12/1/2022

χ^2 tests

We've seen some examples throughout the course where some statistics can be distributed using the χ^2 statistic. These tests are not usually grouped together with tests in this class. In this section of the course, we are going to be looking at categorical data or binned data.

Goodness of Fit Tests

We are going to start with goodness of fit tests for one categorical variable, or for one binned (possibly discrete) numerical variable.

Consider a genetics experiment crossing breeds of peas. The dominant alleles in the experiment were Y= yellow and R=round, resulting in the double dominant YR. Yule examined 269 four seed pods resulting from a dihybrid cross and counted the number of YR seeds in each pod. Let X denote the number of YRs in a randomly selected pod, with possible X values are $\{0, 1, 2, 3, 4\}$. The hypothesis that the Mendelian laws are operative and that genotypes of individual seeds within a pod are independent of one another implies that X has a binomial distribution with $n = 4$ and $p = \frac{9}{16}$. We are testing a series of proportions $p_i = p_{i0}$, where $p_{i0} = \binom{4}{i} p^i (1-p)^{4-i}$, $i = 0, 1, 2, 3, 4$, $p = \frac{9}{16}$. Suppose we obtain the data in the table.

Number of YR	0	1	2	3	4
Observed	16	45	100	82	26

We need observations to compare these to, so we use the probabilities from our formula, multiplied by the total sample size to obtain the number of observations we predict.

When we do that, we obtain a table of expected observations.

Number of YR	0	1	2	3	4
Observed	16	45	100	82	26
Expected	9.86	50.68	97.75	83.78	26.93

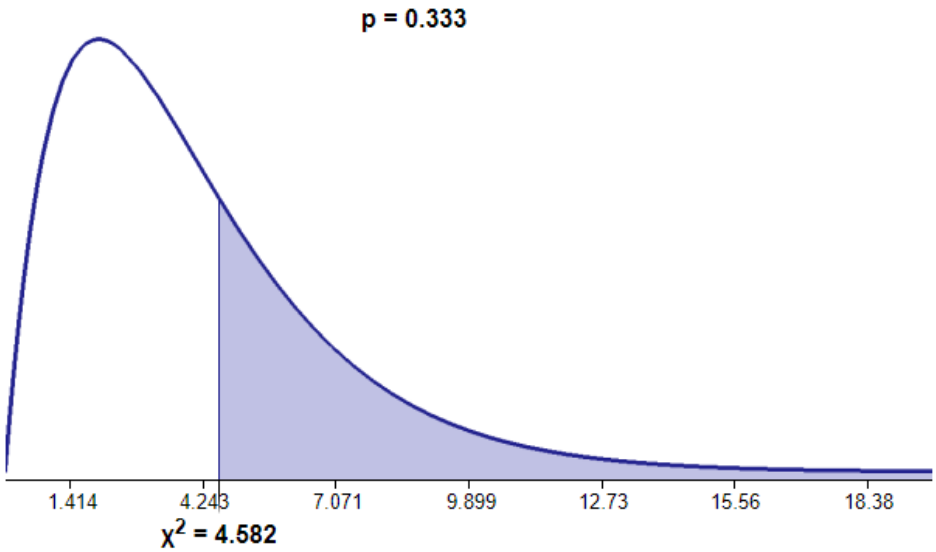
Keep in mind that while these are expected counts, they are averages and therefore do not need to be whole numbers. Even if they were, since the variable is random, we should not expect the match-up to be perfect. Random implies variability. What we want to know is whether or not our data fits the expectations well, or poorly. Our test statistic is going to measure the relative (absolute) differences between what we expected and what we got.

$$\chi^2 = \sum_{i=0}^{4} \frac{(O_i - E_i)^2}{E_i}$$

Where O is the observed values and E is the expected values. In this case, we find that

$$\chi^2 = \frac{(16-9.86)^2}{9.86} + \frac{(45-50.68)^2}{50.68} + \frac{(100-97.75)^2}{97.75} + \frac{(82-83.78)^2}{83.78} + \frac{(26-26.93)^2}{26.93} = 4.582$$

We tend calculate the P-value by finding $P(\chi^2 \geq 4.582)$ using 4 degrees of freedom (there are 5 categories of observations, and so we use 5-1=4 degrees of freedom).



Since the P-value is greater than any commonly used significance level, we can claim that a binomial distribution with $p = \frac{9}{16}$ is a reasonable distribution for this data.

If the underlying distribution is continuous, then we need to bin the data. So, for instance, suppose we grouped continuous women's height data into the following bins:

$$\{< 56, 57\text{-}58", 59\text{-}60", 61\text{-}62" \ 63\text{-}64", 65\text{-}66", 67\text{-}68", >69\}$$

We would calculate the probability for the first category as a normally distributed variable with a mean of 64" and a standard deviation of 3", we would find $P(X \leq 56.5)$. Note that the cutoff is between the endpoints of the categories similar to what we did with our binomial approximation. So, the second category was $P(56.5 \leq X \leq 58.5)$ and so forth. The rest of the test would proceed in the same way, with 8-1=7 degrees of freedom. There are a number of ways to test for normality, and this is just one.

If we are using a discrete distribution for counts, the probabilities in the cells are calculated from the distribution exactly, but in the final cell we include all probabilities that remain (the probabilities must add to 1).

Tests for Independence and Homogeneity for Two-Way (Contingency) Tables

Suppose that we have a two-way table (crosstabs) shown below comparing sports preferences among three sports and gender. We want to know whether gender has an effect on sports preference. Both gender and sport are categorical variables, so the data in the table represent counts in each of the categories.

The display of this table includes the totals for each row and column, but when we consider the size of the table for purposes of calculating degrees of freedom, only the body of the table without the total column or total row are included there.

		Sport Preference			
		Archery	Boxing	Cycling	
Gender	Female	35	15	50	100
	Male	10	30	60	100
		45	45	110	200

When we calculate the χ^2 test statistic, we use a similar formula to the goodness-of-fit tests, but the way we calculate the expected value changes.

$$\chi^2 = \sum_{i,j=0}^{m,n} \frac{\left(O_{ij} - E_i i\right)^2}{E_{ij}}$$

For the test of homogeneity asserts that all the probabilities are identical. In other words, p_{ij} is the same for every combination of i and j. If our table is $m \times n$ (m rows and n columns), then every probability is $p_{ij} = \frac{1}{mn}$. In our example above, $p_{ij} = \frac{1}{6}$. To obtain the expected values, we would multiply the probability by the grand total (the number of total observations). The values are shown in the table below.

	Archery	Boxing	Cycling
Female	33.33	33.33	33.33
Male	33.33	33.33	33.33

Our χ^2 test statistic is thus

$$\chi^2 =$$

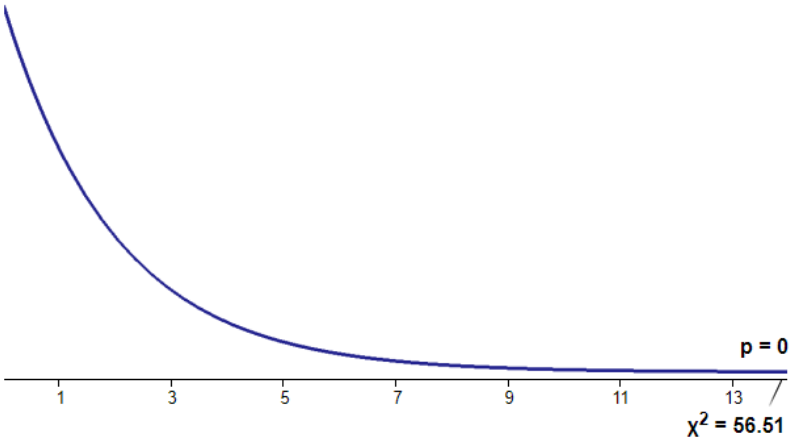
$$\frac{(35-33.33)^2}{33.33} + \frac{(15-33.33)^2}{33.33} + \frac{(50-33.33)^2}{33.33} + \frac{(10-33.33)^2}{33.33} + \frac{(30-33.33)^2}{33.33} + \frac{(60-33.33)^2}{33.33}$$

$$= 56.51$$

The degrees of freedom for our test are $(m-1)(n-1) = (2)(1) = 2$.

Putting that into our χ^2 distribution gives us a P-value which is very nearly zero. I think this was pretty obvious looking at our table this this was unlikely given the very different observations we have in the different categories.

The test for independence is similar except for the way we calculate the expected values to compare with our observations.



To calculate the counts for the test of independence, we want to think about our assumptions for independent probabilities from earlier in the course. Recall that $P(A \text{ and } B) = P(A)P(B)$. From our table, we would expect that $P(female \text{ and } archery) \approx P(female) \times P(archery)$ if they are independent, or in other words $\frac{35}{200} = 0.175 = P(F \text{ and } A) =? P(F)P(A) = \left(\frac{100}{200}\right)\left(\frac{45}{200}\right) = \frac{9}{80} = 0.1125$. As we can see, these probabilities are not identical, but because we are trying to infer to the population and not just describe this table, we need to measure the effect of randomness on this outcome. How far off is this? Is it similar enough to our expectations to infer that the population is independent, or is this good enough evidence conclude that the variables are indeed dependent? Our null hypothesis must be independence because independence is the equality claim. We can disprove equality, not prove equality.

To fill in our expected count, we multiply these probabilities by the grand total. For Female and Archery, we then get the expected count to be $\frac{9}{80} \times 200 = 22.5$. If we go back to the original calculation, we can obtain a formula for the general rule.

$$\left(\frac{100}{200}\right)\left(\frac{45}{200}\right) \times 200 = 22.5$$

$$E_{FA} = \frac{female\ count}{grand\ total} \times \frac{archery\ count}{grand\ total} \times (grand\ total) = (female\ count) \times \frac{archery\ count}{grand\ total}$$

$$E_{ij} = (row\ count) \times \frac{(column\ count)}{grand\ total}$$

Following in this pattern, we can complete the expected table.

	Archery	**Boxing**	**Cycling**
Female	22.5	22.5	55
Male	22.5	22.5	55

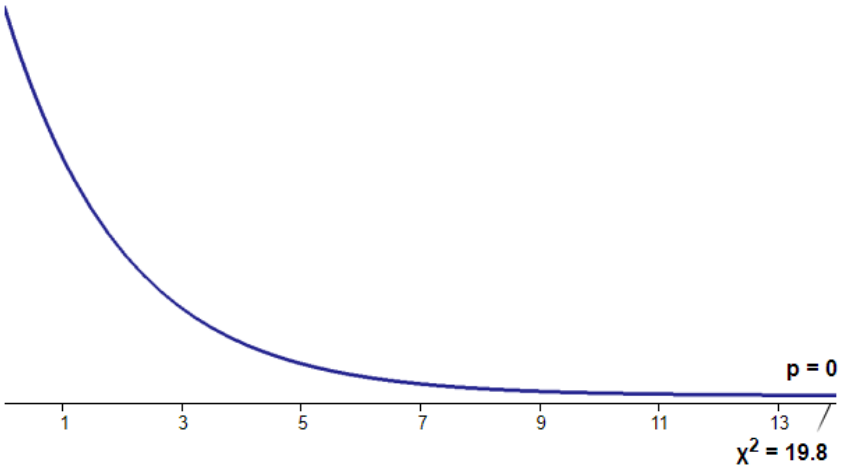
We can calculate our χ^2 test statistic from there.

$$\chi^2 =$$

$$\frac{(35-22.5)^2}{22.5}+\frac{(15-22.5)^2}{22.5}+\frac{(50-22.5)^2}{22.5}+\frac{(10-22.5)^2}{22.5}+\frac{(30-55)^2}{55}+\frac{(60-55)^2}{55}$$

$$=19.80$$

We can then calculate our P-value.



It's not as small as the test of homogeneity, but still highly unlikely that these variables independent. I think this is consistent with our intuition.

To conduct this test in R, you would need to input a summary table like the contingency table above. This can be done directly or by summarizing the raw dataframe in R. We'll examine how to do this in the last lab assignment.

In the next lecture, we'll look at non-parametric tests of the same situation(s).

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
3. http://www.statdistributions.com/chisquare/
4. https://www.datacamp.com/tutorial/contingency-analysis-r