesis (to the effect that the coin is biased with probability p).

What is meant by “too far”?

The central limit theorem suggests that we look at the random variable

$$Z := \frac{(\overline{X} - \mu)\sqrt{N}}{\sigma} = \frac{\overline{X}N - \mu N}{\sigma\sqrt{N}}.$$

Here $\overline{X}$ is the average number of heads when we toss the coin N times and μ is the expected value of the average, namely p. Also, recall that $\sigma^2 = pq$, where $q = 1 - p$ is the suspected probability of getting `T`. Then Z is supposed to be distributed like a standard normal variable, and Z^2 is distributed like the square of a standard normal variable. It's obviously smart to take the square because σ involves a square root.

Note that $\overline{X}N$ is the number of ("observed") heads that we get, while $\mu N = pN$ is the number of "expected" heads.

Thus Z^2 (which is a random variable) measures this quantity:

$$\frac{(\text{number of observed heads} - \text{number of expected heads})^2}{Npq}.$$

With the obvious notation, we can write this fraction more compactly as

$$\frac{(O_H - E_H)^2}{Npq}.$$

Somewhat amazingly, a short calculation shows that this fraction can be written symmetrically as the sum

$$\frac{(O_H - E_H)^2}{E_H} + \frac{(O_T - E_T)^2}{E_T},$$

where the subscript T refers to *tails* (in place of heads).

Although we are counting two quantities (heads and tails), they are linked by a constraint, namely that their sum is given. (The sum is the number of tosses.) Hence there is only one *degree of freedom*, even though there are two quantities being tallied.

If we had a six-sided die and were tallying separately the number of observed 1s, 2s, and so on, there would be six tallies but only five degrees of freedom. (The total number of observed die tops would be the number of rolls.)

The variable for k quantities being tallied is

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

The anticipated distribution of this sum depends on the number of degrees of freedom.

For our original problem,

$$\frac{(O_H - E_H)^2}{E_H} + \frac{(O_T - E_T)^2}{E_T}$$

is supposed to be distributed like the square of a normal variable. The probability that it is “large” will be small. If the probability is too small, then we become unconvinced by the hypothesis that the coin is biased with probability p. Statisticians say that we can reject the hypothesis with increased confidence.

In statistics, there are "p-values." This is a bit unfortunate because p was our alleged probability of getting $\mathtt{H}$. Anyway, the online "textbook" (and everyone else) says:

> *The p-value associated with an possible outcome r is the probability that the test statistic is $\geq r$, assuming the correctness of the null hypothesis H_0.*

Note that large r-values lead to small p-values. The magic p-value (traditionally) is 0.05. For $p \leq 0.05$, we reject H_0.

For $p \geq 0.05$, we sidle up to H_0, but it is incorrect to say that we *accept* it. All I can say is that we do not reject it.

For one degree of freedom, the magic value of the sum is $r = 3.84$ (Wikipedia). We can see this number in action by consulting any number of online web pages. For example, consider

`https://www.colby.edu/biology/BI17x/freq.html`.

The author's null hypothesis is that a coin is fair. (Thus $p = 1/2$ in the original sense of p.) He imagines tossing his coin (or her coin—I have no idea who wrote this page) 200 times. Thus the expected number of heads is 100; same for tails. Now imagine that we observe 108 heads and 92 tails; this is not far off from expectations. The sum

$$\frac{(O_H - E_H)^2}{E_H} + \frac{(O_T - E_T)^2}{E_T}$$

is $\frac{64}{100} + \frac{64}{100} = 1.28$; this is our r-value. The associated p-value is 0.257899. In other words, there's a 26% chance that the coin is fair.

The 26% is not very compelling, but it's a lot larger than 5%. Accordingly, we cannot reject the null hypothesis.

The author goes on to write

> *Because the chi-squared value we obtained in the coin example is greater than 0.05 (0.27 to be precise), we accept the null hypothesis as true and conclude that our coin is fair.*

Just say "no" to this sort of muddy thinking. We do *not* conclude that the coin is fair.

Just to be grumpy about a small point: the "0.27 to be precise" is completely imprecise. The author got to 27% by a crude interpolation. Nowadays, one can just use software to get a value that's as precise as one likes.

Now here's what's really wrong with the author's example. (My wife made this point to me yesterday.) Suppose you had been asked this question two weeks ago:

> *A fair coin is tossed 200 times. What is the probability that heads comes up 108 or more times?*

You would have eaten this question for breakfast. It's a binomial problem. The answer is

$$\frac{1}{2^{200}}\left(\binom{200}{108}+\binom{200}{109}+\cdots+\binom{200}{200}\right).$$

Fifty years ago, this quantity would have been very hard to evaluate in normal practice. Right before lecture, I evaluated it in Sage and got 0.1444. Thus the observed number of heads is significantly less probable than the χ^2 test would indicate.

As several people pointed out in lecture, one should double the 1.444 to get the same probability that's covered by χ^2. That's because χ^2 detects both high and low numbers of heads.

Most treatments of χ^2 discuss *N* rolls of a 6-sided die. The hypothesis to be tested is that the die is fair, meaning that the expected numbers of 1s, 2s and so on are all $N/6$. Wikipedia: Suppose $N = 60$, so $N/6 = 10$. Suppose that the observed frequences are respectively: 5, 8, 9, 8, 10 and 20. Then

$$r = \sum_{i=1}^{6} \frac{(O_i - 10)^2}{10} = 13.4.$$

There are five degrees of freedom. The probability of having $r \geq 13.4$ is $0.019905 < 0.5$. Hence we "reject the null hypothesis" and conclude that the die is biased "at 95% significance level" (because $95 = 100 - 5$).

In Sage, I typed:

```
T = RealDistribution('chisquared', 5)
1-T.cum_distribution_function(13.4)
```

and got the number 0.0199052203347744376 instantaneously.

I found out what to type by conducting a google search, which led me to `http://doc.sagemath.org/html/en/reference/probability/sage/gsl/probability_distribution.html`.

The "textbook" directs you to `http://www.stat.berkeley.edu/~stark/Java/Html/chiHiLite.htm`, which presumably does fine after you figure out how to use it.

Point estimates

When dealing with a random variable that has unkown parameters, we can simple values of the variable and come up with estimates of the mean, variance and standard deviation of the variable. Recall that the true mean is μ; our estimate for the mean is

$$\hat{\mu} := \frac{1}{n}\sum_{k=1}^{n} x_k,$$

where the x_k are the observed values of X.

Similarly, we write $\widehat{\mathrm{Var}}$ for our estimate for the variance of X.

The *standard error* is $\mathrm{SE} := \sqrt{\widehat{\mathrm{Var}}}$.