

Stats:

Categorical Inference

Mick McQuaid

2024-03-04

Week SEVEN

Homework

```
1 x←c(0,1.9,1.7,1.9,2,1.9,1.8,1.7,1.9,0,1.9)
2 stem(x)
```

The decimal point is at the |

```
0 | 00
0 |
1 |
1 | 77899999
2 | 0
```

One homework issue

Difference between \hat{p} and p , where \hat{p} is what we observe and p is what we believe to be true in general

We form a hypothesis about p and observe \hat{p} as evidence

More on the difference

We know \hat{p} so we don't need to form a hypothesis about it. It's the sample proportion we observe. It's a tool to make an inference about p , the actual proportion. The hypothesis is an assertion about something we don't know for certain. We'll use the evidence (\hat{p}) plus what we know about distributions to verify or invalidate the assertion.

Textbook section 6.1 Inference for a single proportion

Is the sample proportion nearly normal?

Note: The book uses p in two ways: as a p -value in a hypothesis test, and as p , a population proportion. Related to the population proportion is the sample proportion, \hat{p} , pronounced p-hat. Too many p s!

The sampling distribution for \hat{p} based on a sample of size n from a population with a true proportion p is nearly normal when:

1. The sample's observations are independent, e.g., are from a simple random sample.
2. We expected to see at least 10 successes and 10 failures in the sample, i.e., $np \geq 10$ and $n(1 - p) \geq 10$. This is called the success-failure condition.

When these conditions are met, then the sampling distribution of \hat{p} is nearly normal with mean p and standard error $SE = \sqrt{p(1 - p)/n}$.

Confidence interval for a proportion

A confidence interval provides a range of plausible values for the parameter p , and when \hat{p} can be modeled using a normal distribution, the confidence interval for p takes the form

$$\hat{p} \pm z^* \times SE$$

where z^* marks the x -axis for the selected confidence interval, e.g., 1.96 for a 95 percent confidence interval.

Prepare, Check, Calculate, Conclude Cycle

The textbook, Diez, Çetinkaya-Rundel, and Barr ([2019](#)), recommends a four step cycle for both confidence intervals and hypothesis tests. It differs a bit from the seven step hypothesis testing method given in Week 06, but achieves the same result.

- Prepare. Identify \hat{p} and n , and determine what confidence level you wish to use.
- Check. Verify the conditions to ensure \hat{p} is nearly normal. For one-proportion confidence intervals, use \hat{p} in place of p to check the success-failure condition.
- Calculate. If the conditions hold, compute SE using \hat{p} , find z^* , and construct the interval.
- Conclude. Interpret the confidence interval in the context of the problem.

Same cycle for hypothesis testing for a proportion

- Prepare. Identify the parameter of interest, list hypotheses, identify the significance level, and identify \hat{p} and n .
- Check. Verify conditions to ensure \hat{p} is nearly normal under H_0 . For one-proportion hypothesis tests, use the null value to check the success-failure condition.
- Calculate. If the conditions hold, compute the standard error, again using p_0 , compute the Z -score, and identify the p -value.
- Conclude. Evaluate the hypothesis test by comparing the p -value to α , and provide a conclusion in the context of the problem.

Choosing sample size when estimating a proportion

This is probably the most important part of this section for practical purposes. The following expression denotes the *margin of error*:

$$z^* \sqrt{\frac{p(1-p)}{n}}$$

You have to choose the margin of error you want to report.
The book gives an example of 0.04. So you want to find

$$z^* \sqrt{\frac{p(1-p)}{n}} < 0.04$$

The problem is that you don't know p . Since the worst-case scenario is $p = 0.5$, you have to use that unless you have some information about p . Recall that z^* represents the z -score for the desired confidence level, so you have to choose that. The book gives an example where you want a 95 percent confidence level, so you choose 1.96. You could find this out in R by saying

```
1 qnorm(0.025, lower.tail=FALSE)
```

```
[1] 1.959964
```

returning the z -score for the upper tail.

The reason for saying that 0.025 instead of 0.05 is that the probability of 0.05 is split between the tails. The complementary function is

`pnorm(1.959964, lower.tail=FALSE)`, which will return 0.025.

Once you have decided on values for p and z^* , solve the preceding inequality for n .

Textbook section 6.2 Difference of two proportions

In this section, we're just modifying the previous section to account for a difference instead of a single proportion.

The difference $\hat{p}_1 - \hat{p}_2$ can be modeled using the normal distribution when

- The data are independent within and between the two groups (random samples or randomized experiment)
- The success-failure condition holds for both groups (at least 10 successes and 10 failures in each sample)

Confidence intervals for $p_1 - p_2$

point estimate $\pm z^* \times \text{SE}$

$$\rightarrow (\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Hypothesis testing for $p_1 - p_2$

When the null hypothesis is that the proportions are equal, use the pooled proportion (\hat{p}_{pooled}) to verify the success-failure condition and estimate the standard error

$$\hat{p}_{\text{pooled}} = \frac{\text{number of “successes”}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

Textbook section 6.3 Testing goodness of fit using χ^2

The χ^2 test, pronounced ki square, is useful in many circumstances. The textbook treats two such circumstances:

1. Suppose your sample can be divided into groups, as can the general population. Does your sample represent the general population?
2. Does your sample resemble a particular distribution, such as the normal distribution?

For the first circumstance, we could divide a sample of people into races or genders and we would like to examine all at once for resemblance to the general population, rather than in pairs. The χ^2 statistic will permit an all-at-once comparison.

The χ^2 statistic is given by the following formula for g groups.

$$\chi^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \dots + \frac{(\text{observed count}_g - \text{null count}_g)^2}{\text{null count}_g}$$

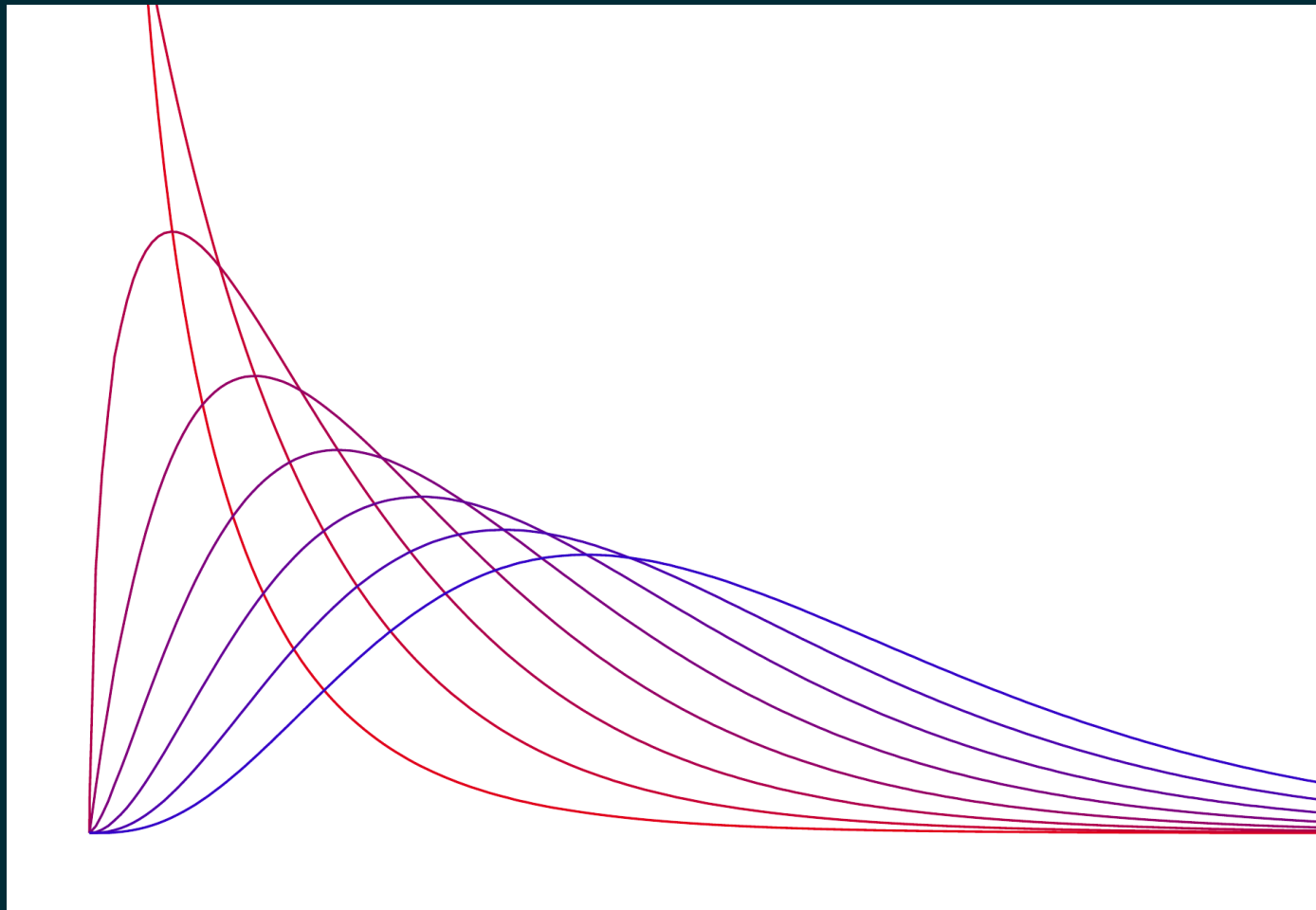
where the expression *null count* refers to the expected number of objects in the group.

You have to be careful about how you determine the null count. For instance, the textbook gives an example of races of jurors. In such a case, the null counts should come from the population who can be selected as jurors. This might be a matter of some dispute since jurors are usually recruited through voting records and these records may not reflect the correct proportions. Statisticians tend to like things they can count, and some people are harder (more expensive) to count than others, particularly *people in marginalized populations*.

The χ^2 distribution

The χ^2 distribution is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall a normal distribution had two parameters – mean and standard deviation – that could be used to describe its exact characteristics. The χ^2 distribution has just one parameter called degrees of freedom (df), which influences the shape, center, and spread of the distribution.

Here is a picture of the χ^2 distribution for several values of df (1–8).



In the jurors example, we can calculate the appropriate p -value in R by using the χ^2 statistic calculated from the sample, 5.89, and the parameter $df = k - 1$ which is the number of groups minus one, using R:

```
1 pchisq(5.89,3,lower.tail=FALSE)  
[1] 0.1170863
```

This is a relatively large p -value given our earlier choices of cutoffs of 0.1, 0.05, and 0.01.

Conclusion

Consequently, we fail to reject the null hypothesis, which is that no racial bias is in evidence in juror selection. Bear in mind that this finding relies on our belief about the proportions in the population, which may have been systematically miscounted in the case of marginalized populations.

χ^2 test

The χ^2 test can be conducted in R for the juror example given in the book as follows.

```
1 o ← c(205,26,25,19)
2 e ← c(198,19.25,33,24.75)/sum(o)
3 chisq.test(o,p=e)
```

Chi-squared test for given probabilities

```
data:  o
X-squared = 5.8896, df = 3, p-value = 0.1171
```

Note that I had to make an adjustment in R. The R variable **p** is supposed to be a vector of probabilities summing to 1. The way the table in the book presented it, it was not a vector of probabilities summing to one. So I divided each element of the input vector for **e** by the sum of the vector **o**.

Textbook section 6.4 Testing independence in 2-way tables

Suppose you have a two way table. Datacamp gives an example of gender and sport as the two ways. The following data frame lists the number of males and females who like the following three sports: archery, boxing, and cycling.

```
1 female <- c(35,15,50)
2 male <- c(10,30,60)
3 df <- cbind(male,female)
4 rownames(df) <- c("archery","boxing","cycling")
5 df
```

	male	female
archery	10	35
boxing	30	15
cycling	60	50

```
1 chisq.test(df)
```

Pearson's Chi-squared test

```
data: df
X-squared = 19.798, df = 2, p-value = 5.023e-05
```

The χ^2 tests suggests that the genders are not independent for the three sports, meaning that the preferences may differ by gender.

END

References

Diez, David, Mine Çetinkaya-Rundel, and Cristopher D Barr. 2019. *OpenIntro Statistics, Fourth Edition*. self-published. <https://openintro.org/os>.

Colophon

This slideshow was produced using **quarto**

Fonts are *Roboto Condensed Bold*, *JetBrains Mono Nerd Font*,
and *STIX2*