χ^2 Tests of Independence

Grinnell College

Warm-up

1. Suppose I flip a fair coin twice:
   - What is the probability that I flip 0 heads?
   - Probability of one heads?
   - Probability of two heads?

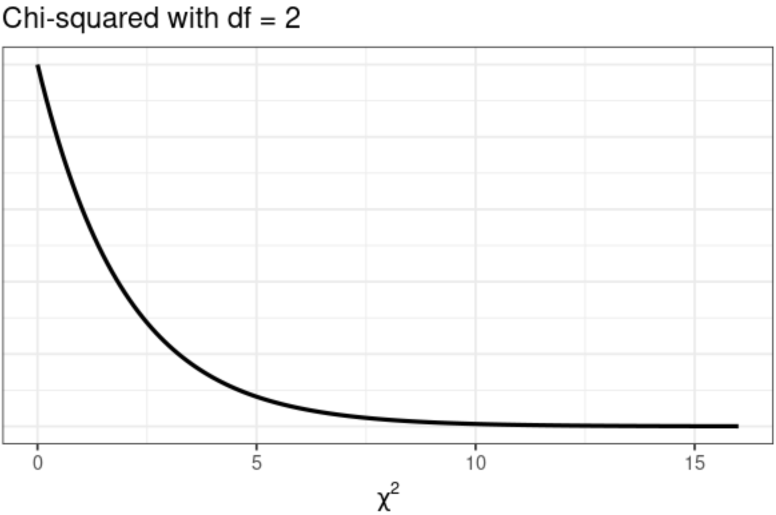
2. Suppose I repeat this twice flipping experiment 100 times. Of these I get 28, 55, and 17 instances of 0, 1, and 2 heads, respectively:
   - Create a table of observed and expected values under the null hypothesis that my coin is fair
   - Using your table, construct a χ^2 test statistic
   - From p-value, what conclusion would you make regarding our null hypothesis?

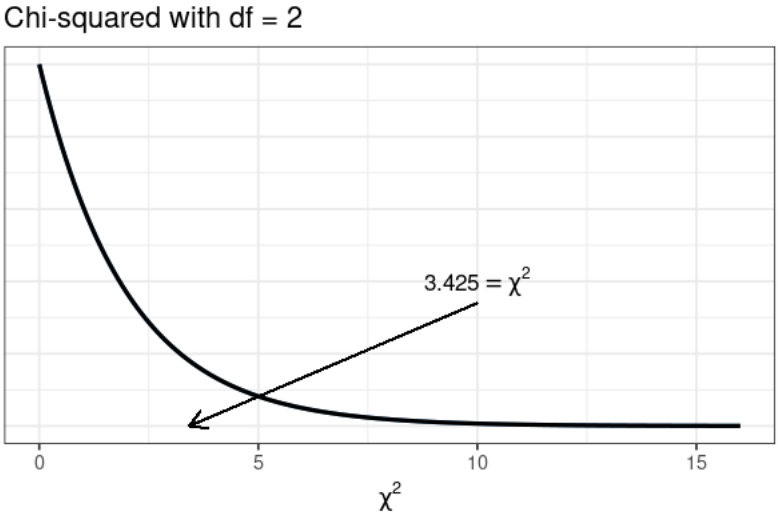
	0H	1H	2H	Total
Expected	25	50	25	100
Observed	28	55	17	100

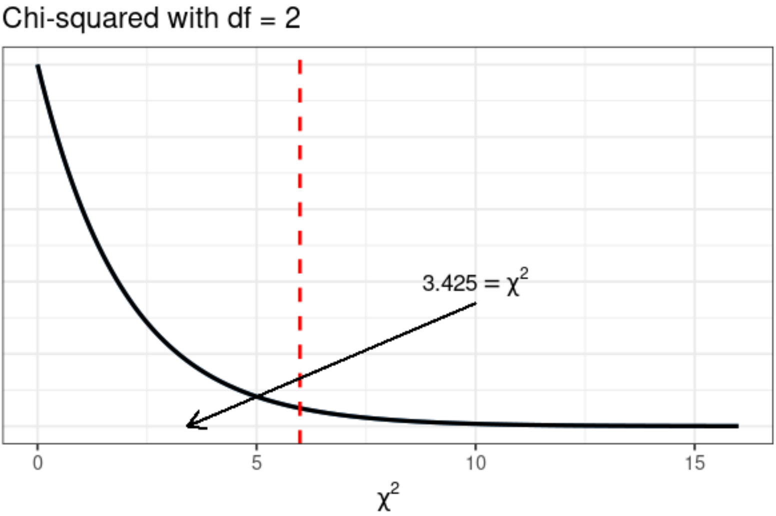
$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Expected}_i - \text{Observed}_i)^2}{\text{Expected}_i}$$

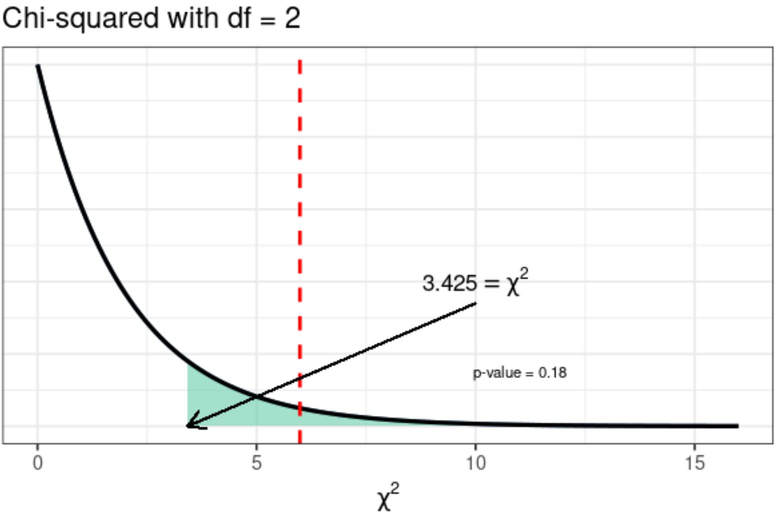
$$= \frac{(25-28)^2}{25} + \frac{(50-55)^2}{50} + \frac{(25-17)^2}{25}$$

$$= 3.42$$


Chi-squared with df = 2
0
5
10
15
χ^2


Chi-squared with df = 2
$3.425 = \chi^2$
0
5
10
15
χ^2


Chi-squared with df = 2
$3.425 = \chi^2$
0
5
10
15
χ^2


Chi-squared with df = 2
3.425 = χ^2
p-value = 0.18
0
5
10
15
χ^2

χ^2 Goodness of Fit

Last class we introduced the χ^2 **Goodness of Fit** test for assessing the goodness of fit for a single categorical variable

- compares Observed data to Expected data (under H_0)
- H_A: proportions are not equal to specified values

We extend this today to the χ^2 **Test of Independence** used to test the independence or lack of association between two categorical variables

- H_0 there is *not* an association
- H_A: there is an association

Calculating the test statistic for both of these is the same, we just need to keep track of the different Hypotheses and different df

Independence and Probability

Recall that, in general, the probability of two events A and B is given as

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$

with indepenedence *if and only if (iff)*

$$P(A \text{ and } B) = P(A)P(B)$$

Next part: How does this translate to a null hypothesis of independence between groups

Suppose we have Cars and Trucks that can be painted either Blue or Red. We could represent these variables as such:

	Red	Blue	Total
Car	n_1	n_2	$n_1 + n_2$
Truck	n_3	n_4	$n_3 + n_4$
Total	$n_1 + n_3$	$n_2 + n_4$	N

The table above gives us the following information (for example):

- There are $n_1 + n_2$ vehicles that are cars
- There are $n_2 + n_4$ blue vehicles
- There are n_3 red trucks

We can use this to establish our null hypothesis

	Red	Blue	Total
Car	n_1	n_2	$n_1 + n_2$
Truck	n_3	n_4	$n_3 + n_4$
Total	$n_1 + n_3$	$n_2 + n_4$	N

If our variables were independent, then our *expected probability* is

$$\begin{aligned} P(\text{Car and Red}) &= P(\text{Car})P(\text{Red}) \\ &= \left(\frac{n_1 + n_2}{N}\right) \times \left(\frac{n_1 + n_3}{N}\right) \end{aligned}$$

To get our expected counts, we would multiply this probability by N, the total number of observations:

$$\begin{aligned} \text{Expected Number of Red Cars} &= N \times \left(\frac{n_1 + n_2}{N}\right) \times \left(\frac{n_1 + n_3}{N}\right) \\ &= \frac{(n_1 + n_3)(n_1 + n_2)}{N} \end{aligned}$$

In other words, our *expected counts* is the product of the row and column margins, divided by the total number of observations

Expected Counts

For example, suppose we had 60 cars, 40 trucks, 50 blue vehicles, and 50 red vehicles. The margins totals would look like this:

	Red	Blue	Total
Car			60
Truck			40
Total	50	50	100

From this, we have the following probabilities:

$$P(\text{Red}) = \frac{50}{100} = 0.5, \quad P(\text{Car}) = \frac{60}{100} = 0.6$$

Under the null hypothesis of independence, the probability of both is

$$P(\text{Car and Red}) = P(\text{Car})P(\text{Red}) = 0.5 \times 0.6 = 0.3$$

Since there are 100 vehicles, and the probability of of a vehicle being a red car is 0.3, the expected number of red cars would be 30

Expected Counts

We could take our expected counts:

	Red	Blue	Total
Car	30	30	60
Truck	20	20	40
Total	50	50	100

And compare them to what we observe:

	Red	Blue	Total
Car	32	28	60
Truck	18	22	40
Total	50	50	100

$$\chi^2 = \frac{(30-32)^2}{30} + \frac{(30-28)^2}{30} + \frac{(20-18)^2}{20} + \frac{(20-22)^2}{20} = 0.735$$

Degrees of Freedom

Just as with the univariate case, the χ^2 test of independence is governed by its degrees of freedom

For a table with k columns and m rows, the total degrees of freedom is $df = (k-1) \times (m-1)$

The degrees of freedom for the car example, then would be $(2-1) \times (2-1) = 1$

The process of finding critical values or p-values then proceeds identically as before

Review

Here are the things to know about the test for independence:

- Expected counts come from products of margin probabilities
- Degrees of freedom for k columns and m rows is $(k-1) \times (m-1)$
- Everything else works the exact same way as the Goodness of Fit test
- The main difference is that the *null hypothesis* comes directly from the assumption of independence