Hypothesis testing and *p*-values

Kenneth A. Ribet

Math 10A
November 22, 2016

Yesterday's pop-in lunch

Announcements

Happy Thanksgiving! No classes Wednesday–Friday.

Breakfast Thursday, December 1 at 9AM.

Breakfast Monday, December 5 at 9AM.

Sign up for breakfasts by sending me email.

Course evaluations

Evaluations for your course(s) have opened to students.

Although students received an invitation email and reminders along the way, previous research demonstrates that a personal reminder from the instructor and an explanation of how evaluations are used to inform your teaching can make a positive impact on response rate and quality.

Please write interesting things when you do the course evaluations.

I will read your comments!!

Prelude: PDF of X^2

Here's an exercise like two that were on the homework: Suppose that X is a continuous random variable with PDF equal to $f(x)$. What is the PDF of X^2?

Let $F(x)$ be the CDF of X; let $g(x)$ and $G(x)$ be the PDF and CDF of X^2.

Then for real numbers d,

$$G(d) = P(X^2 < d).$$

Clearly $G(d) = 0$ for $d \leq 0$, so we are interested only in the values of G and g on *positive* real numbers.

Take $x > 0$. Then

$$G(x^2) = P(X^2 < x^2) = P(-x < X < x) = F(x) - F(-x).$$

Differentiate the equation $G(x^2) = F(x) - F(-x)$ with respect to x, using that $G' = g$ and $F' = f$:

$$2xg(x^2) = f(x) + f(-x).$$

Hence

$$g(x^2) = \frac{f(x) + f(-x)}{2x}, \quad g(t) = \frac{f(\sqrt{t}) + f(-\sqrt{t})}{2\sqrt{t}}.$$

When f is an *even* function,

$$g(t) = \frac{f(\sqrt{t})}{\sqrt{t}}.$$

What is the probability of getting 600 or more heads if you flip a fair coin 1000 times?

The number of outcomes if you flip a coin 1000 times is 2^{1000}, which has 302 digits. The number of ways to get 600 heads in 1000 flips is $\binom{1000}{600}$, a 291-digit number that is more precisely

49652723862542288611507356288962313 26

21341353659827604662932184012645905

732096457382164964136575507417172339 04

20897787519048878570924119105790 77

41240853994820497412977839043739395 4251

67680052468065347826666236435 2619

24418093115402070111198232800077698030 59

55525649501369943202079996789539150.

If the coin is fair, the probability of getting 600 or more heads in 1000 flips is

$$\frac{1}{2^{1000}} \sum_{i=600}^{1000} \binom{1000}{i} \approx 1.364232 \times 10^{-10}.$$

This is a minuscule number! Even the probability of getting 530 or more heads in 1000 tosses is rather small: it's 0.0310.

Another data point: the probability of getting 527 or more heads in 1000 tosses is 0.0468, which is a bit less than 5%.

Now let's change the story. Suppose that someone hands you a coin and asks you if you think it's fair. You flip it 1000 times (or build a machine that flips coins. . .) and observe the results.

- If you get 600 heads, you'll probably conclude that the coin is highly unlikely to be fair—you'll think it's biased toward heads.
- If you get 530 heads, you'll probably think the coin is biased.
- If you're a professional statistician and you get 527 heads, you will reject the hypothesis that the coin is fair (and ask to be paid a consulting fee).
- If you're a professional statistician and get 526 heads, you'll calculate that this is an event that occurs with probability 0.0534 for fair coins and refuse to reject the hypothesis that the coin is fair. You'll ask for a consulting fee.

Now let's change the story. Suppose that someone hands you a coin and asks you if you think it's fair. You flip it 1000 times (or build a machine that flips coins. . .) and observe the results.

- If you get 600 heads, you'll probably conclude that the coin is highly unlikely to be fair—you'll think it's biased toward heads.
- If you get 530 heads, you'll probably think the coin is biased.
- If you're a professional statistician and you get 527 heads, you will reject the hypothesis that the coin is fair (and ask to be paid a consulting fee).
- If you're a professional statistician and get 526 heads, you'll calculate that this is an event that occurs with probability 0.0534 for fair coins and refuse to reject the hypothesis that the coin is fair. You'll ask for a consulting fee.

In this story, the hypothesis that the coin is fair is called the **null hypothesis** (nothing going on). The probabilities that we calculated are called **p-values**.

A *p*-value is a probability calculated under the assumption that the null hypothesis is true. It’s the probability of getting an experimental outcome that is either your outcome or something even more extreme.

Adopted from my March 3, 2016 10B lecture

In statistics, there are "*p*-values. . . ." . . . 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 typically lead to small *p*-values. (Getting 540 heads is less likely than getting 530 heads.) The magic *p*-value (traditionally) is 0.05. For $p \leq 0.05$, we reject H_0.

For $p \geq 0.05$, we warm up to H_0, but it is incorrect to say that we *accept* it. The correct statement is that we cannot reject it.

The Math 10B breakfast on March 3, 2016

The American Statistical Association's Statement on *p*-values:

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Generalization

Suppose that we have a coin and suspect that it comes up $\mathtt{H}$ 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 fair.)

The "p" here is not the same "p" as in "p-values," but that's probably not the end of the world.

To test the null hypothesis, we flip the coin N times (N big), recording the results. We imagine getting around pN heads; 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).

We need to see what is meant by “too far.”

If m is bigger than or equal to pN, we could calculate the actual probability of getting m or more heads, as we did with the first example.

If $m < Np$, the analogue would be to calculate the probability of getting m or fewer heads.

This is great when it is possible. A computational problem could arise if N is enormous. We had $N = 1000$ in the first example. Even if we had a yuge computer, we could have trouble if N were 1,000,000,000 or something.

Here's another sort of issue. You're handed a die and want to test the hypothesis that it's fair. You roll it 600 times and get the six different faces with frequencies

$$82, \quad 103, \quad 95, \quad 88, \quad 102, \quad 130.$$

Does this disturb you? Do you reject the null hypothesis? What computation do you need to make?

The plan is that you'll find out next semester.

Back to coins: let's try the Central Limit Theorem.

Introduce the 0–1 variable X attached to a single flip of the coin. It has expected value $\mu = p$ and variance $p - p^2 = p(1-p) = pq$. As usual, let $X_1, X_2, \ldots$ be copies of X that are indexed by the individual flips and let $\overline{X} = \overline{X}_N$ be the average of the X_i.

Then $\overline{X}$ has mean p and standard deviation $\dfrac{\sigma}{\sqrt{N}}$, where $\sigma = \sqrt{pq}$ is the standard deviation of X.

The random variable

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

is rigged up to have mean 0 and standard deviation 1. It's like the Gaussian in those respects.

The virtue of writing the numerator of Z as $\overline{X}N - \mu N$ is that $\overline{X}N = m$ is the number of heads that we actually observed, whereas $\mu N = pN$ is the number of heads that we were anticipating.

To check that we're doing the right sort of thing, note that if $p = \dfrac{1}{2}$, $N = 100$ and $m = 70$, the fraction is $\dfrac{70 - 50}{\frac{1}{2} \cdot 10} = 4$. This number should be familiar from the HW due yesterday.

You can say that the calculation here is nice mainly because of the accident that the variance $pq = \dfrac{1}{2} \cdot \dfrac{1}{2}$ has a neat-looking square root, namely $\dfrac{1}{2}$, and that N has a nice square root, namely 10.

In general, it might be nicer to look at Z^2, which doesn't involve square roots.

The random variable Z^2 is then easy to write down:

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

With the obvious notation, we can write this fraction symbolically as

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

(So "O" is for "observed" and "E" for "expected," i.e., anticipated.)

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).

To do the calculation, it's useful to know that $\frac{1}{p} + \frac{1}{q} = \frac{1}{pq}$.

A big advantage of Z^2 over Z is that we never have to worry whether Z^2 is positive or negative. It's always positive (or 0). When the head/tail count is wacky, Z^2 is big regardless of whether there are too many heads or too few heads.

A summary of the situation: if $Z^2 = 0$, there were exactly as many heads as we anticipated and we won't reject the null hypothesis. If Z^2 is too big, we'll reject the null hypothesis.

The question is what do we mean by "too big." Of course, the answer is "having probability less than 0.05," but for this to be helpful we need to know how Z^2 is distributed.

Numbers

For reference, in the situation with $p = q = \frac{1}{2}$, $N = 100$, $O_H = 70$,

$$Z^2 = \frac{(O_H - E_H)^2}{Npq} = \frac{400}{25} = 16.$$

If $p = q = \frac{1}{2}$, $N = 1000$, $O_H = 600$,

$$Z^2 = \frac{100^2}{250} = 40.$$

Chi squared

A secret: Z^2 is the simplest example of a Chi-squared distribution. Its PDF is the function $f(x)$ that's 0 for $x \leq 0$ and is given by the formula

$$\frac{e^{-x/2}}{\sqrt{x}\sqrt{2\pi}}$$

for $x > 0$. Note that the PDF approaches ∞ for $x \downarrow 0$ because of the $\sqrt{x}$ in the denominator.

It is not too hard to calculate the PDF. Shall we do that? I want to, but you're going to vote me down, aren't you? The derivation is available in Wikipedia, of course. [Addendum: see the prelude. We did it!]

You can find various applets online that output probabilities (p-values) when you input the statistic Z^2.

HAPPY THANKSGIVING!