Chi-square testing for goodness of fit

March 3, 2022

In 2014, Harvard was sued by a group of Asian-American applicants who were rejected in admissions. They claimed racial discrimination. During the case, Harvard was ordered to release private admissions data. Today, we'll use that data (reconstructed from the plaintiff's report) to explore the discrimination claim using a chi-square test for goodness of fit.

Imagine Harvard claims: "We only accept the top academic applicants and we treat those applicants equally. Our admitted class is as good as a random sample from the pool of top applicants." We would like to test if there is convincing evidence against this claim.

In the pool of top academic applicants for the Class of 2019, 57.5% were Asian-American, 3.1% were Hispanic, 0.8% were African-American, and 38.7% were White. Harvard rates applicants' academics based on high school performance and standardized test scores.
Note: Percentages are only calculated out of these four groups because, even though other racial groups were admitted, they were not substantively discussed in court documents.

1. Let's suppose that Harvard's claim is true. How many students from each group would we expect in the admitted class of 2019, which had a total of 2,023 students admitted from these groups? Fill in the expected counts table below. (Expected counts on the left table, and Observed counts are shown on the right table).

Group	Count
Asian-American	
Hispanic	
African-American	
White	

Group	Count
Asian-American	432
Hispanic	247
African-American	226
White	1118

2. Compare the values. Which groups are overrepresented in Harvard's observed (actual) Class of 2019? Which groups are underrepresented?

Solution: Asian Americans are underrepresented, while the other thee groups are represented.

3. Could the differences between the observed and the expected counts have occurred by chance alone? Answer using your intuition.

Solution: This seems unlikely. The difference between the expected and observed counts is very large for each group.

4. Write down the hypothesis that tests Harvard's claim.

Solution:

H_0 : The racial distribution of admitted students is the same as the claimed distribution.

H_a : The racial distribution of admitted students is not the same as the claimed distribution.

5. Use the table to calculate the test statistic:

	Observed	Expected	(Observed-Expected)	(Observed-Expected)^2	$\frac{\text{(Observed-Expected)}^2}{\text{Expected}}$
Asian-American					
Hispanic					
African-American					
White					

6. Your test statistic is a large number, what does this mean about the size of differences between the observed and expected values? Explain.

Solution: A large test statistic means that the difference between the observed counts and the expected counts were large.

7. What are the conditions to use a chi-square goodness of fitness test.

Solution:

1. Assume admitted students are a random sample from the top academic pool.
2. 10% condition: $2023 \leq (0.10)$(All Top Applicants)
3. large counts: All expected counts are greater than or equal to five.

8. Carry out a chi-square goodness of fitness test.

> **Solution:** PLAN: Name of procedure: chi-square test for goodness of fit.
> DO: Specific formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$
> Test statistic: $\chi^2 = 3861.8$
> p-value: ≈ 0
> CONCLUDE: Assuming H_0 is true (claimed distribution is correct), there is a roughly 0 probability of getting a chi^2 test statistic of 3861.8 or greater purely by chance.
> Because $0 < 0.05$ we reject H_0 and we do have convincing evidence that the racial distribution of admitted students is not the same as the claimed distribution.
> The largest component of the χ^2 statistic is 2717 because there were more African Americans accepted than expected.

9. A traffic light is installed to allow traffic from a seldom used side street to cross a 4-lane highway. Because the side street doesn't get a lot of traffic the light is set to provide a red light for the side street 80% of the time, yellow 5% of the time, and green 15% of the time. A resident who must pass through the light several times per day is suspicious that the light is not functioning according to the claimed distribution. He sets up a trail camera and programs it to snap a picture of the light at 200 randomly selected times throughout the day. Here are the results: Red: 173, Yellow: 13, and Green: 14.

   (a) Do these data provide convincing evidence that the light is not functioning according to the claimed distribution?

> **Solution:** STATE:
>
> $$H_0 : \text{The light is red red 80\%, yellow 5\% , and green \%15}$$
>
> $$H_a : \text{The light in NOT red 80\%, yellow 5\% , and green \%15}$$
>
> $$\alpha = 0.05$$
>
> PLAN: Name of procedure: Chi square test for goodness of fit
> Conditions: Random: "200 randomly selected times" so we may generalize to all times.
> 10%: 200 < 0.10(all times so sampling without replacement is okay.
> Large counts: All expected counts are greater or equal to 50: (160, 10, 30). This means the sampling distribution is approx. chi-square.
>
> DO: Specific formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$
> Work: (Write chi-square formula work)
> Test-Statistic: $\chi^2 = 10.49$.
> p-value: $P(\chi^2_2 > 10.49) = 0.0053$
>
> CONCLUDE: Assuming H_0 is true (claimed distribution is correct), there is a 0.0053. probability of getting a chi^2 of 10.49 or greater purely y chance. Because $0.0053 < 0.05$ we reject H_0 and we have convincing evidence that the claimed distribution of light color is incorrect. The largest component of χ^2 is 8.53 because the number green is significantly less than expected.

10. Are births equally likely across days of the week? A random sample of 150 births give the following sample distribution.

Day	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Count	11	27	23	26	21	29	13

(a) State the appropriate hypothesis.

Solution:

$$H_0 : \text{Births are equally likely across days of the week}$$

$$H_a : \text{Births are not equally likely across days of the week.}$$

(b) Calculate the expected count for each of the possible outcomes.

Solution:

$$\frac{1}{7}(150) = 21.4$$

(c) Which degree of freedom should you use?

Solution: df=7-1=6

(d) Use table C to find the p-value. What conclusion would you make?

Solution: p-value¡0.0005. Because the p-value$< \alpha = 0.05$ we reject H_0. There is convincing evidence that the births are not equally distributed across the days of the week.