Chi-squared Tests

Part 1 - Goodness of Fit Testing

Ryan Miller

Grinnell College
Statistics

Introduction

The hypothesis tests we've studied for categorical data have all involved a *binary* outcome variable:

1) Single proportion (one-sample) tests compared an observed proportion with a hypothesized proportion
2) Difference in proportions (two-sample) tests compared the observed proportions in two groups

This is a limitation, as not all research questions involving categorical data can be framed in a way that uses a binary outcome.

Nominal vs. Binary Categorical Variables

- Binary outcomes are special because knowing the proportion of cases in one category tells us everything we need to know
  - If 85% of the sample survives, we know that the remaining 15% must have died
- For a nominal categorical variable this isn't the case
  - For example, there are four blood types: A, B, AB, and O
    - Knowing that 45% of a sample is type O isn't enough to know the relative frequencies of the other categories

AP Exam Answers

Below is a sample of answers to 400 randomly selected AP Exam questions:

A	B	C	D	E
85	90	79	78	68

1. The developers of AP Exams would like answers to be randomly distributed so that there's no systematic advantage in guessing a particular answer. If answers were randomly distributed, what proportion of this sample would you expect to be A's?
2. Consider a one-sample Z-test suggesting the proportion of A's is consistent with the null hypothesis. Is this sufficient to decide whether or not answers are randomly distributed?

AP Exam Answers

Using our existing toolbox, we could perform 4 hypothesis tests:

1) $H_0 : p_A = 0.2$
2) $H_0 : p_B = 0.2$
3) $H_0 : p_C = 0.2$
4) $H_0 : p_D = 0.2$

If none of these tests provide evidence of a deviation from what is expected, it would be reasonable to believe that answers are randomly distributed; however, this is tedious and it exposes us to the problems associated with multiple testing.

AP Exam Answers

What we'd really like to do is evaluate the hypothesis:

$$H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$$

If this null hypothesis were true, we'd *expect* to observe $400 * 0.2 = 80$ correct answers in each category:

A	B	C	D	E
80	80	80	80	80

AP Exam Answers

Recall that many hypothesis tests use a test statistic of the form:

$$\text{Test Statistic} = \frac{\text{Observed Outcome} - \text{Hypothesized Value}}{SE}$$

Can this form be applied to our data and null hypothesis?

Chi-squared Tests

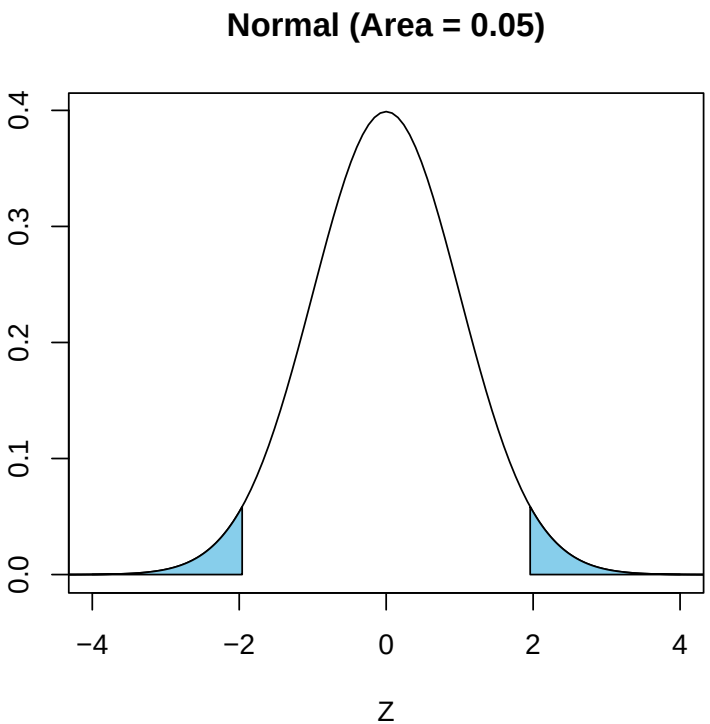
The **Chi-squared test** uses the following *test statistic*:

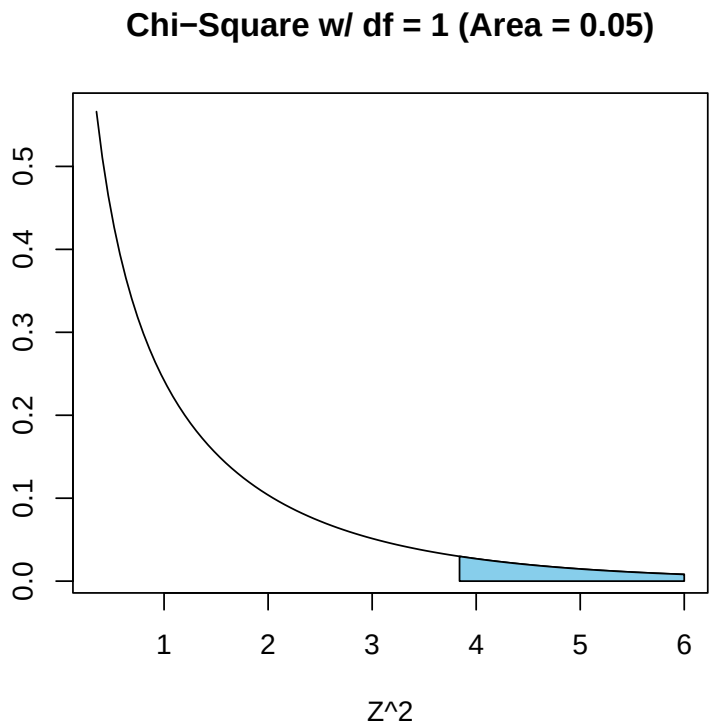
$$X^2 = \sum_{i=1}^{k} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

- Calculating the test statistic requires two tables, a table of *observed frequencies*, and a table of *expected frequencies* that reflect the null hypothesis
- The p-value is found by comparing the test statistic with the probability model it is expected to follow under the null hypothesis
  - This probability model is a **Chi-squared** distribution

The Chi-squared Distribution

The Chi-squared distribution (with $df = 1$) is a squared variant of the Standard Normal curve:





The Chi-squared Distribution

- There are many different χ^2 distributions depending upon how many unique categories we must sum over
- Letting k denote the number of categories of a categorical variable, the χ^2 test statistic for testing a single categorical variable has $k-1$ degrees of freedom
  - This is because the set of k different category proportions are constrained to sum to 1

Revisiting the AP Exam Example

For the AP Exam example:

$$
\begin{aligned}
X^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\
&= \frac{(85-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(79-80)^2}{80} + \frac{(78-80)^2}{80} + \frac{(68-80)^2}{80} \\
&= 3.425
\end{aligned}
$$

For $X^2 = 3.425$ and $k = 5$ (so $df = 4$), the corresponding p-value is 0.49, so this sample provides insufficient evidence to suggest that AP exam answers are not randomly distributed.

Goodness of Fit Testing

- The Chi-squared testing procedure covered in these slides is often referred to as a **Goodness of Fit Test** because it evaluates whether the distribution of the sample data is compatible with a hypothesized distribution
  - The main idea behind goodness of fit tests is to avoid needing to perform several tests of a single proportion
- Our next lecture will cover using Chi-squared tests to assess the independence of two categorical variables