Goodness of Fit Test (Chi-Squared Distribution)

So far, we have mainly worked with continuous variables. Recall, a continuous variable is a variable that can take any real number (integers, fractions, irrational values) within its defined range. E.g. the amount of time it can take to finish a 90 minute exam can be any real number between 0 to 90 minutes. In this section, we begin to work with Categorical variables. Categorical variables are variables that can take on values that are names or labels. E.g. the color of a ball used in an experiment can be red, blue or green. A person’s gender may be male, female or transgender.

A categorical variable that can only take two values is binomial, e.g. marital status can be single or married. (We have worked with binomial categorical variables when working with hypothesis tests of one and two population proportions). A categorical variable that can take more than two values is multinomial. e.g. education attainment can be primary school graduate, secondary school graduate, high school graduate, college/university graduate etc.

A test concerning just one multinomial categorical variable is the Goodness-of-Fit test. This test can be explained using the framework of an experiment. The form of experiment we will use to explain this is called a Multinomial Experiment.

In a Multinomial Experiment, there are:

1. n trials
2. each trial results in one of k possible outcomes (or categories), where k > 2 (It k=2, it is a binomial experiment)
3. the trials are independent
4. probabilities of obtaining each outcome remains constant for each trial

Example:

Consider a fair dice of six sides. Suppose you and your classmates decide to play a game where you each throw the dice 10 times and record your total score. The person with the highest score is the winner. This is an example of a multinomial experiment.

1. n trials = 10 trials (trials are the same as number of attempts)
2. k possible outcomes = 6 possible outcomes
3. the trials are independent = the score you obtain in any one throw does not influence what score you will obtain in your next throw
4. probability of each outcome = since it is a fair dice, the probability of each outcome is $\frac{1}{6}$

From multinomial experiments, we can identify two types of frequencies. Frequency means the number of times an event occurred in an experiment or study.

Observed Frequency: frequencies obtained from the sample / experiment

Expected Frequency: frequency we expect to obtain if the null hypothesis is true

Example: Suppose we are investigating voter popularity among three political candidates running for mayoral election. Voter preference is a multinomial categorical variable with the following categories – Candidate 1, Candidate 2, Candidate 3 and Undecided.

Suppose that according to a national survey, the proportions in favor of each category are given as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Candidate 1 | Candidate 2 | Candidate 3 | Undecided |
| Proportion in favor | 0.35 | 0.3 | 0.2 | 0.15 |

Table 1

Last week, the candidates appeared for a television debate that was watched by all people of the nation. A political pollster wants to see if people’s preferences of mayoral candidate have changed as a result of this debate. They surveyed 500 citizens who have watched the debate and recorded the number of people who preferred each category. The results are presented in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Candidate 1 | Candidate 2 | Candidate 3 | Undecided |
| Proportion in favor | 150 | 200 | 120 | 30 |

Table 2

Test if the debate changed people’s opinions. That means we have to analyze if the proportions in favor of each category prior to the debate is different from the proportions in favor of each category after the debate has taken place based on the evidence provided by the survey data of 500 citizens.

**Step 1: State your hypothesis in words**

The proportions in favor of each category prior to the debate are different from the proportions in favor of each category after the debate for at least one category.

**Step 2: Write your hypothesis in notation**

 $H\_{0}:All of the proportions are the same as before$

 $H\_{1}:At least one of the proportions is different from before$

**Step 3: The test statistic for the goodness-of-fit test is always chi-squared distributed. The test is also always right-tailed.**

**Step 4: Prepare the observed and expected frequencies table to calculate the test statistic**

Recall the observed frequencies are the values that we obtain from the sample/experiment. This means the observed frequencies are the values form Table 2. On the other hand, expected frequencies are the frequencies we expect to obtain if the null hypothesis is true. If the null hypothesis is true, then that means the proportions are the same as before the debate. Hence to calculate expected frequencies, we use the proportions in Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Category | Observed Frequencies | Proportions (p) | Expected Frequencies (np) | (O – E)  | (O – E)2 | $$\frac{(O -E)^{2}}{E}$$ |
| Candidate 1 | 150 | 0.35 | 500 X 0.35 = 175 | 150 – 175 = – 25 | 625 | 3.571 |
| Candidate 2 | 200 | 0.3 | 500 X 0.3 = 150 | 200 – 150 = 50 | 2500 | 16.667 |
| Candidate 3 | 120 | 0.2 | 500 X 0.2 = 100 | 120 – 100 = 20 | 400 | 4 |
| Undecided | 30 | 0.15 | 500 X 0.15 = 75 | 30 – 75 = – 45 | 2025 | 27 |
|  |  |  |  |  |  | Total = 51.238 |

Table 3

The total value we have obtained at the end is the chi-squared test statistic value:

$χ^{2}=51.238$

**Step 5:** Run a **right-tailed test using the chi-squared distribution**. The degrees of freedom is k – 1, where k is the number of categories. State your conclusion.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Contingency Tables

In this section, we will test the relationship between two categorical variables. To do so, we use contingency tables. A contingency table shows the distribution of one variable in rows and another in columns, used to study the correlation between the two variables.



Here we have two categorical variables. The column variable is gender, which takes the categories male and female. The row variable is enrollment status, which takes the categories full-time and part-time. The table shows enrollment status of students disaggregated by gender. e.g. 6768 male students are enrolled full-time.

Using contingency tables, we can run two types of tests to measure the relationship between two categorical variables. These tests are (I) Test of Independence and (II) Test of Homogeneity.

**Test of Independence**

Example: Suppose we want to study measure the relationship between two categories – nationality and the use of a product.

Nationality has three categories: American; Latin; Asian

Use of product: uses the product; does not use the product

We want to find out if the two variables are independent or dependent, i.e. is there any dependency between nationality and the choice to use the product.

The hypothesis can be stated as:

 $H\_{0}:origin and nationality are independent $

 $H\_{1}:origin and nationality are dependent $

Suppose a sample of 200 observations were collected. The survey results are displayed in the contingency table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | American | Latin | Asian |  |
| Use the product | 50 | 30 | 10 | r1 = 90 |
| Do not use the product | 45 | 25 | 40 | r2 = 110 |
|  | c1 = 95 | c2 = 55 | c3 = 50 | n = 200 |

The values in the tables are called observed frequencies.

Using the row and column totals, we have to calculate the expected frequencies.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | American | Latin | Asian |  |
| Use the product | $\left(r\_{1}×c\_{1}\right)/n$ = 42.75 | $\left(r\_{1}×c\_{2}\right)/n$ = 24.75 | $\left(r\_{1}×c\_{3}\right)/n$ = 22.5 | r1 = 90 |
| Do not use the product | $\left(r\_{2}×c\_{1}\right)/n $= 52.25 | $\left(r\_{2}×c\_{2}\right)/n$ = 30.25 | $\left(r\_{2}×c\_{3}\right)/n$ = 27.5 | r2 = 110 |
|  | c1 = 95 | c2 = 55 | c3 = 50 | n = 200 |

The test statistic is the same formula:



$χ^{2}=\frac{(50-42.75)^{2}}{42.75}+\frac{(30-24.75)^{2}}{24.75}+…$

Again, this is always a right-tailed test. Conduct a right-tailed test using chi-squared distribution.