UTAustinX: UT.7.01x Foundations of Data Analysis

KarenWest ▼

Courseware Course Info Discussion Syllabus Download R and RStudio R Tutorials Readings Contact Us
Progress Office Hours Community

CHI-SQUARE TEST-OF-INDEPENDENCE

0:00 / 10:09 1.0x

SPEAKER: MICHAEL J. MAHOMETA, Ph.D.

The chi-square goodness-of-fit test is a great way

to examine the distribution of a single categorical variable.

But what if we wanted compare two categorical variables?

Well sure, we can examine the marginal and conditional distributions.

But can we expand the idea of examining observed and expected counts

Help

into a contingency table?

Can we move into a hypothesis test?

Well sure we can.

All it takes is second categorical variable.

Notice the wording of our question: is one categorical variable INDEPENDENT

of another categorical variable?

That's important.

Because we'll be examining the data with a chi-square test of independence.

Let's go ahead and take a look at the data.

Here's data of respiratory event behaviors

in a table of counts from a recent article.

We see behaviors that college students use during a cough or a sneeze,

such as using a tissue, or using their arm or elbow, using their hand,

or using nothing at all.

Now, we know how to examine this distribution of a single categorical

variable of behavior - a chi-square test for goodness of fit.

We can use the observed counts and the expected counts

from a hypothesized distribution to run the hypothesis test.

But let's expand our data.

Help

In the original article, the researchers also

recorded another categorical variable

at the time of the respiratory event - the gender of the student.

Nice right?

Now, it's an easy transition from a table of counts to a contingency table.

And we can see that things are a little different.

If we use our row percentages, we see that the behavior distribution

looks different for each gender.

Perhaps the behavior is not in fact independent of gender

- it looks like gender might be driving a different behavioral response.

But how do we "test" our question?

Is the distribution of behavior of a respiratory event

independent of gender?

How do we "test the independence" of one categorical variable

from another categorical variable?

Well we use the chi-square "test-of-independence."

If we have categorical data, two categorical variables,

and we want to examine how the

distribution of one categorical

variable might differ when we consider another categorical variable -

we use the chi-square test-of-indepdence.

And we can say thank you to Karl Pearson.

Help

Let's talk about assumptions.

They are pretty typical for the chi-square test: Number ONE:

We need a random sample.

Number TWO: That sample needs to be made up of independent observations.

Number THREE is actually two assumptions rolled into one:

No expected count less than one at any of our cells; and no more than 20%

of our cells with an expected count less than five.

Now, on to the hypothesis test: Step ONE:

the null and alternative hypotheses.

Let's keep them simple: Our Null states that there is simply

no association between the two categorical variables (that they

are independent of one another).

The alternative is the opposite - that there

IS an association (that the two variables are in fact not independent).

Another way to think about this is that for

Help

the Null Hypothesis,

the relative proportions of one variable are independent of the second variable.

Now if that's true, the proportions or percentages of one variable

would be the same for different values - or levels - of the second variable.

Hypothesis step

number TWO: our alpha level.

Again it's 0.05, the standard.

Hypothesis step number THREE: our analysis.

We know that we'll be using the chi-square test -

and here's the formula.

It's the SAME as the Goodness-of-Fit test -

we'll still be using observed and expected counts -

and examining their relationship for each cell of our table.

So, we've got our observed counts...but what about the expected counts?

Where do they come from?

Well - to help out, we can use our marginal distributions.

Remember that idea of the Null - that the proportions or percentages

of one variable would be the same for

different values

of the second variable?

Let's take a look at the contingency table again.

Can we see what the proportions of behavior SHOULD

Help

BE if it was not associated with gender -

by looking at the marginal distribution of behavior?

Let's run with this.

Here's the marginal distribution for the behavior variable

- and here are the corresponding proportions.

So, if all were right with the world, and gender was not related to behavior,

then it stands to reason that we should see

the same proportions for each value of gender.

So for the expected count, if there's a marginal proportion

for "no cover" of 0.2336, then we would "expect"

our females to have that proportion and our males to also have that proportion.

So to get those expected counts - we take the row count

and multiply by the marginal proportion.

Help

In the case of "no cover," those would be 61.67 and 27.33.

Again, it doesn't matter if our expected counts aren't whole numbers.

Now, we can do the same procedure to get the expected

counts for all the remaining cells.

Once that's done, we can move on to using our chi-square formula.

We take the observed count and subtract out the expected count for each cell.

Then we square that difference.

Then, divide out the expected count of the cell.

And, finally, we sum across all of our cells.

With our data, we have a chi-square statistic of 18.25.

Now, let's find our degrees of freedom.

For the chi-square test of independence, the degrees of freedom formula

is: the number of levels of one categorical variable minus one,

times the number of levels in the OTHER categorical variable minus one.

We can think of this as rows minus 1 times columns minus 1.

Think of it as a degree of freedom for Row and a degree of freedom for Column

- then combine the two with multiplication.

For our contingency table, we have 4 minus 1 times 2 minus 1,

so 3 degrees of freedom.

The chi-square critical value at our chosen alpha level

Help

for 3 degrees of freedom is 7.81.

Now on to the fourth step of the hypothesis test: our conclusion.

Since our chi-square statistic is greater than our critical chi-square,

we Reject the Null Hypothesis.

The two variables are NOT independent of one another.

And we again can readily see this if we look at the row percentages:

Use of tissue and elbow are about the same for males and females.

But females tend to use their hand more than males,

and males tend to use nothing at all more than females.

Houston, we have a problem.

Notice the issue - a violation of an assumption.

Not only do we have a cell with an expected count less than 1,

but two cells of eight have an expect count

Help

less than 5.

That violates our 20% rule.

Now, we have two choices here - We can continue on, and tell our reader

that we have actually violated one of our assumptions of chi square

and that our results may be off.

Or we can try and fix it.

Now, when I try a fix, it's usually to combine levels

so that the expected counts are better.

And that combination needs to make sense.

The other option is to loose some data.

Here - I think our best bet is to loose the data of "tissue"

- only 2 people (and they were males) are

in this actual category of behavior.

Now, if we were to do that, we would tell our reader

(we need to provide full disclosure to them), and we re-run our analysis.

So let's do that and run this analysis in R.

Here's the new contingency table, called "events" -

notice there's no "tissue" group here.

And if we run this in R using the chi-square test function, we see that our new - unviolated chi-square is 17.31,

with a p-vlaue that is less than 0.001, which is under our alpha level.

So again, we can reject our null hypothesis.

Help

1. Here is the data from a research study on the relationship between sex and fear of heights.

	Men	Women
Expressed a Fear of Heights	68	109
Did Not Express a Fear of Heights	94	89

$$\chi^2 = 6.09,\ \ p\ <\ 0.05$$

(4 points possible)

1a. What are the expected values for each of the following groups?

Men who expressed a fear of heights: *(Round to 2 decimal places.)*

Help

Answer: 79.65

Women who expressed a fear of heights: *(Round to 2 decimal places.)*

Answer: 97.35

Men who did not express a fear of heights: *(Round to 2 decimal places.)*

Answer: 82.35

Women who did not express a fear of heights: *(Round to 2 decimal places.)*

Answer: 100.65

Hide Answer

Help

(1 point possible)

1b. Which of the following best summarizes how the observed values compare to the expected values?

- There were more women afraid of heights than men.
- Men and women both appear to be less afraid of heights than would be expected.
- Fewer men expressed a fear of height than expected, and more women expected a fear of heights than expected.

Hide Answer

(3 points possible)

1c. The researchers found a significant difference between men and women. It appears that sex is related to a fear of heights. We should see that difference reflected in the conditional probabilities. Fill in those values below.

Proportion of the entire sample that has a fear of heights: *(Round to 3 decimal places.)*

Help

Answer: 0.492

Proportion of the men that have a fear of heights: *(Round to 3 decimal places.)*

Answer: 0.420

Proportion of the women that have a fear of heights: *(Round to 3 decimal places.)*

Answer: .551

Hide Answer

Help