**Test of Independence for Two Categorical Variables**

Tests of Independence show whether or not there is a relationship between two categorical variables measured on the same population.

If two categorical variables are:

• **independent** of one another, the two variables are **not related**

• **not independent** of one another, the two variables are **related** (i.e. dependent)

1. **Formulating the Hypotheses**:

The default assumption, which forms the null hypothesis in this type of test, is that the two variables are independent of one another. If you can reject the null hypothesis, that means the two variables are related.

There is only one set of hypotheses for tests of independence. The two categorical variables are [*Variable 1*] and [*Variable 2*], and you should fill in what those variables are in the context of the problem when doing a test of independence. It doesn’t matter which one is which, but you must maintain consistency in labeling throughout the problem.

|  |
| --- |
| **Hypotheses for Tests of Independence** |
|  |
| **Answers questions about:** |
| Whether or not [*Variable 1*] is related to [*Variable 2*] |
| NOTE:   1. [*Variable 1*] and [*Variable 2*] must both be categorical variables |

1. **The Test Statistic:**

The logic is the similar to a Goodness of Fit test. The expected frequencies are the counts we expect to get for each joint category in the sample if the two variables are independent (that is, if the null hypothesis is true). Observed frequencies are the actual frequencies in the random sample. Some variation between observed and expected frequencies will occur due to random chance, so we use our hypothesis testing techniques and set a threshold (the ) at which we consider the differences between observed and expected counts to be large enough to contradict the null hypothesis. If the differences between the expected and observed frequencies are large, then the sample contradicts the null hypothesis and it is rejected, with the conclusion that is true: the variables are not independent.

The size of the difference between the observed and expected frequencies is quantified in our test statistic, which is calculated according to this formula:

where

where

We will use a contingency table with rows and columns to begin our calculation of this test statistic.

1. **Deciding whether or not to Reject :**

Only large differences between the observed and expected frequencies constitute evidence against the null hypothesis, so all Tests of Independence are **upper tail tests.**

The two possible decisions are:

REMINDER: you can never *accept* the null hypothesis. Remember the swans.

|  |  |
| --- | --- |
| **When to Reject in Tests of Independence:** | |
|  | **Always an Upper Tail Test:** |
| **p-value approach:** | Calculate the upper tail of  If the then reject and accept  If the , then do not reject . is unsupported. |
| **Critical Value: Approach** | Look up the UT Critical Value of , which is  If then reject and accept .  If , then do not reject is unsupported. |
| NOTES:   1. is a Test Statistic 2. is a Critical Value 3. is based on degrees of freedom. The degrees of freedom in Tests of Independence is | |

1. **Interpreting the test:**

(Note: This explanation of interpretation holds for ALL hypothesis tests.) We start every hypothesis test with a question about the parameter of interest, so we must end every hypothesis test with the answer to that question. In other words, we must *interpret* the conclusion of our test in terms of the original question.

Remember: in hypothesis testing you can never prove the null hypothesis. You can only prove the alternative hypothesis: when you reject the null and accept the alternative, then at your given level of significance you may conclude that is true. If you do not reject then you must conclude that is unsupported by the evidence. This gives us a clear guideline for how to *interpret* hypothesis tests: ***always look to the alternative hypothesis!***

In all that follows, you would substitute the actual words and numbers from your hypothesis test for the symbols. Notice that each interpretation simply states the alternative hypothesis in words, and says either that it is true or that it is unsupported by the evidence.

|  |  |
| --- | --- |
| **How to Interpret a Test of Independence:** | |
| **When you:** | **Interpretation:** |
| **Reject** | At the significance level, we can conclude that [*Variable 1*] is not independent of [*Variable 2*]. [*Variable 1*] and [*Variable 2*] are related. |
| **Do not reject** | At the significance level, we cannot conclude that [*Variable 1*] and [*Variable 2*] are related. |
| NOTES:   1. When is rejected, you can:    1. look at the distribution of outcomes in the sample for information about the relationship between [*Variable 1*] and [*Variable 2*].    2. compare the expected frequencies to the observed frequencies to learn about how the two variables are related. | |

**Assumptions Underlying These Hypothesis Tests**

All hypothesis tests use sampling distributions to determine the probability of sample statistics. In order for us to be confident that our choice of sampling distribution for any given test really is the way the sample statistic is distributed, certain assumptions must be met. If the assumptions are not met – that is, if any given assumption is not true – then we cannot rely on the results of the hypothesis tests. They may mislead us, give us the wrong answers, and cause us to draw the wrong conclusions.

For Tests of Independence, the only assumption that must be satisfied is that the expected frequency for each combination of categories must be greater than five:

*Exercise.* An analyst for a chain of coffee shops suspects that a customer’s drink choice is related to the time of day when the customer comes in. The three drink choices the analyst would like to consider are Hot Coffee, Iced Coffee, and Specialty Drinks. The analyst takes a random sample of 503 customers and then splits the customers into two groups: those who came in before 11am and those who came in after 11am. Of the customers who came in before 11am, 90 ordered Hot Coffee, 79 ordered Iced Coffee, and 72 ordered Specialty Drinks. Of the customers who came in after 11am, 67 ordered Hot Coffee, 74 ordered Iced Coffee, and 121 ordered Specialty Drinks.

Conduct a Test of Independence to find out whether drink choice is related to time of visit at the α = .01 significance level.

First, fill in the Observed Frequencies from the sample data, then calculate the Row and Column Totals.

Second, calculate the expected frequency, , for each combination of categories:

where *i* is the row number, *j* is the column number, and n is the sample size.

|  |  |  |  |
| --- | --- | --- | --- |
| **Contingency Table:**  **Observed Frequencies** | | | |
|  | Before 11am | After 11am | **Row Total** |
| Hot Coffee |  |  |  |
| Iced Coffee |  |  |  |
| Specialty Drink |  |  |  |
| **Column Total** |  |  |  |

NOTE: there are three categories of Drink Choice, so r = 3. There are two categories of Time of Day, so c = 2. You will need this information later to calculate degrees of freedom for the test statistic and critical value.

Third, fill in the following table. The sum of the last column is the test statistic:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Drink Choice** | **Time of Visit** | **Observed Frequency** | **Expected Frequency** | **Difference Squared/Expected Frequency** |
|  |  |  |  |  |  |
| 11 | Hot Coffee | Before 11am |  |  |  |
| 12 | Hot Coffee | After 11am |  |  |  |
| 21 | Iced Coffee | Before 11am |  |  |  |
| 22 | Iced Coffee | After 11am |  |  |  |
| 31 | Specialty Drinks | Before 11am |  |  |  |
| 32 | Specialty Drinks | After 11am |  |  |  |
|  |  | | | | |

The test statistic and critical value for a test of independence has (r – 1)(c – 1) degrees of freedom (see note from previous page). Now you are ready to complete your hypothesis test.