Hypothesis testing and *p*-values

Kenneth A. Ribet

Math 10A
November 21, 2017

Announcements

Happy Thanksgiving!
No classes Wednesday–Friday.

Office hours
November 27, 29

Last classes
November 28, 30

RRR Week December 4–8

This class will meet in this room on December 5, 7 for structured reviews by T. Zhu and crew.

Final exam
Thursday evening,
December 14,
7–10PM.

Breakfast Friday, December 1 at 8AM. Slots are still available—sign up!

Pop-in lunch on Friday, December 1 at High Noon (Faculty Club).

Little-known fact: my first Faculty Club student breakfast was held on February 1, 2013. It was for Math 55. I intend to organize a breakfast on February 1, 2018.

Coins, again

Imagine we flip a coin 1000 times.

What are the probabilities of getting 527 or more heads? 530 or more heads? 600 or more heads?

Using technology

The number of outcomes for 1000 flips is $2^{1000} \approx 1.07 \times 10^{301}$.

The probability of getting exactly 527 heads is $\frac{1}{2^{1000}} \binom{1000}{527}$,

where $\binom{1000}{527}$ is a Math 10B-style ratio of factorials:

$$\binom{1000}{527} = \frac{1000!}{527!473!} \approx 6.29 \times 10^{298}.$$

Thus the probability of getting exactly 527 heads is 0.00587, or around 0.6%.

The probability of getting *527 or more heads* is the sum of 474 numbers like this. The sum is 0.0468, which is a bit less than 5%.

Replacing 527 by 530, we get the smaller number 0.0310. (We've subtracted three numbers that are each around $0.5\% or 0.6\%$.)

The probability of getting 600 or more heads is way less than either of these numbers; it's around $\approx 1.364232 \times 10^{-10}$.

Now here's the main point. Suppose that one of the students hands me a coin and asks me whether or not it's a fair coin. Let's say that I flip it 1000 times and get 600 heads.

If I'm a professional statistician, I say to myself: "Getting 600 or more heads would have a probability of roughly $\approx 1.364232 \times 10^{-10}$ if this coin were fair." That's so low that I will proclaim:

"This coin is not fair!!"

In technical terms, I will have rejected the hypothesis that the coin is fair.

The hypothesis that the coin is fair is called the *null hypothesis*. The word “null” means “nothing special going on—this is a generic situation.”

Statisticians reject null hypotheses.

A new beginning

Suppose I flip a coin 1000 times and get exactly 500 heads. What do I do?

I may say to myself: “Wow, this coin seems to be pretty fair” or perhaps “How remarkable that I got exactly the same number of heads as tails.”

I do not *accept* the hypothesis that the coin is fair.

I fail to reject it.

In essence, I do nothing.

A new beginning

Suppose I flip a coin 1000 times and get exactly 500 heads. What do I do?

I may say to myself: “Wow, this coin seems to be pretty fair” or perhaps “How remarkable that I got exactly the same number of heads as tails.”

I do not *accept* the hypothesis that the coin is fair.

I fail to reject it.

In essence, I do nothing.

A new beginning

Suppose I flip a coin 1000 times and get exactly 500 heads. What do I do?

I may say to myself: “Wow, this coin seems to be pretty fair” or perhaps “How remarkable that I got exactly the same number of heads as tails.”

I do not *accept* the hypothesis that the coin is fair.

I fail to reject it.

In essence, I do nothing.

A new beginning

Suppose I flip a coin 1000 times and get exactly 500 heads. What do I do?

I may say to myself: “Wow, this coin seems to be pretty fair” or perhaps “How remarkable that I got exactly the same number of heads as tails.”

I do not *accept* the hypothesis that the coin is fair.

I fail to reject it.

In essence, I do nothing.

Statisticians

They reject hypotheses that assign low probabilities to observed results.

They do not *accept* hypotheses that are compatible with observed results.

Other hypotheses might also be compatible with the observed results.

Statisticians

They reject hypotheses that assign low probabilities to observed results.

They do not *accept* hypotheses that are compatible with observed results.

Other hypotheses might also be compatible with the observed results.

p-values

Statisticians assume the null hypothesis and estimate the probability of getting the observed result or *something more extreme*. This probability is called a **p-value**.

In the story with 1000 flips and 600 observed heads, it is standard to calculate the probability of getting either

600 or more heads *or* 400 or fewer heads.

This is the two-sided view of what is equally as extreme or more extreme than getting exactly 600 heads.

The probability in this case would be twice the probability $\approx 1.364232 \times 10^{-10}$ that was calculated before.

When do we reject the null hypothesis?

We reject the null hypothesis when the *p*-value is too small.

When is the *p*-value too small?

It is traditional in statistics that the magic *p*-value is 0.05: "When $p \leq 0.05$, we reject H_0."

Where does the magic value 0.05 come from? It was invented by someone long ago as a convenient-looking small number.

> *The use of the p-value in statistics was popularized by Ronald Fisher, and it plays a central role in his approach to the subject. In his influential book Statistical Methods for Research Workers (1925), Fisher proposes the level* $p = 0.05$*, or a 1 in 20 chance of being exceeded by chance, as a limit for statistical significance, and applies this to a normal distribution (as a two-tailed test), thus yielding the rule of two standard deviations (on a normal distribution) for statistical significance. . . .*

When is the *p*-value too small?

It is traditional in statistics that the magic *p*-value is 0.05: "When $p \leq 0.05$, we reject H_0."

Where does the magic value 0.05 come from? It was invented by someone long ago as a convenient-looking small number.

> *The use of the p-value in statistics was popularized by Ronald Fisher, and it plays a central role in his approach to the subject. In his influential book Statistical Methods for Research Workers (1925), Fisher proposes the level* $p = 0.05$*, or a 1 in 20 chance of being exceeded by chance, as a limit for statistical significance, and applies this to a normal distribution (as a two-tailed test), thus yielding the rule of two standard deviations (on a normal distribution) for statistical significance. . . .*

According to Table 7.3 on page 566, the probability of being 2 standard deviations to the right of 0 in the standard normal distribution is 0.0228. Thus the probability of being 2 standard deviations from 0 (either to the right or to the left) is about 2×0.0228 or 0.05 (actually more like 0.0456).

The neat circumstance that the probability associated with a round number (2.0) of standard deviations from the mean is a round-looking number (5%) is the driver of the whole story.

Sad!

According to Table 7.3 on page 566, the probability of being 2 standard deviations to the right of 0 in the standard normal distribution is 0.0228. Thus the probability of being 2 standard deviations from 0 (either to the right or to the left) is about 2×0.0228 or 0.05 (actually more like 0.0456).

The neat circumstance that the probability associated with a round number (2.0) of standard deviations from the mean is a round-looking number (5%) is the driver of the whole story.

Sad!

The rest is history. The scientific literature is full of proclamations that something significant has been found because a *p*-value is less than the magic value.

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.

A marginal case

Suppose that we flip a coin 1000 times and obtain 527 heads. Do we reject the null hypothesis that the coin is fair?

The probability of getting 527 or more heads is around 0.0468, which is less than the magic number 0.05.

Do you conclude that something signficiant has been discovered (i.e., that the coin is biased)?

I'd prefer not to draw that particular conclusion. Instead I'll go to the next slide.

A 2016 homework question

A random sample of 1562 undergraduates enrolled in marketing courses was asked to respond on a scale from 1 (strongly disagree) to 7 (strongly agree) to the proposition: "Advertising helps raise our standard of living." The sample mean response was 4.27 and the sample standard deviation was 1.32. Decide whether or not to reject this (null) hypothesis: the mean μ for the full population of 1562 undergrads is 4.

We have to unpack this.

It is helpful to imagine that undergraduates respond to the proposition by rolling a 7-sided biased die, with each side having an associated probability that we don't know. The probability space is then

$$\Omega = \{\, 1, 2, 3, 4, 5, 6, 7 \,\}$$

and there's an implicit random variable $X : \Omega \to \{\, \text{numbers} \,\}$ that takes an outcome i to the number i. We admit as a *null hypothesis* that X has mean $\mu = 4$, and we estimate the standard deviation σ of X to be 1.32 because we have no better idea of what σ might be.

As on November 14, we consider

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

where $n = 1562$ and the variables $X_1, \ldots, X_n$ correspond to the 1562 rolls of the 7-sided die, one roll for each student. The function $\overline{X}$ is a random variable defined on the space of all possible outcomes of 1562 rolls of a 7-sided die; the space has 7^{1562} elements.

The particular value of $\overline{X}$ that we observe from our sample is the *sample mean response*, namely 4.27. Dividing by n in the formula for $\overline{X}$ amounts to taking the average or mean value of the responses of the students who were sampled.

According to the Central Limit theorem, the random variable $(\overline{X} - \mu) \cdot \dfrac{\sqrt{n}}{\sigma}$ is distributed like the standard normal variable.

If the particular value of $(\overline{X} - \mu) \cdot \dfrac{\sqrt{n}}{\sigma}$ is very far from 0 (the mean of the standard normal), then we reject the hypothesis that the mean is μ.

This value is

$$(4.27 - 4) \cdot \frac{\sqrt{1562}}{1.32} \approx 8.08,$$

which is enormous. (According to Table 7.3 on page 566, the probability of being more than 3 standard deviations from the mean is 0.001.)

We reject the null hypothesis!

HAPPY THANKSGIVING!