Stat 102

Introduction to Business Statistics

Class 5

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Association between categorical variables
- Association does not imply causation
- χ^2 tests for independence
- Goodness of fit
- Summary
- Next time

Last time

Main points:

- Comparisons between groups (comparative analytics)
- Confidence intervals and tests for the difference in means
- Confidence interval and a test for the difference in proportions

Big picture

- Thinking about **prediction** via the lack-of-independence between two variables.
- Start by looking at categorical variables.
- Expand from Stine Ch. 5 which measured association (Cramer's V) to hypothesis testing about independence.

How does knowing the outcome of one variable, change our view of the likelihood of the outcome of another?

Examples:

- How does knowing that the roulette wheel wheel came up black this time, impact the probability that it comes up black next time?
- If the stock market has gone up 4 days in a row, how does that impact the probability that it goes up tomorrow?
- How does knowing that a consumer is affluent, impact the probability that they buy a holiday to Greece tomorrow?
- How does knowing how a student did on the mid-term impact our view of how they will do on the final?

Independence

- When knowing the outcome of one event has no impact on the probability of another event, then we call the events *independent*.
- As a probability statement:

$$P(A|B) = P(A),$$

which is equivalent to

$$P(A \cap B) = P(A)P(B).$$

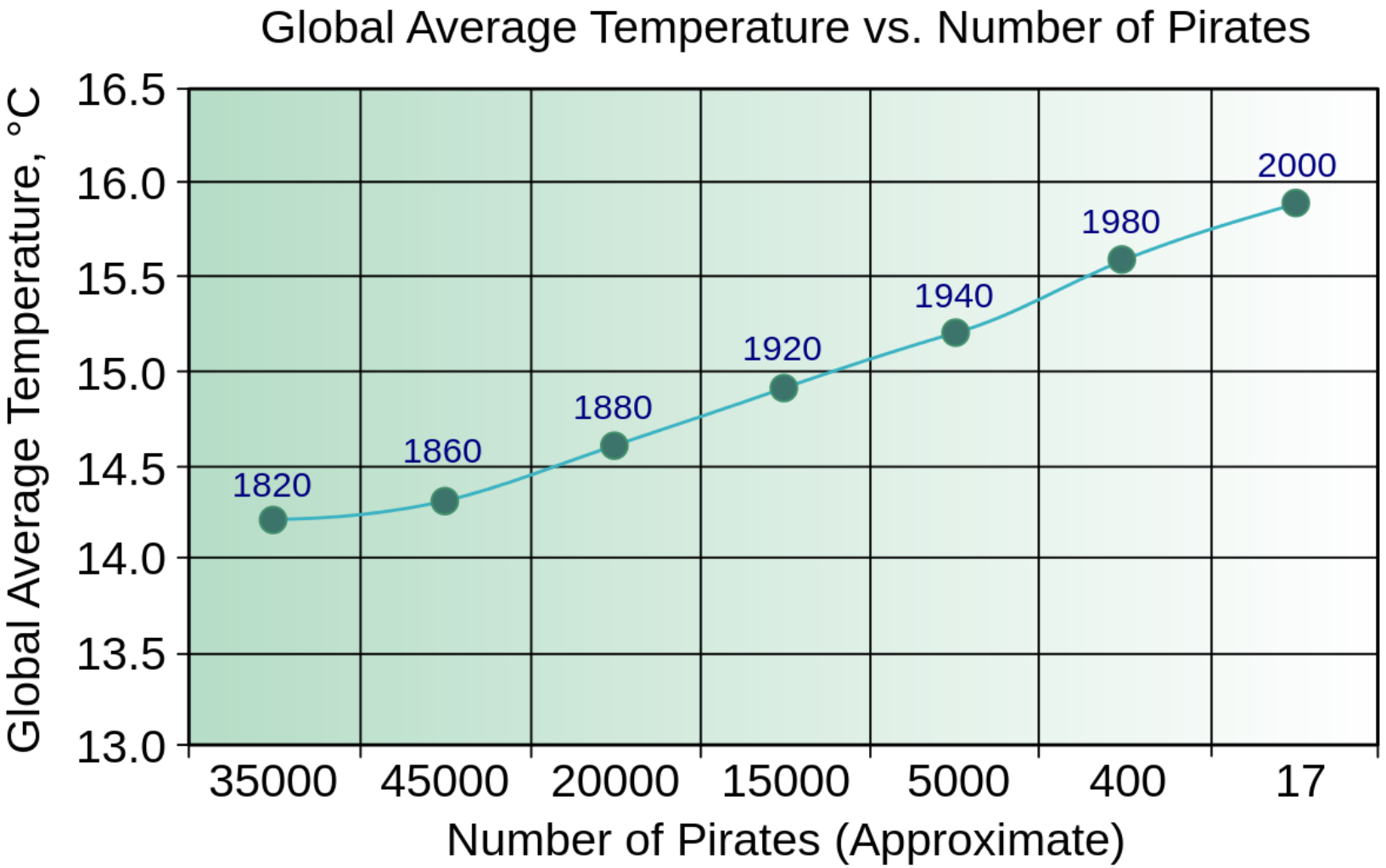
Assuming independence between events makes the probability calculations easier, but in this class we are interested in exploiting dependence to build models that allow us to use knowledge of one variable to improve our forecast of another.

Association and causation

- All because two variables are associated, does not mean that one causes the other to happen.
- Websites that are compromised tend to take longer to load.
- This doesn't mean that the long loading time **causes** the compromise, rather, compromised websites run *malware* which takes a long time to load.
- But, from a prediction perspective, you can still use load time as a valid predictor.
- It is just *proxying* for the presence of *malware*.

Just so you don't forget ...

Figure 1: Association and causation



Contingency tables and independence

- We are considering *count data* – the type of data that we can represented in a contingency table.
- Later in the course, we will formally model this type of data with a *logistic regression* model. The right model for discrete outcome data.
- The standard null hypothesis is that the rows and columns of the contingency table are *independent*.
- Knowing which row you are in, gives no information about which column you are in and *vice versa.*
- This was Chapter 5 of Stine's book, but now with a formal hypothesis test for independence.

Example: purchase intent

- Marketers often measure intent to purchase.
- But intentions do not always equal actions.
- The initial survey question: do you plan to purchase a new vehicle in the next 12 months?
- The follow-up question on a subsequent survey (to the same respondents): did you buy a new vehicle in the last 12 months?
- We want to know if intentions are associated with actions.

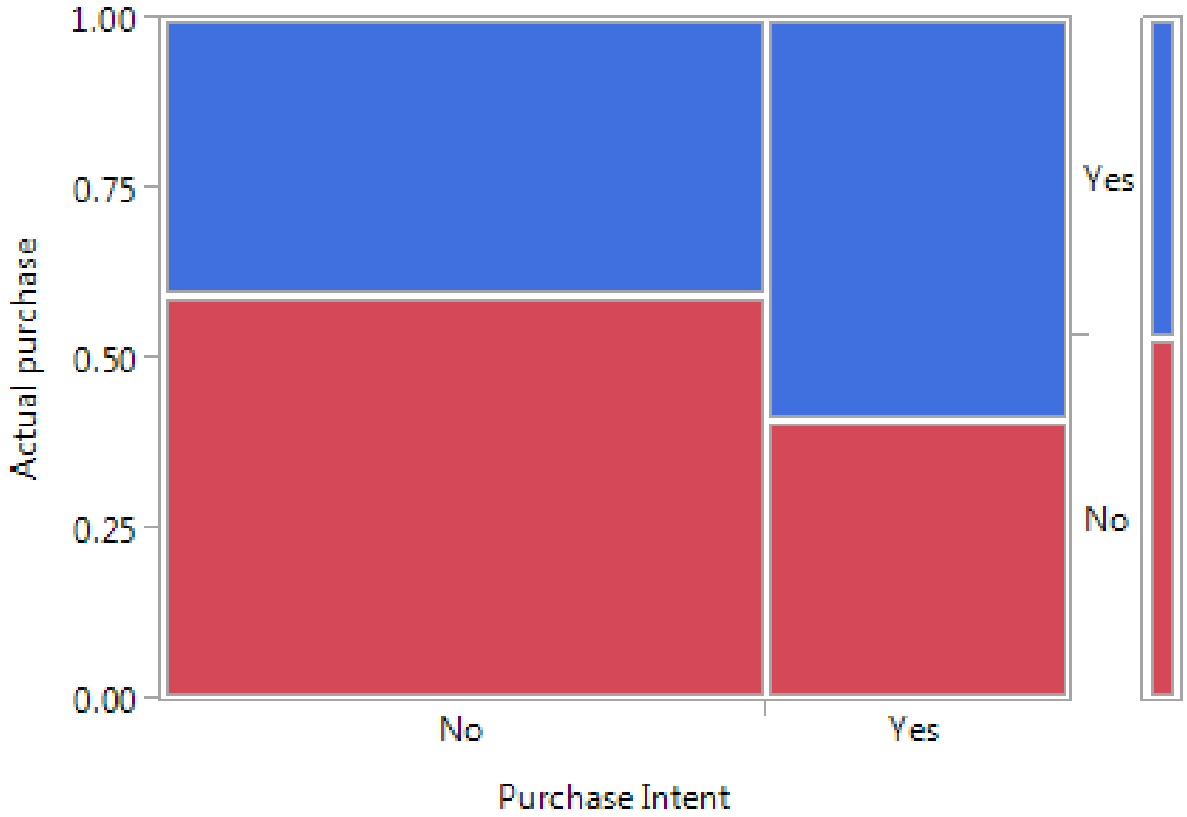
Cross-tabulation of the data from the surveys

Table 1: Crosstabs of the survey purchase data

Intend to purchase	Actual purchase	
	No	Yes
No	390	267
Yes	134	192

The mosaic plot

Figure 2: The mosaic plot for purchase intent



The test will essentially look to see if the heights of the boxes are the same.

Adding percentages: cell percentages

Put in row and column totals:

Table 2: Crosstabs of the survey purchase data

Intend to purchase	Actual purchase		Total
	No	Yes	
No	390	267	657
Yes	134	192	326
Total	524	459	983

Divide the cell counts by 983 to get the cell percentages.

Adding percentages: cell percentages

Table 3: Crosstabs of the survey purchase data with cell percentages

Intend to purchase	Actual purchase		Total
	No	Yes	
No	390 (39.67%)	267 (27.16%)	657 (66.84%)
Yes	134 (13.63%)	192 (19.53%)	326 (33.16%)
Total	524 (53.31%)	459 (46.69%)	983

Divide the individual cell counts by 983 to get the cell percentages. Note that all the cell percentages add up to 100.
Only 19.53% of respondents say, and then do, buy new vehicles.

Marginal distributions

Marginal distributions are obtained by aggregating over either the rows or columns (they are the numbers on the side or bottom of the table).

Table 4: Crosstabs of the survey purchase data with **margins** highlighted

Intend to purchase	Actual purchase		Total
	No	Yes	
No	390 (39.67%)	267 (27.16%)	657 (66.84%)
Yes	134 (13.63%)	192 (19.53%)	326 (33.16%)
Total	524 (53.31%)	459 (46.69%)	983

Approximately one third, 33.16% of respondents say they will buy a new vehicle, but 46.69% actually do.
Marginal distributions do not tell you anything about association.

The hypothesis test

- In English: is actual purchase behavior independent of purchase intent?
- H_0: Intent to purchase and actual purchase are independent.
- H_a: Intent to purchase and actual purchase are not independent.
- Equivalently under independence: all the row-wise conditional probabilities in a column, equal the marginal probability in that column.

$$P(Actual = YES \mid Intent = NO) = P(Actual = YES)$$

$$P(Actual = YES \mid Intent = YES) = P(Actual = YES)$$

- The χ^2 test measures departure from this null hypothesis.

The intent to purchase data

We will build up the test statistic from the independence assumption.
If intent to purchase and actual purchase behavior were independent, then the conditional row probabilities would be the same and equal to the marginal probabilities for actual purchase behavior.

Table 5: Crosstabs of the survey purchase data with **margins** highlighted

Intend to purchase	Actual purchase		Total
	No	Yes	
No			657 (66.84%)
Yes			326 (33.16%)
Total	524 (53.31%)	459 (46.69%)	983

Dropping in the row percentages

Table 6: Crosstabs of the survey purchase data with conditional row probabilities included, assuming they are the same as the marginal percentages (no association)

Intend to purchase	Actual purchase		Total
	No	Yes	
No	(53.31%)	(46.69%)	657 (66.84%)
Yes	(53.31%)	(46.69%)	326 (33.16%)
Total	524 (53.31%)	459 (46.69%)	983

Backing out the expected counts in each cell

Table 7: Crosstabs of the survey purchase data with expected counts included (no association assumption)

Intend to purchase	Actual purchase		Total
	No	Yes	
No	$657 \times 0.5331 = 350.25$	$657 \times 0.4669 = 306.75$	657
Yes	$326 \times 0.5331 = 173.79$	$326 \times 0.4669 = 152.21$	326
Total	524 (53.31%)	459 (46.69%)	983

Backing out the expected counts in each cell

Here's a cleaned up version of the **expected** cell counts, under the independence assumption:

Table 8: Crosstabs of the survey purchase data with expected counts included (independence assumption)

Intend to purchase	Actual purchase		Total
	No	Yes	
No	350.25	306.75	657
Yes	173.79	152.21	326
Total	524	459	983

Recall the original table

Here are the **observed** cell counts:

Table 9: Crosstabs of the survey purchase data with observed counts included (no association assumption)

Intend to purchase	Actual purchase		Total
	No	Yes	
No	390	267	657
Yes	134	192	326
Total	524	459	983

The difference between observed and expected

Table 10: Crosstabs of the survey purchase data with (observed - expected differences included)

Intend to purchase	Actual purchase		Total
	No	Yes	
No	390 - 350.25 = 39.75	267 - 306.75 = -39.75	657
Yes	134 - 173.79 = -39.75	192 - 152.21 = 39.75	326
Total	524	459	983

Two problems:

1. We can't just add up the differences because the pluses and minuses cancel each other out and the sum is zero.
2. The fact that to be off by 39.75, is worse on a relative basis when the observed value is 134, than when it is 390.

The weighted squared differences

We deal with the two problems by calculating for each cell in the table:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} is the observed value in the (i,j) cell, and E_{ij} is the expected value in the (i,j) cell.

$$\begin{aligned}
\chi^2 &= \frac{(390 - 350.25)^2}{350.25} + \frac{(267 - 306.75)^2}{306.75} + \frac{(134 - 173.79)^2}{173.79} + \frac{(192 - 152.21)^2}{152.21} \\
&= \frac{(39.75)^2}{350.25} + \frac{(-39.75)^2}{306.75} + \frac{(-39.75)^2}{173.79} + \frac{(39.75)^2}{152.21} \\
&= 29.13
\end{aligned}$$

In practice you will get this number from the JMP output under the crosstabs and labeled as **Pearson's ChiSquare**.

The χ^2 test

- Take the χ^2 statistic and compare it to the quantile of the appropriate χ^2 distribution.
- χ^2 has a *degrees of freedom*, just like the t.
- For the test of independence in a contingency table the degrees of freedom equals (r - 1) $\times$ (c - 1) where r and c are the number of rows and columns respectively.
- For a 2-by-2 table, the $df = 1$.

Comments on the χ^2 distribution

- Take a Z (standard normal) and square it: $Z^2 = \chi^2_{(1)}$.
- Take n independent standard normals, $Z_1, Z_2, \cdots, Z_n$ and add them up:

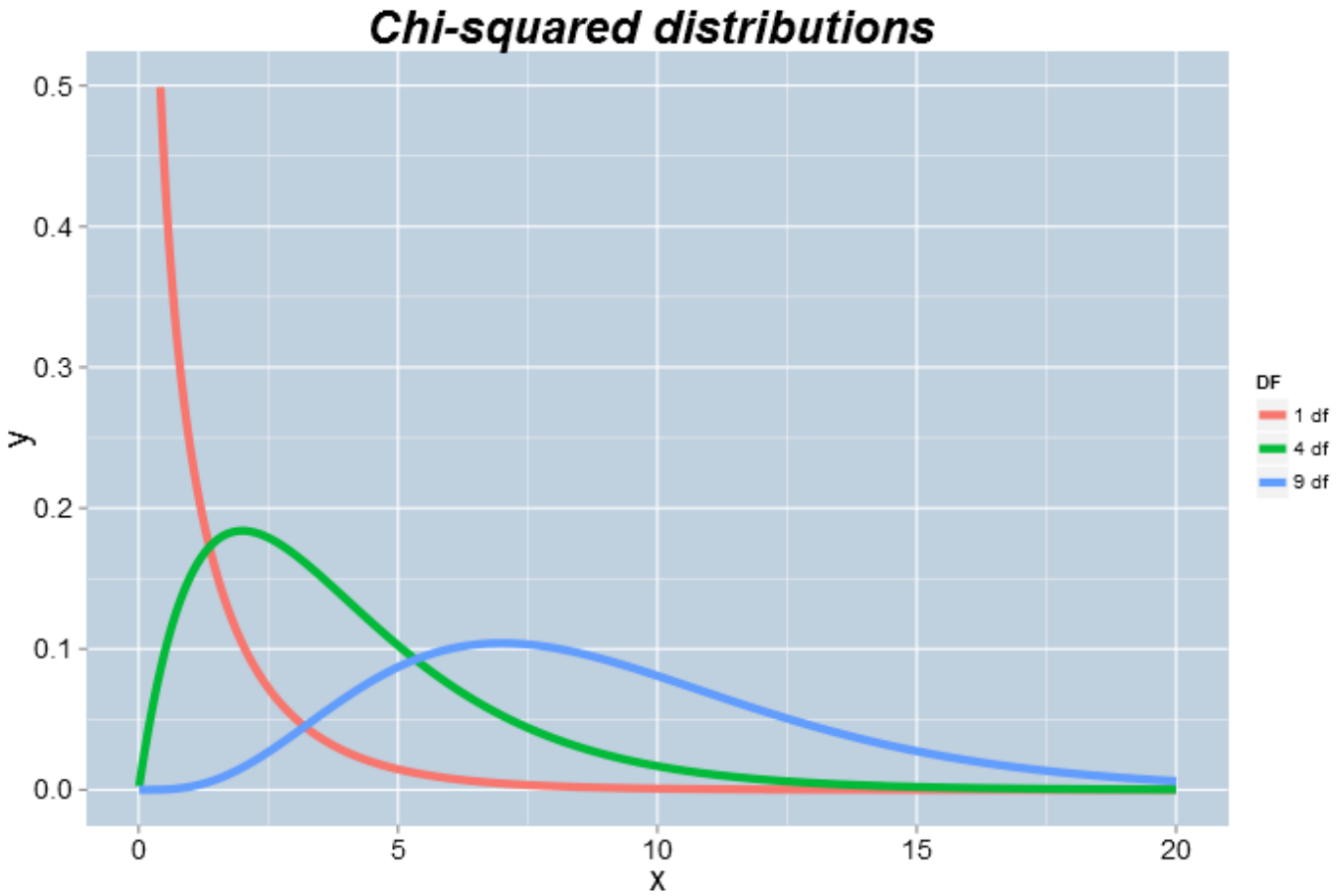
$$\sum_{i=1}^{n} Z_i^2 = \chi^2_{(n)}.$$

- For a χ^2 random variable X_n^2 on n df:

$$E(X_n^2) = n, \qquad Var(X_n^2) = 2\,n, \qquad SD(X_n^2) = \sqrt{2\,n}$$

Pictures of the χ^2

Figure 3: Various chi-squared distributions



Finding the cut-off value

- When you have a Z, then the cut-off is about 2 (1.96).
- Therefore for a $\chi^2_{(1)}$ it is $1.96^2 = 3.8416$.
- There are tables at the back of the book (p. 775) (use the $\alpha = 0.05$ column).
- With software, we will get a p-value and can compare the p-value to α.

Output for the χ^2 test for independence

Figure 4: The χ^2 test output

Tests

N	DF	-LogLike	RSquare (U)
983	1	14.618394	0.0215

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	29.237	<.0001*
Pearson	29.176	<.0001*

Fisher's Exact Test	Prob	Alternative Hypothesis
Left	1.0000	Prob(Actual purchase=Yes) is greater for Purchase Intent=No than Yes
Right	<.0001*	Prob(Actual purchase=Yes) is greater for Purchase Intent=Yes than No
2-Tail	<.0001*	Prob(Actual purchase=Yes) is different across Purchase Intent

Conditions for the test:

- Expected cell frequencies are at least 10 for each cell.
- Or, 5 per cell, if the test has 4 or more df.

Conclusion

- This is a highly significant result.
- There is a significant association between intent and actual purchase behavior.

A second example with the Auto Show data

Who rates the Show higher: those who attend frequently or those that are less frequent attendees?

Table 11: Crosstabs of attendance frequency against Show rating

Attendance regularity	Rating			
	Fair	Good	Excellent	Outstanding
First time	2	34	135	93
Occasionally	3	40	84	34
Most years	10	57	144	44
Every year	16	101	310	166

There was a rating of "Poor" but nobody checked that.

JMP summary of the attendance data

Table 12: Attendance by rating

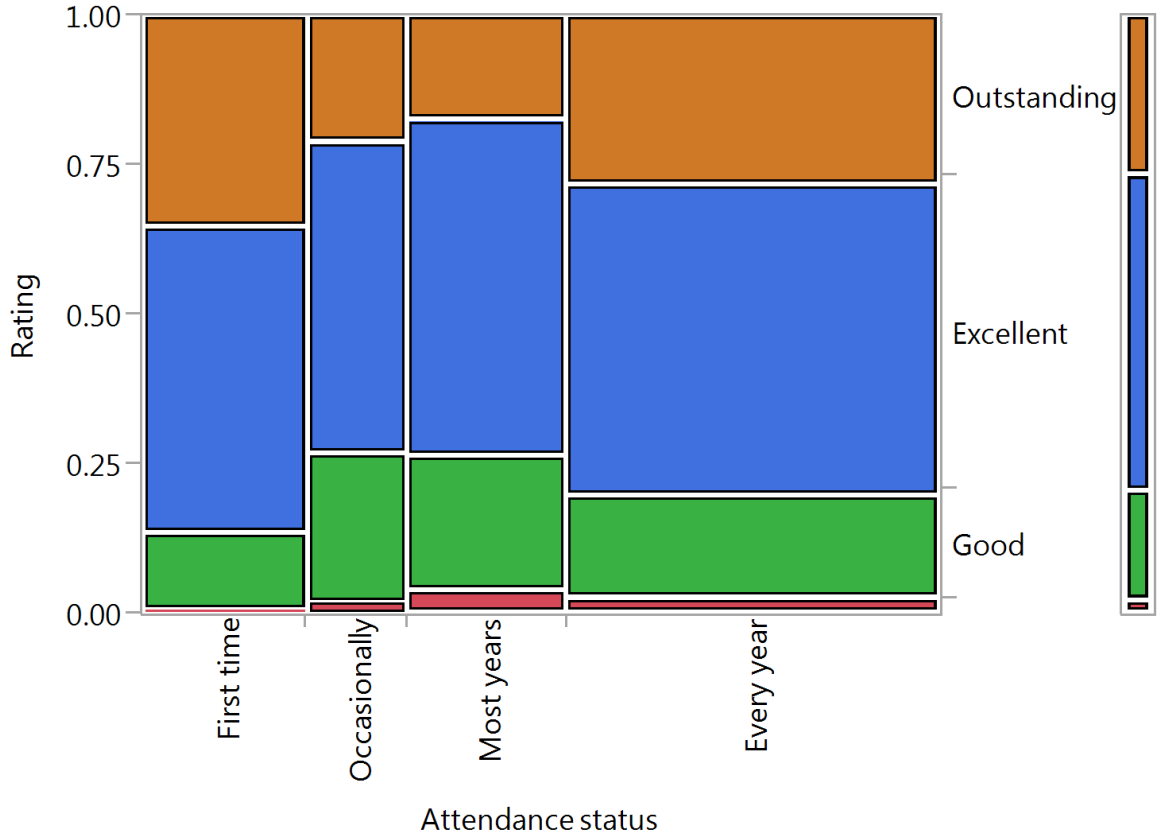
Attendance status	Rating				
Count Total % Col % Row %	Fair	Good	Excellent	Outstanding	
First time	2	34	135	93	264
	0.16	2.67	10.60	7.31	20.74
	6.45	14.66	20.06	27.60	
	0.76	12.88	51.14	35.23	
Occasionally	3	40	84	34	161
	0.24	3.14	6.60	2.67	12.65
	9.68	17.24	12.48	10.09	
	1.86	24.84	52.17	21.12	
Most years	10	57	144	44	255
	0.79	4.48	11.31	3.46	20.03
	32.26	24.57	21.40	13.06	
	3.92	22.35	56.47	17.25	
Every year	16	101	310	166	593
	1.26	7.93	24.35	13.04	46.58
	51.61	43.53	46.06	49.26	
	2.70	17.03	52.28	27.99	
	31	232	673	337	1273
	2.44	18.22	52.87	26.47	

The frequency representation of data

- If you are putting table counts into JMP it can be easier to use the **frequency** representation of the data set.
- Each row in the data table represents one of the possible cross-classifications of all the categorical variables.
- In this example there are 16 (4×4) rows.
- There is an extra column called *Frequency* which represents the counts in the original table.
- Give the Frequency column a role of *frequency* in the analysis.
- Other useful JMP features are the properties a variable can have. In this case **value labels** and **value ordering** have been used to control the output form.
- See the JMP data table *Attendance by rating* for the details.

Mosaic plot of the attendance data

Figure 5: Mosaic plot of attendance by rating



Description

- Hardly anyone (2.44%) gave the Show a *fair rating*.
- Most of the ratings are either an *Excellent* (53%) or *Outstanding* (26%).
- It is the first timers and those who attend every year that most often give the highest rating.
- χ^2 for the table is 35.527 (from the JMP output).

Figure 6: Pearson χ^2 and the p-value

Tests

N	DF	-LogLike	RSquare (U)
1273	9	18.391642	0.0133

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	36.783	<.0001*
Pearson	35.527	<.0001*

Conclusion: rating and attendance frequency are not independent.

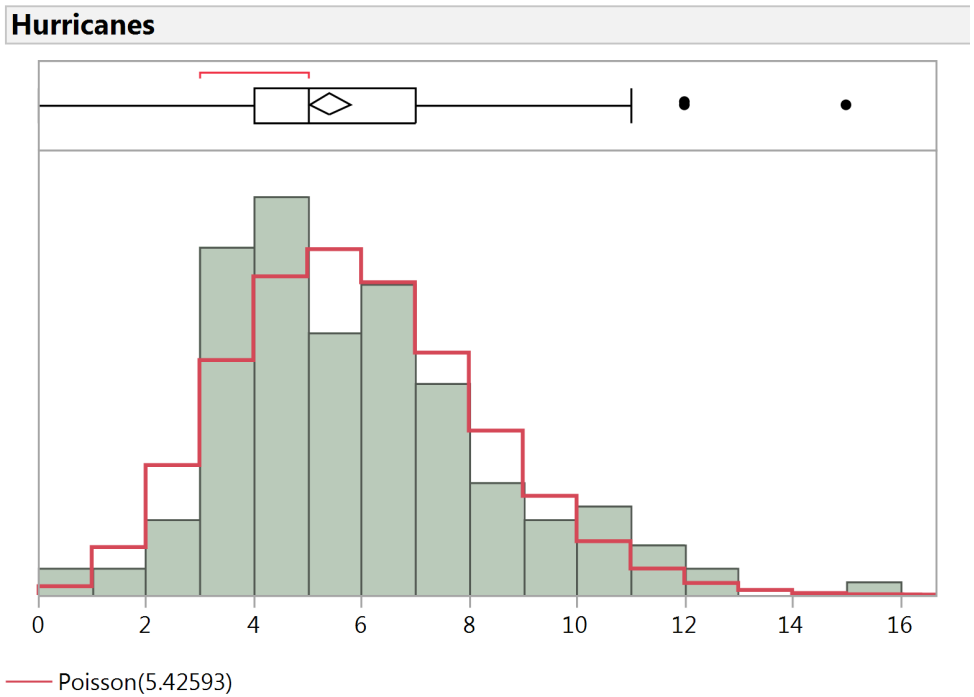
The χ^2 goodness-of-fit statistic

- This is used when we have a single categorical variable and we want to test whether the frequencies are in accordance with some pre-specified distribution.
- Often these frequencies come from some well known distribution, for example:
  - Does it look like the data comes from a Poisson distribution?
  - Does it look like the data comes from a Binomial distribution?
- So this test is useful in assessing the realism of many modeling assumptions
- When the outcome is potentially Normally distributed, we use a Normal Quantile Plot.
- This is the same idea, a diagnostic, but we are using a formal test statistic instead.

Hurricanes

Does it look like the annual number of hurricanes in the North Atlantic basin follows a Poisson distribution?

Figure 7: Annual hurricane distribution with Poisson overlaid



Process

- Find the expected counts under the Poisson assumption.
- Recall: $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$ where λ is the Poisson mean.
- Find the χ^2 statistic.
- Find the cut-off/p-value for the χ^2 statistic.
- DF for the χ^2 = #cells - #constraints.
- In the Poisson example we constrain:
  1. The number of counts to equal 162
  2. And use the mean $\bar{x} = 5.426$ to calculate the mean of the Poisson.
- We may need to collapse cells so no category has less than 5 observations.

The counts

Number	0-2	3	4	5	6	7	8	9	10+	Total
Observed	10	28	32	21	25	17	9	6	14	162
Expected	15.1	19.0	25.7	27.9	25.3	19.6	13.3	8.0	8.1	162
χ^2	1.72	4.26	1.54	1.71	0.00	0.34	1.39	0.5	4.29	15.77

- Where for example does the 19.0 come from?
  $162 \times \frac{5.426^3}{3!} e^{-5.426} = 19.0$ (with a little rounding).
- Where does the 1.54 come from? $\frac{(32-25.7)^2}{25.7} = 1.54$.
- The χ^2 statistic $= 15.77$. It is on 9 - 2 $= 7$df.
- From Stine's table (p. 775) : $\chi^2_{(7),0.05} = 14.067$.
- So we have evidence (test stat $= 15.77 \geq 14.067$) to reject the Poisson null hypothesis.
- From other software the p-value is 0.027.

Some comments

- Notice that it is the tails of the distribution that contribute the most to the lack-of-fit. That is, we observe more variability in practice than the Poisson predicts.
- For this data, the mean $\bar{x} = 5.426$ and $var(x) = 6.445$. In an ideal Poisson the mean and variance are equal.
- We say that the observed data is *over-dispersed* with respect to the Poisson distribution.

Module summary

1. χ^2 test for independence.
2. χ^2 test for goodness-of-fit.

Next time

- Regression models