Randomization Tests

Stats 102A

Miles Chen

Department of Statistics

Week 9 Monday

UCLA

Copyright Miles Chen. For personal use only. Do not distribute.

Section 1

Comparing Proportions

Copyright Miles Chen. For personal use only. Do not distribute.

A Simple Example

Is the proportion of international students different for Stats Majors and Applied Math Majors?

(I'm completely making these numbers up just for the purpose of illustration.)
Let's say I get a random sample of 150 students majoring in Statistics. We find that 75 of them are international students. We get a random sample of 160 students majoring in Applied Math and 64 of them are international students.
Does this data provide evidence that the proportion of international students is different for students majoring in Statistics and students majoring in Applied Math?

Copyright Miles Chen. For personal use only. Do not distribute.

Z-proportion test

In this problem, we are comparing the proportions of two samples.

This is our observed data:

- Statistics: 75 international, 150 students total. Observed proportion: $\hat{p}_{stats} = 0.5$
- Applied Math: 64 international, 160 students total. Observed proportion: $\hat{p}_{math} = 0.4$

The observed difference between proportions is $\hat{p}_{stats} - \hat{p}_{math} = 0.1$

Copyright Miles Chen. For personal use only. Do not distribute.

Z-proportion test: Hypotheses

The null hypothesis: Both majors have the same proportion of international students.
i.e. There is zero difference between the proportion of international students for the two majors.

$$H_0 : p_{stats} - p_{math} = 0$$

The alternative hypothesis: The difference between the proportions of international students is not zero.

$$H_A : p_{stats} - p_{math} \neq 0$$

Where p is the proportion of students who are International.

If the null hypothesis were true, we would expect the observed difference between proportions to be 0. We observed a difference of 0.1. Is that difference statistically significant?

Copyright Miles Chen. For personal use only. Do not distribute.

The Z-proportion test: The Sampling Distribution

Because of the central limit theorem, we can approximate the sampling distribution of a difference between sample proportions with a normal distribution.

$$(\hat{p}_{stats} - \hat{p}_{math}) \sim N\left(\mu = 0, \sigma = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Where $\hat{p} = \frac{y_1+y_2}{n_1+n_2}$

Copyright Miles Chen. For personal use only. Do not distribute.

The Z-proportion test: P-value interpretation

p-value = What is the probability that two populations with the same proportion of international students produce two random samples where the difference is 0.1 or something more extreme?

$$\text{P-value} = \Pr(|\hat{p}_{stats} - \hat{p}_{math}| > 0.1)$$

We take the absolute value of the difference because we are interested in a positive or negative difference. It would be interesting to us if there were more international students majoring in statistics than math and it would also be interesting to us if there were more international students majoring in math than statistics.

Copyright Miles Chen. For personal use only. Do not distribute.

The Z-proportion test: calculations

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{75 + 64}{150 + 160} \approx 0.4484$$

Test statistic Z:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.5 - .4}{\sqrt{0.4484(1-0.4484)(1/150 + 1/160)}} \approx \frac{.1}{0.0565} \approx 1.7692$$

P-value $= \Pr(|\hat{p}_{stats} - \hat{p}_{math}| > 0.1) = \Pr(Z > 1.769) + \Pr(Z < -1.769) = 2 \times 0.0384 = 0.0768$

Copyright Miles Chen. For personal use only. Do not distribute.

The Z-proportion test: Conclusion

Our p-value is 0.0768

If $\alpha = 0.05$, our p-value is larger than α. We do not reject the null hypothesis.

The data we have does not provide evidence that the proportion of international students is different for statistics majors and applied math majors.

If the proportion of international students majoring in math and majoring in statistics is the same, we expect samples drawn from each population to have the same proportion. The random sampling process can still produce two random samples where there is a difference of 0.1. This will happen by random chance with a probability of about 0.0768. While not a very large probability, it is not small enough for us to completely dismiss it.

We cannot say that the difference of 0.1 that we observed was definitely caused by something other than random chance.

Copyright Miles Chen. For personal use only. Do not distribute.

Conditions to use the Z-proportion test

The Z-proportion test is a parametric test that relies on the Central Limit Theorem.

To use the Z-proportion test, the samples must be "large enough." The consensus is that "large enough" means there were at least 10 "yes" and at least 10 "no" in both samples.

Our samples in the previous example were large enough to meet this criteria so we could use the Z-proportion test.

If the conditions for sample size are not met, using the Z-proportion test will give us the wrong p-value and could lead us to making the wrong conclusion. In this case, we may need to use a **randomization test**

Copyright Miles Chen. For personal use only. Do not distribute.

Section 2

Randomization Tests

Copyright Miles Chen. For personal use only. Do not distribute.

Mythbusters Yawning

Let's watch a video!

https://youtu.be/LuRB_OoplAw?t=1222

(I had to purchase this video.)

Copyright Miles Chen. For personal use only. Do not distribute.

Recap for those who missed the video:

Mythbusters TV show wanted to see if yawning is contagious.

They select 50 people to participate. Each participant is told to sit alone into a small room and monitored for several minutes.

Before they enter the room, some of the people see a person yawn (seed yawn). Other people do not (control).

Mythbusters team records who yawns and who does not yawn.

Copyright Miles Chen. For personal use only. Do not distribute.

The data

The sample of people was not randomly selected.

The treatments are randomly assigned. (The assignment process is systematic - every third person is control, but the order of people is random.)

"seed yawn" group: 34 participants. 10 yawned. observed proportion: $\hat{p}_{seed} = 10/34 = 0.29412$

control group: 16 participants. 4 yanwed. observed proportion: $\hat{p}_{control} = 4/16 = 0.25$

Mythbusters says: "There's little doubt, it does seem to be contaigous"

Copyright Miles Chen. For personal use only. Do not distribute.

A Hypothesis Test

Notation: $p_{seed} = \Pr(\text{Yawning}|\text{Seed})$, $p_{control} = \Pr(\text{Yawning}|\text{Control})$

Null Hypothesis: yawning is not contagious.

$H_0 : p_{seed} - p_{control} = 0$

Alternative Hypothesis: Yawning is contagious. Receiving a seed yawn makes it more likely for the person to yawn.

$H_A : p_{seed} - p_{control} > 0$

Our observed data:

$\hat{p}_{seed} - \hat{p}_{control} = \frac{10}{34} - \frac{4}{16} \approx 0.044$

If the null hypothesis were true, we would expect the observed difference between proportions to be 0. We observed a difference of 0.044. Is that difference statistically significant?

Copyright Miles Chen. For personal use only. Do not distribute.

The Null Hypothesis

The Null Hypothesis states "yawning is not contagious." $H_0 : p_{seed} - p_{control} = 0$

If the null hypothesis were true, then whether a person yawns or doesn't is independent of whether the person got a 'seed-yawn' or not. This means that the observed difference in our results are just a result of the randomness in the assignment of treatments.

Another way to think of it: 14 people in our sample were 'destined' to yawn anyway. The fact that we observed 29% yawn in one group and only 25% yawn in the other group is merely a result of the random assignment. Randomness alone could produce the difference between proportions of 0.044.

Copyright Miles Chen. For personal use only. Do not distribute.

The P-value in this context

The p-value is the probability of observing our difference or something more extreme if the null hypothesis were true.

If the null hypothesis were true, we would expect the proportion of yawners to be the same for both groups. We would expect to see a difference of 0 between the two proportions. However, we observed a difference of 0.044.

p-value = If we assume that the two populations (seed vs control) have the same proportion of yawners, what is the probability that random assignment could produce samples of data where the difference is 0.044 or something more unusual?

Copyright Miles Chen. For personal use only. Do not distribute.

Randomization Tests

Our samples are too small to use a Z-proportion test, so we will use a randomization test which has no distributional assumptions.

The big idea behind a randomization test is to **simulate the sampling distribution of outcomes when the Null Hypothesis is true**. The null hypothesis states that the random assignment of treatments is the source of variation between sample proportions.

We will simulate data where 14 people are "destined" to yawn. We will randomly assign them into two groups: Group A with 34 people, and Group B with 16 people. We will calculate the proportion of 'yawners' in each group and find the difference between the group proportions. *Because we have done the random assignment ourselves, we know that any difference between the group proportions is from the random assignment process*.

We will repeat the randomization process many times. Each time we record the difference between the group proportions. After many iterations, we will have built a sampling distribution of proportion differences where the only source of variation is randomization.

To get the p-value associated with our observed difference of 0.044, we find the proportion of times that the random assignment process produces a difference of 0.044 or greater.

Copyright Miles Chen. For personal use only. Do not distribute.

Exchangeability

The one assumption of a randomization test is that there is **exchangeability** in the outcomes.

Exchangeability is the idea that under the null hypothesis, all possible permutations of the data are equally likely. It also means that any permutation of the data is exchangeable with any other permutation of the data.

Practically speaking, this means that **exchangeability applies only to experiments where random assignment has been used**. If there was no random assignment in the gathering of the data, do not use a randomization test.

If the patients are assigned treatments randomly, any arrangement of patients is equally likely. In cases where treatments were not assigned randomly (observational study), the permutations of groups are not equally likely and we cannot use the randomization test.

Copyright Miles Chen. For personal use only. Do not distribute.

Inputting our Data

```
yawning <- c( rep(TRUE, 10), rep(FALSE, 24), rep(TRUE, 4), rep(FALSE, 12))
treatment <- yawning[ 1:34]
control   <- yawning[35:50]
mean(treatment)  # 10 out of 34
```

```
## [1] 0.2941176
```

```
mean(control)    #  4 out of 16
```

```
## [1] 0.25
```

```
obs_dif <- mean(treatment) - mean(control)
obs_dif
```

```
## [1] 0.04411765
```

Copyright Miles Chen. For personal use only. Do not distribute.

sample()

`sample()` applied to a vector shuffles the contents of the vector.

All of the values remain in the vector, but the order is randomized.

Every time we call `shuffle(yawning)`, the output vector will have 50 values: 14 TRUE and 36 FALSE. Only the order of the values will change.

Copyright Miles Chen. For personal use only. Do not distribute.

Randomize

```
set.seed(1) # for reproducibility
randomized <- sample(yawning)  # who yawns is now randomized
groupA <- randomized[1:34]
groupB <- randomized[35:50]
mean(groupA)
```

```
## [1] 0.2941176
```

```
mean(groupB)
```

```
## [1] 0.25
```

```
rand_dif <- mean(groupA) - mean(groupB)
rand_dif
```

```
## [1] 0.04411765
```

14 out of 50 people are 'destined' to yawn. We randomly assigned 34 people to group A, the other 16 to group B. Approx 29% of group A were yawners, and 25% in group B were yawners. Our random assignment process happened to produce a difference of 0.044 just by randomness. Maybe this was just a conicidence. To find out, we'll need to run some more trials.

Copyright Miles Chen. For personal use only. Do not distribute.

Another randomization trial

```
set.seed(5) # for reproducibility
randomized <- sample(yawning)
groupA <- randomized[1:34]
groupB <- randomized[35:50]
mean(groupA)
```

```
## [1] 0.2352941
```

```
mean(groupB)
```

```
## [1] 0.375
```

```
rand_dif <- mean(groupA) - mean(groupB)
rand_dif
```

```
## [1] -0.1397059
```

This time approx 23.5% of group A were yawners, and 37.5% in group B were yawners. This time the process produced a difference of about -0.14 from randomness alone.

Copyright Miles Chen for personal use only. Do not distribute.

Many randomizations

```r
set.seed(1) # for reproducibility
differences <- rep(NA, 10000)
for(i in seq_along(differences)){
  randomized <- sample(yawning)
  groupA <- randomized[1:34]
  groupB <- randomized[35:50]
  differences[i] <- mean(groupA) - mean(groupB)
}
summary(differences)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.599265 -0.047794  0.044118 -0.001985  0.136029  0.411765
```

Copyright Miles Chen. For personal use only. Do not distribute.

Empirical p-value

What is the probability of observing a difference as large as our data's difference?

```
obs_dif
```

```
## [1] 0.04411765
```

We observed a difference of 0.04411765. We calculate our P-value: What is the probability that the randomization process produces a difference of 0.044 or more?

Our alternative Hypothesis is $H_A : p_{seed} - p_{control} > 0$, meaning we are only interested in differences of 0.044 or something more positive.

$$\text{P-value} = \Pr(\hat{p}_{seed} - \hat{p}_{control} > 0.044)$$

```
mean(differences >= obs_dif) # empirical p-value
```

```
## [1] 0.507
```

Copyright Miles Chen. For personal use only. Do not distribute.

Conclusion

Our empirical p-value is 0.507.

The Randomization process has a high probability of producing the data that we observed in the Mythbusters experiment.

We have no reason to believe that the data we observed was caused by something other than randomness.

We would not reject the null hypothesis. *The data does not provide reason to believe that getting a seed yawn makes a person more likely to yawn.*

We cannot say conclusively that yawning is not contagious. Our data simply does not provide evidence that it is contagious. (It is possible that a seed yawn increases the probability of a yawn, but maybe we did not collect enough data, or maybe we happened to get some weird data that leads us to believe it is not.)

Copyright Miles Chen. For personal use only. Do not distribute.

A two-sided alternative

If the Alternative hypothesis said something like: "Does seeing someone yawn have any effect (positive or negative) on a person yawning?", we would use a two-sided alternative. We would be interested in the possibility of a difference that is more extreme (both positive an negative) than the observed difference.

$$\text{P-value} = \Pr(|\hat{p}_{seed} - \hat{p}_{control}| > 0.044)$$

```
mean(abs(differences) >= obs_dif)
```

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

In our case, we get a p-value of 1, meaning that every single randomization resulted in a difference that was equal to or greater than our observed difference.

Copyright Miles Chen. For personal use only. Do not distribute.