Categorical Data Analyses

Module 11

Statistics 251: Statistical Methods

Updated 2021

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

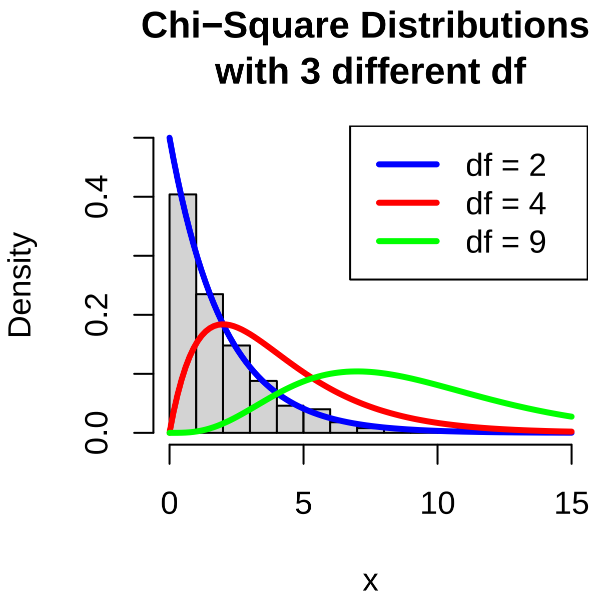
Chi-square distribution

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

$$H_0 : \text{The data follows <specified> distribution}$$

$$H_a : H_0 \textit{ not true } \text{(the data does not follow <specified> distribution)}$$

GoF formulas

Expected value

$$E_i = np_i$$

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
Reject H_0 iff $pvalue \leq \alpha$

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 exists 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

$$H_0 : \text{The row <context> and column <context> are independent}$$

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

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 iff $pvalue \leq \alpha$

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

$$H_0 : \text{The row <context> is distributed the same over the column <context>}$$

$$H_a : H_0 \textit{ not true} \text{ (the distribution is not the same for all categories)}$$

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?

Tomato data

```
tomatoes
```

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

```
n1
```

```
[1] 1611
```

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

Tomato expected counts

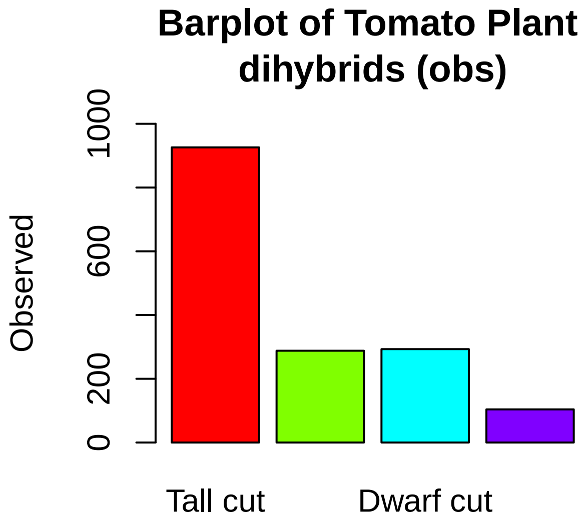
$E = np_i$

```
plantexp
```

```
[1] 906.19 302.06 302.06 100.69
```

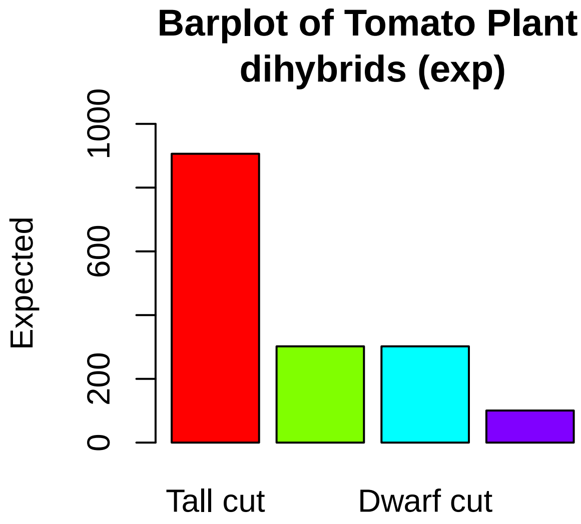
Tomato graphs

```
barplot(plants,names.arg=types1,col=rainbow(4),ylab="Observed",ylim=c(0,1000))
title('Barplot of Tomato Plant \ndihybrids (obs)')
```



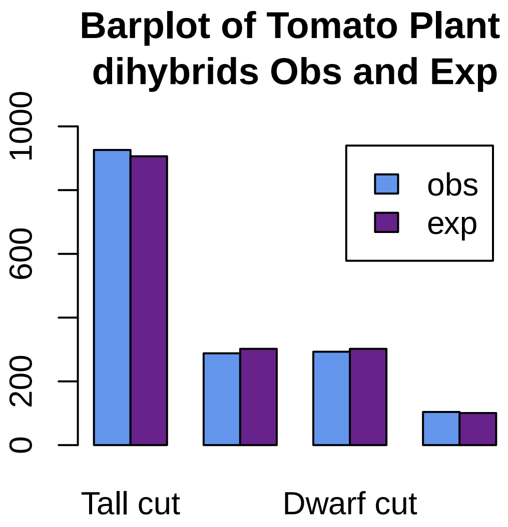
Tomato graphs

```
barplot(plantexp,names.arg=types1,col=rainbow(4),ylab="Expected",ylim=c(0,1000))
title('Barplot of Tomato Plant \ndihybrids (exp)')
```



Tomato graphs

```
barplot(counts,col=colors,legend=plantnames,beside=T,ylim=c(0,1000))
title('Barplot of Tomato Plant \ndihybrids Obs and Exp')
```



Tomato setup

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

$$H_a : H_0 \textit{ not true} \text{ (data does not follow Mendel's Law)}$$

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

Organization of information:
$n = 1611$
4 categories
$p_1 = 9/16$, $p_2 = p_3 = 3/16$, $p_4 = 1/16$
$\alpha = 0.05$ (assumed because not specifically stated otherwise)

Tomato analysis output

```
chisq.test(plants,p=probs)
```

```

    Chi-squared test for given probabilities

data:  plants
X-squared = 1.4687, df = 3, p-value = 0.6895
```

Tomato conclusion

Test statistic: $\chi^2 = 1.4687$ with $df = 3$, and $pvalue = 0.6895$

Results: $pvalue = 0.6895 \nleq \alpha(0.05) \therefore$ (therefore) H_0 is not rejected

Conclusion: since the null is not rejected, that means that the phenotypes of tomato plants do follow Mendel's Law

Error: since H_0 was not rejected, a Type II error (not rejecting null when null is false) could have been made; we think the plants do follow Mendel's Law but they do not

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?

Petrol data

```
gas
```

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

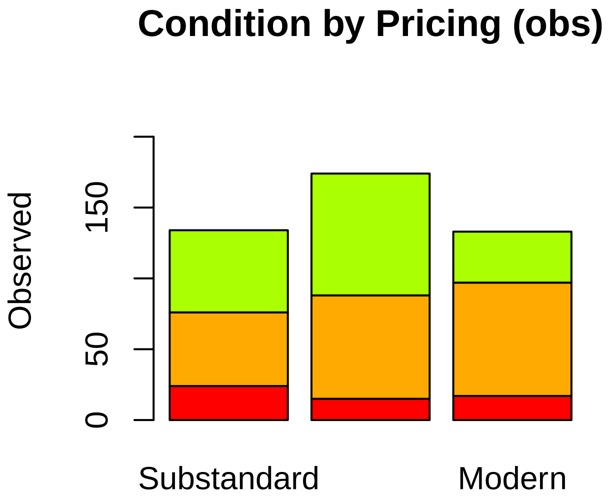
Petrol expected counts

```
gasexp
```

```
             Pricing.Policy
Condition      Aggressive Neutral Nonaggressive
  Substandard       17.02   62.29         54.69
  Standard          22.10   80.88         71.02
  Modern            16.89   61.83         54.29
```

Petrol graphs

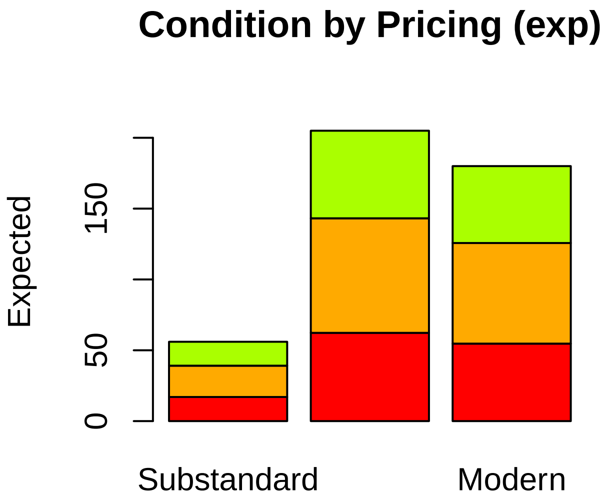
```
barplot(gas,names.arg=rownames(gas),col=rainbow(9),ylab="Observed",ylim=c(0,225))
title('Condition by Pricing (obs)')
```



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

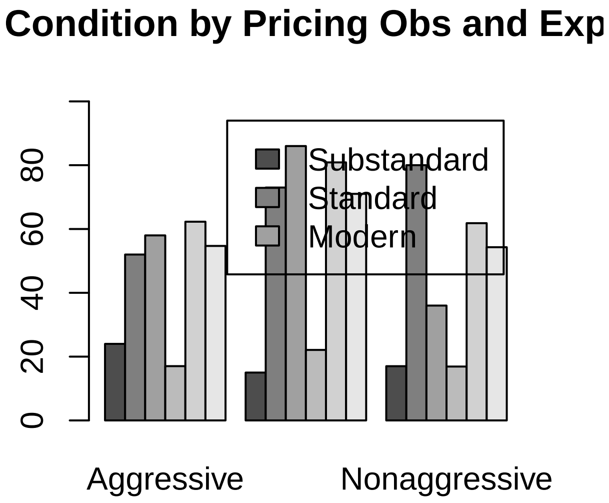
Petrol graphs

```
barplot(gasexp,names.arg=rownames(gas),col=rainbow(9),ylab="Expected",ylim=c(0,225))
title('Condition by Pricing (exp)')
```



Petrol graphs

```
barplot(counts2,legend=rownames(gas),beside=T,ylim=c(0,100))
title('Condition by Pricing Obs and Exp')
```



Petrol setup

$$H_0 : \text{pricing and conditions in gasoline are independent}$$

$$H_a : H_0 \; not \; true \text{ (pricing and conditions in gasoline are dependent)}$$

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

Organization of information:
$n = 441$
$r = 3, c = 3$ rows,columns
$\alpha = 0.05$ (assumed because not specifically stated otherwise)

Petrol analysis output

```
chisq.test(gas)
```

```

	Pearson's Chi-squared test

data:  gas
X-squared = 22.476, df = 4, p-value = 0.0001611
```

Petrol conclusion

Test statistic: $\chi^2 = 22.476$ with $df = 4$, and $pvalue = 0.0001611$

Results: $pvalue = 0.0001611 \leq \alpha(0.05) \therefore$ (therefore) H_0 is rejected

Conclusion: since the null is rejected; knowledge of a station'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 non-aggressive policy

Error: H_0 is rejected a Type I error could have been made (rejecting a true null hypothesis); we think there is a relationship between pricing and condition at gas stations when they are independent

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 nonconformance categories are not the same for the three lines?

Blemish data

```
fail
```

```
              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
```

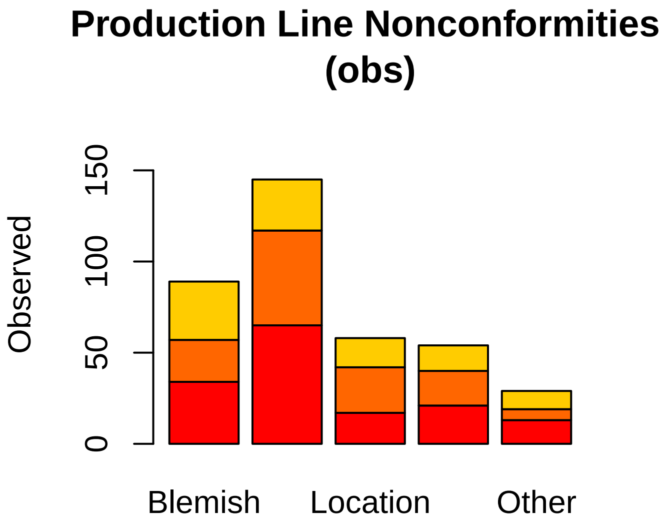
Blemish expected counts

```
defectexp
```

```
               Nonconformity
Production.Line Blemish Crack Location Missing Other
              1   35.60 58.00    23.20    21.6 11.60
              2   29.67 48.33    19.33    18.0  9.67
              3   23.73 38.67    15.47    14.4  7.73
```

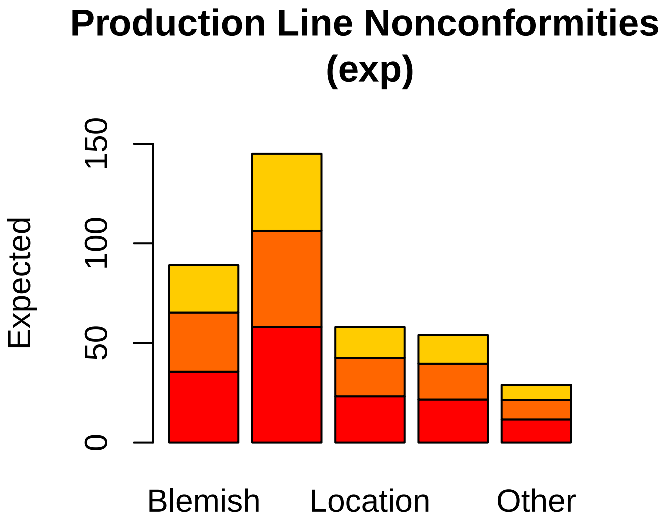
Blemish graphs

```
barplot(defects,names.arg=colnames(defects),col=rainbow(15),ylab="Observed",ylim=c(0,175))
title('Production Line Nonconformities \n(obs)')
```



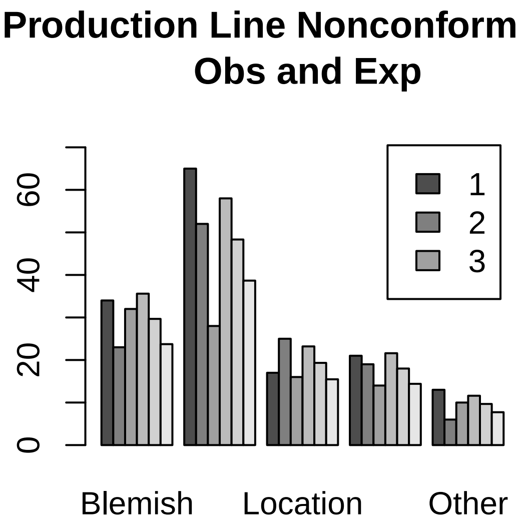
Blemish graphs

```
barplot(defectexp,names.arg=colnames(defects),col=rainbow(15),ylab="Expected",ylim=c(0,160))
title('Production Line Nonconformities \n(exp)')
```



Blemish graphs

```
barplot(counts3,legend=rownames(defects),beside=T,ylim=c(0,75))
title('Production Line Nonconformities \n Obs and Exp')
```



Blemish setup

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

$$H_a : H_0 \ not \ true$$

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

Organization of information:
$n = 375$
$r = 3, c = 5$ rows,columns
$\alpha = 0.05$ (assumed because not specifically stated otherwise)

Blemish analysis output

```
chisq.test(fail)
```

```

    Pearson's Chi-squared test

data:  fail
X-squared = 14.159, df = 8, p-value = 0.07772
```

Blemish conclusion

Test statistic: $\chi^2 = 14.159$ with $df = 8$, and $pvalue = 0.07772$

Results: $pvalue = 0.07772 \nleq \alpha(0.05) \therefore$ (therefore) H_0 is not rejected

Conclusion: since the null is not rejected, production lines have the same rates of different types of nonconformities, so no one production line produces more of a specific type of nonconformity

Error: H_0 was not rejected a Type II error could have been made (not rejecting a false null hypothesis); we think defects are the same across all production lines when they are not (some lines produce more defective parts than others)