Statistics 301: Probability and Statistics

Categorical Data Analyses

Module 11

2018

Testing categorical data

For most of the analyses that you have learned about, all are analyzing quantitative data. But that leaves out a large portion of data, categorical data. Now we can see how to analyze things like:

(1) making sure a sample follows a specific distribution
(2) exploring whether or not two or more categories have a relationship
(3) analyzing data to see how one category is distributed over another

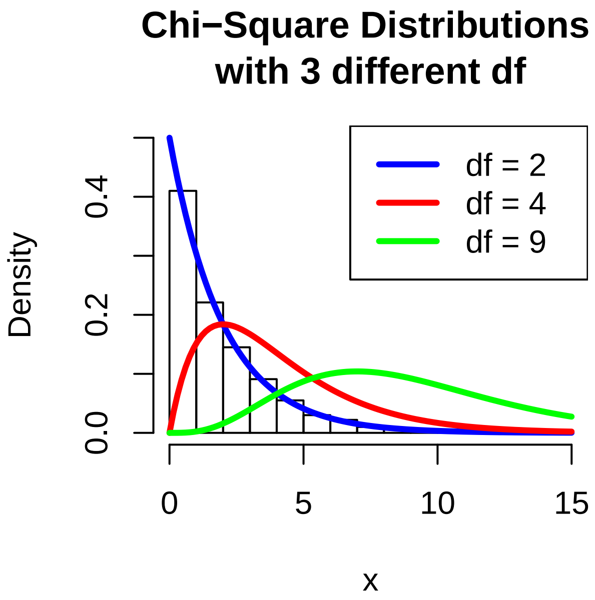
Chi-square distribution

While we have analyses for comparing more than 2 means, we cannot use them when trying to compare more than one proportion. However, there is a distribution that is related to the standard normal distribution (z) that works for comparing more than two proportions. Rather than a test statistic for each pair of proportions, we'd rather like to use just one to prevent the Type I error from inflating. What we do is measure the distance each sample value is from the average (from the "norm"). If we had a z-score for each pair, the sum of the squared z-scores would be a new (new to you) distribution called Chi-square (pronounced "ky" as in "sky"), denoted by χ^2. The distribution is a skewed distribution (skewed right) so it is not a symmetric distribution like z or t, until $df \to \infty$.

$$\chi^2 = \sum_{i=1}^{n} z_i^2 = z_1^2 + z_2^2 + \cdots + z_n^2$$

χ^2 with varying df

The following graph illustrates how the χ^2 distribution changes shape with increasing df.



Assumptions of any Chi-square test

(1) The data must be counts from categories
(2) Independence of observations
(3) $E_i \geq 5$; each individual expected value (E_i) must be at least 5

Test statistic (for all 3 tests), *df*

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} = \sum \frac{(O-E)^2}{E}$$

df for GoF is $df = k - 1$, where $k =$ number of categories

df for Independence and Homogeneity is $df = (r-1)(c-1)$

$(r =$ number of rows, $c =$ number of columns$)$

Goodness-of-Fit (GoF)

Chi-square for a one-way table (a table that has categories and counts for each category): In evaluating whether there is sufficient evidence that a set of observed counts, $O_1, O_2, \cdots, O_k$ in k categories are unusually different from what would be expected under a null hypothesis. The expected values under the null hypothesis, called $E_1, E_2, \ldots, E_k$.

GoF hypotheses

$$H_0 : p_1 = p_2 = \cdots = p_k = p_0$$

Or H_0 : The data follows <specified> distribution

$$H_a : \text{At least one } p_i \text{ differs}$$

Or $H_a : H_0$ is not true (the data does not follow <specified> distribution)

GoF formulas

Expected value

$$E_i = np_i$$

You will need to find the probabilities associated with the null hypothesized distribution (given), then multiply each category value by the probability to get the expected value.

GoF H_0 rejection

Rejection region

(1) Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $df = k - 1$, where $k =$ number of categories or
(2) Reject H_0 iff $pvalue \leq \alpha$ where $pvalue = P(\chi^2 \geq \chi^2_{calc})$

Conclusion (in context)
When the null hypothesis is rejected, in terms of the context of the data, it means that we think that the data does not follow the theoretical (specified) distribution. When we fail to reject the null hypothesis, we are maintaining that the data does follow the theoretical (specified) distribution

Test of Independence

The test of Independence explores whether two categorical random variables are independent or whether some level of dependency existst between them. Each dataset will be constructed into a table with I rows and J columns. Let n_{ij} denote the number of individuals in the sample falling in the $(i, j)^{th}$ cell (of row i, column j) of the table. The following is a prototype of a general table that displays the counts (n_{ij}) and is called a *two-way contingency table.* I and J (capital I,J) are the row and column totals, respectively.

Data organization

	1	2	$\dots$	j	$\dots$	J
1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1J}
2	n_{21}					$\vdots$
$\vdots$						
i	n_{i1}	$\dots$		n_{ij}	$\dots$	
$\vdots$						
I	n_{I1}	$\dots$				$n_{IJ} = n$

Independence test hypotheses

$$H_0 : p_{ij} = (p_{i\cdot})(p_{\cdot j}) \ {}_{i=1,2,\dots,I \text{ and } j=1,2,\dots,J}$$

Or H_0 : The row context and column <context> are independent

$$H_a : H_0 \text{ is not true (meaning that rows and columns are dependent)}$$

Independence test formulas

Expected values

$$E_{ij} = \frac{n_i n_j}{n} = \frac{(rtotal)(ctotal)}{grandtotal}$$

Independence test rejection

Rejection region
Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $df = (r-1)(c-1)$ (r = number of rows, c = number of columns). Or reject H_0 iff $pvalue \leq \alpha$ where $pvalue = P(\chi^2 \geq \chi^2_{calc})$

Conclusion (in context)
When the null hypothesis is rejected, in terms of the context of the data, it means that we think that the context of the rows and context of the columns are dependent (there is a dependency). When we fail to reject the null hypothesis, we are maintaining that the context of the rows and context of the columns are dependent (there is no relationship).

Homogeneous Test

We are assuming that each individual in every one of the I populations belongs in exactly one of J categories. An example would be to see if voting habits are the same over regions.

Homogeneous test hypotheses

$$H_0 : p_{1j} = p_{2j} = \ldots = p_{Ij} \quad {}_{i=1,2,\ldots,I}\ ,\ {}_{j=1,2,\ldots,J}$$

Or

H_0 : The row <context> is distributed the same over the column <context>

H_a : H_0 is not true (the distribution is not the same for all categories)

Homogeneous test formulas+

Test statistic
Same as Independence Test

Expected values
Same as Independence Test

Rejection region
Same as Independence Test

Conclusion (in context)
When the null hypothesis is rejected, in terms of the context of the data, it means that we think that the context of the rows are distributed differently across the context of the columns. When we fail to reject the

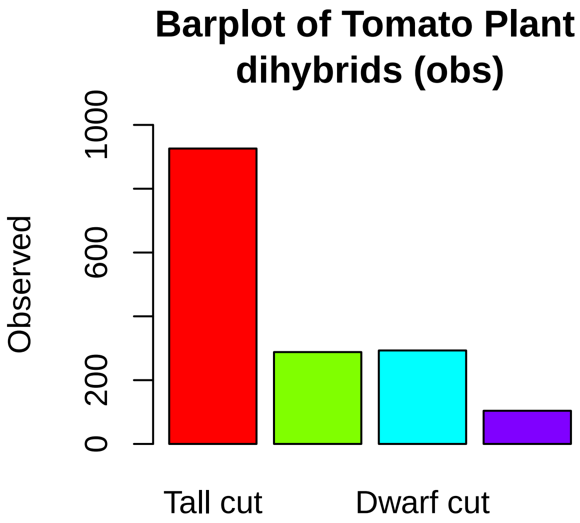
null hypothesis, we are maintaining that the context of the rows are distributed similarly across the context of the columns.

GoF example

A paper[1] about an experiment reported the following data on phenotypes resulting from crossing tall cut-leaf tomatoes with dwarf potato-leaf tomatoes. We wish to investigate if the frequencies below are consistent with Mendel's laws of inheritance which implies that the phenotypes should occur in a 9:3:3:1 ratio. A 9:3:3:1 ratio means the probabilities are 9/16, 3/16 (×2) and 1/16 (the 16 comes from the sum of all the numbers in the ratio). Is there sufficient evidence that the tomato plants follow Mendel's Law?

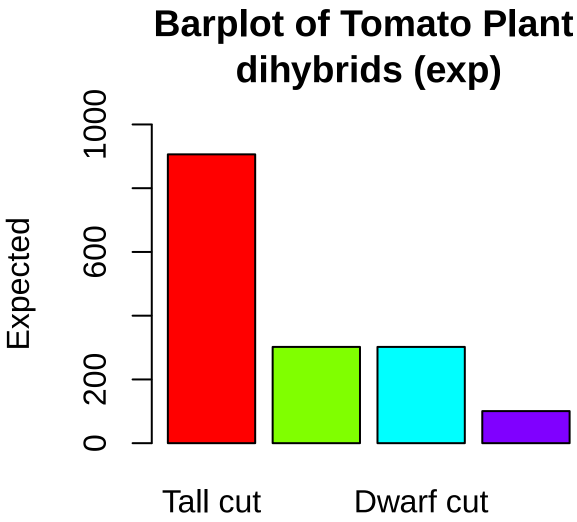
```
        types1 plants plantexp  probs
1     Tall cut    926   906.19 0.5625
2  Tall potato    288   302.06 0.1875
3    Dwarf cut    293   302.06 0.1875
4 Dwarf potato    104   100.69 0.0625
```

Gof example con't

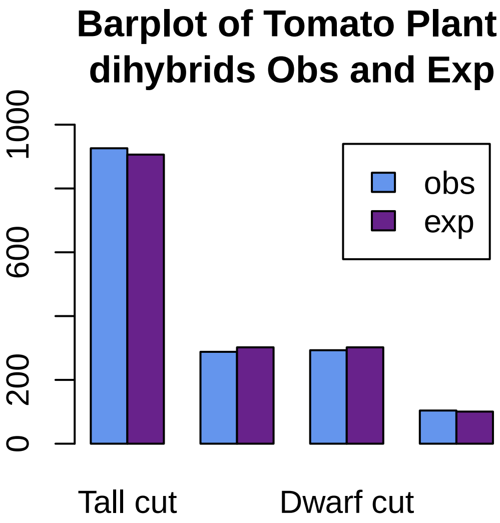


[1] "Linkage Studies of the Tomato" (*Transactions of the Royal Canadian Institute*, 1931: 1-19)

Gof example con't



Gof example con't



Gof example con't

Hypotheses
$H_0 : p_1 = \frac{9}{16}, p_2 = p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$ (data follows Mendel's Law)

H_a : At least one p_i differs (data does not follow Mendel's Law)

Gof example con't

Expected values

$$E_i = n(p_i)$$

$E_1 = 1611(9/16) = 906.19$
$E_2 = 1611(3/16) = 302.06$
$E_3 = 1611(3/16) = 302.06$
$E_4 = 1611(1/16) = 100.69$

Here we can check to see all $E_i \geq 5$

Gof example con't

Test Statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \sum \frac{(observed - expected)^2}{expected}$$

$= \frac{(926-906.19)^2}{906.19} + \frac{(288-302.06)^2}{302.06} + \frac{(293-302.6)^2}{302.6} + \frac{(104-100.69)^2}{100.69} = 1.468$
$df = k - 1$ where $k = 4$ so $df = 4 - 1 = 3$

Rejection Region

(1) *Critical Value approach*: Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $\chi^2_{\alpha,df} = \chi^2_{.05,3} = 7.815$
(2) *pvalue approach*: Reject H_0 if $pvalue \leq \alpha$

Gof example con't

Results
We are doing the critical value approach in the "by-hand" example. So $\chi^2_{.05,3} = 7.815$. $1.468 \ngeq 7.815$ so we will fail to reject H_0.

Conclusion (in context)
We failed to reject H_0 so that tells us that the plants do follow Mendel's law (maintaining the 9:3:3:1 ratio).

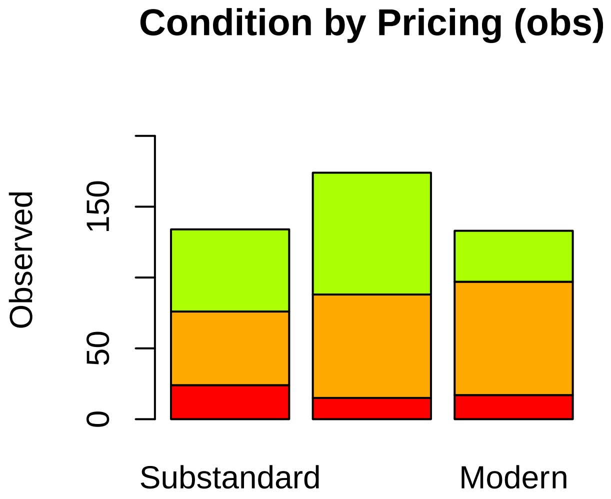
Independence Test example

A study of the relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline[2] reports the accompanying data based on a sample of $n = 441$ stations. Does the data suggest that facility conditions and pricing policy are independent of one another?

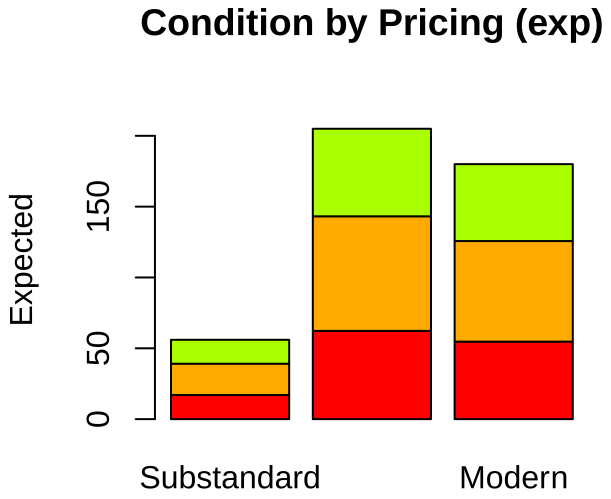
```
             Pricing.Policy
Condition     Aggressive Neutral Nonaggressive
  Substandard         24      15            17
  Standard            52      73            80
  Modern              58      86            36
```

[2] "An Analysis of Price Aggressiveness in Gasoline Marketing", (*Journal Marketing Research*, 1970: 36-42)

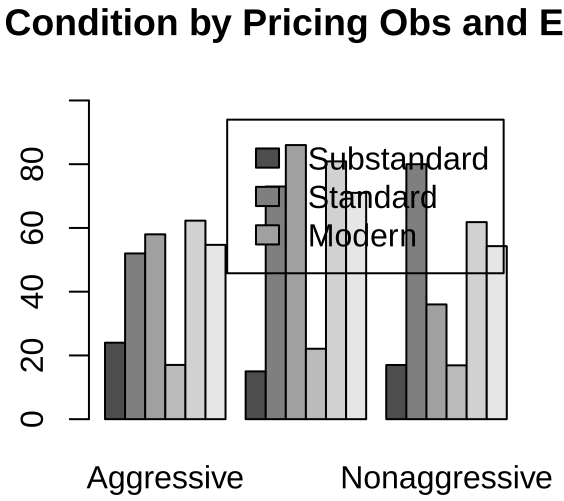
Independence Test example con't



Independence Test example con't



Independence Test example con't



Independence Test example con't

Null Hypothesis
$H_0 : p_{ij} = (p_{i\cdot})(p_{\cdot j})$ $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$ (pricing and conditions in gasoline marketing are independent)

Alternative hypothesis
$H_a : H_0$ is not true (pricing and conditions in gasoline are dependent)

Independence Test example con't

Expected values

$$E_{ij} = \frac{n_i n_j}{n} = \frac{(rtotal)(ctotal)}{grandtotal}$$

$E_{11} = (56 * 134/441) = 17.02$
$E_{12} = (56 * 174/441) = 62.29$
$E_{13} = (56 * 133/441) = 54.69$
$E_{21} = (205 * 134/441) = 22.1$
$E_{22} = (205 * 174/441) = 80.88$
$E_{23} = (205 * 133/441) = 71.02$
$E_{31} = (180 * 134/441) = 16.89$
$E_{32} = (180 * 174/441) = 61.83$
$E_{33} = (180 * 133/441) = 54.29$

Here we can check to see all $E_{ij} \geq 5$

Independence Test example con't

Test Statistic

$$\chi^2 = \sum_{i=1}^{n_i}\sum_{j=1}^{n_j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{r\times c}^{all}\frac{(observed - expected)^2}{expected}$$

$= \frac{(24-17.02)^2}{17.02} + \frac{(15-22.1)^2}{22.1} + \cdots + \frac{(36-54.29)^2}{54.29} = 22.473$

$df = (r-1)(c-1)$ where $r, c = (3,3)$ so $df = (3-1)(3-1) = 4$

Independence Test example con't

Rejection Region

(1) *Critical Value approach*: Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $\chi^2_{\alpha,df} = \chi^2_{.05,4} = 9.488$
(2) *pvalue approach*: Reject H_0 if $pvalue \leq \alpha$

Results
We are doing the critical value approach in the "by-hand" example. So $\chi^2_{.05,4} = 9.488$. $22.473 \geq 9.488$ so we will reject H_0.

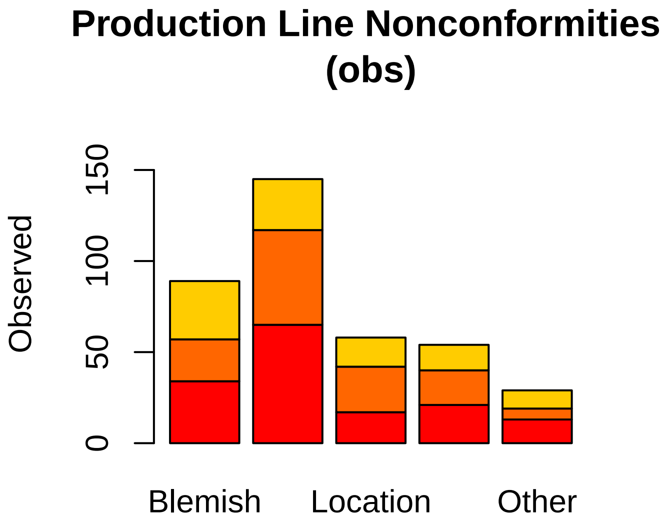
Conclusion (in context)
We rejected H_0 so that tells us that knowledge of a stations's pricing policy does give information about the condition of facilities at the station. Stations with an aggressive pricing policy appear to have more substandard facilities than stations with a neutral or nonaggressive policy.

Homogeneous example

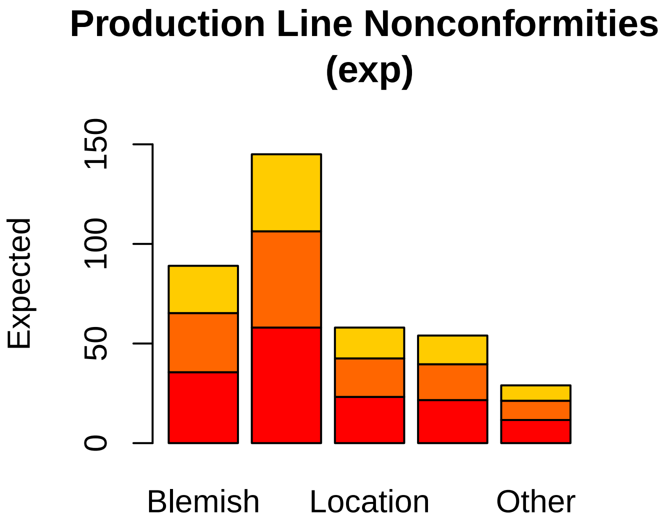
A company packages a particular product in cans of three different sizes, each one using a different production line. Most cans conform to specifications, but a quality control engineer has identified blemish on a can, crack in the can, improper pull tab location, pull tab missing or some other as main reasons for nonconformities. A sample of nonconforming units is selected from each of the three lines, and each unit is categorical according to reason for nonconformity. Do the data suggest that the proportions failing in the various nonoconformance categories are not the same for the three lines?

```
              Nonconformity
ProductionLine Blemish Crack Location Missing Other
             1      34    65       17      21    13
             2      23    52       25      19     6
             3      32    28       16      14    10
```

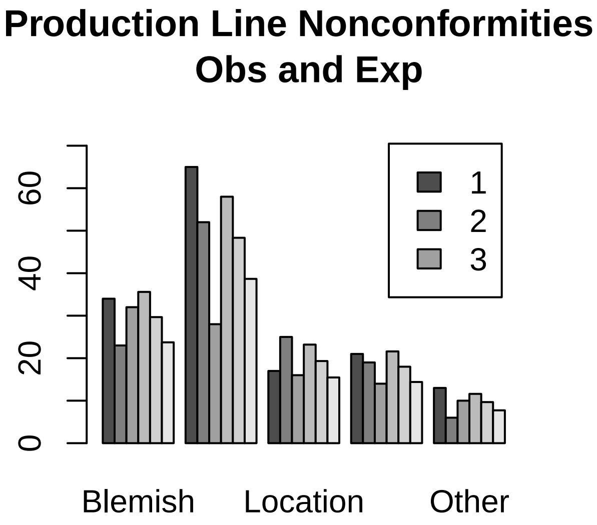
Homogeneous example con't



Homogeneous example con't



Homogeneous example con't



Homogeneous example con't

$$H_0 : p_{1j} = p_{2j} = \ldots = p_{Ij} \quad {}_{i=1,2,\ldots,I \; ; \; j=1,2,\ldots,J}$$

Or
H_0 : Blemish types have a homogeneous distribution over the prouction lines (no one production line has more types of blemishes than any other production line)

$$H_a : H_0 \textit{ is not true}$$

Homogeneous example con't

Expected values

$$E_{ij} = \frac{(n_i)(n_j)}{n} = \frac{(rtotal)(ctotal)}{grandtotal}$$

$E_{11} = \frac{(158)(97)}{383} = 35.6$, $E_{12} = \frac{(158)(145)}{383} = 58$
$E_{13} = \frac{(158)(58)}{383} = 23.2$, $E_{14} = \frac{(158)(54)}{383} = 21.6$
$E_{15} = \frac{(158)(29)}{383} = 11.6$, $E_{21} = \frac{(125)(97)}{383} = 29.67$
$E_{22} = \frac{(125)(145)}{383} = 48.33$, $E_{23} = \frac{(125)(58)}{383} = 19.33$
$E_{24} = \frac{(125)(54)}{383} = 18$, $E_{25} = \frac{(125)(29)}{383} = 9.67$
$E_{31} = \frac{(100)(97)}{383} = 23.73$, $E_{32} = \frac{(100)(145)}{383} = 38.67$
$E_{33} = \frac{(100)(58)}{383} = 15.47$, $E_{34} = \frac{(100)(54)}{383} = 14.4$
$E_{35} = \frac{(100)(29)}{383} = 7.73$

Here we can check to see all $E_{ij} \geq 5$

Homogeneous example con't

Test Statistic

$$\chi^2 = \sum_{i=1}^{n_i}\sum_{j=1}^{n_j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{r\times c}^{all}\frac{(observed - expected)^2}{expected}$$

$= \frac{(34-35.6)^2}{35.6} + \frac{(65-58)^2}{58} + \cdots + \frac{(10-7.73)^2}{7.73} = 14.171$

$df = (r-1)(c-1)$ where $r, c = (3,5)$ so $df = (3-1)(5-1) = 2*4 = 8$

Rejection Region

(1) *Critical Value approach*: Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $\chi^2_{\alpha,df} = \chi^2_{.05,8} = 15.507$
(2) *pvalue approach*: Reject H_0 if $pvalue \leq \alpha$

Homogeneous example con't

Results
We are doing the critical value approach in the "by-hand" example. So $\chi^2_{.05,8} = 15.507$. $14.171 \ngeq 15.507$ so we will fail to reject H_0.

Conclusion (in context)
We failed to reject H_0 so that tells us that the production lines have the same rates of different types of nonconformities, so no one production line produces more of a specific type of nonconformity.