Lecture 23 **Contingency Tables**

STAT 102

- Stat 102 has concentrated on analysis of explanatory and predictive models with one dependent variable (Y) and one or more predictive variables or factors (often denoted by X).
- Data Structures examined include situations described as:

- Ordinary linear regression:
  Y **continuous**; One **continuous** *x*-variable.
- Multiple linear regression:
  Y **continuous**; Several **continuous** *x*-variables
- Polynomial regression:
  Like multiple regression, but with some polynomial terms
- ANOVA:
  Y **continuous**; One or more **qualitative** X-factors
- Regression and ANOVA combined:
- Logistic Regression
  Y **qualitative**; One or more **continuous** X-factors
  (We also included a qualitative X-factor (sex) in one example)

We haven't yet talked about ….

Contingency Tables

- In these, **both** *Y* and *X* are **qualitative**.
- The focus is on understanding the ***relation*** between *Y* and *X*;
  We don't *necessarily* think of one (some) as causing the other, or as being useful to predict the other.
- Such situations are often analyzed in Contingency Tables;
  In these, the null hypothesis, H_0, is [usually] Independence of the various factors.

- We'll give two examples
- We'll describe how to compute the test statistic, with complete formulas
- We'll also show where to get the tests in JMP

Example

Wine and Music:

Do shoppers buy more wine when appropriate music is played?

Data: A supermarket in N. Ireland played no background music, French music, or Italian music at random times of the day. Here is the number of purchases for different combination of type of music and type of wine: **2-way table of results**

Type of Wine By **Type of Music**

Count	French	Italian	None	Total
French	39	30	30	99
Italian	1	19	11	31
Other	35	35	43	113
Total	75	84	84	243

Taken from Moore and McCabe, *Introduction to the Practice of Statistics*, *5th edition*. Original source, Ryan, Northrup-Clewes, Knox and Thurnham (1998). The effect of in-store music on consumer choice of wine. *Proc. Nutrition Soc*. *57*. p1069.

Test of Independence

Null Hypothesis:

In words -- H_0: Row effects are independent of Column effects

Symbolically –

The (population) probabilities for the row effects (Wine) are denoted by

$$\begin{array}{l} \rho_1 = P(BuyingFrenchWine), \\ \textit{Similar}\ \text{def for}\quad \rho_2, \rho_3 \end{array}.$$

The probabilities for the column effects (Music) are

$$\begin{array}{l} \kappa_1 = P(FrenchMusic) \\ \kappa_2, \kappa_3\ \text{are similarly defined} \end{array}.$$

H_0 is that the probability of a count in the ***i, j*** cell is the product, for example

$$\Pr(\text{French\&None}) = \Pr(\text{count in cell } 1,3) = \rho_1 \kappa_3 \triangleq P_{13}.$$

H_a is that H_0 is not true.

Computations for Test

Let N_{ij} denote the number of observations in the *i*, *j* cell. Let N_{i+} denote the sum of observations in the *i*-th row, etc. The table now looks like

Count	French	Italian	None	Total
French	$N_{11}=$ **39**	$N_{12}=$ **30**	$N_{13}=$ **30**	$N_{1+}=99$
Italian	$N_{21}=$ **1**	$N_{22}=$ **19**	$N_{23}=$ **11**	$N_{2+}=31$
Other	$N_{31}=$ **35**	$N_{32}=$ **35**	$N_{33}=$ **43**	$N_{3+}=113$
Total	$N_{+1}=75$	$N_{+2}=84$	$N_{+3}=84$	$N_{++}=243$

We can estimate the ρ and κ terms in the obvious way – *eg*,

$$\hat{\rho}_i = N_{i+}/N_{++}, \; \hat{\kappa}_j = N_{+j}/N_{++}.$$

Under H_0 the cell probabilities then have estimates

$$\hat{P}_{ij} = \hat{\rho}_i \hat{\kappa}_j$$

Computations (cont)

We can put this together to get estimates for what the cell counts should be **if H_0 is true.**

The expected (under H_0) cell counts would be

$$\text{Expected}_{ij} \triangleq E_{ij} = N_{++}\hat{P}_{ij}.$$

Pearson's Chi-Square statistic for testing H_0 is

$$\chi^2 = \sum_{ij} \frac{\left(N_{ij} - E_{ij}\right)^2}{E_{ij}}.$$

Under H_0 it has a Chi-Square distribution with

$$df = (I-1)(J-1)$$

In our example, $df = 2 \times 2 = 4$.

Your book has tables of the Chi-Square distribution, or use JMP.

We REJECT when $\chi^2 > C$, with C as found in the table.

Computations (summary)

- Get the table. It has I rows and J columns and entries N_{ij}.
- Calculate E_{ij}, the Expected (under H_0) cell counts. Looking back we see they have the formula

$$E_{ij} = \frac{N_{i+}}{N_{++}} \frac{N_{+j}}{N_{++}} N_{++} = \frac{N_{i+} N_{+j}}{N_{++}}.$$

- Calculate

$$\chi^2 = \sum_{ij} \frac{\left(N_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum \frac{\left(\text{Observed}_{ij} - \text{Expected}_{ij}\right)^2}{\text{Expected}_{ij}}.$$

- Use $df = (I-1)(J-1)$. Get the critical value, C, from the Chi-Square table with this *df*. Reject whenever $\chi^2 > C$.

Analysis in JMP

JMP performs the necessary calculations via "Fit Y by X".

(Music as Y Wine as X; number as Frequency)

Wine By Music

Count Expected Cell Chi^2	French	Italian	None	
French	39 30.5556 2.3337	**30** **34.2222** **0.5209**	30 34.2222 0.5209	**99**
Italian	1 9.5679 7.6724	19 10.716 6.4038	11 10.716 0.0075	31
Other	35 34.8765 0.0004	35 39.0617 0.4223	43 39.0617 0.3971	113
	75	**84**	84	243

For example, $34.2222 = \dfrac{84 \times 99}{243}$ and $0.5209 = \dfrac{(30 - 34.2222)^2}{34.2222}$.

Analysis in JMP (cont)

JMP also gives the value of χ^2, and the resulting P-value:

Test	**ChiSquare**	**Prob>ChiSq**
Likelihood Ratio	21.875	0.0002
Pearson	**18.279**	**0.0011**

If you calculated by hand you would get this from the table on *p*9 as

$$18.279 = 2.3337 + 7.6724 + .0004 + .5209 + \dots \ .$$

(*You can ignore the Likelihood Ratio statistic. It belongs to a different statistical method. It comes from a different set of formulas than Pearson's Chi-Square. The P-values from the two tests are nearly always very similar.*)

The result of this calculation is that we Reject H_0.

(We do so at any reasonable level of significance since P-value = .0011).

This should be followed by a summary of the nature of the effects we have detected.

Nature of the Effects

We have Rejected H_0.

Therefore we believe **choice of wine** and **nature of music** are related to each other. What is the relation?

There are formal ways to investigate this *[see the Appendix to this lecture]*, but for Stat 102 we recommend looking at the numbers and use common-sense.

You can look at the cell counts (as on p6), and/or at the cell Chi-Square values (as on p9) and/or at row or column % (as below).

Used carefully, any of these looks should yield similar conclusions.

Here is the table of counts and column %:

Nature of Effects (cont)

Wine By Music

Count Col %	French	Italian	None	
French	**39** **52.00**	**30** **35.71**	30 35.71	99
Italian	**1** **1.33**	**19** **22.62**	11 13.10	31
Other	35 46.67	35 41.67	43 51.19	113
	75	84	84	243

Interesting conclusions are visible in the colored **2×2** section:

"When French music plays customers seem much more likely to purchase French wine than they are to purchase French wine when Italian music is playing."

Also

“When French music is playing customers seem **very much less** likely to purchase Italian wine than they are to purchase Italian wine when Italian music is playing.”

Also (from the **last two columns**)
“There doesn’t seem to be much difference in distribution of wine choice between times when Italian music is playing and when no music is playing. (There’s a slightly greater % of people choosing Italian wine when Italian music is playing, but it’s not a large difference.)”

Second Example
Income and Job Satisfaction

About 50 years ago, a classic sociological study looked at the relationship between a worker's salary and their job satisfaction.

One two-way table from this study looked at the results from a survey of worker's income and a questionnaire judging their job satisfaction.

Annual job **INCOME** was divided into 4 categories

<$6,000, $6,000 - $15,000, $15,000 - $25,000, >$25,000.

Their **SATISFACTION** was graded into four categories

1 = very dissatisfied, **2** = somewhat dissatisfied,

3 = somewhat satisfied, **4** = very satisfied.

[By accident, both examples in this lecture have square contingency tables. But that's not necessary; rectangular tables are also common].

Data table and results of the analysis follow:

Income and Job Satisfaction Data Table

Income By Satisfaction

Count Row %	1	2	3	4	Total
(1) 0-6	20 9.71	24 11.65	80 38.83	82 39.81	206
(2) 6-15	22 7.61	38 13.15	104 35.99	125 43.25	289
(3) 15-25	13 5.53	28 11.91	81 34.47	113 48.09	235
(4) 25+	7 4.09	18 10.53	54 31.58	92 53.80	171
Total	62	108	319	412	901

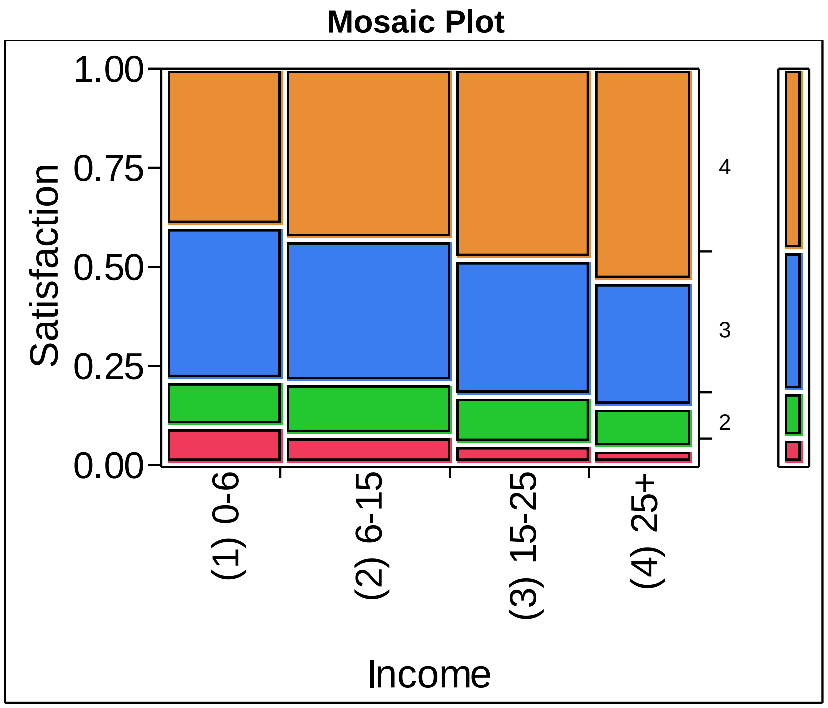
The level of satisfaction does seem to increase somewhat with income. For example 9.7% of the lowest income people are very dissatisfied, whereas only 4.1% of the highest income people are dissatisfied. **BUT**

Are the differences we see in the row % **statistically significant** to show that Income and job satisfaction are not independent?

Data taken from Agresti, *Categorical Data Analysis*, which contains the original citation.

An Aside

The Mosaic Plot in JMP provides a very nice visualization for this data:



You can clearly see how the % of ***very dissatisfied*** *workers goes down as earnings go up, and the % of* ***very satisfied*** *workers increases with income.*

Pearson Chi-Square Test of Independence

Test of H_0: Satisfaction is independent of Income

df = $3 \times 3 = 9$.

Results from JMP:

Test	**ChiSquare**	**Prob>ChiSq**
Likelihood Ratio	12.037	0.2112
Pearson	**11.989**	**0.2140**

The P-value is 0.2140. **We cannot reject the null hypothesis.**

[We have to regretfully conclude that numbers like those in the table might well have occurred by chance.

BUT note that not only do we have numbers like these, but they occur in a very obvious pattern. Our test isn't looking for any special pattern. Might there be a better test to use on this data ***? See the Appendix*** *.]*

Appendix
Going Beyond the Test of Independence

This Material is ***OPTIONAL***. It will **not** be on our exam.

Our other methods have provided regression coefficients or estimates of factor effects that can be used to understand the nature of factor effects. These are also very useful to predict out-of-sample observations, especially when the overall null hypothesis is false.

One method called a **G**eneral **L**inear **M**odel (with a “log”-link) corresponds to a model of the form:

$$Log_e\left[E\left(N_{ij}\right)\right] = \mu + \alpha_i + \beta_j + \gamma_{ij} \ .$$

This has the same character as our previous Two-way ANOVA, except that it models $Log_e\left[E\left(N_{ij}\right)\right]$ rather than $E\left(N_{ij}\right)$.

GLM (log-link)
Wine and Music Data

Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	38.99	77.98	8	<.0001
Full	21.72			
Reduced	60.71			

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
Wine	2	68.33	<.0001
Music	2	11.31	0.0035
Music*Wine	4	21.88	0.0002

These tables show that various things we can see in the table are statistically significant. In particular, the interaction is significant.

It is this interaction term that describes the dependence of Wine and Music. The significance of this term is another confirmation of the result of our earlier Pearson Chi-Square test.

Parameter Estimates

This analysis yields a parameter estimates table that one can use to estimate the values of relevant quantities.

For example one can estimate a quantity like $E\left[Log\left(N_{12}/N_{+2}\right)\right]$, and hence also $E\left[N_{12}/N_{+2}\right]$. This corresponds to the proportion of Italian Music listeners who purchase French wine.

Parameter Estimates

Term	**Estimate**	**Std Error**	**L-R ChiSquare**	**Prob>ChiSq**
Intercept	2.96	0.13	51.04	<.0001
Wine[French]	0.52	0.14	20.25	<.0001
Wine[Italian]	-1.18	0.24	67.05	<.0001
Music[French]	-0.56	0.24	10.64	0.0011
Music[Italian]	0.34	0.14	7.05	0.0079
Music[French]*Wine[French]	0.73	0.25	15.21	<.0001
Music[French]*Wine[Italian]	-1.22	0.46	16.53	<.0001
Music[Italian]*Wine[French]	-0.42	0.17	7.82	0.0052
Music[Italian]*Wine[Italian]	0.83	0.26	15.21	<.0001

A Peculiar Feature

The preceding tables do have one peculiar feature. We can rationalize this as understandable and not important, but we don't fully understand it.

Note that on p19 the Effect Test for "Music" is significant. This suggests that the three different types of music were played to very different total numbers of wine buyers. [More precisely, one would think this is a test of $E\left(N_{+1}\right)=E\left(N_{+2}\right)=E\left(N_{+3}\right)$*. However,* $N_{+1}=75, N_{+2}=84, N_{+3}=84$*. These numbers aren't very different, so it's very unclear why a test of this null hypothesis should Reject.*

The explanation must lie in the fact that this test is after controlling for the effect of wine ***and also the effect of the interaction terms****. There must be something about the presence of the interaction terms that's driving this peculiarity.*

As confirmation of that conjecture, we note that the test of a model without interactions gives NO SIGNIFICANCE to the Music factor ---

Effect Tests

Source	**DF**	**L-R ChiSquare**	**Prob>ChiSq**
Wine	2	55.43	<.0001
Music	2	0.68	0.7134

GLM (log-link)

Income and Job Satisfaction Data

For this data we could use a model in which Income levels are still treated as categories, BUT in which the satisfaction numbers (1, 2, 3, 4) are treated numerically. This allows for a model that has a monotone trend. (More precisely, a linear trend for $Log_e\left[E\left(N_{ij}\right)\right]$.)

In the model without interactions we have

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
Income	3	32.92	<.0001
Satisfaction	1	370.40	<.0001

[Note: There is one DF for Satisfaction because there is exactly one regression coefficient for this factor.]

Both income and satisfaction are significant in this model as one might expect from looking at the tables. [The cell numbers change as income level changes, and they ***increase*** as satisfaction increases.]

The Interaction Term
(A Measure of Dependence)

Here is the table for the model with an interaction term:

Effect Tests

Source	**DF**	**L-R ChiSquare**	**Prob>ChiSq**
Income	3	32.91	<.0001
Satisfaction	1	369.64	<.0001
Satisfaction*Income	3	8.65	0.0343

Our earlier test of independence yielded P-value = .21 – not significant.

But now, note that the p-value = .0343 < 0.05. So at the conventional level this interaction is significant – although not dramatically so.

Looking at the data this way enables us to be (weakly) convinced that there is some sort of interaction (= dependence) between income and satisfaction in the worker population.