**Activity 7.6.1 – Evaluating the Fit of a Categorical Variable**

**Are You Concerned about the Cost of College?**

A high school counselor claims to know how high school seniors in his high school feel about the cost of college. He claims that high school seniors’ levels of concern fall into three categories and are distributed in the following way.

|  |  |
| --- | --- |
| Level of Concern about the Cost of College | |
| Highly Concerned | 20% |
| Somewhat Concerned | 50% |
| Not Concerned | 30% |

1. Suppose you want to test whether the distribution above accurately reflects the *actual* distribution of high school seniors’ feelings in this school. What could you do to test this claim?
2. Suppose the distribution above is accurate. If you randomly survey 100 seniors in this school, how many would you expect to respond “highly concerned”, “somewhat concerned”, and “not concerned”?
3. Suppose a random sample of 100 high school seniors from the school was asked to indicate their feelings about the cost of college. The results are shown in the table below.

|  |  |
| --- | --- |
| Level of Concern | Observed Count |
| Highly Concerned | 25 |
| Somewhat Concerned | 63 |
| Not Concerned | 12 |

Create a numerical measure that describes how well these observed results from the sample align with the expected results based on the counselor’s claim.

**Chi-Square Statistic**

The *chi-square statistic* measures the difference between observed frequencies from a random sample and expected frequencies based on a hypothesized distribution. It is calculated by finding the sum of the squared differences between observed and expected frequencies divided by expected frequencies.

The table below displays the hypothesized percentages from the counselor’s claim and the observed frequencies from the random sample.

|  |  |  |
| --- | --- | --- |
| Level of Concern | Hypothesized Percentages | Observed Frequencies (O) |
| Highly Concerned | 20% | 25 |
| Somewhat Concerned | 50% | 63 |
| Not Concerned | 30% | 12 |

1. Complete the table below. Calculate the expected frequencies (*E*) by applying the hypothesized percentages to the sample size *n* = 100.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Level of Concern |  |  |  |  |  |
| Highly Concerned | 25 |  |  |  |  |
| Somewhat Concerned | 63 |  |  |  |  |
| Not Concerned | 12 |  |  |  |  |

1. Calculate by finding the sum of the last column.
2. The chi-square statistic was computed based on three categories. Which category had the greatest impact on the statistic?

**What’s Your Favorite Season?**

A psychologist claims to have insight on teenagers’ favorite wintertime hobbies. She asserts that the their favorite hobbies are distributed in the following way (left table). Suppose you randomly select 200 teenagers and ask them to describe their favorite wintertime hobby. The survey results are shown below (right table).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Distribution of Favorite Winter Hobby | |  | Survey Results *n* = 200 | |
| Watch movies | 20% |  | Watch movies | 60 |
| Shopping at mall | 15% |  | Shopping at mall | 20 |
| Playing winter sports | 40% |  | Playing winter sports | 92 |
| Listening to music | 25% |  | Listening to music | 28 |

1. Complete the table below. Calculate the expected frequencies (*E*) by applying the hypothesized percentages to the sample size *n* = 200.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Favorite Winter Hobby |  |  |  |  |  |
| Watch movies | 60 |  |  |  |  |
| Shopping at mall | 20 |  |  |  |  |
| Playing winter sports | 92 |  |  |  |  |
| Listening to music | 28 |  |  |  |  |

1. Calculate by finding the sum of the last column.
2. The chi-square statistic was computed based on four categories. Which category had the greatest impact on the statistic?
3. Do small or large chi-square statistics provide evidence that the claim about the distribution is false? Explain why.
4. What is the smallest possible value for ? When would this minimum value occur?