**Chi-Square**

**The Chi-Square Distribution**

To analyze patterns between distinct categories, such as genders, political candidates, locations, or preferences, we use the chi-square goodness-of-fit test.

This test is used when estimating how closely a sample matches the expected distribution (also known as the goodness-of-fit test) and when estimating if two random variables are independent of one another (also known as the test of independence).

The **chi-square distribution** can be used to perform the **goodness-of-fit test,** which compares the observed values of a categorical variable with the expected values of that same variable.

**The Chi-Square Statistic**

The value that indicates the comparison between the observed and expected frequency is called the chi-square statistic. The idea is that if the observed frequency is close to the expected frequency, then the chi-square statistic will be small. On the other hand, if there is a substantial difference between the two frequencies, then we would expect the chi-square statistic to be large.

To calculate the chi-square statistic, χ2, we use the following formula:

**χ2=Σ**((𝐎−𝐄)𝟐)/𝑬

where:

**χ2** - is the chi-square test statistic.

**O**- is the observed frequency value for each event.

**E**- is the expected frequency value for each event.

We compare the value of the test statistic to a tabled chi-square value to determine the probability that a sample fits an expected pattern.

**Testing Hypotheses**

Let’s use our original example to create and test a hypothesis using the goodness-of-fit chi-square test. First, we will need to state the **null and alternative hypotheses** for our research question. Since our research question asks, “Do 11th grade students prefer a certain type of lunch?” our null hypothesis for the chi-square test would state that there is no difference between the observed and the expected frequencies. Therefore, our alternative hypothesis would state that there is a significant difference between the observed and expected frequencies.

**Null Hypothesis**

H0:O=E (There is no statistically significant difference between observed and expected frequencies.)

**Alternative Hypothesis**

Ha:O≠E (There is a statistically significant difference between observed and expected frequencies.)

Also, the number of degrees of freedom for this test is 3.

Using an alpha level of 0.05, we look under the column for 0.05 and the row for degrees of freedom, which, again, is 3. According to the standard chi-square distribution table, we see that the critical value for chi-square is 7.815. Therefore, we would reject the null hypothesis if the chi-square statistic is greater than 7.815.

Note that we can calculate the chi-square statistic with relative ease.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Lunch** | **Observed Frequency** | **Expected Frequency** | ((𝐎−𝐄)𝟐)/𝑬 |
| Salad | 21 | 25 | 0.64 |
| Sub Sandwich | 29 | 25 | 0.64 |
| Daily Special | 14 | 25 | 4.84 |
| Brought Own Lunch | 36 | 25 | 4.84 |
| Total (chi-square) | | 10.96 | |

Since our chi-square statistic of 10.96 is greater than 7.815, we reject the null hypotheses and accept the alternative hypothesis. Therefore, we can conclude that there is a significant difference between the types of lunches that 11th grade students prefer.

**Using the Chi-Square Goodness of Fit Test**

A game involves rolling 3 dice. The winnings are directly proportional to the number of fives rolled. Suppose someone plays the game 100 times with the following observed counts:

|  |  |
| --- | --- |
| **Number of Fives** | **Observed Number of rolls** |
| 0 | 48 |
| 1 | 35 |
| 2 | 15 |
| 3 | 2 |

Someone becomes suspicious and wants to determine whether the dice are fair.

If the dice are fair the probability of rolling a 5 is 1/6. If we roll 3 dice, independently then the number fives in three rolls is distributed as a Binomial (3,1/6).

a. Determine the probability of 0, 1, 2 and 3 fives under this distribution.

Since we have a binomial distribution with 3 independent trials and probability of success 1/6 on each trial, we can compute the probabilities using either the TI Calculator binompdf(3,1/6, k) where k represents the particular value in which we are interested or we can use the formula

**P(k)(** 𝟑/𝒌**)(**𝟏/𝟔**)k(**𝟓/𝟔**)3−k**

**K 0 1 2 3**

**P(k) 0.58 0.345 0.07 0.005**

b. Determine if the dice are fair (Use a chi-square goodness of fit test).

First you must find the expected number of rolls for each category. To do this, multiply the probability of each category by 100. For example, the expected number of rolls where you observe zero 5's is 0.5787⋅100=57.87. The formula for the chi-square goodness of fit test is Σ((𝐎−𝐄)𝟐)/𝑬 where O represents the observed and E represents the expected. You can do this calculation on the TI Calculator by putting the observed values in List 1, the Expected values in List 2, and in List 3 put (((L1−L2))𝟐)/L2

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of Fives** | **Observed Number of rolls** | **Expected Number of rolls** | (O−E)2E |
| 0 | 48 | 58 | 1.72 |
| 1 | 35 | 34.5 | 0.007 |
| 2 | 15 | 7 | 9.14 |
| 3 | 2 | 0.5 | 4.5 |

You then sum the values in List 3. This will be the value of your chi-square statistic: χ2=1.72+0.007+9.14+4.5=15.367

In the previous example, we saw that the critical value for a chi-squared statistic at the 0.05 level of significance is 7.815. Since χ2=15.367>7.815, at the .05 level of significance, we can reject the null hypothesis and conclude that the dice are not fair.

**Drawing Data from Contingency Tables Needed to Perform Calculations when Running a Chi-Square Test**

**Contingency tables** can help us frame our hypotheses and solve problems. Often, we use contingency tables to list the variables and observational patterns that will help us to run a chi-square test. For example, we could use a contingency table to record the answers to phone surveys or observed behavioral patterns.

We would use a contingency table to record the data when analyzing whether women are more likely to vote for a Republican or Democratic candidate when compared to men. In this example, we want to know if voting patterns are independent of gender. Hypothetical data for 76 females and 62 males from the state of California are in the contingency table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Frequency of California Citizens voting for a Republican or Democratic Candidate** | | | | | |
| **Democratic** | | **Republican** | | **Total** | |
| Female | 48 | | 28 | | 76 |
| Male | 36 | | 26 | | 62 |
| Total | 84 | | 54 | | 138 |

Similar to the chi-square goodness-of-fit test, the *test of independence* is a comparison of the differences between observed and expected values. However, in this test, we need to calculate the expected value using the row and column totals from the table. The expected value for each of the potential outcomes in the table can be calculated using the following formula:

**Expected Frequency =**((𝑹𝒐𝒘 𝑻𝒐𝒕𝒂𝒍)(𝑪𝒐𝒍𝒖𝒎𝒏 𝑻𝒐𝒕𝒂𝒍))/(𝑻𝒐𝒕𝒂𝒍 𝑵𝒖𝒎𝒃𝒆𝒓 𝒐𝒇 𝑶𝒃𝒔𝒆𝒓𝒗𝒂𝒕𝒊𝒐𝒏)

In the table above, we calculated the row totals to be 76 females and 62 males, while the column totals are 84 Democrats and 54 Republicans. Using the formula, we find the following expected frequencies for the potential outcomes:

The expected frequency for female Democratic outcome is

𝟕𝟔∗𝟖𝟒/𝟏𝟑𝟖**= 46.26**

The expected frequency for female Republican outcome is

𝟕𝟔∗𝟓𝟒/𝟏𝟑𝟖**= 29.74**

The expected frequency for male Democratic outcome is

𝟔𝟐∗𝟖𝟒/𝟏𝟑𝟖**= 37.74**

**Tests of Single Variance**

We have learned how the chi-square test can help us assess the relationships between two variables. In addition to assessing these relationships, the chi-square test can also help us test hypotheses surrounding variance, which is the measure of the variation, or scattering, of scores in a distribution. There are several different tests that we can use to assess the variance of a sample. The most common tests used to assess variance are the chi-square test for one variance, the F-test, and the Analysis of Variance (ANOVA). Both the chi-square test and the F-test are extremely sensitive to non-normality (or when the populations do not have a normal distribution), so the ANOVA test is used most often for this analysis. However, in this Concept, we will examine in greater detail the testing of a single variance using the chi-square test.

**Testing a** Single **Variance Hypothesis Using the Chi-Square Test**

Suppose that we want to test two samples to determine if they belong to the same population. The test of variance between samples is used quite frequently in the manufacturing of food, parts, and medications, since it is necessary for individual products of each of these types to be very similar in size and chemical make-up. This test is called the test for one variance.

To perform the test for one variance using the chi-square distribution, we need several pieces of information. First, as mentioned, we should check to make sure that the population has a normal distribution. Next, we need to determine the number of observations in the sample. The remaining pieces of information that we need are the standard deviation and the hypothetical population variance. For the purposes of this exercise, we will assume that we will be provided with the standard deviation and the population variance.

Using these key pieces of information, we use the following formula to calculate the chi-square value to test a hypothesis surrounding single variance:

X2 = df(s2) / σ2

where:

**χ2** is the chi-square statistical value.

**Df** = n−1, where n is the size of the sample.

**s2** is the sample variance.

**σ2** is the population variance.

We want to test the hypothesis that the sample comes from a population with a variance greater than the observed variance.

Testing the Randomness of a Sample with Respect to Variance

Suppose we have a sample of 41 female gymnasts from Mission High School. We want to know if their heights are truly a random sample of the general high school population with respect to variance. We know from a previous study that the standard deviation of the heights of high school women is 2.2.

To test this question, we first need to generate **null and alternative hypotheses.** Our null hypothesis states that the sample comes from a population that has a variance of less than or equal to 4.84 (σ2 is the square of the standard deviation).

Null Hypothesis

H0:σ2≤4.84 (The variance of the female gymnasts is less than or equal to that of the general female high school population.)

Alternative Hypothesis

Ha:σ2>4.84 (The variance of the female gymnasts is greater than that of the general female high school population.)

Using the sample of the 41 gymnasts, we compute the standard deviation and find it to be s=1.2. Using the information from above, we calculate our chi-square value and find the following:

**X2 =**((𝟒𝟎)(𝟏.𝟐))/(𝟒.𝟖𝟒) **= 11.9**

Therefore, since 11.9 is less than 55.758 (the value from the chi-square table given an alpha level of 0.05 and 40 degrees of freedom), we fail to reject the null hypothesis and, therefore, cannot conclude that the female gymnasts have a significantly higher variance in height than the general female high school population.