Checking for Independence in Samples

An advertising agency in a large city is conducting a survey of adults to investigate whether there is an association between highest level of educational achievement and primary source for news. The company takes a random sample of 2,500 adults in the city. The results are shown in the table below.

	Highest Level of Educational Achievement			
Primary Source For News	Not a high school graduate	High school graduate but not a college graduate	College graduate	Total
Newspaper	49	205	188	442
Local Television	90	170	75	335
Cable Television	113	496	147	756
Internet	41	401	245	687
None	77	165	38	280
Total	370	1437	693	2500

Know how to find a probability, a union probability (A or B), and a conditional probability using a two way table, (11, 2a; 10b, 5ab; 03b, 2ab; M7, 18)

If an adult is selected at random from this sample, what is the probability that the selected adult obtains news primarily from the internet?

$$P(Internet) = \frac{687}{2500} = 0.2748$$

If an adult is selected at random from this sample, what is the probability that the selected adult is a college graduate or obtains news primarily from the internet?

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(College \cup Internet) = P(College) + P(Internet) - P(College \cap Internet)$$

$$= \frac{693}{2500} + \frac{687}{2500} - \frac{245}{2500} = \frac{1135}{2500} = 0.454$$

If an adult who is a college graduate is to be selected at random from this sample, what is the probability that the selected adult obtains news primarily from the internet?

We are only looking at college graduates

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(Internent|College) = \frac{P(Internet \cap College)}{P(College)} = \frac{\frac{245}{2500}}{\frac{693}{2500}} = \frac{245}{693} = 0.3535$$

Know the two ways to check for independence (11, 2b; 10b, 5c; 03b, 2c)

When selecting adults at random from the sample of 2500 adults are the events "is a college graduate" and "obtains news primarily from the internet" independent?

Method 1

$P(A|B) = P(A) \rightarrow P(Internet|College) = P(Internet) \rightarrow 0.3535 \neq 0.2748$

Since these are not equal, they are not independent. It appears that level of education does impact the likelihood of using the internet for news

Method 2

$$P(A \cap B) = P(A) \cdot P(B) \rightarrow P(Intenet \cap College) = P(Internet) \cdot P(College) \rightarrow \frac{245}{2500} \neq \frac{687}{2500} \cdot \frac{693}{2500}$$

Since these are not equal, they are not independent. It appears that level of education does impact the likelihood of using the internet for news

Chi-Squared Test of Independence of a population

An advertising agency in a large city is conducting a survey of adults to investigate whether there is an association between highest level of educational achievement and primary source for news. The company takes a random sample of 2,500 adults in the city. The results are shown in the table below.

Highest Level of Educational Achievement

Primary Source For News	Not a high school graduate	High school graduate but not a college graduate	College graduate	Total
Newspaper	49	205	188	442
Local Television	90	170	75	335
Cable Television	113	496	147	756
Internet	41	401	245	687
None	77	165	38	280
Total	370	1437	693	2500

Do the data provide convincing statistical evidence that there is an association between primary source for news and highest level of educational achievement for adults in this city?

Know when to use a chi-squared test of independence and how to find the degrees of freedom.

This is a chi-squared test of independence, because…

1) Categorical data
2) Two variables with multiple choices (Two-way table)
3) Determining if there is an association (not independent)

Know how to write a null and alternative hypothesis for a chi-square test of independence (13, 4; 11b, 4a; 09, 1c; 04, 5a; 03, 5)

H_0: There is no association between source for news and highest level of education (Source and education are independent)

H_a: There is an association between sources of news and highest level of education (Source and education are not independent)

Know how to find the expected outcomes for a chi squared test of independence, (M2, 11)

The null hypothesis claims that the two variables are independent. If that is the case, then…

$$expected\ cell\ value = \frac{row \cdot column}{n}$$

For example, the expected value for non high school graduate who use the newspaper for their news would be

$$expected = \frac{(442)(370)}{2500} = 65.416$$

If these variable are independent, then we would expect 65.416 people in this sample to be non graduates who use the paper for their primary news source

The table below shows the expected counts for each cell. Notice the top left has the 65.416 that we calculated on the previous page

Highest Level of Educational Achievement

Primary Source For News	Not a high school graduate	High school graduate but not a college graduate	College graduate	Total
Newspaper	65.416	254.06	122.52	442
Local Television	49.58	192.56	92.862	335
Cable Television	111.89	434.55	209.56	756
Internet	101.68	394.89	190.44	687
None	41.44	160.94	77.616	280
Total	370	1437	693	2500

Know how to check the conditions for a chi-squared test of independence, (13, 4; 11b, 4b; 04, 5a, 03, 5)

The two conditions for a chi-squared test of independence are

1) Random sample
2) All expected counts > 1 and at least 80% of expected counts > 1

For our problem, we were told that it was a "random sample." Looking at our table on the previous page, all of the expected counts are larger than 5 with the smallest being 41.44, so we have met both conditions.

Know how to find a chi squared statistic, the degrees of freedom, and the P-value for a chi squared test of independence, (13, 4; 04, 5a; 03, 510b, 5d)

$$X^2 = \sum \frac{(Observed-Expected)^2}{Expected} = \frac{(49-65.416)^2}{65.416} + \frac{(205-253.06)^2}{253.06} + \frac{(188-122.52)^2}{122.52} + \cdots + \frac{(38-77.616)^2}{77.616}$$

$$= \quad 4.20 \quad + \quad 9.13 \quad + \quad 35.00 \quad +\ldots+ \quad 20.22$$

$$= 217.77 \text{ (from calculator)}$$

The degrees or freedom are

Df = (r – 1)(c – 1) = (5 – 1)(3 – 1) = 8

P value: $1.14\ x\ 10^{-42} \approx 0$ (from calculator)

Know how to draw a conclusion in context for a chi squared test of independence, (14, 1c; 13, 4; 11b, 4c)

We reject the null hypothesis at the 1% significance level, because 0 < 0.01. There is evidence that there is a relationship between level of education and an individual's source for news in this city. There is statistical evidence to say that the two variables are NOT independent.

THE SIX STEPS ON THE PREVIOUS TWO PAGES WOULD HAVE TO BE DONE FOR ANY CHI SQUARED TEST OF INDEPENDENCE. YOU MAY USE THE MATRIX FUNCTION AND CHI SQUARED TEST FUNCTION ON YOUR CALCULATOR INSTEAD OF DOING WORK

Know how to interpret a P-value for a chi squared test of independence, (04, 5a; 03, 5)

General: If the null hypothesis is true, there is a (p-value)% chance of randomly getting the data that we got or more extreme.

Specific: If level of education and source for news are independent, there is a 0% chance that we would have gotten the data that we got in our sample or data more extreme.

Chi Squared Goodness of Fit Test

The distribution of Algebra 1 grades in a large school district with several high schools is shown below.

Grade	A	B	C	D	F
Proportion of students with this grade	0.12	0.32	0.26	0.18	0.12

55 Randomly chosen students from one of the high schools showed the distribution below

Grade	A	B	C	D	F
Number of students with this grade	20	15	13	4	3

The district claims that all high schools have the same grade distributions in Algebra 1. Is this data consistent with this belief? Perform an appropriate statistical test to support your conclusion.

This is a chi-squared goodness of fit test

Know how to write a null and alternative hypothesis for a goodness of fit test, (08, 5a; 03b, 5c)

H_0: The distribution of grades at this school is the same as the district

H_a: The distribution of grades at this school is not the same as the district

Know how to find expected counts for a GOF test if each possibility has a different proportion and check conditions, (08, 4a; 03b, 5c; M2, 19)

The null hypothesis says the schools grades should match the district. If that were the case, we should have 12% of our 55 get an A. So, the expected count for A is (0.12)(55) = 6.6. All expected counts are shown below

Grade	A	B	C	D	F
Expected number of students	6.6	17.6	14.3	9.9	6.6

Our conditions are

1) Random-"random sample"
2) All expected > 1, at least 80% of expected > 5- Both conditions are met. The smallest expected is 6.6

Know how to find a chi squared statistic, the degrees of freedom, and the P-value for a GOF test, (08, 4a; 06, 6c; 03b, 5c; M7, 17; M2, 19)

$$X^2 = \sum \frac{(Observed-Expected)^2}{Expected} = \frac{(20-6.6)^2}{6.6} + \frac{(15-17.6)^2}{17.6} + \frac{(13-14.3)^2}{14.3} + \frac{(4-9.9)^2}{9.9} + \frac{(3-6.6)^2}{6.6}$$

$$= 27.21 + 0.38 + 0.12 + 3.52 + 1.96$$

$$= 33.19$$

Degrees of Freedom = k – 1 = 5 – 1 = 4

$P\,Value \approx 0$

Know how to draw a conclusion in context for a GOF test, (08, 4a; 03b, 5c)

We reject the null hypothesis in favor of the alternative at the 1% significance level, because 0 < 0.01. There is evidence that this school's distribution of Algebra 1 grades is different than the district's.

THE FOUR STEPS ABOVE MUST BE DONE FOR EVERY CHI SQUARED GOODNESS OF FIT TEST. YOU MAY USE YOUR CALCULATOR TO FIND THE STATISTIC AND P-VALUE, BUT NOT ALL CALCULATORS HAVE THAT FUNCTION.

Know how to interpret a P-value for a GOF test in the context of the problem, (06, 6f)

Using the context form the previous problem: Assuming that the distribution of grades at this school is the same as the district, there is a 0% chance that we would have randomly gotten the data that we got or more extreme.

What does an individual chi squared value represent? Does it tell you if the expected was higher or lower? (08, 4b)

Looking at our calculation in the problem on the previous page, we can see that the biggest difference in grades occurs with the A's. One way you can see that, is that it has the largest contributor (27.21) to the chi squared statistic of 33.19. So, that is where the biggest difference occurs. The chi squared statistic though, does not necessarily tell you the direction of the difference because it is a square number. We can see that the observed is 13.4 above the expected, but you would have gotten the same statistic (27.21) even if it was 13.4 below the expected. So, it tells us there is a difference, it just doesn't tell us the direction.

OTHER STUFF

Know how to find union probabilities of disjoint events, (09b, 2bc;08, 3cd; 04, 4a)

The ELISA tests whether a patient has contracted HIV. The ELISA is said to be positive if it indicates that HIV is present in a blood sample, and the ELISA is said to be negative if it does not indicate that HIV is present in a blood sample. Because of variability in antibody levels among human patients, the ELISA does not always indicate the correct result.

As part of a training program, the staff applied the ELISA to 500 blood samples known to not contain HIV. The ELISA was positive for 37 of those blood samples and negative for the other 463 samples.

When a blood sample yields a positive ELISA result, two more ELISAs are performed on the same blood sample. If at least one of the two additional ELISAs is positive, the blood sample is subjected to a more expensive and more accurate test to make a definitive determination of whether HIV is present in the sample. Under the assumption of independence and using the data above, what is the probability that a new blood sample that comes into the lab will be subjected to the more expensive test if that sample does not contain HIV?

$P(positive) = \frac{37}{500} = 0.074$

We will have to run the more expensive test for one of the two situations listed below

1) PP (2 positive tests in a row)
2) PNP (First positive, then a negative, but third test positive)

Since we are assuming independence, a positive results will always have 0.074 probability.

The probability of option 1, positive on first test and then another positive test is

P(positive then positive) = (0.074)(0.074) = 0.005476

The probability of option 2, positive on the first test, negative on second test, positive on third test is

P(positive then negative then positive) = (0.074)(0.926)(0.074) = 0.005070776

So, the probability of having to run the more expensive test if you are HIV negative is

= 0.005476 + 0.005070776
= 0.010546776

Know how to find a joint probability (A and B) of two or more independent events, (08, 3b; 04, 3b; 04, 4a; 03b, 5a; M7, 38; M2, 36)

In the problem above, each test is independent

$P(Positive \cap Postivie) = 0.074 \cdot 0.074$ is the joint probability of two independent events

Know how to interpret a probability in the context of a problem and deem if it is likely to occur or not, (04, 3c)

If I have a 0.0001% chance of winning a game, you would determine that this probably would not happen randomly. You would call me a cheater if I won.

If I have a 25% chance of winning a game, you would determine that this could happen randomly. You would not have strong evidence to call me a cheater if I won.

Know what it means for events to be mutually exclusive and how that affects their joint probability and their union probability, (M7, 36; M2, 23)

If two events are mutually exclusive, they cannot happen at the same time. For example, I cannot be in ninth grade and tenth grade at the same time. If two events are mutually exclusive, $P(A \cup B) = P(A) + P(B)$. You do not have to worry about $P(A \cap B)$ because $P(A \cap B) = 0$ for disjoint events. They cannot occur simultaneously.

Know how to set up a simulation and perform it using a table of random digits, (M2, 4)

If 80% of population is female and 20% male, you could use a RNG and make

0,1→Male 2, 3, 4, 5, 6, 7, 8, 9→Female