Statistical Tests

Patrick Lam

Outline

Outline

The Chi-Square Test for Independence of Two Variables

The Chi-Square Test for Independence of Two Variables

Suppose you have two discrete variables X and Y, where X has m possible values and Y has n possible values.

The Chi-Square Test for Independence of Two Variables

Suppose you have two discrete variables X and Y, where X has m possible values and Y has n possible values.

We can create an $m \times n$ table of the frequencies of the observations.

The Chi-Square Test for Independence of Two Variables

Suppose you have two discrete variables X and Y, where X has m possible values and Y has n possible values.

We can create an $m \times n$ table of the frequencies of the observations.

We can then conduct a chi-square test by comparing the observed frequencies to the expected frequencies under the null hypothesis of independence.

Let O_i denote the observed frequency in position i of the table and E_i denote the expected frequency in position i.

Let O_i denote the observed frequency in position i of the table and E_i denote the expected frequency in position i.

Then we have a test statistic T, where

Let O_i denote the observed frequency in position i of the table and E_i denote the expected frequency in position i.

Then we have a test statistic T, where

$$T = \sum_{i=1}^{m \times n} \frac{(O_i - E_i)^2}{E_i}$$

Let O_i denote the observed frequency in position i of the table and E_i denote the expected frequency in position i.

Then we have a test statistic T, where

$$T = \sum_{i=1}^{m \times n} \frac{(O_i - E_i)^2}{E_i}$$

T is distributed χ^2 with $(m-1) \times (n-1)$ degrees of freedom.

Let O_i denote the observed frequency in position i of the table and E_i denote the expected frequency in position i.

Then we have a test statistic T, where

$$T = \sum_{i=1}^{m \times n} \frac{(O_i - E_i)^2}{E_i}$$

T is distributed χ^2 with $(m-1) \times (n-1)$ degrees of freedom.

We calculate the test statistic and then see whether we can reject the null hypothesis of independence.

An Example

Suppose we want to know whether the `Parreg` and `Parcomp` variables in the POLITY dataset are independent (Treier and Jackman, 2008).

An Example

Suppose we want to know whether the `Parreg` and `Parcomp` variables in the POLITY dataset are independent (Treier and Jackman, 2008).

We have the following observed frequencies in a 5×6 table since `Parreg` takes on 5 possible values and `Parcomp` takes on 6 possible values.

An Example

Suppose we want to know whether the `Parreg` and `Parcomp` variables in the POLITY dataset are independent (Treier and Jackman, 2008).

We have the following observed frequencies in a 5×6 table since `Parreg` takes on 5 possible values and `Parcomp` takes on 6 possible values.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	487	0	0	0	10	0	497
2	96	0	0	740	583	0	1419
3	0	0	299	3509	76	0	3884
4	0	3878	1811	0	0	0	5689
5	0	0	0	0	116	2336	2452
Total	583	3878	2110	4249	785	2336	13941

Expected Frequencies

Expected Frequencies

We have the observed frequencies, but how do get the expected frequencies under independence?

Expected Frequencies

We have the observed frequencies, but how do get the expected frequencies under independence?

We can estimate the expected frequencies from the data by doing the following:

Suppose we want to calculate the expected frequency for the top-left position in the table where $\texttt{Parcomp} = 0$ and $\texttt{Parreg} = 1$.

Suppose we want to calculate the expected frequency for the top-left position in the table where Parcomp $= 0$ and Parreg $= 1$.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	E_1	E_2	E_3	E_4	E_5	E_6	**497**
2	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	1419
3	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	3884
4	E_{19}	E_{20}	E_{21}	E_{22}	E_{23}	E_{24}	5689
5	E_{25}	E_{26}	E_{27}	E_{28}	E_{29}	E_{30}	2452
Total	**583**	3878	2110	4249	785	2336	**13941**

Suppose we want to calculate the expected frequency for the top-left position in the table where Parcomp $= 0$ and Parreg $= 1$.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	E_1	E_2	E_3	E_4	E_5	E_6	**497**
2	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	1419
3	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	3884
4	E_{19}	E_{20}	E_{21}	E_{22}	E_{23}	E_{24}	5689
5	E_{25}	E_{26}	E_{27}	E_{28}	E_{29}	E_{30}	2452
Total	**583**	3878	2110	4249	785	2336	**13941**

1. First calculate the proportion of the data where Parreg $= 1$:

Suppose we want to calculate the expected frequency for the top-left position in the table where Parcomp $= 0$ and Parreg $= 1$.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	E_1	E_2	E_3	E_4	E_5	E_6	**497**
2	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	1419
3	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	3884
4	E_{19}	E_{20}	E_{21}	E_{22}	E_{23}	E_{24}	5689
5	E_{25}	E_{26}	E_{27}	E_{28}	E_{29}	E_{30}	2452
Total	**583**	3878	2110	4249	785	2336	**13941**

1. First calculate the proportion of the data where Parreg $= 1$: $\frac{497}{13941} \approx 0.036$

Suppose we want to calculate the expected frequency for the top-left position in the table where $\texttt{Parcomp} = 0$ and $\texttt{Parreg} = 1$.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	E_1	E_2	E_3	E_4	E_5	E_6	**497**
2	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	1419
3	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	3884
4	E_{19}	E_{20}	E_{21}	E_{22}	E_{23}	E_{24}	5689
5	E_{25}	E_{26}	E_{27}	E_{28}	E_{29}	E_{30}	2452
Total	**583**	3878	2110	4249	785	2336	**13941**

1. First calculate the proportion of the data where $\texttt{Parreg} = 1$: $\frac{497}{13941} \approx 0.036$
2. Then find the expected frequency by taking that proportion and multiplying it by the number of observations with $\texttt{Parcomp} = 0$:

Suppose we want to calculate the expected frequency for the top-left position in the table where $\texttt{Parcomp} = 0$ and $\texttt{Parreg} = 1$.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	E_1	E_2	E_3	E_4	E_5	E_6	**497**
2	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	1419
3	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	3884
4	E_{19}	E_{20}	E_{21}	E_{22}	E_{23}	E_{24}	5689
5	E_{25}	E_{26}	E_{27}	E_{28}	E_{29}	E_{30}	2452
Total	**583**	3878	2110	4249	785	2336	**13941**

1. First calculate the proportion of the data where $\texttt{Parreg} = 1$: $\frac{497}{13941} \approx 0.036$
2. Then find the expected frequency by taking that proportion and multiplying it by the number of observations with $\texttt{Parcomp} = 0$: $E_1 \approx 0.036 \times 583 \approx \mathbf{21}$

Suppose we want to calculate the expected frequency for the top-left position in the table where Parcomp $= 0$ and Parreg $= 1$.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	E_1	E_2	E_3	E_4	E_5	E_6	**497**
2	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	1419
3	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	3884
4	E_{19}	E_{20}	E_{21}	E_{22}	E_{23}	E_{24}	5689
5	E_{25}	E_{26}	E_{27}	E_{28}	E_{29}	E_{30}	2452
Total	**583**	3878	2110	4249	785	2336	**13941**

1. First calculate the proportion of the data where Parreg $= 1$: $\frac{497}{13941} \approx 0.036$
2. Then find the expected frequency by taking that proportion and multiplying it by the number of observations with Parcomp $= 0$: $E_1 \approx 0.036 \times 583 \approx \mathbf{21}$

We can also switch the order of Parreg $= 1$ and Parcomp $= 0$.

We can then fill out the rest of the expected frequencies table using the same method.

We can then fill out the rest of the expected frequencies table using the same method.

	Parcomp						
Parreg	0	1	2	3	4	5	Total
1	21	138	75	152	28	83	**497**
2	59	395	215	432	80	238	1419
3	162	1080	588	1184	219	651	3884
4	238	1583	861	1734	320	953	5689
5	103	682	371	747	138	411	2452
Total	**583**	3878	2110	4249	785	2336	**13941**

We then calculate our test statistic T:

We then calculate our test statistic T:

$$
\begin{aligned}
T \quad = \quad & \frac{(487-21)^2}{21} + \frac{(0-138)^2}{138} + \frac{(0-75)^2}{75} + \frac{(0-152)^2}{152} + \frac{(10-28)^2}{28} + \frac{(0-83)^2}{83} + \\
& \frac{(96-59)^2}{59} + \frac{(0-395)^2}{395} + \frac{(0-215)^2}{215} + \frac{(740-432)^2}{432} + \frac{(583-80)^2}{80} + \frac{(0-238)^2}{238} + \\
& \frac{(0-162)^2}{162} + \frac{(0-1080)^2}{1080} + \frac{(299-588)^2}{588} + \frac{(3509-1184)^2}{1184} + \frac{(76-219)^2}{219} + \frac{(0-651)^2}{651} + \\
& \frac{(0-238)^2}{238} + \frac{(3878-1583)^2}{1583} + \frac{(1811-861)^2}{861} + \frac{(0-1734)^2}{1734} + \frac{(0-320)^2}{320} + \frac{(0-953)^2}{953} + \\
& \frac{(0-103)^2}{103} + \frac{(0-682)^2}{682} + \frac{(0-371)^2}{371} + \frac{(0-747)^2}{747} + \frac{(116-138)^2}{138} + \frac{(2336-411)^2}{411}
\end{aligned}
$$

We then calculate our test statistic T:

$$
\begin{aligned}
T &= \frac{(487-21)^2}{21} + \frac{(0-138)^2}{138} + \frac{(0-75)^2}{75} + \frac{(0-152)^2}{152} + \frac{(10-28)^2}{28} + \frac{(0-83)^2}{83} + \\
&\quad \frac{(96-59)^2}{59} + \frac{(0-395)^2}{395} + \frac{(0-215)^2}{215} + \frac{(740-432)^2}{432} + \frac{(583-80)^2}{80} + \frac{(0-238)^2}{238} + \\
&\quad \frac{(0-162)^2}{162} + \frac{(0-1080)^2}{1080} + \frac{(299-588)^2}{588} + \frac{(3509-1184)^2}{1184} + \frac{(76-219)^2}{219} + \frac{(0-651)^2}{651} + \\
&\quad \frac{(0-238)^2}{238} + \frac{(3878-1583)^2}{1583} + \frac{(1811-861)^2}{861} + \frac{(0-1734)^2}{1734} + \frac{(0-320)^2}{320} + \frac{(0-953)^2}{953} + \\
&\quad \frac{(0-103)^2}{103} + \frac{(0-682)^2}{682} + \frac{(0-371)^2}{371} + \frac{(0-747)^2}{747} + \frac{(116-138)^2}{138} + \frac{(2336-411)^2}{411} \\
&= \mathbf{40290.79}
\end{aligned}
$$

We can then test the null hypothesis of independence by seeing the area to the right of our T in a χ^2_{20} distribution.

We can then test the null hypothesis of independence by seeing the area to the right of our T in a χ^2_{20} distribution.

```
> pchisq(40290.79, df = 20, lower.tail = F)

[1] 0
```

We can then test the null hypothesis of independence by seeing the area to the right of our T in a χ^2_{20} distribution.

```
> pchisq(40290.79, df = 20, lower.tail = F)

[1] 0
```

We can conclude that the two variables are clearly not independent since the probability of getting a T that is as extreme as our T given independence is 0.