Chapter 6 Categorical data

Categorical variables

- **Review:** Categorical variables place individuals into one of several groups or categories.
- The values of a categorical variable are labels for the different categories.
- The distribution of a categorical variable lists the percent of individuals who fall into each category.
- When a dataset involves two categorical variables, we begin by examining the percents in various categories for the two variables.

  A two-way table describes two categorical variables, organizing counts according to a row variable and a column variable.

Proportion

- Indicator variable

For an event A, X=$\mathrm{I}(A) = \begin{cases} 1, & If\ A\ occurs \\ 0, & if\ A\ no\ occur \end{cases}$

p(X=1) is the probability that A occurs,

p(X=0)=1-p(X=1) is the probability that A doesn’t occur

Review of Binomial Experiments

- Recall that a binary random variable (RV) is a specific type of categorical RV with two possible categories.
  - Convention: One category is “success” (S) and the other is “failure” (F).
- **Binomial distribution** can often be used to describe the probability distribution for an RV that corresponds to the number of “successes” in a fixed number of trials.
- An RV that follows a binomial distribution is known as a **binomial RV** and comes from a **binomial experiment**. In a binomial experiment, the following must be satisfied:
  - There are a fixed number of trials (n).
  - Trials are independent.
  - There are only two possible outcomes for each trial: success or failure.
  - Each trial has the same probability of success (p)
  - Notation: X~Binom(n, p), $\mu = np, \sigma^2 = np(1-p)$

Multinomial Experiments

- In many biomedical and public health applications, we are interested in settings where more than two possible outcomes may occur.
  - Example 1: blood type—A, B, AB, O
  - Example 2: type of health insurance—private/employer, Medicare, Medicaid, none
- Specifically, we might be interested in the number of people/observations that fall in each category for some group of interest.
- A **multinomial experiment** is one that has the following properties:
  - There are a fixed number of trials (n)
  - Trials are independent
  - The outcome for each trial can be classified as one of several different categories
  - The probability of falling in each category is the same across all trials ((p1, p2, …pk)

  - Notation: X ~ Multinomial(n, p1, p2, … pk)

Multinomial Experiments and One-Way Frequency Tables

- You can organize data from a multinomial experiment into a **one-way frequency table**.
- A **one-way frequency table** lists **observed** frequencies for each category in one row or one column.
  - Example: Suppose that data on blood type are collected on a sample of individuals.

Blood Type	A	B	AB	O
Observed frequency	O_1	O_2	O_3	O_4

- In a multinomial experiment setting, we often want to **test the claim that the observed frequencies for each category agree with some claimed distribution**.
- We can use **chi-squared goodness-of-fit test** to test this claim.

Goodness of fit

- A **one-way frequency table** lists **observed** frequencies for each category in one row or one column.

  - Example: Suppose that data on blood type are collected on a sample of individuals.

Blood Type	A	B	AB	O
Observed frequency	O_1	O_2	O_3	O_4

- The **chi-squared goodness-of-fit test** is used to test the *null hypothesis* (H_0) that an observed frequency distribution fits some claimed distribution.
  - For example: Are these four types of blood type profiles *equally likely to occur*?
  - In our example, the claimed distribution is that each category should have 25% of the observations.
  - H_0: There is a 25% chance of observing each of the four different blood types? If H_0 is true, then in a sample of 280 patients, we would have expected 280/4 = 70 observations.
  - The **chi-squared goodness-of-fit test** compares the observed frequency in each category to the expected frequency (assuming H_0 is true) in each category and determines if there is a statistically significant difference.

Some notation

- $n =$ total number of trials
- $K =$ number of different categories
- Suppose that categories are indexed as $k = 1, 2, \ldots, K$
- $O_k =$ observed frequency for the kth category
- $E_k =$ expected frequency for the kth category *if the null hypothesis is true*
- $p_k =$ probability of falling in the kth category

Chi-Squared Goodness-of-Fit Test

- How do we determine expected frequencies (E_k) for each category?

  - If the null hypothesis is that the probabilities of falling in categories $1, 2, \dots, K$ are $p_1, p_2, \dots, p_K$ respectively, then:

$$E_1 = np_1,\ E_2 = np_2,\ \dots\ ,\ E_K = np_K$$

Chi-Squared Goodness-of-Fit Test: Assumptions

- Data have been randomly sampled.
- The sample data consist of frequency counts for each category.
- The sample data come from a multinomial experiment.
- For each category, the expected frequency is at least 5.

Chi-Squared Goodness-of-Fit Test

- Test statistic:

$$X^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_K - E_K)^2}{E_K}$$

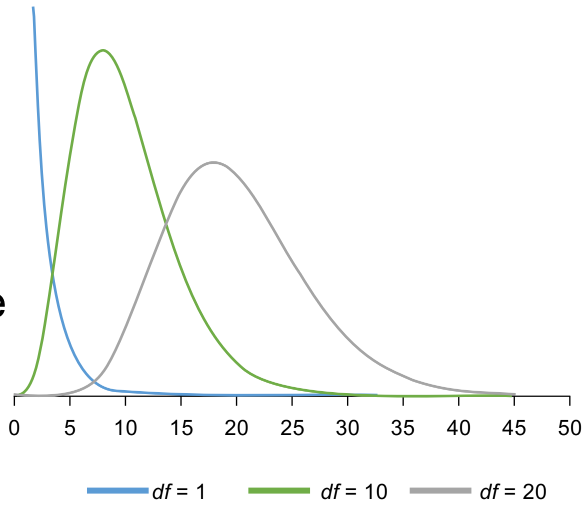
  - The test statistic measures the discrepancy between the frequencies that are actually observed and the frequencies that are expected **if the null hypothesis is true**.

  - If X^2 is **big**, then this means that the observed and expected frequencies are very different.
    - So, we should **reject** the null hypothesis.

  - If X^2 is small, then this means that the observed and expected frequencies are **not** very different.
    - So, we should **fail to reject** the null hypothesis.

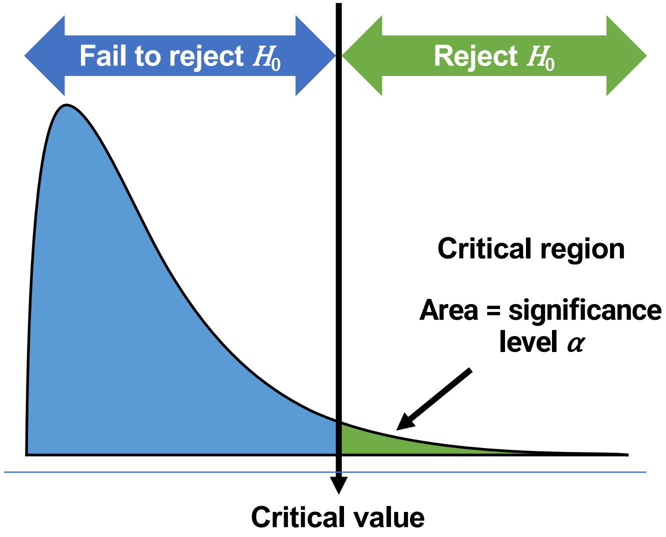
Chi-Squared Distribution

- How big is big enough for X^2?
- Assuming H_0 is true, then X^2 follows a **chi-squared (χ^2) distribution** with $K - 1$ **degrees of freedom (*df*)**.
- We can use this distribution to find **critical values** for the test.
- Facts about **chi-squared distributions** are as follows:
  - Chi-squared distributions are skewed
  - Random variables that follow a chi-squared distribution can be 0 or positive, never negative
  - Chi-squared distributions are indexed by a parameter called the **degrees of freedom (*df*)**
  - Notation: χ^2_{df}
  - As ***df*** increases, the chi-squared distributions become more symmetric



Chi-Squared Distribution Critical Value

- We use the chi-squared distribution with *df* = $K - 1$ to determine the **critical region** for a chi-squared goodness-of-fit test with K categories.
- The **critical region** is at the far right of the distribution.
- If $X^2 >$ critical value, then we **reject** H_0.
- If $X^2 \leq$ critical value, then we **fail to reject** H_0.
- Or use software reported p-value to make decisions.
- P-value=p(test statistic takes value as extreme or more extreme than the one observed under H0)
- When p-value<=0.05, there is sufficient evidence to reject H0.



Eg: The Chi-Square Test for Goodness of Fit

Mars, Inc. makes milk chocolate candies. Here's what the company's Consumer Affairs Department says about the color distribution of its M&M's candies:

On average, the new mix of colors of M&M's milk chocolate candies will contain 13 percent of each of browns and reds, 14 percent yellows, 16 percent greens, 20 percent oranges, and 24 percent blues.

The **one-way table** below summarizes the data from a sample bag of M&M's. In general, one-way tables display the distribution of a categorical variable for the individuals in a sample.

Color	Blue	Orange	Green	Yellow	Red	Brown	**Total**
Count	9	8	12	15	10	6	**60**

Eg: The Chi-Square Test for Goodness of Fit

We can write the hypotheses in symbols as follows:

H_0: $p_{blue} = 0.24$, $p_{orange} = 0.20$, $p_{green} = 0.16$, $p_{yellow} = 0.14$, $p_{red} = 0.13$, $p_{brown} = 0.13$,

H_a: At least one of the proportions is different than claimed

where p_{color} = the true population proportion of M&M's of that color.

The idea of the chi-square test for goodness of fit is this: We compare the **observed counts** from our sample with the counts that would be expected if H_0 is true. The more the observed counts differ from the **expected counts**, the more evidence we have against the null hypothesis.

In general, the expected counts can be obtained by multiplying the proportion of the population distribution in each category by the sample size.

Eg: The Chi-Square Test for Goodness of Fit

Assuming that the color distribution stated by Mars, Inc. is true, 24% of all M&M's produced are blue.

For random samples of 60 candies, the average number of blue M&M's should be (0.24)(60) = 14.40. This is our expected count of blue M&M's.

Using this same method, we can find the expected counts for the other color categories:

Blue:	**(0.24)(60) = 14.40**
Orange:	**(0.20)(60) = 12.00**
Green:	**(0.16)(60) = 9.60**
Yellow:	**(0.14)(60) = 8.40**
Red:	**(0.13)(60) = 7.80**
Brown:	**(0.13)(60) = 7.80**

Eg: The Chi-Square Test for Goodness of Fit

To calculate the chi-square statistic, use the same formula as you did earlier in the chapter.

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

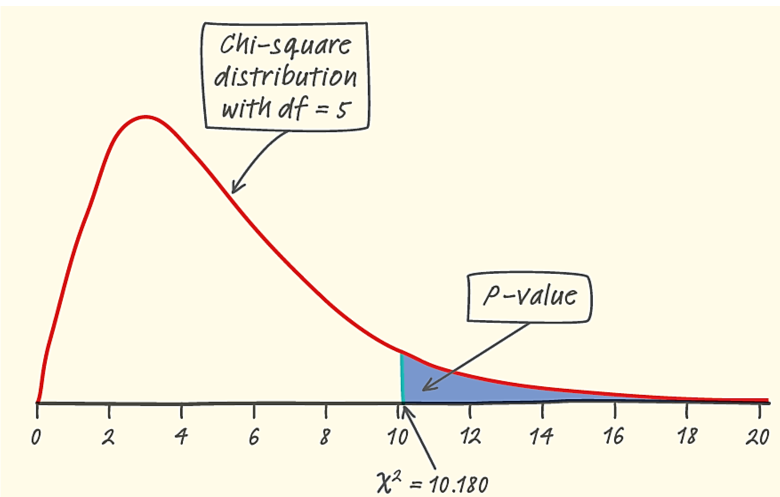
Color	Observed	Expected
Blue	9	14.40
Orange	8	12.00
Green	12	9.60
Yellow	15	8.40
Red	10	7.80
Brown	6	7.80

$$\chi^2 = \boxed{\frac{(9-14.40)^2}{14.40}} + \boxed{\frac{(8-12.00)^2}{12.00}} + \boxed{\frac{(12-9.60)^2}{9.60}} + \boxed{\frac{(15-8.40)^2}{8.40}} + \boxed{\frac{(10-7.80)^2}{7.80}} + \boxed{\frac{(6-7.80)^2}{7.80}}$$

$$\begin{aligned}\chi^2 &= 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415 \\ &= 10.180\end{aligned}$$

Eg: The Chi-Square Test for Goodness of Fit

We computed the chi-square statistic for our sample of 60 M&M's to be $\chi^2 = 10.180$. Because all of the expected counts are at least 5, the χ^2 statistic will follow a chi-square distribution with df=6-1=5 reasonably well when H_0 is true.



		P	
df	**.15**	**.10**	**.05**
4	6.74	7.78	9.49
5	8.12	9.24	11.07
6	9.45	10.64	12.59

Using $\alpha = 0.05$, df=5, the chi-square **critical value is 11.07**. $X^2=10.18 < 11.07$, fail to reject H0.

Or since our *P*-value is between 0.05 and 0.10, it is greater than $\alpha = 0.05$. Therefore, we fail to reject H_0.

We don't have sufficient evidence to conclude that the company's claimed color distribution is incorrect.

JMP steps for GOF test

- 1. Open existing data or enter new data
- 2. Click Analyze->distribution, select Color into Y, Count into Frequency, then click ok
- 3. Click the red triangle next to Color, click “test probabilities”, enter hypothesized probability for each color, then choose “fix hypothesized values, rescale omitted”. Click “done”.

Frequency

Level	Count	Prob
Blue	9	0.15000
Orange	8	0.13333
Green	12	0.20000
Yellow	15	0.25000
Red	10	0.16667
Brown	6	0.10000
Total	60	1.00000

Test probabilities

Level	Estim Prob	Hypoth Prob
Blue	0.15000	0.24
Orange	0.13333	0.2
Green	0.20000	0.16
Yellow	0.25000	0.14
Red	0.16667	0.13
Brown	0.10000	0.13

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	9.6233	5	0.0866
Pearson	10.1799	5	0.0703

Exercise: Car accidents and day of the week

A study of 667 drivers who were using a cell phone when they were involved in a collision on a weekday examined the relationship between these accidents and the day of the week.

Number of collisions by day of the week					
Day of the week					
Mon.	Tue.	Wed.	Thu.	Fri.	Total
133	126	159	136	113	667

Are the accidents equally likely to occur on any day of the working week?

H_0 specifies that all 5 days are equally likely for car accidents ➔ each $p_i = 1/5$.

Two-way tables

Two-way tables: subjects are cross-classified by two categorical variables X, Y, each has level I and J. Also called I by J table.

A Contingency Table analysis can be performed to test the hypothesis H0 : no association between X and Y.

Two-way tables

- We call education the row variable and age group the column variable.
- Each combination of values for these two variables is called a cell.
- For each cell, we can compute a proportion by dividing the cell entry by the total sample size. The collection of these proportions would be the joint distribution of the two variables.

TABLE 6.1 Years of school completed, by age (thousands of persons)

	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

Joint distribution Education By Age (Total %)

	25 to 34	**35 to 54**	**55 and over**	**total**
Did not complete high school	2.54	5.24	8.12	15.90
Completed high school	6.60	15.10	11.45	33.14
College, 1 to 3 years	6.10	12.92	6.35	25.38
College, 4 or more years	6.32	13.22	6.05	25.58
total	21.56	46.47	31.96	100%

Marginal distribution

- The **marginal distribution** of one of the categorical variables, in a two-way table of counts, is the distribution of values of that variable among all individuals described by the table.
- *Note:* Percents are often more informative than counts, especially when comparing groups of different sizes.

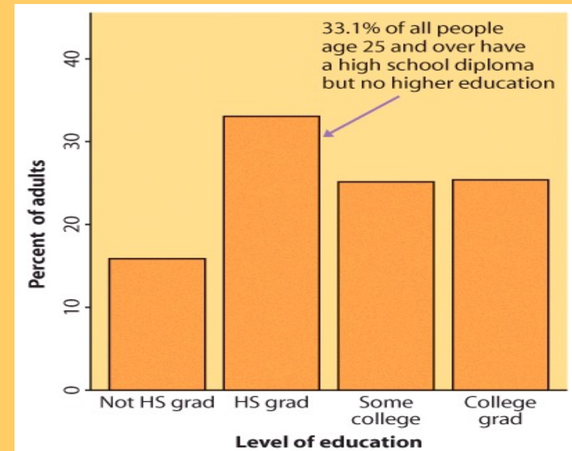
To examine a marginal distribution:

1. Use the data in the table to calculate the marginal distribution (in percents) of the row or column totals.
2. Make a graph to display the marginal distribution.

The marginal distributions can then be displayed on separate bar graphs, typically expressed as percents instead of raw counts. Each graph represents only one of the two variables, completely ignoring the second one.

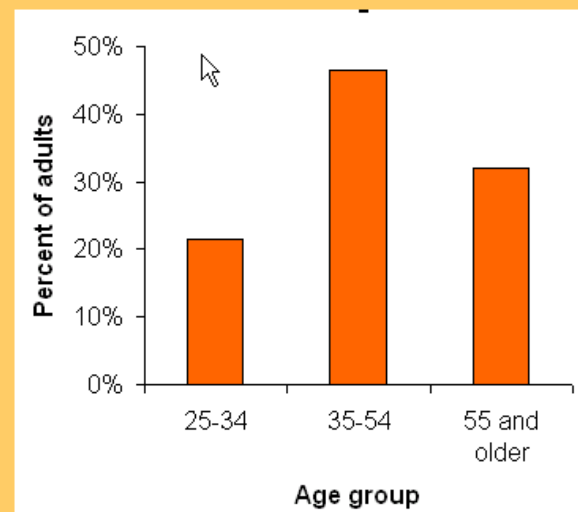
TABLE 6.1 Years of school completed, by age (thousands of persons)

Education	Age group: 25 to 34	Age group: 35 to 54	Age group: 55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230



Joint distribution Education By Age (Total %)

	25 to 34	35 to 54	55 and over	total
Did not complete high school	2.54	5.24	8.12	15.90
Completed high school	6.60	15.10	11.45	33.14
College, 1 to 3 years	6.10	12.92	6.35	25.38
College, 4 or more years	6.32	13.22	6.05	25.58
total	21.56	46.47	31.96	100%

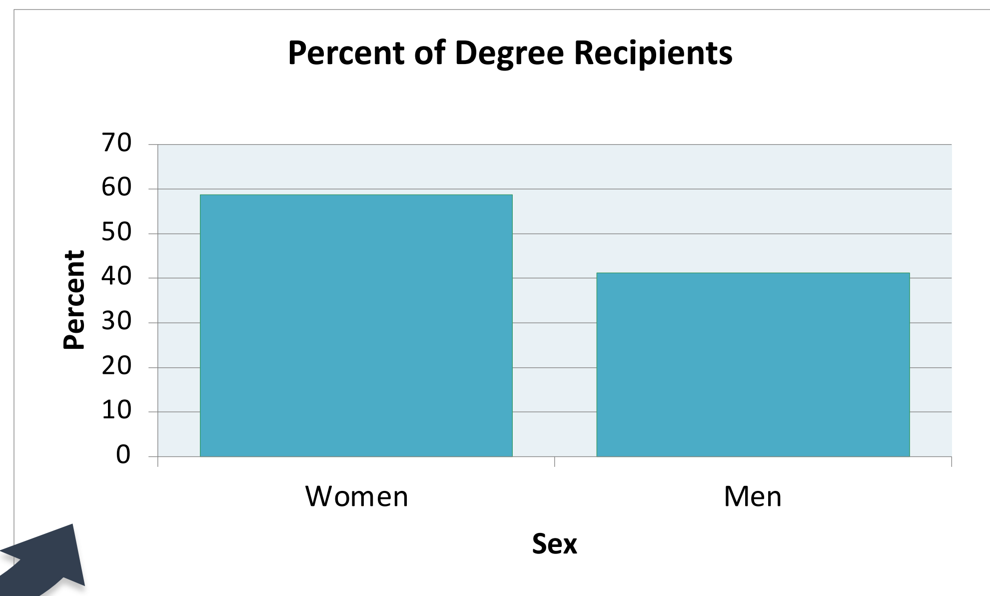


Marginal distribution

Sex	Degrees Conferred (thousands) Associate	Degrees Conferred (thousands) Bachelor's	Degrees Conferred (thousands) Master's	Degrees Conferred (thousands) [illegible]octorate	Total
Women	673	10[illegible]0	481	95	2299
Men	401	780	342	89	1612
Total	1074	1830	823	184	3911

Examine the **marginal distribution** of gender.

Response	Pe[illegible]t
Women	2299/3911= 58.8%
Men	1612/3911 = 41.2%



Conditional distribution

- Marginal distributions tell us nothing about the relationship between two variables.
- A conditional distribution of a variable describes the values of that variable among individuals who have a given value of another variable.
- Condition on the value of one variable and calculate the distribution of the other variable.
- Use software to generate a **side-by-side bar graph**, or a **mosaic plot** to compare distributions.

Conditional Distribution

- **Conditional distribution**: Condition on the value of one variable and calculate the distribution of the other variable.
- In the table below, the 25 to 34 age group occupies the first column. To find the complete distribution of education in this age group, look only at that column. Compute each count as a percent of the column total.
- These percent should add up to 100% because all persons in this age group fall into one of the education categories. These four percent together are the conditional distribution of education, given the 25 to 34 age group.

Years of school completed, by age (thousands of persons)

	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

2000 U.S. census

Conditional distributions

The percent within the table represent the **conditional distributions**. Comparing the conditional distributions allows you to describe the "relationship" between both categorical variables.

Years of school completed, by age (thousands of persons)

Education	Age group: 25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

	25 to 34	35 to 54	55 up	All
1:NotHS	4459	9174	14226	27859
	11.80	11.27	25.40	15.90
2:HSgrad	11562	26455	20060	58077
	30.60	32.49	35.82	33.14
3:SomeCo	10693	22647	11125	44465
	28.30	27.81	19.86	25.38
4:CollGr	11071	23160	10597	44828
	29.30	28.44	18.92	25.58
All	37785	81436	56008	175229
	100.00	100.00	100.00	100.00

ell Contents-
Count
% of Col

Here the percents are calculated by age range (columns).

$$29.30\% = \frac{11071}{37785} = \frac{\text{cell total}}{\text{column total}}$$

Music and wine purchase decision

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

What is the relationship between type of music played in supermarkets and type of wine purchased?

We want to compare the conditional distributions of the response variable (wine purchased) for each value of the explanatory variable (music played). Therefore, we calculate column percents.

$\frac{30}{84} = 35.7\%$

$= \frac{\text{cell total}}{\text{column total}}$

Calculations: When no music was played, there were 84 bottles of wine sold. Of these, 30 were French wine. 30/84 = 0.357 ➔ 35.7% of the wine sold was French when no music was played.

Column percents for wine and music

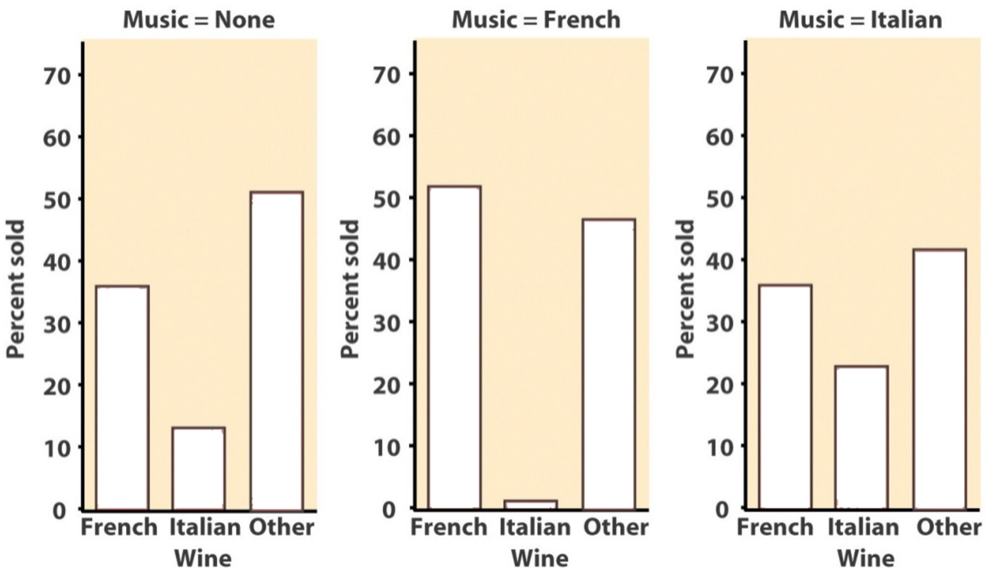
	Music			
Wine	None	French	Italian	Total
French	35.7	52.0	35.7	40.7
Italian	13.1	1.3	22.6	12.8
Other	51.9	46.7	41.7	46.5
Total	100.0	100.0	100.0	100.0

We calculate the column conditional percents similarly for each of the nine cells in the table:

For every two-way table, there are two sets of possible conditional distributions.

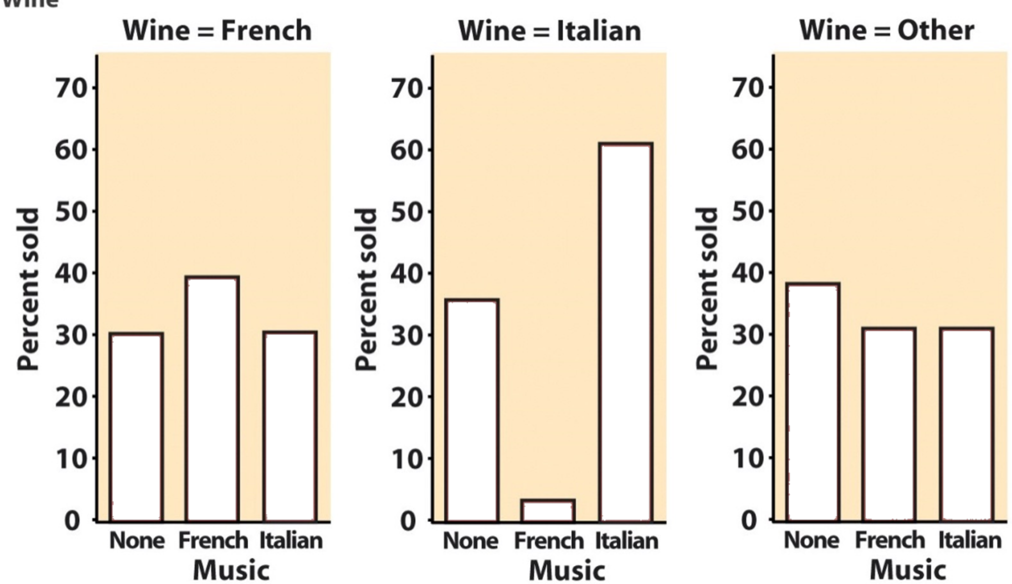
	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Does background music in supermarkets influence customer purchasing decisions?



Wine purchased for each kind of music played (column percents)

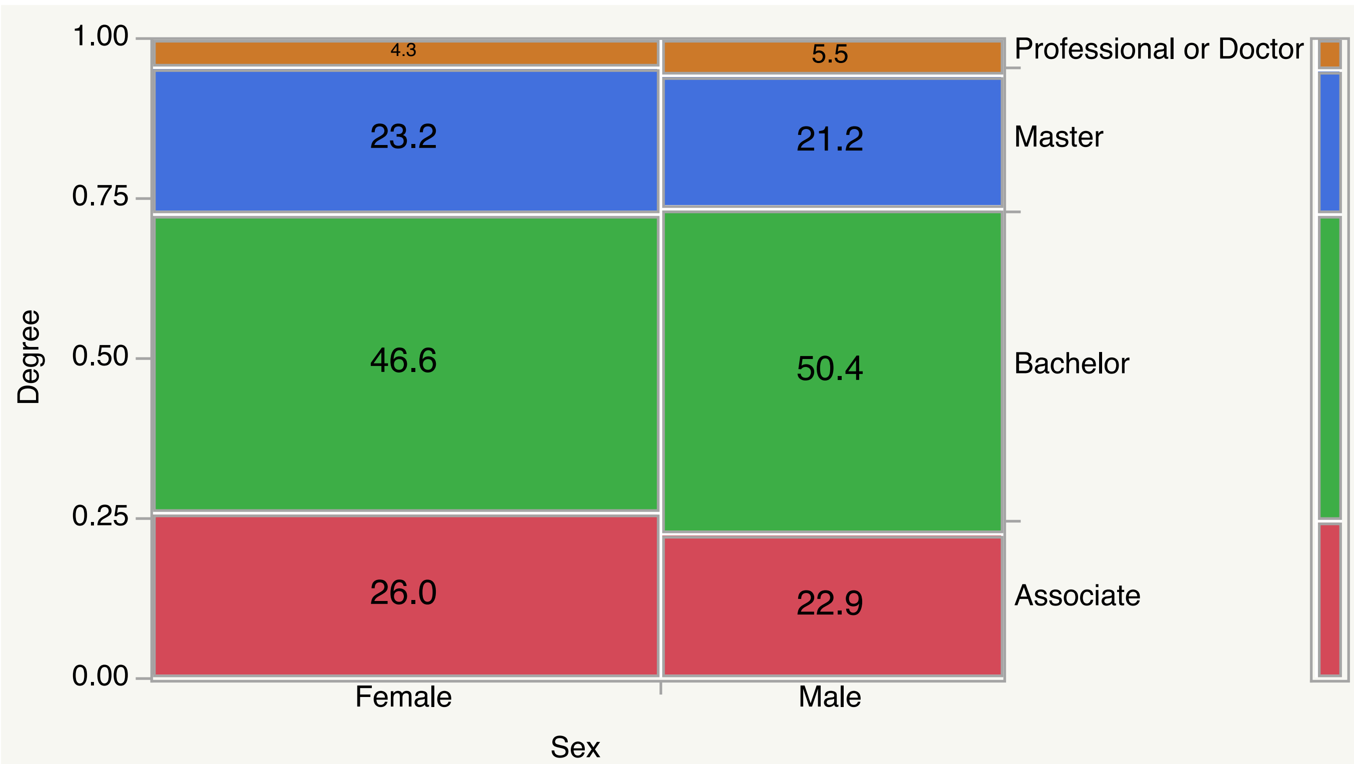
Music played for each kind of wine purchased (row percents)



P176. Ex 6.29:
Contingency Table
Sex By Degree

Count	Associate	Bachelor	Master	Professional or Doctor	Total
Female	646	1160	576	106	2488
Male	383	844	354	92	1673
Total	1029	2004	930	198	4161

Conditional distribution of degree given sex
Mosaic plot



Chi-Squared Test for two-way tables

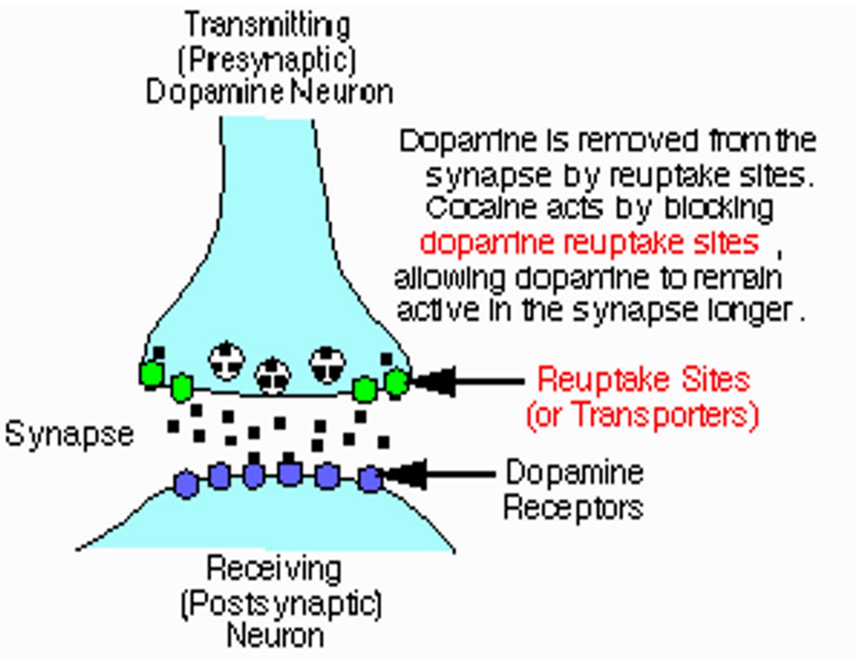
- The **chi-squared test for two-way tables** tests the null hypothesis that there is no association between the row and column variables in a contingency table.
  - H_0: Row and column variables are independent
  - H_1: Row and column variables are dependent
- Also called **Pearson's chi-squared** test.
- This test is similar to the chi-squared goodness-of-fit test in that we need to compute the **expected cell frequencies**, assuming that the null hypothesis is true, and then compare them to the observed cell frequencies.

Example: Cocaine addiction

Cocaine produces short-term feelings of physical and mental well being. To maintain the effect, the drug may have to be taken more frequently and at higher doses. After stopping use, users will feel tired, sleepy and **depressed.**

The pleasurable high followed by unpleasant after-effects encourage repeated compulsive use, which can easily lead to dependency.

Desipramine is an **antidepressant** affecting the brain chemicals that may become unbalanced and cause depression. It was thus tested for recovery from cocaine addiction.



Treatment with desipramine was compared to a standard treatment (lithium, with strong anti-manic effects) and a placebo.

Two-way tables

Treatment	Relapse		
	No	Yes	Total
Desipramine	15	10	25
Lithium	7	19	26
Placebo	4	19	23
Total	26	48	74

Hypothesis: no association

Want to know if the differences in sample proportions are likely to have occurred just by chance due to random sampling.

Use the **chi-square (χ^2) test** to assess

***H*$_0$** : no relationship between the row variable and column variable.

Computing Expected Cell Frequencies

- if events A and B are independent, then: $P(A \; and \; B) = P(A)P(B)$
- For the chi-squared test for independence, the null hypothesis is that the row and column variables are **independent**. We can use the margins to compute the probability of falling in a given row and given column:

$$P(Being \; in \; row \; \mathrm{i} \; and \; column \; j) = P(Being \; in \; row \; i)P(Being \; in \; column \; \mathrm{j})$$

$$= \left(\frac{Row \; i \; total}{Table \; total}\right)\left(\frac{Column \; j \; total}{Table \; total}\right)$$

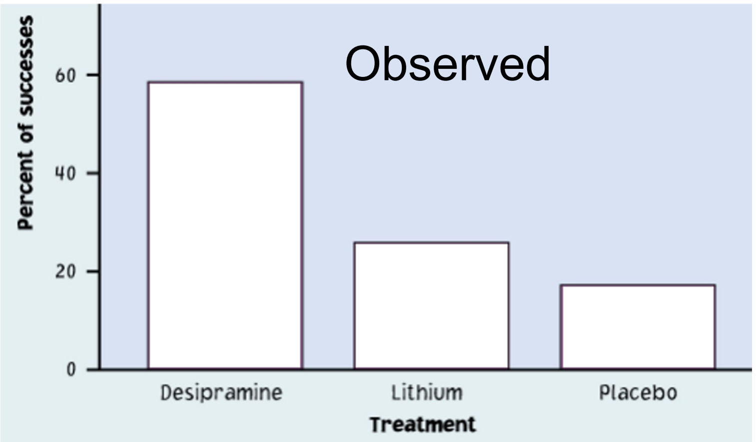
- the expected cell frequency in row *i* and column *j* (E_{ij}), assuming independence is

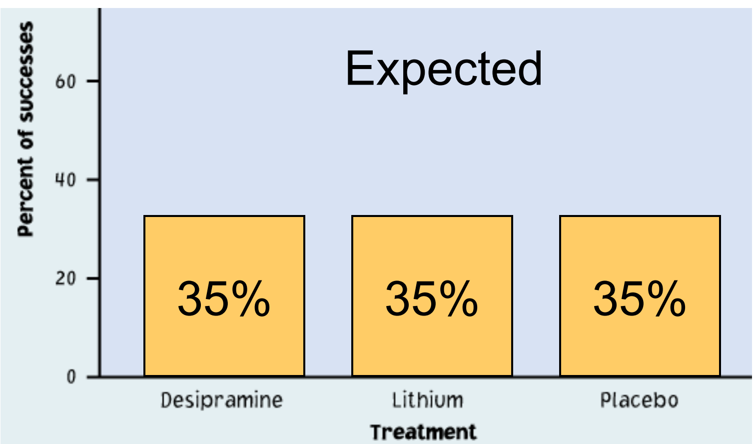
$$E_{ij} = (Table \; total)\left(\frac{Row \; i \; total}{Table \; total}\right)\left(\frac{Column \; j \; total}{Table \; total}\right) = \frac{(Row \; i \; total)(Column \; j \; total)}{Table \; total}$$

$=\frac{r_{i+}c_{+j}}{n}$, r_{i+} represents the ith row total, c_{+j} represents the jth column total

Cocaine addiction

H_0: there is no association between treatments and relapse.
H_a: there is some association between treatments and relapse.





	Relapse: No	Relapse: Yes	Total
Desipramine	15	10	25
Lithium	7	19	26
Placebo	4	19	23
Total	26	48	74

Expected relapse counts

	No	Yes
Desipramine	(25)(26)/74 ≈ 8.78	(25)(48)/74=16.22
Lithium	=9.14	=16.86
Placebo	=8.08	=14.92

The chi-square test

The chi-square statistic (χ^2) measures how much the observed cell counts in a two-way table diverge from the expected cell counts.

χ^2 statistic: (summed over all *r* x *c* cells in the table)

$$\boxed{\begin{aligned} &\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}} = \sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \\ &df = (r-1)(c-1) \end{aligned}}$$

Large values for χ^2 represent strong deviations from the expected distribution under the H_0 and provide evidence against H_0.

Cocaine addiction

Table of counts:
"actual / **expected**," with three rows and two columns:

df = (3−1)*(2−1) = 2

	No relapse	Relapse
Desipramine	15 **8.78**	10 **16.22**
Lithium	7 **9.14**	19 **16.86**
Placebo	4 **8.08**	19 **14.92**

$$X^2 = \frac{(15-8.78)^2}{8.78} + \frac{(10-16.22)^2}{16.22} + \frac{(7-9.14)^2}{9.14} + \frac{(19-16.86)^2}{16.86} + \frac{(4-8.08)^2}{8.08} + \frac{(19-14.92)^2}{14.92} = 10.74$$

χ^2 components:

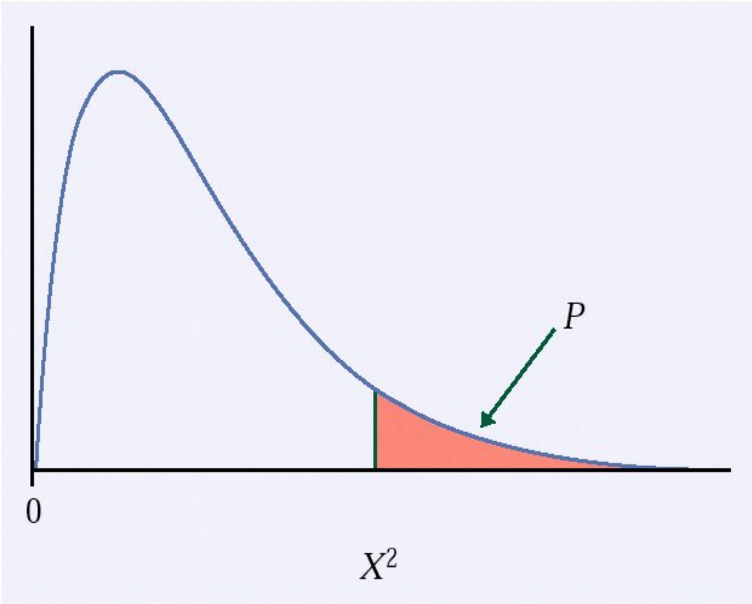
4.41	2.39
0.50	0.27
2.06	1.12

For the chi-square test, (in a two-way table)

H_0 : there is no association between the row and column variables

H_a : these variables are related.

If H_0 is true, the chi-square test has approximately a **χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom**.



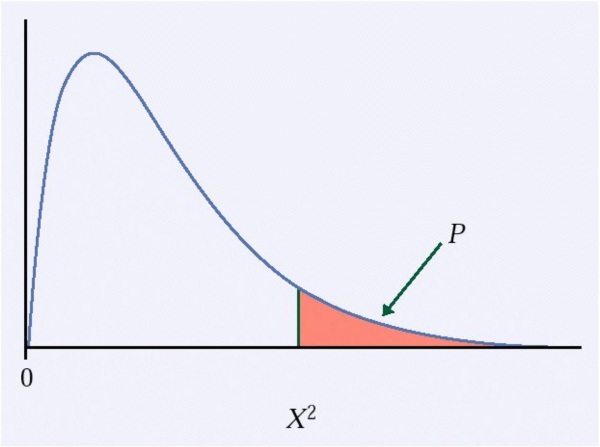
The P-value for the chi-square test is the area to the right of X^2 under the χ^2 distribution with df =(r−1)(c−1): $P(\chi^2 \geq X^2)$.

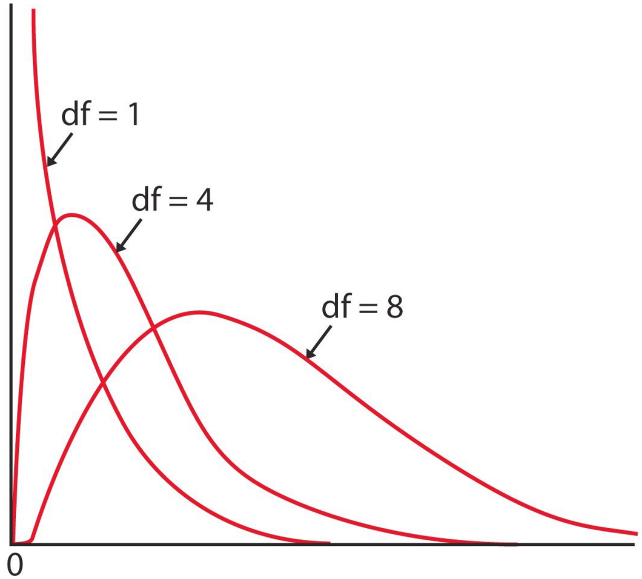
Given significance level alpha, we can find chi-square critical value $x^2_{\alpha,(r-1)(c-1)}$, if is $X^2 \geq x^2_{\alpha,(r-1)(c-1)}$, we reject H0.

Finding the p-value with Table F

The χ^2 distributions are a family of distributions that can take only positive values, are skewed to the right, and are described by a specific degrees of freedom.

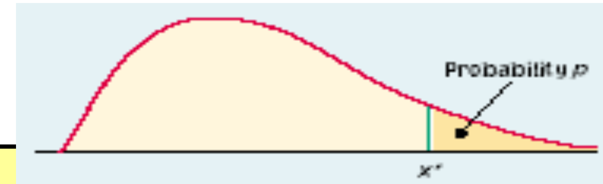
Table F gives upper critical values for many χ^2 distributions.





Cocaine addiction: Table F

H_0: There is no relationship between treatments and proportion of success.



	p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11

$X^2 = 10.71$ and df = 2

$10.60 < X^2 < 11.98$ ➔ $0.0025 < p < 0.005$ ➔ reject the H_0

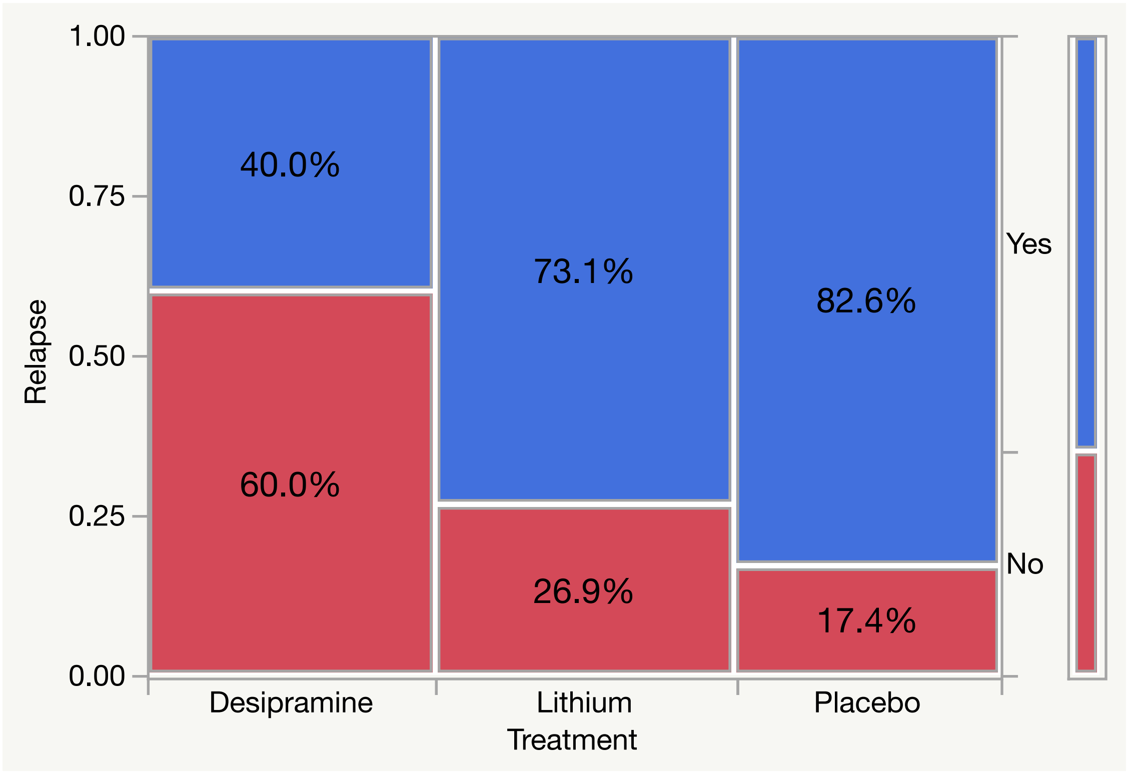
Or with 5% significance level, df=2, Chi-square critical value=5.99, test-statistic X^2=10.71>5.99, reject H0

➔ There is a relationship between treatment and relapse of cocaine.

As to what kind of relationship, we need to look at the conditional distribution of relapse given treatment.

JMP output

Mosaic Plot



Contingency Table
Treatment By Relapse

Count Total % Col % Row %	No	Yes	Total
Desipramine	15 20.27 57.69 60.00	10 13.51 20.83 40.00	25 33.78
Lithium	7 9.46 26.92 26.92	19 25.68 39.58 73.08	26 35.14
Placebo	4 5.41 15.38 17.39	19 25.68 39.58 82.61	23 31.08
Total	26 35.14	48 64.86	74

test

N	DF	-LogLike	RSquare (U)
74	2	5.3757195	0.1121

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.751	0.0046*
Pearson	10.729	0.0047*

JMP two-way table analysis example

- ex 6.28: “It is right to use animals for medical testing if it might save human lives.” The General Social Survey asked 1152 adults to react to this statement. Two-way table summarized their responses:

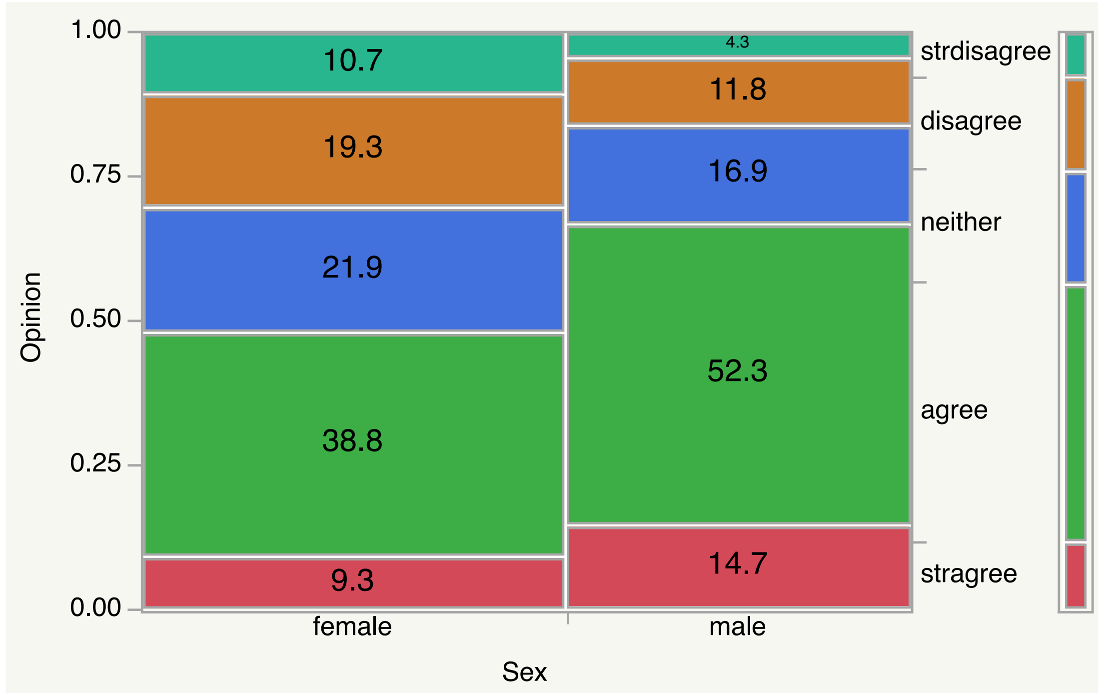
- Right click “Opinion” column, click “column property”, uncheck “customer order” and “numerical order”, check “row order levels”, then click ok. This is to use the order of the category shown in the Opinion column, not by alphabetical order.

- Analyze-> fit y by x-> click Opinion to y, Sex to x, count to freq, click ok, you will see the result.

- Right click the mosaic plot, click cell labelling by percent, you will see the conditional distribution of opinion given sex.

- The contingency table shows the joint distribution in count and total %, conditional distribution of Opinion given sex (col %), and the conditional distribution of sex given Opinion (row %).

Mosaic plot



Contingency Table

Sex By Opinion

Count Total % Col % Row %	stragree	agree	neither	disagree	strdisagree	Total
female	59 5.12 43.70 9.28	247 21.44 47.78 38.84	139 12.07 61.50 21.86	123 10.68 66.85 19.34	68 5.90 75.56 10.69	636 55.21
male	76 6.60 56.30 14.73	270 23.44 52.22 52.33	87 7.55 38.50 16.86	61 5.30 33.15 11.82	22 1.91 24.44 4.26	516 44.79
Total	135 11.72	517 44.88	226 19.62	184 15.97	90 7.81	1152

test

N	DF	-LogLike	RSquare (U)
1152	4	24.342593	0.0149

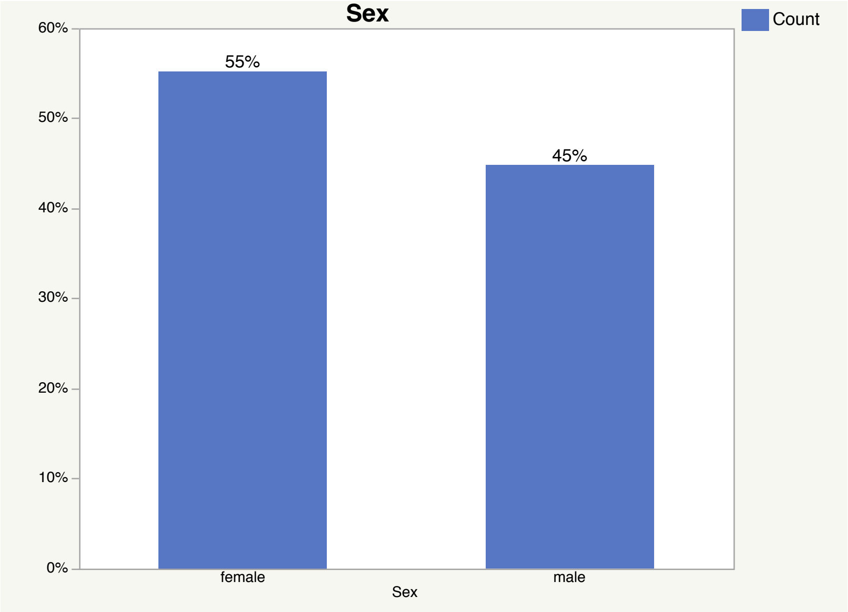
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	48.685	<.0001*
Pearson	47.547	<.0001*

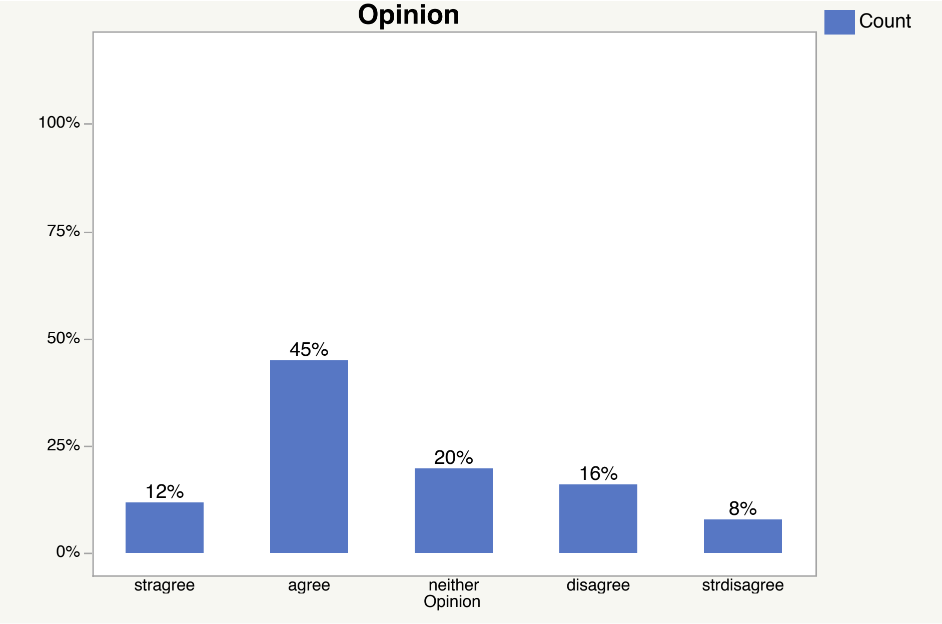
X^2 statistics is 47.547, df=4, p-value<0.0001, we reject H0, conclude There is significant relationship between gender and opinion.

From the conditional distribution, there are more males than females towards agree and strong agree, more females than males towards to disagree and strong disagree.

Ex6.28: Marginal distribution by JMP

Graph builder->bar chart, select sleep to x, count to freq, summary statistics use % of total, label by value, click ok. That's the marginal distribution of sleep.
Similarly you can get the marginal distribution of exercise.





Ex6.28 Conditional distribution of Opinion given sex by JMP

Graph builder->bar chart, select Opinion to x, Sex to Overlay, count to freq, summary statistics use % of total, label by value, click ok. That' s the conditional distribution of Opinion given sex.