**Module 4: Association and Dependence**

# Association

The value of one variable often influences the value of another variable:

Price of item and quantity sold

Type of a home and purchase price

Nationality and happiness over World Cup

Risk taking and insurance premium

Association simply means that the value taken on by one variable is anticipated by that of another, opening the door to prediction.

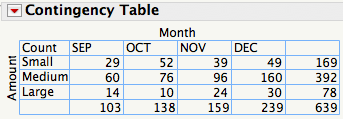
Caution: Association ≠ Causation, as in the example of   
price and quantity or risk and premium.

Understanding association is important in statistics and requires that we consider two or more variables *simultaneously*. In this module, we introduce association in data and the corresponding notion between random variables.[[1]](#footnote-1)

# Example: amazon\_purchases.jmp

The following *contingency table* summarizes *n* = 639 visits to amazon.com that resulted in purchases during the fall season.[[2]](#footnote-2)

The purchases are organized by two categorical characteristics: the amount of the purchase and the month of the purchase. Small purchases are for up to $20, medium run from $20 to $100, and large are for more than $100.



column totals

row totals

The margins of the table in the last row and column show the total counts.

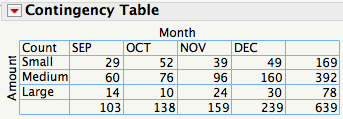
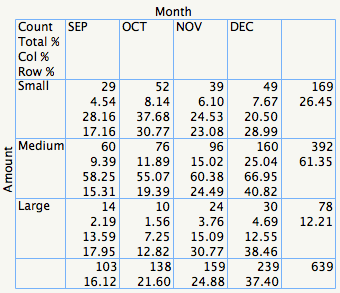
The volume of purchases rose steadily over these four months and most purchases are in the range $20 to $100.

Are the variables Month and Amount associated?

# Percentages in Tables

By adding the right percentages to the table, we see that Month and Amount are associated. We have to decide which percentages to add.

By default, JMP shows all of the percentages.

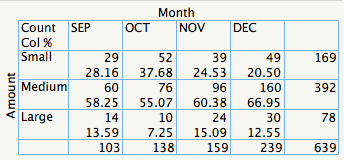
For example, the count 29 in the upper left cell of the table is  
 4.54% of all of these purchases (29/639, % of total)  
 28.16% of purchases in September (29/103, % of column)  
 17.16% of small purchases (29/169, % of row)

All of these are possibly interesting, but too much of a good thing.

Which of these choices for the percentages best shows that the proportion of large purchases changed over these months?

# Percentages and Association

This table shows the counts and the column percentages.



The column percentages show that the proportion of purchases that were Large, for instance, vary over these months

14% in September,  
 7% in October,   
15% in November and  
13% in December.

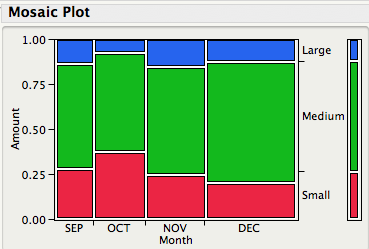
Because these percentages change from column to column, we conclude that   
Month and Amount are associated.[[3]](#footnote-3)

Association: column percentages depend on the column.

No association: column percentages match the marginal percentages.

# Mosaic Plot

A mosaic plot displays a contingency table and visually reveals association.[[4]](#footnote-4)



The areas of the colored rectangles are proportional to the counts in the cells of the contingency table, obeying the *area principle* for statistical graphics. The widths of the columns are proportional to the marginal totals by Month. The segmented bar at the right summarizes the row margins and color scheme.

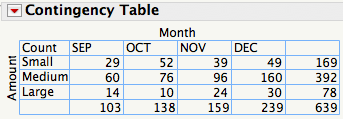
How does the mosaic plot show us that Amount and Month are associated?

What would the mosaic plot look like if there were no association?[[5]](#footnote-5)

# Chi-Squared: Strength of Association

We can see the association between Amount and Month in the contingency table and in the mosaic plot, but it is helpful to have a numerical statistic that quantifies the strength of association.

The chi-squared statistic is the most common measure of association in a table. Chi-squared measures the deviation of the observed counts from those in a contingency table with the same marginal totals and *no* association.

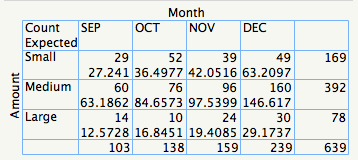


If there were *no* association, what values would fill the cells in this table?

Recall no association means, for example, that the column percentages are the same in every column. Hence, use the margins to fill in the hidden cells.

Example: For small purchases in September:   
 (# in Sept) (% small purchases overall) = 103 × (169/639) = 27.241

The lower number in each cell of the following table shows the “expected count” that we’d find were there no association.



The chi-squared statistic, sometimes written as χ2,  
 - subtracts the numbers in each cell (but not the margins),  
 - squares the differences,  
 - divides the squared differences by the expected counts, and lastly  
 - sums the ratios.

JMP will do this for us. It calls the statistic the “Pearson chi-square.”

For these data, χ2 = 16.5. Is that a lot?[[6]](#footnote-6)

Chi-squared is tricky to interpret because it depends on  
 the number of cases summarized in the table *n*   
as well as   
 the number of rows *r* and columns *c* in the table.

To quantify the degree of association, we can use Cramer’s V which we calculate easily from chi-squared as



Cramer’s V lies between 0 (no association) and 1 (“perfect” association).

For the contingency table of Amount and Month,



We’d call that weak association. The proportions of small, medium, and large purchases change over these months, but not dramatically.

Another way to measure the size of chi-squared is to use a statistical hypothesis test. We’ll cover those in Module 8.

# Quick Summary: Association in Contingency Tables

*Contingency table.* Mutually exclusive cells in the table count the number of cases that have the row and column attributes. Contingency tables are usually made from two categorical variables.

*Mosaic plot.* Graphical display of a contingency table.

*Association between categorical variables.* In the contingency table, row percentages or column percentages change from row-to-row or column-to-column. You cannot reproduce the cells of the table from the margins alone.

*Chi-squared.* Numerical measure of the amount of association. Cramer’s V normalizes the chi-squared statistic to the interval 0 to 1.

# Joint Random Variables

Percentages in a contingency table resemble probabilities. The following table gives probabilities for two random variables that describe auto purchases. X denotes the base price of the purchased car model; Y denotes the number of option packages added by the customer.

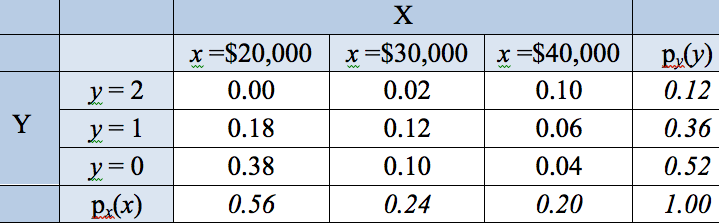
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | X | | |  |
|  |  | *x* =$20,000 | *x* =$30,000 | *x* =$40,000 | p*y*(*y*) |
| Y | *y* = 2 | 0.00 | 0.02 | 0.10 | *0.12* |
| *y* = 1 | 0.18 | 0.12 | 0.06 | *0.36* |
| *y* = 0 | 0.38 | 0.10 | 0.04 | *0.52* |
|  | p*x*(*x*) | *0.56* | *0.24* | *0.20* | *1.00* |

Analogous to the percentages defined by a contingency table (total, marginal, row, and column), this table defines several probabilities:[[7]](#footnote-7)

joint probabilities within the table

marginal probabilities along the bottom and right

conditional probabilities.

Joint probabilities give the probability   
associated with two (or more) r.v.   
*simultaneously* taking on specified values.

P(X = 20000, Y = 0) = *px,y*(20000,0) = 0.38

P(X = 30000, Y = 2) = *px,y*(30000,2) = 0.02

Marginal probabilities p*x*(*x*) and p*y*(*y*) are familiar. Each defines a random variable.[[8]](#footnote-8)

P(X = 20000) = *px*(20000) = 0.56

P(Y= 2) = *py*(2) = 0.12

The marginal probabilities sum the joint probabilities in each row or column, as in a contingency table. For example, the marginal probability for the first column is

P(X = 20000) = P(X=20000,Y=0) + P(X=20000,Y=1) + P(X=20000,Y=2)  
 = 0.38 + 0.18 + 0 = 0.56

Can we find the joint probability from the marginal probabilities? Is this possible? (*Hint*: Think back to association.)

# Conditional Probability and Dependence

In Module 3, we used several *independent*   
random variables to represent rolling dice.   
The defining property of independent random variables is that “probabilities multiply”, as in

P(Green=0.8, Red=3) = P(Green=0.8) × P(Red=3)

The random variables X and Y that describe the car purchases are not independent because, for example,

P(X = 20000, Y = 0) ≠ P(X=20000) × P(Y=0)   
 = 0.38 = 0.56 × 0.52 = 0.2912

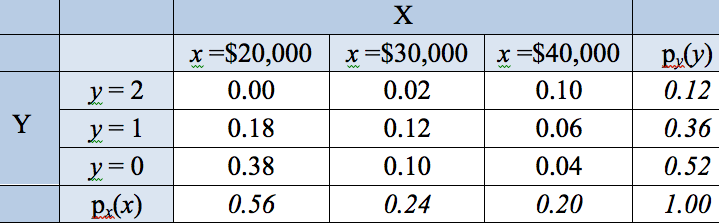
and

P(X = 40000, Y = 2) ≠ P(X=40000) × P(Y=2)   
 = 0.10 = 0.20 × 0.12 = 0.024

Because X and Y are *dependent* (*i.e.*, not independent), we cannot reproduce the joint probability from the marginal probabilities. We have to know more.

Conditional probabilities capture dependence.   
We define the conditional probability of Y given  
X, written P(Y=*y* | X=*x*), in a manner analogous to

column percentages in a contingency table.

P(Y=*y*|X=*x*) = P(X=*x*, Y=*y*)/P(X=*x*) [as in (cell count)/(column total)]

Read the vertical bar | as “given”.[[9]](#footnote-9)

Conditioning limits us to a row or column, so we make sure the conditional probabilities within that row or column sum to 1 by dividing by P(X=*x*).

If X and Y are independent, the conditional probability equals the marginal probability

P(Y=*y*|X=*x*) = P(X=*x*,Y=*y*)/P(X=*x*) = [P(Y=*y*)P(X=*x*)]/P(X=*x*) = P(Y=*y*).

Dependence implies that the conditional probability P(Y|X) depends on X.

P(Y = 2 | X = 20000) = P(Y = 2, X = 20000) / P(X = 20000)  
 = 0/0.56 = 0

P(Y = 2 | X = 40000) = P(Y = 2, X = 40000) / P(X = 40000)  
 = 0.10/0.20 = 0.50

# Condintional Mean

The conditional mean of Y given X, abbreviated E[Y|X] is the mean value of Y defined by the conditional distribution of Y given X. Rather than weight the possible values by p*y*(*y*), the conditional mean weights the values by p(Y|X).

For example, expected number of option packages ordered for cars that list for $20,000 is

E[Y|X=20000] = 0 P(Y=0|X=20000) + 1 P(Y=1|X=20000) + 2 P(Y=2|X=20000)  
 = 0 (0.38/0.56) + 1 (0.18/0.56) + 2 (0/0.56)  
 = 0.32

The expected number of option packages for cars that list for $40,000 is more than twice as large:

E[Y|X=40000] = 0 P(Y=0|X=40000) + 1 P(Y=1|X=40000) + 2 P(Y=2|X=40000)  
 = 0 (0.10/0.24) + 1 (0.12/0.24) + 2 (0.02/0.24)  
 = 0.67

Regression analysis, the methodology of Statistics 621, concerns the analysis of conditional means.

# Bayes Rule

The order of conditioning in a conditional probability is important. For example

P(Y = 2 | X = 40000) = 0.50

This is the probability that a customer orders 2 option packages given that the list price of the car is $40,000 (the third column of the table)

In comparison, the conditional probability with the random variables reversed is very different, both in value and interpretation.

P(X = 40000 | Y = 2) = 0.10/0.12 = 0.083

This is the probability that the car costs $40,000, given that two options were ordered (the top row of the table).

*Bayes’ rule* is an expression that allows us to reverse the order of conditioning.[[10]](#footnote-10)



# Example of Reversing Conditioning

An important use of Bayes rule and conditional probability occurs in medical applications. For example, suppose that a woman tests positive for breast cancer when screened by mammography. What’s the probability that she has cancer?

For mammography scans in younger women, it is known that

P(test + | cancer) = 0.85 (sensitivity of test)

P(test – | healthy) = 0.925 (specificity)

The marginal rate of cancer in this group is P(cancer) = 0.003.

You can find the needed probability P(cancer | test +) by using formulas or by filling in the associated probabilities in a table of the joint distribution like this. Start by filling in what you know…[[11]](#footnote-11)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Outcome | |  |
|  | + | – |  |
| Cancer |  |  | 0.003 |
| Healthy |  |  | 0.997 |
|  |  |  | 1 |

The sensitivity and specificity determine two values within the table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Outcome | |  |
|  | + | – |  |
| Cancer | (0.85)(0.003)=0.00255 |  | 0.003 |
| Healthy |  | (0.925)(0.997)=0.922225 | 0.997 |
|  |  |  | 1 |

Now use the fact that the sum of the probabilities within a row equals the marginal probability.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Outcome | |  |
|  | + | – |  |
| Cancer | (0.85)(0.003)=0.00255 | 0.00045 | 0.003 |
| Healthy | 0.074775 | (0.925)(0.997)=0.922225 | 0.997 |
|  |  |  | 1 |

All we need to do now is add the probabilities within each column.

Here is the completed table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Outcome | |  |
|  | + | – |  |
| Cancer | 0.002550 | 0.000450 | 0.003 |
| Healthy | 0.074755 | 0.922225 | 0.997 |
|  | 0.077325 | 0.922675 | 1 |

Once the table is filled in, it’s easy to find the conditional probability of cancer given a positive test:

P(cancer | test +) = 0.00255/0.077325 = 0.033

Even though the test has good sensitivity and specificity, the marginal rate is rather small and implies many “false positive” results.[[12]](#footnote-12) [[13]](#footnote-13)

# Summary

- *Contingency tables* count the simultaneous occurrences of two variables.

- If percentages within the rows (or columns) of the table do not match the corresponding marginal percentages, the variables are *associated*.

- The *chi-squared statistic* measures the amount of dependence in a contingency table. A *mosaic plot* displays the relative frequencies and conveys a visual impression of the degree of association.

- The analogous probability concept is the *joint distribution* of two random variables. From the joint distribution, we can form   
 *marginal* probabilities  
by summing the joint probabilities in a row or column and we can form  
 *conditional* probabilities  
by normalizing the joint probabilities within a row or column.

If the conditional probabilities within rows or columns differ from the corresponding marginal probabilities, the two random variables are *dependent*.

Bayes rule and probability tables allow us to manipulate conditional probabilities.

# Next Module

Portfolios and covariance: how to mix investments to reduce risk while preserving returns. The portfolio simulation in Module 3 had three artificial features:

(1) We made up Red.

(2) The returns on each investment were independent of each other.

(3) We knew the means and variances of the returns on the assets.

We are now ready to consider the properties of portfolios made from real stock returns.

1. For further motivation and examples, see Chapters 5-6 (data) and Chapters 8 and 10 (probability models) of SF. [↑](#footnote-ref-1)
2. Use Analyze > Fit Y by X to get this display. JMP reverses the table axes from the input dialog. To get the contingency table with only counts, uncheck the percentage items by clicking on the red button beside the title “Contingency Table”. The JMP data file contains over properties of these purchases. Chapter 5 of SF introduces contingency tables using a related data set that summarizes web purchases. Output in JMP depends on the “types” of the two variables. The output in this module is for the case where “X” and “Y” are both categorical. [↑](#footnote-ref-2)
3. Similarly, though not shown, row percentages differ from the marginal row percentages as well. [↑](#footnote-ref-3)
4. To get the plot oriented like this, we reversed the order of the variables in the Fit Y by X dialog (Amount as Y and Month as X). A pain, but not that hard! Segmented bar charts (SF, p. 80-81) are similar, but the areas are not proportional to the overall counts. [↑](#footnote-ref-4)
5. An example of a mosaic plot with no association is on page 181 of SF. [↑](#footnote-ref-5)
6. To find the value of chi-squared in the JMP output, look below the contingency table in the JMP window. Choose the version called Pearson (though both versions are similar). We could also answer this question using a hypothesis test, but we’ve not covered those yet. You’ll be able to do that at the end of Stat 603. For further explanation of the steps used to compute chi-squared, see SF, pp 89-91. [↑](#footnote-ref-6)
7. For more details, see page 176 of SF for a definition of joint and marginal probability. Conditional probability is defined on page 177. [↑](#footnote-ref-7)
8. They’re called marginal probabilities because they are defined by the margin of the table. [↑](#footnote-ref-8)
9. This definition works if P(X=*x*)>0. If P(X=*x*)=0 then P(Y=*y*|X=*x*) doesn’t make a lot of sense in this situation. [↑](#footnote-ref-9)
10. This equation says is that given the conditional probabilities in one direction (say col %) and the marginal probabilities for the columns one can find the joint table (analogous to the contingency table) and hence the conditional probabilities in the other direction (say row %). The importance of order in conditional probabilities is the topic of section 8.4 of SF (page 185); Bayes rule is on page 188. [↑](#footnote-ref-10)
11. If you are good with manipulating formulas, you can often finish more quickly than building up a table. Although building the full table may require a little extra calculation, however, a table allows you to check your calculations along the way. This example begins on page 186 of SF. [↑](#footnote-ref-11)
12. Here’s some intuition for what seems an odd result. The test is accurate, but breast cancer is so rare that most positive events are the result of the imperfect specificity of the test. Among those without cancer, the test correctly indicates no cancer 92.5% of the time. The remaining 7.5%, however, amounts to a lot of anxious women. This problem is even more acute for tests of HIV. [↑](#footnote-ref-12)
13. For a less medical example of Bayes rule, consider Example 8.2 of SF that shows how this concept can be used to customize a spam filter to the habits of each user. Also, exercise 5 in Assignment 2 explores the use of Bayes rule in making decisions. [↑](#footnote-ref-13)