Empirical p-values Notes

Lt Col Ken Horton Professor Bradley Warner

14 July, 2020

Objectives

1) Know and properly use the terminology of a hypothesis test.
2) Conduct a hypothesis test, all four steps, using randomization.
3) Discuss and explain the ideas of decision errors, one-sided versus two-sided, and choice of statistical signficance.

Hypothesis testing using probability models

As a lead into the central limit theorem and mathematical sampling distributions, we will look at a class of hypothesis testing where the null hypothesis specifies a probability model. In some cases we can get an exact answer and in others we will use simulation to get an empirical p-value. By the way, the permutation test is an exact test, since the complete enumeration of all permuations is difficult we approxiamte it with randomization.

Let's use three examples to illustrate.

Tappers and listeners

Here's a game you can try with your friends or family: pick a simple, well-known song, tap that tune on your desk, and see if the other person can guess the song. In this simple game, you are the tapper, and the other person is the listener.

A Stanford University graduate student named Elizabeth Newton conducted an experiment using the tapper-listener game.[1] In her study, she recruited 120 tappers and 120 listeners into the study. About 50% of the tappers expected that the listener would be able to guess the song. Newton wondered, is 50% a reasonable expectation?

Step 1- State the null and alternative hypotheses

Newton's research question can be framed into two hypotheses:

H_0: The tappers are correct, and generally 50% of the time listeners are able to guess the tune. $p = 0.50$

H_A: The tappers are incorrect, and either more than or less than 50% of listeners will be able to guess the tune. $p \neq 0.50$

[1] This case study is described in http://www.openintro.org/redirect.php?go=made-to-stick&redirect=simulation_textbook_pdf_preliminary Made to Stick by Chip and Dan Heath.

Exercise: Is this a two-sided or one-sided hypothesis test? How many variables are in this model?

The tappers think that listeners will guess the song 50% of the time, so this is a two-sided test since we don't know before hand if listeners will be better or worse than this value.

There is only one variable, if the listener is correct.

Step 2 - Compute a test statistic.

In Newton's study, only 3 out of 120 listeners ($\hat{p} = 0.025$) were able to guess the tune! From the perspective of the null hypothesis, we might wonder, how likely is it that we would get this result from chance alone? That is, what's the chance we would happen to see such a small fraction if H_0 were true and the true correct-guess rate is 0.50?

Now before we use simulation, let's frame this as a probability model. The random variable X is the number of correct our of 120. If the observations are independent and the probability of success is constant then we could use a binomial model. We can't answer the validity of these assumptions without knowing more about the experiment, the subjects, and the data collection. For educational purposes, we will assume they are valid. Thus our test statistic is the number of successes. The observed value is 3.

Step 3 - Determine the p-value.

We now want to find the p-value from $P(X \leq 3)$ given the null hypothesis is true, that the probability of success is 0.50. We will use `R` to get the one-sided value and then double.

```
2*pbinom(3,120,prob=0.5)
```

```
## [1] 4.334862e-31
```

That is a small p-value.

Step 4 - Draw a conclusion

Based on our data, if the listeners were guessing correct 50% of the time, there is less than a 4.3×10^{-31} probability that only 3 or less or 117 or more listeners would get it right. This is much less than 0.05, so we reject that the listeners are guessing correctly half of the time.

Repeat using simulation

We will repeat the analysis using an empirical p-value. Step 1 is the same.

Step 2 - Compute a test statistic.

We will use the proportion of listeners that get the song correct.

```
obs<-3/120
obs
```

```
## [1] 0.025
```

Step 3 - Determine the p-value.

To simulate 120 games under the null hypothesis where $p = 0.50$, we could flip a coin 120 times. Each time the coin came up heads, this could represent the listener guessing correctly, and tails would represent the listener guessing incorrectly. For example, we can simulate 5 tapper-listener pairs by flipping a coin 5~times:

H	H	T	H	T
Correct	Correct	Wrong	Correct	Wrong

After flipping the coin 120 times, we got 56 heads for $\hat{p}_{sim} = 0.467$. As we did with the randomization technique, seeing what would happen with one simulation isn't enough. In order to evaluate whether our originally observed proportion of 0.025 is unusual or not, we should generate more simulations. Here we've repeated this simulation 10000 times:

```
results <- rbinom(10000, 120, 0.5) / 120
```

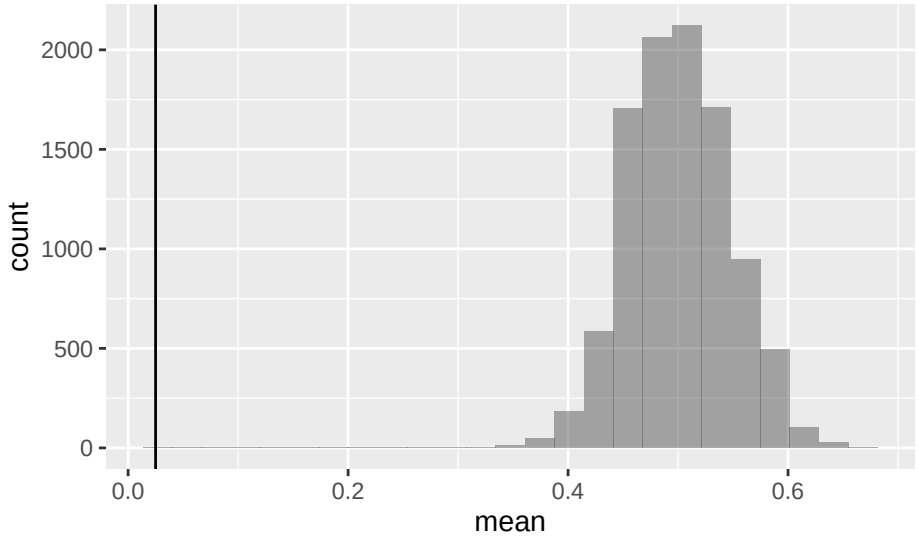
Note, we could simulate it a number of ways. Here is a way using `do` that will look like how we have coded for other randomization tests.

```
set.seed(604)
results<-do(10000)*mean(sample(c(0,1),size=120,replace = TRUE))
```

```
head(results)
```

```
##        mean
## 1 0.4250000
## 2 0.5250000
## 3 0.5916667
## 4 0.5000000
## 5 0.5250000
## 6 0.5083333
```

```
results %>%
  gf_histogram(~mean) %>%
  gf_vline(xintercept =obs)
```



Notice how the sampling distribution is centered at 0.5 and looks symmetrical.

The p-value is found using the `prop1` function, in this problem we really need the observed case added back in to prevent a p-value of zero.

```
prop1(~(mean<=obs),data=results)
```

```
## prop_TRUE
## 9.999e-05
```

Step 4 - Draw a conclusion

In these 10,000 simulations, we don't see any results close to 0.025.

> **Exercise**: In the context of the experiment, what is the p-value for the hypothesis test?[2]

> **Exercise**:
> Do the data provide statistically significant evidence against the null hypothesis? State an appropriate conclusion in the context of the research question.[3]

Cardiopulmonary resuscitation (CPR)

Let's return to the CPR example from last lesson. As a reminder, we will repeat the backbground material.

Cardiopulmonary resuscitation (CPR) is a procedure used on individuals suffering a heart attack when other emergency resources are unavailable. This procedure is helpful in providing some blood circulation to keep a person alive, but CPR chest compressions can also cause internal injuries. Internal bleeding and other injuries that can result from CPR complicate additional treatment efforts. For instance, blood thinners may be used to help release a clot that is causing the heart attack once a patient arrives in the hospital. However, blood thinners negatively affect internal injuries.

Here we consider an experiment with patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.[4] Each patient was randomly assigned to either receive a blood thinner (treatment group) or not receive a blood thinner (control group). The outcome variable of interest was whether the patient survived for at least 24 hours.

Step 1- State the null and alternative hypotheses

We want to understand whether blood thinners are helpful or harmful. We'll consider both of these possibilities using a two-sided hypothesis test.

H_0: Blood thinners do not have an overall survival effect, experimental treatment is independent of survival rate. $p_c - p_t = 0$.

H_A: Blood thinners have an impact on survival, either positive or negative, but not zero. $p_c - p_t \neq 0$.

[2]The p-value is the chance of seeing the data summary or something more in favor of the alternative hypothesis given that guessing has a probability of success of 0.5. Since we didn't observe anything even close to just 3 correct, the p-value will be small, around 1-in-10,000 or smaller.

[3]The p-value is less than 0.05, so we reject the null hypothesis. There is statistically significant evidence, and the data provide strong evidence that the chance a listener will guess the correct tune is less than 50%.

[4]"Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial." The Lancet, 2001.

```
thinner <- read_csv("data/blood_thinner.csv")
```

```
head(thinner)
```

```
## # A tibble: 6 x 2
##   group     outcome
##   <chr>     <chr>
## 1 treatment survived
## 2 control   survived
## 3 control   died
## 4 control   died
## 5 control   died
## 6 treatment survived
```

Let's put it in a table.

```
tally(~group+outcome,data=thinner,margins = TRUE)
```

```
##            outcome
## group       died survived Total
##   control     39       11    50
##   treatment   26       14    40
##   Total       65       25    90
```

Step 2 - Compute a test statistic.

In this case the data is from a **hypergeometric** distribution, this is really a binomial from a finite population. We can calculate the p-value using this probability distribution. The random variable is the number of control patients that survived from a population of 50 control patients, 40 treatment patients, where a total of 25 survived.

Step 3 - Determine the p-value.

In this case we want to find $\mathrm{P}(X \leq 11)$ and double it since it is a two-sided test.

```
2*phyper(11,50,40,25)
```

```
## [1] 0.2581356
```

Or since `R` has a built in function, `fisher.test`, we could use that.

```
fisher.test(tally(~group+outcome,data=thinner))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tally(~group + outcome, data = thinner)
## p-value = 0.2366
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
##  0.6794355 5.4174460
## sample estimates:
## odds ratio
##   1.895136
```

The p-value is slightly different since the `fisher.test` is not symetric and finds and adds all probabilities less than or equal to $P(X = 11)$.

The randomization test in the last lesson yielded a p-value of 0.257 so all tests are consistent.

Step 4 - Draw a conclusion

Since this p-value is larger than 0.05, we do not reject the null hypothesis. That is, we do not find statistically significant evidence that the blood thinner has any influence on survival of patients who undergo CPR prior to arriving at the hospital. Once again, we can discuss the causal conclusion since this is an experiment.

Notice that in these first two examples, we had a test of a single proportion and a test of two proportions. The single proportion test did not have an equivalent randomization test since there is not a second variable to shuffle. We were able to get answers since we found a probability model that we could use in each case.

File Creation Information

- File creation date: 2020-07-14
- Windows version: Windows 10 x64 (build 17763)
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.6.0
- `tidyverse` package version: 1.3.0