Chapter 11: Inference for Distributions of Categorical Data

1. Which of the following statements about the chi-square (χ2) distribution is *true*?

A. The density curve of the χ2 distribution is symmetric.

AR. Incorrect. For any number of degrees of freedom, the density curve of the χ2 distribution will be skewed to the right.

\*B. As the number of degrees of freedom increases, the mean of the χ2 distribution increases.

BR. Correct. In fact, the mean of the χ2 distribution is equal to the degrees of freedom.

C. The mean of any χ2 distribution is less than the median.

CR. Incorrect. Since all χ2 distributions are skewed right, the mean is always *greater* than the median.

2. From experience, the owner of an ice-cream shop has found that 60% of all sales of ice-cream cone are for one-scoop cones, 30% are for two-scoop cones, and the remaining 10% are for three-scoop cones. Recently, the shop added frozen-yogurt cones to its menu. A random sample of 250 frozen yogurt-cone sales revealed the following distribution:

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of scoops** | one | two | three |
| **Number of sales** | 154 | 84 | 12 |

The owner wishes to know whether the pattern of sales for frozen-yogurt cones differs from that of ice-cream cones. Let *p*1, *p*2, and *p*3 be the proportions of sales of one-scoop, two-scoop, and three-scoop frozen-yogurt cones, respectively. Which one of the following pairs of null and alternative hypotheses are appropriate for this test?

\*A. *H*0: The distribution of the variable “number of scoops” is the same for frozen yogurt cones as it is for ice cream cones.

 *H*a: The distribution of the variable “number of scoops” is different for frozen yogurt cones than it is for ice cream cones.

AR. Correct. We are comparing the distribution of “number of scoops” for the population of all frozen yogurt cones to the given distribution of “number of scoops” for the population of all ice cream cones.

B. *H*0: The observed counts for “number of scoops” are the same as the expected counts.

 *H*a: At least one of the observed counts for “number of scoops” is different from the corresponding expected count.

BR. Incorrect. We are testing a characteristic of the *population* of all frozen yogurt cones, not a characteristic of the sample (the observed counts). Neither hypothesis should refer to the counts form this particular sample.

C. *H*0: Ice-cream cone sales and frozen-yogurt cone sales are independent.

 *H*a: Ice-cream cone sales and frozen-yogurt cone sales are dependent.

CR. Incorrect. These hypotheses are for a goodness-of-fit test, not for a test of independence.

3. From experience, the owner of an ice-cream shop has found that 60% of all sales of ice-cream cones are for one-scoop cones, 30% are for two-scoop cones, and the remaining 10% are for three-scoop cones. Recently, the shop added frozen-yogurt cones to its menu. A random sample of 250 sales of frozen-yogurt cones revealed the following distribution:

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of scoops** | one | two | three |
| **Number of sales** | 154 | 84 | 12 |

The owner wishes to know whether or not the pattern of sales for frozen-yogurt cones differs from that of ice-cream cones. For a goodness-of-fit test based on the χ2 distribution, which of the following is the expected count for the cell corresponding to two scoops?

A. 84.

AR. Incorrect. This value is the observed count for two scoops.

\*B. 75.

BR. Correct. The expected count is given by *np*2, where *n* = the sample size and *p*2 = the value of *p*2 hypothesized under *H*0. In this case, *n* = 250 and *p*2 = 0.3, so the expected count is (250)(0.3) = 75.

C. 83.33.

CR. Incorrect. The expected count is given by *np*2, where *n* = the sample size and *p*2 = the value of *p*2 hypothesized under *H*0. You used the value *p*2 = 1/3 rather than the correct value of 0.3. It is not always the case that outcomes are equally likely under *H*0 in a goodness-of-fit test!

4. From experience, the owner of an ice-cream shop has found that 60% of all sales of ice-cream cones are for one-scoop cones, 30% are for two-scoop cones, and the remaining 10% are for three-scoop cones. Recently, the shop added frozen-yogurt cones to its menu. A random sample of 250 sales of frozen-yogurt cones revealed the following distribution:

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of scoops** | one | two | three |
| **Number of sales** | 154 | 84 | 12 |

The owner wishes to know whether or not the pattern of sales for frozen-yogurt cones differs from that of ice-cream cones and performs a 2 goodness-of-fit test on the data. which of the following is the correct value of the 2 statistic?

A. 0.2855

AR. Incorrect. If you got this value, you used the formula, which is incorrect.

\*B. 7.947

BR. Correct. The calculation is 

C. 15.141

CR. Incorrect. If you got this value, you used the formula, which is incorrect.

5. A fast-talking salesman offers you a “random digits” smartphone app that comes with a “guarantee” of randomness. You don’t trust him, so you insist on testing the app by generating a sample of 100 digits. Here are the resulting frequencies of the ten digits 0 through 9 in your sample of size 100:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Digit** | 0 | 1 | 2 | 3 | 4 |
| **Frequency** | 12 | 11 | 11 | 4 | 9 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Digit** | 5 | 6 | 7 | 8 | 9 |
| **Frequency** | 9 | 12 | 10 | 10 | 12 |

With the salesman getting more nervous by the minute, you then conduct a goodness-of-fit test of the null hypothesis that the app’s digits are indeed random (that is, that the probability of seeing any of the ten digits 0 through 9 in a single observation is equal to 1/10 or 0.10) against the alternative that they are not random. Which one of the following analyses of the conditions for performing this test is correct?

A. Since the data do not come from a simple random sample of possible digits produced by the app, the conditions for performing the test have not been met.

AR. Incorrect. Under the assumption that the null hypothesis is true, any 100 digits produced by the app constitute a simple random sample.

 B. Since all the observed counts are not greater than 5, the conditions for performing the test have not been met.

BR. Incorrect. This is not a condition for performing this test. It’s the *expected* counts that should all be at least 5.

\*C. All conditions for performing this test have been satisfied.

CR. Correct. We have taken a random sample, and all the expected counts are at least 5. (The 10% condition is not relevant, since the population in this case—all digits the app produces—is essentially infinite.

6. A fast-talking salesman offers you a “random digits” smartphone app that comes with a “guarantee” of randomness. You don’t trust him, so you insist on testing the app by generating a sample of 100 digits. Here are the resulting frequencies of the ten digits 0 through 9 in your sample of size 100:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Digit** | 0 | 1 | 2 | 3 | 4 |
| **Frequency** | 15 | 7 | 11 | 4 | 6 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Digit** | 5 | 6 | 7 | 8 | 9 |
| **Frequency** | 9 | 12 | 10 | 14 | 12 |

With the salesman getting more nervous by the minute, you then conduct a goodness-of-fit test of the null hypothesis that the app’s digits are indeed random (that is, that the probability of seeing any of the ten digits 0 through 9 in a single observation is equal to 1/10 or 0.10) against the alternative that they are not random. The 2 statistic for this test is 11.2. Which of the following conclusions regarding the null hypothesis is appropriate?

A. Reject *H­­0* at either the .

AR. Incorrect. For a 2 distribution with 9 degrees of freedom, the *P-*value corresponding to is greater than 0.25 (specifically, 0.622), so we fail to reject the null hypothesis at either significance level.

B. Reject *H­­0* at ; fail to reject *H­­0* at 

BR. Incorrect. Are you using the correct degrees of freedom? (Number of categories – 1). Your *P*-value should be greater than 0.25 (specifically, 0.2622).

\*C. Fail to reject *H­­0* at either the .

CR. Correct. Your *P*-value should greater than 0.25 (specifically, 0.2622). Fail to Reject.

7. A study was done to examine the personal goals of elementary school children. A random sample of 500 students was selected from schools in the state of Georgia. The students received a questionnaire regarding personal goals. They were asked what they would most like to do at school: get good grades, be popular, or be good at sports. The results are presented in the table, classified by the gender of the child.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Get good grades** | **Be popular** | **Be good at sports** |
| **Boys** | 96 | 32 | 94 |
| **Girls** | 193 | 45 | 40 |

Which one of the following tables of conditional distributions is most appropriate for determining if gender has an impact on personal goals?

|  |  |  |  |
| --- | --- | --- | --- |
| A. | **Get good grades** | **Be popular** | **Be good at sports** |
| **Boys** | 0.192 | 0.064 | 0.188 |
| **Girls** | 0.386 | 0.090 | 0.080 |

AR. Incorrect. These are each cells proportions of the total sample, and are not as effective for comparison as conditional distributions for each value of the explanatory variable.

|  |  |  |  |
| --- | --- | --- | --- |
| B. | **Get good grades** | **Be popular** | **Be good at sports** |
| **Boys** | 0.332 | 0.416 | 0.701 |
| **Girls** | 0.668 | 0.584 | 0.299 |

BR. Incorrect. These are conditional distributions for each value of the response variable (personal goal), which are not as effective for comparison as conditional distributions for each value of the explanatory variable.

|  |  |  |  |
| --- | --- | --- | --- |
| \*C. | **Get good grades** | **Be popular** | **Be good at sports** |
| **Boys** | 0.432 | 0.144 | 0.423 |
| **Girls** | 0.694 | 0.162 | 0.144 |

CR. Correct. This table presents conditional distributions for each value of the explanatory variable (gender), the most effect way to compare differences between genders.

8. A study was done to examine the personal goals of elementary school children. A random sample of 500 students was selected from schools in the state of Georgia. The students received a questionnaire regarding personal goals. They were asked what they would most like to do at school: get good grades, be popular, or be good at sports. The results are presented in the table, classified by the gender of the child.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Get good grades** | **Be popular** | **Be good at sports** |
| **Boys** | 96 | 32 | 94 |
| **Girls** | 193 | 45 | 40 |

We wish to test the null hypothesis that gender and personal goal are independent. Under the null hypothesis, the expected number of boys that would select “be good at sports” is

\*A. 59.5.

AR. Correct. The formula for the expected count of a given cell in the table is  or for the boys/be good at sports cell, 

B. 67.

BR. Incorrect. You have incorrectly assumed that under the null hypothesis, the 134 children who chose “be good at sports” must be equally divided between boys and girls. Under that assumption, the expected cell count is 134/2 = 67. The correct formula for the expected cell count of a given cell is .

C. 74.

CR. Incorrect. You have incorrectly assumed that under the null hypothesis, the 222 boys must be equally divided between those who chose “get good grades,” “be popular,” and “be good at sports.” Under that assumption, the expected cell count is 222/3 = 74. The correct formula for the expected cell count of a given cell is .

9. A study was done to examine the personal goals of elementary school children. A random sample of 500 students was selected from schools in the state of Georgia. The students received a questionnaire regarding personal goals. They were asked what they would most like to do at school: get good grades, be popular, or be good at sports. The results are presented in the table, classified by the gender of the child.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Get good grades** | **Be popular** | **Be good at sports** |
| **Boys** | 96 | 32 | 94 |
| **Girls** | 193 | 45 | 40 |

We wish to test the null hypothesis that gender and personal goal are independent. Which one of the following is the correct 2 component for the “Boys/Be good at sports” cell?

\*A. 20

AR. Correct. The 2 component is the value of for any specific cell. For the “Boys/Be good at sports” cell, 

B. 34.5

BR. Incorrect. This is the difference between the observed and expected counts for indicated cell, not the component of the  statistic.

C. 50.88

CR. Incorrect. This is the  statistic—the sum of the components form all 6 cells.

10. In a study of the responsiveness of small businesses in different locations, questionnaires were sent to 200 randomly chosen small businesses in rural and small-town locations, 200 randomly chosen small businesses in suburban areas, and 200 randomly chosen small businesses in urban areas. The numbers of businesses of each type that responded to the questionnaire were recorded. Here is a table of the results.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Rural or small-town** | **Suburban** | **Urban** |
| **Responded** | 115 | 96 | 79 |
| **Did not respond** | 85 | 104 | 121 |

We wish to know whether the data provide sufficient evidence to conclude that the distributions of response rates for small businesses in these three locations are not all the same. The value of the X2 test statistic for the appropriate 2 test for homogeneity is 12.988. The *P*-value of this test is

\*A. between 0.001 and 0.0025. AR. Correct. The test statistic in this case has (*r* – 1)(*c* – 1) degrees of freedom, where *r* = the number of rows in the table and *c* = the number of columns in the table. Here, *r* = 2 and *c* = 3, so the number of degrees of freedom is (2 – 1)(3 – 1) = 2. Since the value 12.988 falls between 11.98 and 13.82, the upper 0.0025 and 0.001 critical values of the χ2 distribution with 2 degrees of freedom, the *P*-value is between 0.001 and 0.0025.

B. between 0.01 and 0.025

BR. Incorrect. The test statistic in this case has (*r* – 1)(*c* – 1) degrees of freedom, where *r* = the number of rows in the table and *c* = the number of columns in the table. You may have used the formula *rc* – 1 when calculating the number of degrees of freedom.

C. between 0.025 and 0.05.

CR. Incorrect. The test statistic in this case has (*r* – 1)(*c* – 1) degrees of freedom, where *r* = the number of rows in the table and *c* = the number of columns in the table. You may have used the formula *rc* when calculating the number of degrees of freedom.

11. A study was conducted to determine whether or not there was a relationship between a child’s birth order and his or her chances of exhibiting delinquent behavior. The subjects were a random sample of girls enrolled in public high schools in a large city. Each subject filled out a questionnaire that measured whether or not she had shown evidence of delinquent behavior, as well as her birth order. The resulting data are given in the table.

|  |  |
| --- | --- |
|  | **Delinquent behavior?** |
|  | **Yes** | **No** |
| **Oldest**  | 24 | 285 |
| **In-between** | 29 | 247 |
| **Youngest** | 35 | 211 |
| **Only child** | 23 | 70 |

Which one of the following is the appropriate null hypothesis for this situation?

A. The distribution of delinquent behavior is the same for each birth order.

AR. Incorrect. A test of this hypothesis requires that we select independent simple random samples from the population of girls in each birth order and classify them according to the variable of delinquent behavior (yes/no). We have not selected SRSs from each birth order, but have instead taken a single SRS from the population of high-school girls and classified the subjects according to two categorical variables.

\*B. delinquent behavior and birth order are independent.

BR. Correct. We have taken a single SRS from the population of high-school girls and classified the subjects according to two categorical variables. In this case, a test of independence is the appropriate test to use.

C. The distribution of birth order is the same for those subjects exhibiting delinquent and those who do not.

CR. Incorrect. This answer requires that we select independent simple random samples from the population of girls in each category of delinquent behavior (yes/no) and classify them according to the variable of birth order. We have not selected SRSs from each category of delinquent behavior, but have instead taken a single SRS from the population of high-school girls and classified the subjects according to two categorical variables.

12. A study was conducted to explore the relationship between a child’s birth order and his or her chances of becoming a juvenile delinquent. The subjects were a random sample of girls enrolled in public high schools in a large city. Each subject filled out a questionnaire that measured whether or not she had shown evidence of delinquent behavior, as well as her birth order. The resulting data are given in the table.

|  |  |
| --- | --- |
|  | **Delinquent behavior?** |
|  | **Yes** | **No** |
| **Oldest**  | 24 | 285 |
| **In-between** | 29 | 247 |
| **Youngest** | 35 | 211 |
| **Only child** | 23 | 70 |

 A 2 test of independence between these two variables produced the following table of 2 components:

|  |  |
| --- | --- |
|  | **Delinquent behavior?** |
|  | **Yes** | **No** |
| **Oldest**  | 4.637 | 0.633 |
| **In-between** | 0.521 | 0.071 |
| **Youngest** | 1.004 | 0.137 |
| **Only child** | 12.522 | 1.710 |

Which statement below is supported by the information the two tables?

A. The cell with the largest difference between observed and expected counts is “In‑between/No”

AR. Incorrect. This cell has the lowest 2 component, which means the observed and expected counts are very close to equal.

\*B. Only children exhibit more delinquency that expected under the null hypothesis.

BR. Correct. This cell has the highest 2 component, and the expected count (11.2) is much lower than the observed count (23).

C. The relatively high component for the “Oldest/Yes” cell means that the observed count for this cell is considerably higher than the expected count.

CR. Incorrect. While it is true that the 2 component for this cell is the second highest in the table, this is because the observed count for this cell is much *lower* than expected.