Chapter 6.3-6.4: Contingency Table Analysis

Univariate Categorical Analysis	Bivariate Categorical Analysis
One-Way Table	Two-way Table
I. Goodness-of-Fit Test	II. Test of Independence
The individuals in the sample are categorized according to one categorical variable.	The individuals in the sample are classified according to two categorical variables.

I. GOODNESS-OF-FIT TEST (6.3)

- There is one random sample of size n from a single population. Each individual in the sample gives a response on a categorical variable which has k categories.
- The true proportion of the population who is categorized by category i is p_i. Recall that a proportion p_i is also a probability,

$$p_i = P(\text{Category } i).$$

  Thus, there are k parameters of interest, $p_1,\ p_2, ..., p_k$ and $\sum_{i=1}^{k} p_i = 1$.
- This test is called a "goodness of fit test" because when one hypothesizes what each proportion p_i is, one is really hypothesizing a probability distribution for the categorical variable:

Categorical Variable	Category 1	Category 2	...	Category k
Probabilities	p_1	p_2	...	p_k

  The test determines how good this probability distribution fits the categorical variable.

1. Hypotheses:

   H_0: $p_1 = p_{1_0},\ p_2 = p_{2_0},\ \ldots,\ p_k = p_{k_0}$

   H_a: $p_i \neq p_{i_0}$ for some i (at least one proportion differs from its hypothesized value)

   The values p_{1_0}, p_{2_0}, ..., p_{k_0} are hypothesized values under H_0, and so $\sum_{i=1}^{k} p_{i_0} = 1$.

2. Assumptions:
   - The data is from a SRS.
   - If the sample was chosen without replacement, then $0.05N > n$.
   - Each *expected count* ≥ 5 (this is a new check for "is this a big sample"). By definition, (expected count)$_i = np_{i_0}$.

Steps 3 and 4 are performed assuming that H_0 is true!

3. Test Statistic:

$$\chi^2 = \sum_i \frac{((\text{observed count})_i - (\text{expected count})_i)^2}{(\text{expected count})_i}$$

- *observed counts* are the frequencies actually observed in the sample data
- *expected counts* are the frequencies you would expect to see if H_0 was true,

$$(\text{expected count})_i = np_{i_0}.$$

EXAMPLE: If $n = 100$ and $p_{i_0} = \frac{1}{4}$, then $(\text{expected count})_i = 100\left(\frac{1}{4}\right) =$ _____

4. p-value:

The p-value $= P(\chi^2 > x^2)$ is a right-tail probability from a "chi-square" distribution.

- The test statistic has a chi-square distribution, $\chi^2 \dot{\sim} \chi^2(df = k-1)$, when H_0 is true.
- $\chi^2(df)$ has a single parameter df called the *degrees of freedom*.
- For small df, $\chi^2(df)$ is right-skewed. For larger values of df, $\chi^2(df)$ becomes more symmetric.n See Figure 6.8 on p290 of your textbook.
- p-values are always right-tails of $\chi^2(df)$. Right tail probabilities can be found in Table B.3 on p432 of your textbook. Or, you can use R's **pchisq(χ^2, df =#,lower.tail=FALSE)** function.

5. and 6. Make a Decision and give a Conclusion.

7. Follow-up Analysis:

If H_0 is rejected, then perform an "ad hoc" follow-up analysis:

- This follow-up analysis is not a hypothesis test and does not give significant differences among the $p_i's$!
- If the i^{th} category has the highest chi-square contribution, then:
  - p_i is the largest proportion if the $(\text{observed count})_i > (\text{expected count})_i$.
  - p_i is the smallest proportion if the $(\text{observed count})_i < (\text{expected count})_i$.

GOODNESS-OF-FIT TEST EXAMPLE

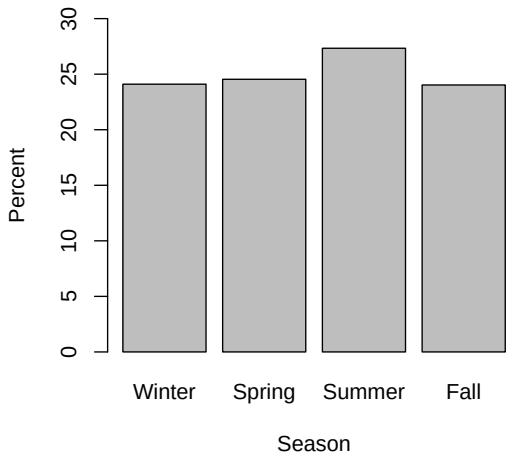
Criminologists have long debated whether there is a association between weather and violent crime. The author of the article "Is There a Season for Homicide?" (Criminology [1988]: 287-296) classified 1361 homicides according to season, resulting in the following data. Do these data support the theory that the homicide rate is not the same over the four seasons? Test the relevant hypotheses using a significance level of 0.05.

Winter	Spring	Summer	Fall
328	334	372	327

Summarize Your Data (graphically and numerically)

```
> freq = c(328,334,372,327)
> perc = prop.table(freq)*100
# Display precentages for the 4 seasons
> rbind(c("Winter","Spring","Summer","Fall"),perc)
     [,1]               [,2]               [,3]               [,4]
     "Winter"           "Spring"           "Summer"           "Fall"
perc "24.0999265246143" "24.5407788390889" "27.3328434974284" "24.0264511388685"

> barplot(perc,names=c("Winter",
  "Spring","Summer","Fall"),xlab="Season",ylab="Percent")
```



Goodness-of-Fit Test

```
> hyp.p=c(0.25,0.25,0.25,0.25)
> season.csq=chisq.test(freq,p=hyp.p)
> season.csq

        Chi-squared test for given probabilities

data:  freq
X-squared = 4.0345, df = 3, p-value = 0.2578
```

Check Your Assumptions

Are all of the expected counts at least 5?

```
> season.csq$expected
[1] 340.25 340.25 340.25 340.25
```

Goodness-of-Fit Test by hand

1. Hypotheses:

2. Test statistic value:

Table of Expected Counts

Winter	Spring	Summer	Fall

Table of χ^2 contributions

Winter	Spring	Summer	Fall

So, χ^2=_____

3. Check the Assumptions:

4. Distribution of the test statistic and the *p*-value:

5. Decision at $\alpha = .05$:

6. Conclusion:

7. Follow-up:

II. TEST OF INDEPENDENCE (6.4)

- There is one random sample where each individual in the sample is classified according to two categorical variables. The *row* variable has r categories and the *column* variable has c categories.
- Among all individuals in the population who are categorized in Category i of the row variable, let p_{ij} be the proportion of individuals who are categorized in category j of the column variable. Thus, p_{ij} is a conditional probability

$$p_{ij} = P(\text{Column Category } j \mid \text{Row Category } i).$$

Among individuals in the population who are categorized in Category i of the row variable, $p_{i1}, p_{i2}, ...p_{ic}$ is a probability distribution of the c column categories and so $\sum_j p_{ij} = 1$. Across all r rows, there are rc parameters of interest:

	Column $j=1$	Column $j=2$	...	Column $j=c$
Row $i=1$	p_{11}	p_{12}	...	p_{1c}
Row $i=2$	p_{21}	p_{22}	...	p_{2c}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
Row $i=r$	p_{r1}	p_{r2}	...	p_{rc}

- The Test of Independence is used to determine if the row and column variables are *dependent* (i.e. not independent). Thus, the question of interest is whether knowledge of one variable's category provides any information about the category of the other variable. Dependent variables are said to have an *association.*

1. Hypotheses:

H_0: The two categorical variables are independent.
H_a: The two categorical variables are not independent.

OR

H_0: $p_{11} = p_{21} = ...p_{r1}$ and $p_{12} = p_{22} = ...p_{r2}$ and ... $p_{1c} = p_{2c} = ...p_{rc}$
(all of the proportions in a given column are equal)
H_a: For some column k, $p_{ik} \neq p_{jk}$ for some rows i and j.
(at least two of the proportions in some column are different)

2. Assumptions:

- A SRS is chosen.
- If the sample was without replacement, then $0.05N > n$.
- Each expected count ≥ 5.

Steps 3 and 4 are performed assuming that H_0 is true!

3. Test Statistic:

$$\chi^2 = \sum_{i,j} \frac{((\text{observed count})_{ij} - (\text{expected count})_{ij})^2}{(\text{expected count})_{ij}}$$

- *observed counts* are the frequencies actually observed in the sample data
- *expected counts* are the frequencies you would expect to see if H_0 was true,

$$(\text{expected count})_{ij} = \frac{(\text{row total})_i (\text{column total})_j}{\text{grand total}}$$

EXAMPLE: Consider the following table of observed counts where $r = 3$ and $c = 2$:

	Male	Female	
	(30%)	(70%)	
East	30	70	$n_1 = 100$
	(75%)	(25%)	
West	150	50	$n_2 = 200$
	(20%)	(80%)	
South	20	80	$n_3 = 100$
	200	200	$N = 400$

The table of expected counts (given that H_0 is true) is:

	Male	Female	
East			$n_1 = 100$
West			$n_2 = 200$
South			$n_3 = 100$
	200	200	$N = 400$

4. *p*-value:
   The *p*-value = $P(\chi^2 > x^2)$
   - The test statistic has a "chi-square" distribution, $\chi^2 \dot{\sim} \chi^2((r-1)(c-1))$.
   - $\chi^2(df)$ has a single parameter *df* called the *degrees of freedom*.
   - For small *df*, $\chi^2(df)$ is right-skewed. For larger values of *df*, $\chi^2(df)$ becomes more symmetric.
   - *p*-values are always right-tails of $\chi^2(df)$. Right tail probabilities can be found in Table B.3 on p290 of your textbook. Or, you can use R's **pchisq(χ^2, *df* =#,lower.tail=FALSE)** function.

5. and 6. Make a Decision and give a Conclusion.

7. Follow-up Analysis:
   If H_0 is rejected, then perform an "ad hoc" follow-up analysis:
   - This follow-up analysis is not a hypothesis test and does not give significant differences among the $p'_{ij}s$!
   - Choose the two largest chi-square contributions. If the i^{th} row and the j^{th} column has one of the two highest chi-square contributions, then:
     - Row i has the highest proportion in column j (p_{ij} is the largest in the j^{th} column) if the (observed count)$_{ij}$ > (expected count)$_{ij}$.
     - Row i has the smallest proportion in column j (p_{ij} is the smallest in the j^{th} column) if the (observed count)$_{ij}$ < (expected count)$_{ij}$.

Cramer's V^2: $V^2 = \frac{\chi^2}{n(\min(r-1,c-1))}$ is the proportion of variability in the counts that can be explained by the association between the row variable and the column variable. This is similar to the R^2 value that we learned about for an ANOVA.

TEST OF INDEPENDENCE EXAMPLE

The following table is based on data from 21,876 male physicians, age 40-84, who are participating in the Physicians Health Study ("Light-to-Moderate Alcohol Consumption and Risk of Stroke Among U.S. Male Physicians," *New England Journal of Medicine* [1999]: 1557-1564). The male physicians were categorized according to the smoking status and the alcohol consumption.

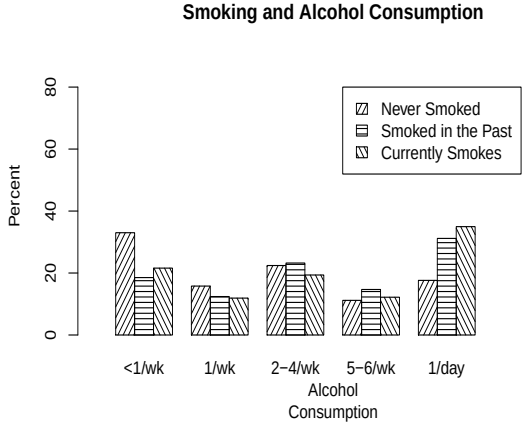
	Alcohol Consumption (number of drinks)					
	< 1/wk	1/wk	2-4/wk	5-6/wk	1/day	Total
Never Smoked	3577	1711	2430	1211	1910	10839
Smoked in the Past	1595	1068	1999	1264	2683	8609
Currently Smokes	524	289	470	296	849	2428
Total	5696	3068	4899	2771	5442	21,876

Summarize Your Data (graphically and numerically)

```
> freq=rbind(c(3577,1711,2430,1211,1910),c(1595,1068,1999,1264,2683),c(524,289,470,296,849))
> colnames(freq) = c("<1/wk","1/wk","2-4/wk","5-6/wk","1/day")
> rownames(freq) = c("Never Smoked","Smoked in the Past ","Currently Smokes")
> freq
                    <1/wk 1/wk 2-4/wk 5-6/wk 1/day
Never Smoked         3577 1711   2430   1211  1910
Smoked in the Past   1595 1068   1999   1264  2683
Currently Smokes      524  289    470    296   849

> perc = prop.table(freq,1)*100
> perc
                       <1/wk     1/wk   2-4/wk   5-6/wk    1/day
Never Smoked        33.00120 15.78559 22.41904 11.17262 17.62155
Smoked in the Past  18.52712 12.40562 23.21989 14.68231 31.16506
Currently Smokes    21.58155 11.90280 19.35750 12.19110 34.96705

> c(60,180,120)
> barplot(perc,beside=TRUE,angle=ang,density=20,col="black",xlab="Alcohol
  Consumption",xlim=c(0,25),ylab="Percent",ylim=c(0,90),main=
  "Smoking and Alcohol Consumption")
> legend(13,80,fill=TRUE,legend=rownames(freq),angle=ang,density=20,merge=TRUE,
  bg="white")
```



Test of Independence

```
> smoking.csq = chisq.test(freq)
> smoking.csq    # Show chi-square test results

        Pearson's Chi-squared test
data:  freq
X-squared = 980.0679, df = 8, p-value < 2.2e-16
```

Check Your Assumptions

Are all of the expected counts at least 5?

```
> smoking.csq$expected  # Show expected counts
                          <1/wk      1/wk    2-4/wk    5-6/wk     1/day
Never Smoked          2822.2227 1520.1157 2427.3295 1372.9598 2696.3722
Smoked in the Past    2241.5827 1207.3694 1927.9343 1090.4891 2141.6245
Currently Smokes       632.1946  340.5149  543.7361  307.5511  604.0033
```

Follow-up Analysis

```
> obs = smoking.csq$observed
> obs                 <1/wk 1/wk 2-4/wk 5-6/wk 1/day
Never Smoked           3577 1711   2430   1211  1910
Smoked in the Past     1595 1068   1999   1264  2683
Currently Smokes        524  289    470    296   849

> exp = smoking.csq$expected
> exp                      <1/wk      1/wk    2-4/wk    5-6/wk     1/day
Never Smoked          2822.2227 1520.1157 2427.3295 1372.9598 2696.3722
Smoked in the Past    2241.5827 1207.3694 1927.9343 1090.4891 2141.6245
Currently Smokes       632.1946  340.5149  543.7361  307.5511  604.0033

> contributions = (obs - exp)^2/exp
> contributions           <1/wk      1/wk      2-4/wk     5-6/wk     1/day
Never Smoked          201.85819 23.969753 0.002937945 19.1054265 229.33823
Smoked in the Past    186.50627 16.087717 2.619556092 27.6078338 136.85285
Currently Smokes       18.51655  7.793448 9.999371400  0.4338403  99.37593
```

Test of Independence by hand

1. Hypotheses:

2. Test statistic value:

3. Distribution of the test statistic and the p-value:

4. Decision:

5. Conclusion:

6. Follow-up:

Exercises

Goodness of fit (1-way tables) on p321: 6.39 - 6.43 (odd)
Independence (2-way tables) on p324: 6.45 - 6.49 (odd)