# Chapter 22

Two Categorical Variables:
The Chi-Square Test



## Relationships: Categorical Variables

- Chapter 20: compare proportions of successes for two groups
  - "Group" is explanatory variable (2 levels)
  - "Success or Failure" is outcome (2 values)
- Chapter 22: "***is there a relationship between two categorical variables?***"
  - may have 2 or more groups (one variable)
  - may have 2 or more outcomes (2nd variable)



## Two-Way Tables

- (from Chapter 6:)
  - When there are two categorical variables, the data are summarized in a *two-way table*
  - The number of observations falling into each combination of the two categorical variables is entered into each *cell* of the table
  - Relationships between categorical variables are described by calculating appropriate ***percents*** from the counts given in the table



## Case Study

Health Care: Canada and U.S.

Mark, D. B. et al., "Use of medical resources and quality of life after acute myocardial infarction in Canada and the United States," *New England Journal of Medicine*, **331** (1994), pp. 1130-1135.

Data from patients' own assessment of their quality of life relative to what it had been before their heart attack (data from patients who survived at least a year)



## Case Study

Health Care: Canada and U.S.

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 75 | 541 |
| Somewhat better | 71 | 498 |
| About the same | 96 | 779 |
| Somewhat worse | 50 | 282 |
| Much worse | 19 | 65 |
| Total | 311 | 2165 |



## Case Study

Health Care: Canada and U.S.

Compare the Canadian group to the U.S. group in terms of feeling much better:

| Quality of life | Canada | United States |
|---|---|---|
| Much better | $\boxed{75}$ | $\boxed{541}$ |
| Somewhat better | 71 | 498 |
| About the same | 96 | 779 |
| Somewhat worse | 50 | 282 |
| Much worse | 19 | 65 |
| Total | $\boxed{311}$ | $\boxed{2165}$ |

We have that 75 Canadians reported feeling much better, compared to 541 Americans.

The groups appear greatly different, but look at the group totals.

# Case Study

## Health Care: Canada and U.S.

Compare the Canadian group to the U.S. group in terms of feeling much better:

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 24% | 25% |
| Somewhat better | 23% | 23% |
| About the same | 31% | 36% |
| Somewhat worse | 16% | 13% |
| Much worse | 6% | 3% |
| Total | 100% | 100% |

**Change the counts to percents**

Now, with a fairer comparison using percents, the groups appear very similar in terms of feeling much better.



# Case Study

## Health Care: Canada and U.S.

Is there a relationship between the explanatory variable (*Country*) and the response variable (*Quality of life*)?

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 24% | 25% |
| Somewhat better | 23% | 23% |
| About the same | 31% | 36% |
| Somewhat worse | 16% | 13% |
| Much worse | 6% | 3% |
| Total | 100% | 100% |

Look at the ***conditional distributions*** of the response variable (Quality of life), given each level of the explanatory variable (Country).



# Conditional Distributions

- If the conditional distributions of the second variable are nearly the same for each category of the first variable, then we say that there is not an association between the two variables.
- If there are significant differences in the conditional distributions for each category, then we say that there is an association between the two variables.



# Hypothesis Test

- In tests for two categorical variables, we are interested in whether a relationship observed in a single sample reflects a real relationship in the population.
- Hypotheses:
  - Null: the percentages for one variable are the same for every level of the other variable (no difference in conditional distributions). (No real relationship).
  - Alt: the percentages for one variable vary over levels of the other variable. (Is a real relationship).



# Case Study

## Health Care: Canada and U.S.

Null hypothesis:
The percentages for one variable are the same for every level of the other variable.
(No real relationship).

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 24% | 25% |
| Somewhat better | 23% | 23% |
| About the same | 31% | 36% |
| Somewhat worse | 16% | 13% |
| Much worse | 6% | 3% |
| Total | 100% | 100% |

For example, could look at differences in percentages between Canada and U.S. for each level of "Quality of life":

24% vs. 25% for those who felt '*Much better*',
23% vs. 23% for '*Somewhat better*', etc.

Problem of *multiple comparisons*!



# Multiple Comparisons

- Problem of how to do many comparisons at the same time with some overall measure of confidence in all the conclusions
- Two steps:
  - overall test to test for any differences
  - follow-up analysis to decide which parameters (or groups) differ and how large the differences are
- Follow-up analyses can be quite complex; we will look at only the overall test for a relationship between two categorical variables

# Hypothesis Test

- $H_0$: no real relationship between the two categorical variables that make up the rows and columns of a two-way table
- **To test $H_0$, compare the *observed counts* in the table (the original data) with the *expected counts* (the counts we would expect if $H_0$ were true)**
  - if the observed counts are far from the expected counts, that is evidence against $H_0$ in favor of a real relationship between the two variables



# Case Study

## Health Care: Canada and U.S.

For the observed data to the right, find the expected value for each cell:

| Quality of life | Canada | United States | Total |
|---|---|---|---|
| Much better | 75 | 541 | $\boxed{616}$ |
| Somewhat better | 71 | 498 | 569 |
| About the same | 96 | 779 | 875 |
| Somewhat worse | 50 | 282 | 332 |
| Much worse | 19 | 65 | 84 |
| Total | $\boxed{311}$ | 2165 | $\boxed{2476}$ |

For the expected count of *Canadians* who feel '*Much better*' (expected count for Row 1, Column 1):

$$\text{expected count} = \frac{616}{2476} \times 311 = 77.37$$



# Expected Counts

- The expected count in any cell of a two-way table (when $H_0$ is true) is

$$\text{expected count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

- The development of this formula is based on the fact that the number of expected successes in *n* independent tries is equal to *n* times the probability *p* of success on each try (expected count = $n \times p$)
  - Example: find expected count in certain row and column (cell): *p* = proportion in row = (row total)/(table total); *n* = column total; expected count in cell = *np* = (row total)(column total)/(table total)



# Case Study

## Health Care: Canada and U.S.

Observed counts:

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 75 | 541 |
| Somewhat better | 71 | 498 |
| About the same | 96 | 779 |
| Somewhat worse | 50 | 282 |
| Much worse | 19 | 65 |

Compare to see if the data support the null hypothesis

Expected counts:

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 77.37 | 538.63 |
| Somewhat better | 71.47 | 497.53 |
| About the same | 109.91 | 765.09 |
| Somewhat worse | 41.70 | 290.30 |
| Much worse | 10.55 | 73.45 |



# Chi-Square Statistic

- To determine if the differences between the observed counts and expected counts are statistically significant (to show a real relationship between the two categorical variables), we use the ***chi-square statistic***:

$$X^2 = \sum \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}}$$

where the sum is over all cells in the table.



# Chi-Square Statistic

- The chi-square statistic is a measure of the distance of the observed counts from the expected counts
  - is always zero or positive
  - is only zero when the observed counts are exactly equal to the expected counts
  - large values of $X^2$ are evidence against $H_0$ because these would show that the observed counts are far from what would be expected if $H_0$ were true
  - the chi-square test is one-sided (any violation of $H_0$ produces a large value of $X^2$)

## Case Study

### Health Care: Canada and U.S.

Observed counts

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 75 | 541 |
| Somewhat better | 71 | 498 |
| About the same | 96 | 779 |
| Somewhat worse | 50 | 282 |
| Much worse | 19 | 65 |

Expected counts

| Canada | United States |
|---|---|
| 77.37 | 538.63 |
| 71.47 | 497.53 |
| 109.91 | 765.09 |
| 41.70 | 290.30 |
| 10.55 | 73.45 |

$$X^2 = \sum \left[ \frac{(75 - 77.37)^2}{77.37} + \frac{(541 - 538.63)^2}{538.63} + \cdots \right]$$

$$= 0.073 + 0.010 + \cdots$$

$$= 11.725$$

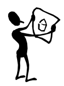


## Chi-Square Test

- Calculate value of chi-square statistic
  - by hand (cumbersome)
  - using technology (computer software, etc.)
- Find *P*-value in order to reject or fail to reject $H_0$
  - use **chi-square table** for **chi-square distribution** (later in this chapter)
  - from computer output
- If significant relationship exists (small *P*-value):
  - compare appropriate percents in data table
  - compare individual observed and expected cell counts
  - look at individual terms in the chi-square statistic



## Case Study



### Health Care: Canada and U.S.

Using Technology:

```
Chi-Square Test: Canada, USA

Expected counts are printed below observed counts

                   Canada       USA    Total
Much better            75       541      616
                    77.37    538.63

Somewhat better        71       498      569
                    71.47    497.53

About the same         96       779      875
                   109.91    765.09

Somewhat worse         50       282      332
                    41.70    290.30

Much worse             19        65       84
                    10.55     73.45

Total                 311      2165     2476

Chi-Sq =  0.073 +  0.010 +
          0.003 +  0.000 +
          1.759 +  0.253 +
          1.652 +  0.237 +
          6.766 +  0.972 = 11.725
DF = 4, P-Value = 0.020
```



## Chi-Square Test: Requirements

- The chi-square test is an approximate method, and becomes more accurate as the counts in the cells of the table get larger
- The following must be satisfied for the approximation to be accurate:
  - No more than 20% of the <u>expected</u> counts are less than 5
  - All individual <u>expected</u> counts are 1 or greater
- If these requirements fail, then two or more groups must be <u>combined</u> to form a new ('smaller') two-way table



## Uses of the Chi-Square Test

- Tests the null hypothesis

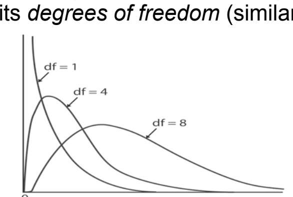
  *$H_0$: no relationship between two categorical variables*

  when you have a two-way table from either of these situations:
  - Independent SRSs from each of several populations, with each individual classified according to one categorical variable [Example: Health Care case study: two samples (Canadians & Americans); each individual classified according to "Quality of life"]
  - A single SRS with each individual classified according to both of two categorical variables [Example: Sample of 8235 subjects, with each classified according to their "Job Grade" (1, 2, 3, or 4) and their "Marital Status" (Single, Married, Divorced, or Widowed)]



## Chi-Square Distributions

- Distributions that take only positive values and are skewed to the right
- Specific chi-square distribution is specified by giving its *degrees of freedom* (similar to *t* distn)



## Chi-Square Test

- Chi-square test for a two-way table with $r$ rows and $c$ columns uses critical values from a chi-square distribution with **($r$ - 1)($c$ - 1)** degrees of freedom
- $P$-value is the area to the right of $X^2$ under the density curve of the chi-square distribution
  - use *chi-square table*

BPS - 5th Ed. Chapter 22 25

## Table D: Chi-Square Table

- See page 694 in text for Table D (*"Chi-square Table"*)
- The process for using the chi-square table (Table D) is identical to the process for using the $t$-table (Table C, page 693), as discussed in Chapter 17
- For particular degrees of freedom (*df*) in the left margin of Table D, locate the $X^2$ *critical value* ($x^*$) in the body of the table; the corresponding probability ($p$) of lying to the right of this value is found in the top margin of the table (this is how to find the P-value for a chi-square test)

BPS - 5th Ed. Chapter 22 26

## Case Study

Health Care: Canada and U.S.

**$X^2$ = 11.725**

***df*** = ($r$-1)($c$-1)
= (5-1)(2-1)
= **4**

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 75 | 541 |
| Somewhat better | 71 | 498 |
| About the same | 96 | 779 |
| Somewhat worse | 50 | 282 |
| Much worse | 19 | 65 |

Look in the *df*=4 row of Table D; the value $X^2$ = 11.725 falls between the 0.02 and 0.01 critical values.

Thus, the ***P*-value** for this chi-square test is **between 0.01 and 0.02** (is actually 0.019482).

** $P$-value < .05, so we conclude a significant relationship **

BPS - 5th Ed. Chapter 22 27

## Chi-Square Test and *Z* Test

- If a two-way table consists of $r$=**2** rows (representing 2 groups) and the columns represent "success" and "failure" (so $c$=**2**), then we will have a 2×2 table that essentially compares two proportions (the proportions of "successes" for the 2 groups)
  - this would yield a chi-square test with 1 df
  - we could also use the $z$ test from Chapter 20 for comparing two proportions
  - ** these will give **identical** results **

BPS - 5th Ed. Chapter 22 28

## Chi-Square Test and *Z* Test

- For a 2×2 table, the $X^2$ with df=1 is just the square of the $z$ statistic
  - $P$-value for $X^2$ will be the same as the two-sided $P$-value for $z$
  - should use the $z$ test to compare two proportions, because it gives the choice of a one-sided or two-sided test (and is also related to a confidence interval for the difference in two proportions)

BPS - 5th Ed. Chapter 22 29

## Chi-Square Goodness of Fit Test

- A variation of the Chi-square statistic can be used to test a different kind of null hypothesis: that *a single categorical variable has a specific distribution*
- The null hypothesis specifies the probabilities ($p_i$) of each of the $k$ possible outcomes of the categorical variable
- The chi-square goodness of fit test compares the observed counts for each category with the expected counts under the null hypothesis

BPS - 5th Ed. Chapter 22 30

## Chi-Square <u>Goodness of Fit</u> Test

- $H_o$: $p_1 = p_{1o}$, $p_2 = p_{2o}$, …, $p_k = p_{ko}$
- $H_a$: proportions are not as specified in $H_o$
- For a sample of *n* subjects, observe how many subjects fall in each category
- Calculate the expected number of subjects in each category under the null hypothesis: expected count = $n \times p_i$ for the $i^{th}$ category



## Chi-Square <u>Goodness of Fit</u> Test

- Calculate the chi-square statistic (same as in previous test):

$$X^2 = \sum \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}}$$

- The degrees of freedom for this statistic are $df = k$-1 (the number of possible categories minus one)
- Find *P*-value using Table D



## Chi-Square <u>Goodness of Fit</u> Test

**THE CHI-SQUARE TEST FOR GOODNESS OF FIT**

A categorical variable has $k$ possible outcomes, with probabilities $p_1$, $p_2$, $p_3$, …, $p_k$. That is, $p_i$ is the probability of the $i$th outcome. We have $n$ independent observations from this categorical variable.

To test the null hypothesis that the probabilities have specified values

$$H_0: p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$$

use the **chi-square statistic**

$$X^2 = \sum \frac{(\text{count of outcome } i - np_{i0})^2}{np_{i0}}$$

The $P$-value is the area to the right of $X^2$ under the density curve of the chi-square distribution with $k - 1$ degrees of freedom.



## Case Study

Births on Weekends?

National Center for Health Statistics, "Births: Final Data for 1999," *National Vital Statistics Reports*, Vol. 49, No. 1, 1994.

A random sample of 140 births from local records was collected to show that there are fewer births on Saturdays and Sundays than there are on weekdays



## Case Study

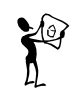
Births on Weekends?
Data

| Day | Sun. | Mon. | Tue. | Wed. | Thu. | Fri. | Sat. |
|---|---|---|---|---|---|---|---|
| Births | 13 | 23 | 24 | 20 | 27 | 18 | 15 |

Do these data give significant evidence that local births are not equally likely on all days of the week?



## Case Study



Births on Weekends?
Null Hypothesis

| Day | Sun. | Mon. | Tue. | Wed. | Thu. | Fri. | Sat. |
|---|---|---|---|---|---|---|---|
| Probability | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ |

$H_o$: probabilities are the same on all days

$H_o$: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$

## Case Study

Births on Weekends?

Expected Counts

Expected count = $n \times p_i$ =140×(1/7) = 20 for each category (day of the week)

| Day | Sun. | Mon. | Tue. | Wed. | Thu. | Fri. | Sat. |
|---|---|---|---|---|---|---|---|
| Observed births | 13 | 23 | 24 | 20 | 27 | 18 | 15 |
| Expected births | 20 | 20 | 20 | 20 | 20 | 20 | 20 |

BPS - 5th Ed. Chapter 22 37

## Case Study

Births on Weekends?

Chi-square statistic

$$X^2 = \sum_{i=1}^{7} \frac{(\textit{observed count} - 20)^2}{20}$$

$$= \sum \left[ \frac{(13-20)^2}{20} + \frac{(23-20)^2}{20} + \cdots + \frac{(15-20)^2}{20} \right]$$

$$= 2.45 + 0.45 + \cdots + 1.25$$

$$= \underline{\mathbf{7.60}}$$

BPS - 5th Ed. Chapter 22 38

## Case Study

Births on Weekends?

*P*-value, Conclusion

***X*²** = 7.60

***df*** = *k*-1 = 7-1 = 6

***P*-value** = Prob($X^2$ > 7.60):
$X^2$ = 7.60 is smaller than smallest entry in *df*=6 row of Table D, so the *P*-value is > 0.25.

**Conclusion:** Fail to reject $H_o$ – there is not significant evidence that births are not equally likely on all days of the week

BPS - 5th Ed. Chapter 22 39