Activity #11: Chi-square goodness-of-fit test and tests for independence Resources: sports.sav

1. Read the article on *Benford's Law*. Suppose we gather 315 numbers from a tax return. How many 1's do we expect to find as leading digits? How many 2's? Complete the table:

Leading Digit	1	2	3	4	5	6	7	8	9	
Expected Relative Frequency	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	
Expected Frequency										N = 315

2. The actual leading digits from our observed tax return have the following frequencies. Do you believe this tax return is fraudulent? What display can you use to visualize the comparison between the observed leading digits and Benford's Law?

Leading Digit	1	2	3	4	5	6	7	8	9
Observed Frequency	82	38	32	30	50	22	20	21	20

3. Before we determine if this tax return is fraudulent, let's look at a simpler example. Suppose we toss a die 120 times and observe the following results. We're interested in determining whether or not this die is fair. What are the expected frequencies and relative frequencies for each side of the die?

	1	2	3	4	5	6	
Expected Relative Frequency							
Expected Frequency							
Observed	16	23	28	19	19	15	N = 120

4. We need to derive a method for determining how "far off" our observations were from our expectations. How would you determine how far off from the expectations this die is?

5. Use your new method to compute a “distance from what we expect” index for the following two tables. Are these tables the same “distance” from their expectations? Should they be?

	1	2	3	4	5	6	
Expected Frequency							
Observed	4	6	5	5	4	6	N = 30

	1	2	3	4	5	6	
Expected Frequency							
Observed	400	600	500	500	400	600	N = 3000

To test the distribution of a categorical variable, we use a Chi-square goodness-of-fit test.

The test statistic is: $\chi^2_{r-1} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ for rows i and columns j.

6. Run a chi-square goodness-of-fit test to determine whether or not the die in question #3 is fair.

7. Run a chi-square goodness-of-fit test to determine whether or not the tax return numbers are faked.

Oftentimes in statistics, we must make assumptions about the underlying distribution of a random variable. For example, to run an ANOVA, we must assume the data are normally distributed. We can use the chi-square goodness-of-fit test to check the underlying distributional assumptions.

Note: These goodness-of-fit tests are not the best tools to use for this purpose. We can also use normal probability plots, quantile plots, Kolmogorov-Smirnoff tests, and simple histograms to check the validity of the underlying distributional assumptions.

Recall the following properties of the **exponential distribution**:

It is often used to model waiting times

The cumulative distribution function is: $1 - e^{-\lambda y}$

$$E(X) = \frac{1}{\lambda} \text{ and } Var(X) = \frac{1}{\lambda^2}$$

8. The following data represent the time intervals between breakdowns of a machine. Is the exponential distribution an appropriate model for this data?

0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.5	0.6	0.6
0.6	0.6	0.7	0.7	0.8	0.9	0.9	1.0	1.1	1.2	1.3
1.3	1.6	1.7	1.8	1.9	1.9	2.0	2.0	2.1	2.2	2.2
2.4	2.4	2.4	2.7	2.8	2.8	2.8	3.0	3.1	3.5	3.7
3.7	3.7	4.1	4.1	4.2	4.2	4.3	4.3	4.5	4.9	4.9
4.9	5.0	5.2	5.3	5.3	5.7	5.7	5.9	6.1	6.2	6.2
6.2	6.3	7.4	7.5	7.5	7.7	8.1	8.6	9.2	9.5	10.0
10.3	10.6	10.9	12.3	13.9	13.7	13.9	14.8	14.9	17.3	17.6

Step #1: Calculate the distribution parameters.

In this case, we need to calculate λ. We know $E(X) = \frac{1}{\lambda}$, so we can find the average of the observations.

$\overline{X} = 4.69 = \frac{1}{\lambda}$, so $\lambda = \frac{1}{4.69} = 0.213$

Step #2: The chi-square test required categorical data, so we must break our continuous data into bins of equal width.

	0-3	3-6	6-9	9-12	12-15	15-18	Total
Observed	40	23	11	6	6	2	88

Step #3: We now must calculate the expected frequencies in each bin

Bin	Expected Relative Frequency	Probability	Expected Frequency
0-3	P(X<3)	$1 - e^{-.213(3)} = 0.472$	$= 0.472(88) = 41.55$
3-6	P(3<x<6)	$\left[1 - e^{-.213(6)}\right] - \left[1 - e^{-.213(3)}\right] = 0.249$	$0.249(88) = 21.95$
6-9	P(6<x<9)	$\left[1 - e^{-.213(9)}\right] - \left[1 - e^{-.213(6)}\right] = 0.1315$	$0.1315(88) = 11.58$
9-12	P(9<x<12)	$\left[1 - e^{-.213(12)}\right] - \left[1 - e^{-.213(9)}\right] = 0.069$	$0.069(88) = 6.11$
12-15	P(12<x<15)	$\left[1 - e^{-.213(15)}\right] - \left[1 - e^{-.213(12)}\right] = 0.037$	$0.037(88) = 3.23$
15-18	P(15<x<18)	$\left[1 - e^{-.213(18)}\right] - \left[1 - e^{-.213(15)}\right] = 0.041$	$0.041(88) = 3.61$

Step #4: Now we calculate the chi-square statistic:

	0-3	3-6	6-9	9-12	12-15	15-18
Observed	40	23	11	6	6	2
Expected	41.55	21.95	11.58	6.11	3.23	3.61
Obs-Exp	-1.55	1.05	-0.58	-0.11	2.77	-1.61
$(obs - \exp)^2$	2.4025	1.1025	0.3364	0.0121	7.6729	2.5921

Sum = 14.1185

Compare to a chi-square with (rows – estimated parameters – 1) degrees of freedom = (6 – 1 – 1) = 4.

According to our table, the critical value is 13.277 at a 0.01 level of significance and 14.860 at a 0.005 level of significance.

Therefore, the p-value for this study is: $0.005 < p < 0.01$

We've seen how to use the chi-square statistic to test the distribution of a single variable. We can also use the chi-square statistic to see if two categorical variables are independent. This is called the *chi-square test for independence*.

9. Explain in plain language what it means if two variables are independent. Then, give the definition of independence using probability notation.

10. Suppose variables A and B are independent, P(A) = 0.2, and P(B) = 0.4. Find $P(A \cup B)$ and $P(A \cap B)$.

11. Suppose we're interested in the relationship between two variables: (a) gender and (b) the outcome of a coin toss. We get 50 men and 50 women to toss a coin and record the result. Should gender and the outcome of a coin toss be independent of each other? Why or why not?

12. The chi-square test for independence is very similar to the goodness-of-fit test. Our first step is to write out our expectations for the variables. Complete the following 2x2 contingency table by writing the expected frequencies in each cell. Also write out the expected marginal frequencies for each row and column. How did you determine these probabilities?

	Male	Female	Total
Heads			
Tails			
Total			

13. Suppose we observe the following. Compute a chi-square statistic to determine if gender and coin toss result are independent.

	Male	Female	Total
Heads	30	10	40
Tails	20	40	60
Total	50	50	100

14. The chi-square test for independence is useful in many situations where we have two categorical variables. In a study of heart disease in male federal employees, researchers classified 356 volunteer subjects according to their socioeconomic status (SES) and their smoking habits. What can you conclude by examining the data?

	High SES	Mid SES	Low SES	Total
Current Smoker	51	22	43	116
Former Smoker	92	21	28	141
Never Smoked	68	9	22	99
Total	211	52	93	356

15. Before running a chi-square test, it's often worthwhile to calculate conditional probabilities. Since we're interested in the effect of SES on smoking habits, let's calculate the probability that an individual is a current smoker given he is either High, Middle, or Low SES. Compute similar conditional probabilities for the other smoking habits.

	High SES	Mid SES	Low SES	Total
Current Smoker				
Former Smoker				
Never Smoked				
Total				

16. We're interested in determining whether or not SES is related to smoking habits. Write out the null and alternate hypotheses for this study.

17. Now we need to calculate the expected cell frequencies. Recall that when two variables are independent, P(A∩B)=P(A)P(B). Using this, fill-in the table with your expected frequencies. Derive a general formula for calculating these expected frequencies.

	High SES	Mid SES	Low SES	Total
Current Smoker	E_{11} =	E_{12} =	E_{13} =	$E_{1\bullet}$ =
Former Smoker	E_{21} =	E_{22} =	E_{23} =	$E_{2\bullet}$ =
Never Smoked	E_{31} =	E_{32} =	E_{33} =	$E_{3\bullet}$ =
Total	$E_{\bullet 1}$ =	$E_{\bullet 2}$ =	$E_{\bullet 3}$ =	$E_{\bullet\bullet} = 356$

18. We can now run the test.

	High SES	Mid SES	Low SES	Total
Current Smoker	$O_{11} = 51$ $E_{11} = 68.75$	$O_{12} = 22$ $E_{12} = 16.94$	$O_{13} = 43$ $E_{13} = 30.31$	$O_{1\bullet} = 116$ $E_{1\bullet} = 116$
Former Smoker	$O_{21} = 92$ $E_{21} = 83.57$	$O_{22} = 21$ $E_{22} = 20.6$	$O_{23} = 28$ $E_{23} = 36.83$	$O_{2\bullet} = 141$ $E_{2\bullet} = 141$
Never Smoked	$O_{31} = 68$ $E_{31} = 56.68$	$O_{32} = 9$ $E_{32} = 14.46$	$O_{33} = 22$ $E_{33} = 25.86$	$O_{3\bullet} = 99$ $E_{3\bullet} = 99$
Total	$O_{\bullet 1} = 211$ $E_{\bullet 1} = 211$	$O_{\bullet 2} = 52$ $E_{\bullet 2} = 52$	$O_{\bullet 3} = 93$ $E_{\bullet 3} = 93$	$O_{\bullet\bullet} = 356$ $E_{\bullet\bullet} = 356$

	High SES	Mid SES	Low SES	Total
Current Smoker				
Former Smoker				
Never Smoked				
Total				

19. We can calculate additional statistics when we examine 2x2 contingency tables. In order to see if seat belts help prevent fatalities, records of the last 100 automobile accidents to occur along a particular highway were examined. These 100 accidents involved 242 persons. Each person was classified as using or not using seat belts when the accident occurred and as injured fatally or a survivor.

	Fatally Injured	Survived	Total
Seat Belt	7	89	96
No Seat Belt	24	122	146
Total	31	211	242

We'll begin by running a chi-square test for independence.

	Fatally Injured	Survived
Seat Belt	$O_{11} = 7$ $E_{11} = \frac{96(31)}{242} = 12.3$ $\chi^2 = 2.28$	$O_{12} = 89$ $E_{12} = \frac{96(211)}{242} = 83.7$ $\chi^2 = 0.34$
No Seat Belt	$O_{21} = 24$ $E_{21} = \frac{146(31)}{242} = 18.7$ $\chi^2 = 1.50$	$O_{22} = 122$ $E_{22} = \frac{211(146)}{242} = 127.3$ $\chi^2 = 0.22$

20. Another statistic we can calculate for 2x2 tables is the *phi coefficient*. This is an index that measures the degree of dependency between the two variables. A phi-coefficient is calculated as:

$\phi = \frac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}}$, where the letters represent cell frequencies:

	Variable #1		
Variable #2	a	b	r_1
	c	d	r_2
	c_1	c_2	

$$\phi = \frac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}} = \frac{(7)(122) - (24)(89)}{\sqrt{31(211)(146)(96)}} = \frac{-1282}{9574.897} = -0.134$$

21. We'll discuss the phi-coefficient more in the next activity. Now, let's look at another measure we can use: *odds ratios*.

The odds of event A occurring are defined as: $\frac{P(A)}{1-P(A)}$.

Calculate the odds of surviving for individuals who do and do not wear seatbelts.

	Fatally Injured	Survived	Total
Seat Belt	7	89	96
No Seat Belt	24	122	146
Total	31	211	242