**STA 2183W**

**Chapter 5 – Inference on One or More Qualitative Variables**

Throughout Chapters 2 and 4, we performed inferential methods on one- and two-sample means and medians. However, such metrics are restricted to only ***quantitative data***. What if we wished to perform inference on ***qualitative data***? In Chapter 5, we’ll explore inference on qualitative data from one or more samples, and use our results to demonstrate properties such as *independence*, *heterogeneity*, and *likelihood*.

**5.1 Tables and Graphics with Qualitative Data**

Before we jump into the statistical analysis of qualitative variables, let us briefly pause to discuss and review the use of tables and graphics that will prove to be useful for qualitative data.

One-Way Tables

Take the following example: *“A random sample of 250 nursing applications taken from a large university included 30 from men and 220 from women.”*

|  |  |  |  |
| --- | --- | --- | --- |
| ***Gender*** | ***Man*** | ***Woman*** | ***Total*** |
| ***Count*** | 30 | 220 | **250** |

Now, we could have this information provided to us by one of two options: through a dataset of some sort, or through a *table*. We could have a dataset that has 250 observations, with each observation responding to the *gender* variable, or simply have a table that produces the summaries of the information, as on our right.

In Chapter 1, we discussed performing statistical analysis using qualitative data, we focused on the case when we have a full dataset to work with. When using the entries from a full dataset, we can obtain our *one-way table* through the following commands.

Using SAS Code to Obtain Frequencies, Relative Frequencies, Percentages, and Tables:

**PROC** **FREQ** DATA=**DATA\_SET**;  
TABLES **VARIABLE\_DESIRED**;

**RUN**;

Using R Code to Obtain Frequencies, Relative Frequencies, and Percentages:

**table(x)**

**prop.table(table(x))**

**100\*prop.table(table(x))**

* To get a table presentable in the way it is in SAS, we can use the package **gmodels** and use the **CrossTable** command. Keep **prop.chisq** as false.

**install.packages("gmodels")**

**library(gmodels)**

**CrossTable(x,** **prop.r=T/F, prop.c=T/F, prop.t=T/F, prop.chisq=F)**

[Note: For the *prop.\_\_* options, r stands for row, c stands for column, and t stands for joint.]

If we do not have a data set, but just the table provided as above, we need to determine how to code the data so that we can perform the correct analyses. SAS and R both provide you with ways to analyze qualitative variables when all you have are the frequencies:

Using SAS to Code A Frequency Table:

* To create a dataset from a table, we will include *two* variables the categorical variable we are measuring and a variable ***number***, which records the number of individuals that exist within each category. Below, we consider a case where we have categories.
* Furthermore, note that SAS will order each variable in *alphabetical order*. So, if you wish to have a certain order, it would be advised to assign a *numerical value* to our categories, and then to use   
  **PROC** **FORMAT** to format your responses to get what you wish to see.

**DATA** **DATA\_SET**;

INPUT **CATEGORICAL\_VAR** $ **number**;

DATALINES;

CAT\_1 VALUE\_1

CAT\_2 VALUE\_2

...

CAT\_C VALUE\_C

;

**RUN**;

Using R to Code A Frequency Table:

* To create a dataset from a table, we have to begin working with what is called a ***matrix***. In the sense of computing, think of it as a rectangular data set with number of rows and number of columns. For one variable, we will have a matrix with row and columns. We will specify with **ncol** as number of columns, and include **byrow=TRUE** (which will prove useful soon).
* Label your column names in the order that you put them in your matrix.

**x <- matrix(c(VALUE\_1, VALUE\_2, … , VALUE\_C), ncol=c, byrow=TRUE )**

**colnames(x) <- c(“CAT\_1”, “CAT\_2”,…, “CAT\_C”)**

**rownames(x)<- "Count"**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Gender*** | ***Man*** | ***Woman*** | ***Total*** |
| ***Count*** | 30 | 220 | **250** |

**Example 5.1:**

Code the following table into SAS/R:

|  |
| --- |
| SAS Code:  **DATA** Nursing;  INPUT Gender $ number;  DATALINES;  Woman 220  Man 30  ;  **RUN**;  **PROC** **FREQ** Data=Nursing;  WEIGHT number;  TABLES Gender;  **RUN**; |
| R Code:  **nursing <- matrix(c(220, 30), ncol=2, byrow=TRUE )**  **colnames(nursing) <- c("Woman", "Man")**  **rownames(nursing)<- "Count"**  **CrossTable(nursing, prop.r=F, prop.c=F, prop.t=T, prop.chisq=F)** |

Contingency Tables & Two-Way Tables  
When we simultaneously group data from two or more variables into a frequency distribution, this frequency distribution is called a **contingency table**.

|  |
| --- |
| Introductory Statistics Review:  **Univariate data** – data from one variable of a population   * We used frequency distributions to group univariate data above.   **Bivariate data** – data from two variables of a population   * **Two-way contingency** table – frequency distribution for bivariate data * Each box in the table is a **cell.** Each cell represents a **joint event.** * **Joint proportion** – proportion of a joint event occurring * **Marginal proportion** – proportion of the events represented in the margins of the table. |

Example:   
The following data was taken from the General Social Survey (GSS) in 2010. In each survey, the GSS asks, “Taken all together, would you say that you are very happy, pretty happy, or not too happy? They also asked “Compared with American families in general, would you say that your family income is

below average, average, or above average?” Here are the results to those questions:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Happiness** | | | | |
| **Income** | ***Not too Happy (NH)*** | ***Pretty Happy (PH)*** | ***Very Happy (VH)*** | **Total** |
| ***Above Average (AA)*** | 31 | 241 | 140 | **412** |
| ***Average (A)*** | 96 | 506 | 248 | **850** |
| ***Below Average (BA)*** | 187 | 425 | 142 | **754** |
| **Total** | **314** | **1172** | **530** | **2016** |

Note that this contingency table has three categories for the variable ***Happiness***, and three categories for the variable ***Income***, leading us to have a total of cells. We can use this table to find our marginal probabilities/percentages

When using the entries from a full dataset, we can obtain our *contingency table* through the following commands.

Using SAS Code to Obtain Contingency Tables:

**PROC** **FREQ** DATA=**DATA\_SET**;  
TABLES **VAR\_1\*VAR\_2**;

**RUN**;

* Note that below the frequency counts, it also lists the Percentage, Row Percentage, and Column Percentages, respectively.
* If we wish to do *more than two variables (say three)*, we can expand the TABLES command to **VAR\_1\*VAR\_2\*…\*VAR\_3**. Note that it will produce separate two-way tables of each category in **VAR\_1**.

Using R Code to Obtain Contingency Tables:

**table(X1, X2)**

**table(X1, X2, X3)**

* If we wish to do *more than two variables (say three)*, we can use **table(X1, X2, X3)**. Note that it will produce separate two-way tables of each category in **X3**.
* To get a two-way table presentable in the way it is in SAS, we can use the **CrossTable** command.

**CrossTable(X1, X2, prop.r=T/F, prop.c=T/F, prop.t=T/F, prop.chisq=F)**

Again, if we do not have a data set, but just the table provided as above, we need to determine how to code the data so that we can perform the correct analyses. SAS and R both provide you with ways to analyze qualitative variables when all you have are the frequencies:

Using SAS to Code a Contingency Table:

* To create a dataset from a table, we will include *two* variables the categorical variable we are measuring and a variable ***number***, which records the number of individuals that exist within each category. Below, we consider a case where we have categories.
* Furthermore, note that SAS will order each variable in *alphabetical order*. So, if you wish to have a certain order, it would be advised to assign a *numerical value* to our categories, and then to use   
  **PROC** **FORMAT** to format your responses to get what you wish to see.

**DATA** **DATA\_SET**;

INPUT **CATEGORICAL\_VAR1** $ **CATEGORICAL\_VAR2** $ **number**;

DATALINES;

COLCAT\_1 ROWCAT\_1 COMBO#

COLCAT\_1 ROWCAT\_2 COMBO#

...

COLCAT\_1 ROWCAT\_R COMBO#

COLCAT\_2 ROWCAT\_1 COMBO#

...

COLCAT\_2 ROWCAT\_R COMBO#

...

COLCAT\_C ROWCAT\_1 COMBO#

...

COLCAT\_C ROWCAT\_R COMBO#

;

**RUN**;

* Once coded, to read using the frequency table, we use the line “WEIGHT ***number***;” in **PROC** **FREQ**.

**PROC** **FREQ** DATA=**DATA\_SET**;

WEIGHT ***number***;  
TABLES **VAR\_1\*VAR\_2**;

**RUN**;

Using R to Code A Contingency Table:

* To create a dataset from a table, we have to begin working with another way to input data, called a ***matrix***. In the sense of computing, think of it as a rectangular data set with number of rows and number of columns. For two variables, we will have a matrix with rows and columns. We will specify with **ncol** as number of columns, and include **byrow=TRUE** (which will prove useful soon).
* When inputting your matrix of values, be sure to go **left-to-right**, **top-to bottom**.  
  So, looks like **c(4,5,6,7)**.

**x <- matrix(c(COMBO\_1, COMBO\_2, … , COMBO\_CxR), ncol=c, nrow=r, byrow=TRUE )**

**colnames(x) <- c(“COLCAT\_1”, “COLCAT\_2”,…, “COLCAT\_C”)**

**rownames(x) <- c(“ROWCAT\_1”, “ROWCAT\_2”,…, “ROWCAT\_R”)**

* To get the *marginal* variables separately, you can use the following commands:  
  **Row Variable: rowSums(GSS)**

**Column Variable: colSums(GSS)**

* We can turn that matrix into a dataset of sorts through the ***melt*** command as learned in Chapter 3.

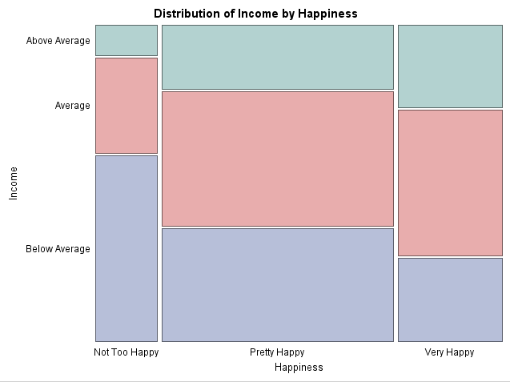
**Example 5.2:**

Code the contingency table for the GSS data above into SAS / R.

|  |
| --- |
| SAS Code:  **DATA** GSS;  INPUT Happiness $ Income $ number;  DATALINES;  1NH 1AA 31  1NH 2A 96  1NH 3BA 187  2PH 1AA 241  2PH 2A 506  2PH 3BA 425  3VH 1AA 140  3VH 2A 248  3VH 3BA 142  ;  **RUN**;  **PROC** **FORMAT**;  VALUE $Happ "1NH"="Not Too Happy"  "2PH"="Pretty Happy"  "3VH"="Very Happy";  VALUE $Inc "1AA" ="Above Average"  "2A" ="Average"  "3BA" ="Below Average";  **RUN**;  **PROC** **FREQ** DATA=GSS;  FORMAT Happiness $Happ. Income $Inc.;  WEIGHT number;  TABLES Income\*Happiness;  **RUN**; |
| R Code:  **GSS <- matrix(c(31,241,140,**  **96,506,248,**  **187,425,142),**  **ncol=3, nrow=3, byrow=TRUE )**  **colnames(GSS) <- c("Not too Happy", "Pretty Happy", "Very Happy")**  **rownames(GSS)<- c("Above Average", "Average", "Below Average")**  **CrossTable(GSS,prop.r=T, prop.c=T, prop.t=T, prop.chisq=F)** |

Graphics with Qualitative Data

We’ll finish this section with an overview on various graphics with qualitative data.



In Chapter 1, we covered bar graphs and pie charts. We also mentioned how we can separate these graphs by various groups.   
A third type of plot we can use is known as the *mosaic plot*. What this plot does is create a plot akin to the *stacking* effect to our bar graphs, which allows us to compare the distribution of one qualitative variable with respect to another.

Here is an overview of constructing these charts, along with how to construct them using data from a provided contingency table.

Using SAS Code to Create Pie Charts:

* We can use PROC GCHART to obtain a pie chart of your data.

**PROC** **GCHART** DATA=**DATA\_SET**;  
PIE **VARIABLE\_DESIRED**/TYPE=percent SLICE=outside VALUE=inside;

**RUN**;

* *Note that for* TYPE*,* SLICE*, and* VALUE*, you can choose freq/percent and outside/inside, respectively.*
* *If you coded your contingency table, include the command* FREQ=***number*** *into your options line.*
* *You can also add* GROUP=**VAR\_2** *into your options line in order to make multiple pie charts for each response for your second variable.*

Using SAS Code to Create Bar Charts:

* We can use PROC SGPLOT to obtain a bar chart of your data.

**PROC** **SGPLOT** DATA= **DATASET**;  
VBAR **VARIABLE\_DESIRED**/STAT=freq;

**RUN**;

* *Note that for horizontal bars, use* **HBAR** *instead of* **VBAR**. *For* STAT, *you can choose freq/percent.*
* *If you coded your contingency table, include the command* FREQ=***number*** *into your options line.*
* *You can also add* GROUP=**VAR\_2** *into your options line in order to make multiple pie charts for each response for your second variable.* 
  + *The default appearance is a stacked set of bar graphs. To make a clustered set of bar graphs, we use the* GROUPDISPLAY=cluster *command in the options line.*

Using SAS Code to Create Mosaic Charts:

* We can use **PROC** **FREQ** to obtain a mosaic chart of your data.

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VAR\_1\*VAR\_2** / PLOTS=MOSAIC;

**RUN**;

* *If you coded your contingency table into the data,* *use the line* “WEIGHT ***number***;” *in* **PROC** **FREQ**.

When performing these tasks in R, we can rely on similar former methods, but we will shift some focus to using the command **ggplot** (Grammar of Graphics) from the famous package **ggplot2**. The creator won the young statistician’s version of the Nobel Prize for this package.

Pie Charts are arguably the most difficult to get decent graphics in R.

Using R Code to Create Pie Charts Using Matrices:

* We will use the commands **paste** and **pie** to obtain a pie chart of your data.
* In this case, we will also get the *percentages* for our values, and get a pie chart on the *rows* variable.

**lbls <-paste(rownames(DATAMATRIX),   
 round(rowSums(DATAMATRIX )/sum(DATAMATRIX)\*100,1), "%", sep="")**

**pie(rowSums(DATAMATRIX),labels=lbls, main="TITLE", radius=1)**

* If you wish to use the *columns* variable, you can use **colSums** instead of **rowSums**, as well as **colnames** instead of **rownames**.

Using GGPlot with pie charts… is not very fun. With two-way tables, you have to be *very* sneaky into the code in order to get everything to work. Below is an idea of the process you need to follow:

1. Install the packages and open the libraries for **ggplot2** and **reshape2**.
2. **melt** your matrix into a dataset.
   * Provide column names in the form **c("Row\_Var", "Col\_Var", "Number")**
   * If you have multiple variables and only want a graph for one of them you need to use the   
     **aggregate** command.
3. We then run the ggplot command to get our values.

Using R Code to Create a One-Variable Pie Chart Using GGPlot:

**library(ggplot2)**

**library(reshape2)**

**Matrix.data<- melt(MATRIX)**

**colnames(Matrix.data)<- c("Row\_Var", "Col\_Var", "Number")**

**Matrix.VAR<-aggregate(Number~VAR\_DESIRED, Matrix.data, sum)**

**ggplot(data= Matrix.VAR, aes(x="", y=Number, fill= VAR\_DESIRED))+**

**geom\_bar(stat = "identity", position=position\_fill())+**

**geom\_text(aes(label=round(Number/sum(Number)\*100,1)),**

**position=position\_fill(vjust=0.5)) +**

**coord\_polar(theta="y")+**

**theme(axis.title.x=element\_blank(), axis.title.y=element\_blank())+**

**theme(legend.position = "bottom") +**

**ggtitle("MAIN TITLE")**

Using R Code to Create a Two-Variable Pie Chart Using GGPlot:

**library(ggplot2)**

**library(reshape2)**

**Matrix.data<- melt(MATRIX)**

**colnames(Matrix.data)<- c("Row\_Var", "Col\_Var", "Number")**

**ggplot(data= Matrix.data, aes(x="", y=Number, fill=VAR1))+**

**geom\_bar(stat = "identity", position=position\_fill())+**

**geom\_text(aes(label=Number), position=position\_fill(vjust=0.5)) +**

**coord\_polar(theta="y")+**

**facet\_wrap(~VAR2)+**

**theme(axis.title.x=element\_blank(), axis.title.y=element\_blank())+**

**theme(legend.position = "bottom") +**

**ggtitle("MAIN TITLE")**

Like I said – a pain, ya?

In general, most folks tend to prefer bar charts, as they can often provide greater insight on the differences among groups. For most assignments, we will focus on such procedures.

Using R Code to Create a One-Variable Bar Chart Using GGPlot:

**Matrix.data<- melt(MATRIX)**

**colnames(Matrix.data)<- c("Row\_Var", "Col\_Var", "Number")**

**ggplot(data= Matrix.data, aes(x=VAR\_DESIRED, y=Number))+**

**geom\_bar(width = 0.75, fill="COLOR", stat = "identity")+**

**ggtitle("MAIN TITLE")**

Using R Code to Create a Two-Variable Bar Charts Using GGPlot:

**Matrix.data<- melt(MATRIX)**

**colnames(Matrix.data)<- c("Row\_Var", "Col\_Var", "Number")**

**ggplot(data= Matrix.data, aes(x=VAR\_1, y=Number, fill=VAR\_2))+**

**geom\_bar(width = 0.75, position="stacked", stat = "identity")+**

**ggtitle("MAIN TITLE")**

* If you wish to use percentages instead of raw counts, change **y=Number** to **y=Number/sum(Number)**, and add in the line **+ scale\_y\_continuous(label=scales::percent)** .
* The above command gives us *stacked* bars. To get *clustered* bars, change **position="stacked",** to **position="dodge",**.

We can adjust the bar graphs to even give us Mosaic Charts.

Using R Code to Create Mosaic Charts:

**Matrix.data<- melt(MATRIX)**

**colnames(Matrix.data)<- c("Row\_Var", "Col\_Var", "Number")**

**ggplot(data= Matrix.data, aes(x=VAR\_1, y=Number/sum(Number), fill=VAR\_2))+**

**geom\_bar(width = 1, color="black", position="fill", stat = "identity")+**

**scale\_y\_continuous(label=scales::percent)+**

**ggtitle("MAIN TITLE")**

**Example 5.3:**

Use the contingency table you coded for the GSS data in Example 5.2 to perform the following:

1. A pie chart for Income.
2. A pie chart for Income, broken down by levels for Happiness.
3. A bar graph for Income.
4. A bar graph for Income, broken down by levels for Happiness.
5. A mosaic chart for Income, broken down by levels for Happiness.

|  |
| --- |
| SAS Code:  **PROC** **GCHART** DATA=GSS;  FORMAT Income $Inc.;  PIE Income/FREQ=number TYPE=percent SLICE=outside VALUE=inside;  TITLE 'Perception of Income';  **RUN**;  **PROC** **GCHART** DATA=GSS;  FORMAT Happiness $Happ. Income $Inc.;  PIE Income/FREQ=number GROUP=Happiness TYPE=percent SLICE=outside VALUE=inside;  TITLE 'Perception of Income and Happiness';  **RUN**;  **PROC** **SGPLOT** DATA=GSS;  FORMAT Income $Inc.;  VBAR Income/STAT=freq FREQ=number;  TITLE 'Perception of Income';  **RUN**;  **PROC** **SGPLOT** DATA=GSS;  FORMAT Happiness $Happ. Income $Inc.;  VBAR Happiness/STAT=percent FREQ=number GROUPDISPLAY=cluster GROUP=Income;  TITLE 'Perception of Income and Happiness';  **RUN**;  **PROC** **FREQ** DATA=GSS;  FORMAT Happiness $Happ. Income $Inc.;  WEIGHT number;  TABLES Income\*Happiness / PLOTS=MOSAIC;  TITLE 'Perception of Income and Happiness';  **RUN**; |

|  |
| --- |
| R Code:  **library(ggplot2)**  **library(reshape2)**  **GSS.data<-melt(GSS)**  **colnames(GSS.data)<- c("Income", "Happiness", "Number")**  **GSS.Income<-aggregate(Number~Income, GSS.data, sum)**  **ggplot(data=GSS.Income, aes(x="", y=Number, fill=Income))+**  **geom\_bar(stat = "identity", position=position\_fill())+**  **geom\_text(aes(label=round(Number/sum(Number)\*100,1)),**  **position=position\_fill(vjust=0.5)) +**  **coord\_polar(theta="y")+**  **theme(axis.title.x=element\_blank(), axis.title.y=element\_blank())+**  **theme(legend.position = "bottom") +**  **ggtitle("Perception of Income")**  **ggplot(data=GSS.data, aes(x="", y=Number, fill=Happiness))+**  **geom\_bar(stat = "identity", position=position\_fill())+**  **geom\_text(aes(label=Number), position=position\_fill(vjust=0.5)) +**  **coord\_polar(theta="y")+**  **facet\_wrap(~Income)+**  **theme(axis.title.x=element\_blank(), axis.title.y=element\_blank())+**  **theme(legend.position = "bottom") +**  **ggtitle("Perception of Happiness Based on Income")**  **ggplot(data=GSS.data, aes(x=Income, y=Number))+**  **geom\_bar(width = 0.75, fill="red", stat = "identity")+**  **ggtitle("Perception of Income")**  **ggplot(data=GSS.data, aes(x=Happiness, y=Number/sum(Number),**  **fill=Income))+**  **geom\_bar(width = 0.75, position="dodge", stat = "identity")+**  **ggtitle("Perceptions of Income and Happiness")+**  **ylab("Percentages")+**  **scale\_y\_continuous(label=scales::percent)**  **ggplot(data=GSS.data, aes(x=Happiness, y=Number/sum(Number),**  **fill=Income))+**  **geom\_bar(width=1,color="black", position="fill", stat="identity")+**  **ggtitle("Perceptions of Income and Happiness")+**  **ylab("Percentages")+**  **scale\_y\_continuous(label=scales::percent)** |

With these tables and graphics developed, we can now begin to develop methods to perform inference on qualitative data. About darn time, eh?

**5.2 Large-Sample Inference for a Proportion**

In this section, we will study the ***large-sample inference on proportions***.

Suppose you wish to estimate the proportion of voters in favor of a bill, the percentage of store customers who use a credit card for purchases, or the fraction of company CEOs who defer their annual pay.

In each case, the data of interest are categorical in natural, where each member either has a specified attribute or does not. Then, having said attribute could be seen as the same as being labeled as a “Success”, like in binomial experiments. Thus, all inference made in this setting would be considered ***inference about a binomial proportion,*** , where **.**

**Note:**

In the text and in some other places, you may see the use of instead of .   
We’ll stick to for this course.

We will use the following notation and terminology:

**Definitions:**

1. The **population proportion**, written as *,* is the proportion of the entire *population* that has the specified attribute.
2. The **sample proportion**, written as , is the proportion of a *sample* from the population that has the specified attribute.

|  |
| --- |
| **Point Estimate for :**  , where is the number of members in the sample  that have the specified attribute and is the sample size. |

**Notes:**

* is a **point estimator** for .
* is the number of successes and so is the number of failures.

When is unknown, we can use based on a random sample to make inferences about .

In order to make these inferences, we must know the sampling distribution of .

Recall when we discussed ***“Using a Normal Distribution to Approximate Binomial Probabilities”***. We mentioned that when is large enough, the Binomial distribution begins to look more and more similar to the Normal distribution. Here, we will require *that the number of successes and failures are at least 5*.

Also, since is the sum of successes, is then just the *average success rate*, and much like , the ***Central Limit Theorem*** suggests that the distribution of is approximately Normal when is large.

**The Sampling Distribution of**

|  |  |
| --- | --- |
| For samples of size ***n***, | |
| 1. . (*So, is an unbiased estimate of .*) |  |
| 1. The sampling distribution of is approximately Normal for large *n*.  * Large *n* means here that both and . | |

**Notes:**

* When the sample size is large, we may use in place of in .
  + - We will call the *estimated standard deviation the standard error of* .
* Like with the sampling distribution of the mean, we will ignore the ***finite population correction factor*** and assume that the population is large relative to the sample (*20 times larger or more*).

Example:   
Toss a coin 10 times. Let *X* be the number of “Heads”. Suppose we get 9 “Heads”.  
Based on the following information, would we consider this a fair (50-50) coin?

* Such skepticism provides us the fuel to use *inferential methods* to analyze and prove/disprove claims based on one-sample proportion, and the idea behind our *nonparametric procedures.*

Large-Sample Confidence Intervals for a Population Proportion

Previously, we had been constructing confidence intervals for population means/medians.

What if we want a confidence interval for a population proportion?

|  |
| --- |
| **Large-Sample Confidence Interval (CI) for a Population Proportion** |
| **Assumptions:**   1. A simple random sample of size *n* is drawn from the target population. 2. When the sample size is large: both and . (Number of successes and number of failures are at least 5). |
|  |

|  |
| --- |
| **Caution:** Unless is extremely large, the large-sample procedure presented in this section performs poorly when (or **)** is near or near 1. |

The example below illustrates a likely example of such issue occurring:

“Suppose you want to estimate the proportion of executives who parish from work-related injuries, using a sample of size . The proportion is likely to be near zero, say . If so, then

, and so a confidence interval for based on a sample of will probably be misleading.”

To combat this problem, you *could* get an *extremely large sample*, but often, such sampling becomes difficult to attain (due to limited resources). Instead, statisticians have suggested a simple modification we can make that improves our accuracy: this is the **Wilson-Agresti-Couli (WAC)** or **“Plus Four”** method for confidence intervals. This method has shown that this confidence interval works well for any value of , even when the sample size is small. In this course, we will focus solely on the WAC intervals.

|  |
| --- |
| **Adjusted WAC Confidence Interval for a Population Proportion,**    where  (1) is the adjusted sample proportion of observations with the characteristic of interest,  (2) is the number of successes in the sample, and  (3) is the sample size.   Thus, you add four imaginary observations, two successes and two failures to “balance” the proportion. |

Using SAS to Obtain a WAC Confidence Interval for One-Sample Proportions

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the BINOMIAL option as well as use the **AC** component to specify the WAC CI.

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VARIABLE\_DESIRED** / BINOMIAL(level="CATEGORY\_TO\_TEST" AC);

**RUN**;

* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).

Using R to Obtain a WAC Confidence Interval for One-Sample Proportions

* You will first need to install the **binom** package to use the **binom.confint** procedure, selecting the method **ac** to get a WAC confidence interval.
* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.

**install.packages("binom")**

**library(binom)**

**binom.confint(x,n, conf.level=CONFIDENCE\_LEVEL, methods ="ac")**

**Example 5.4** – Using the WAC Confidence Interval

According to the Bureau of Labor Statistics, the probability of injury while working at a jewelry store is very unlikely. Suppose that in a random sample of 200 jewelry store workers, 3 were injured on the job. Estimate the true proportion of jewelry store workers who are injured on the job, using 95% confidence.

|  |
| --- |
| SAS Code: (Create Data Set Jeweler)  **PROC** **FREQ** DATA=jeweler;  WEIGHT number;  TABLES Injured / BINOMIAL(level="Yes" AC p=**0.05**);  **RUN**; |
| R Code:  **library(binom)**  **binom.confint(3,200, conf.level=0.95, methods ="ac")** |

Large-Sample Test of a Hypothesis about a Population Proportion

Now it is time to explore hypothesis testing on the population proportion .

Recall the following information about the sampling distribution of :

|  |  |  |
| --- | --- | --- |
| (1) | (2) | (3) The sampling distribution of is approximately   Normal for large *n* [both and ]. |

Suppose we want to test  . Assuming the null hypothesis is true, then the quantity:

**=**  follows a normal distribution with mean and standard deviation .

Thus, we will use the test statistic to perform our hypothesis test.

|  |
| --- |
| Note With R:  At this point, you’d expect there to be a hiccup, right? When performing tests for proportion in R, it does not directly provide us with , but instead “X-Squared” a *Chi-Squared test statistic* ().  **Fact:**  If we square a Z-score test statistic, we get a Chi-Square test statistic with 1 degree of freedom.  (For more information, see STA 4158: Mathematical Statistics II – with inference!)  Both test statistics will provide you with the correct p-value. You will be provided a function to obtain the z-score, but for folks who use solely R, both values will count for when stating your test statistic. |

**Large-Sample Test for**

|  |  |  |
| --- | --- | --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | | |
| (1) A simple random sample from a binomial population.  (2) The sampling size is is sufficiently large [both and ]. | | |
| ***Step 2: State the null and alternative hypotheses.*** | | |
| **Right-sided Test** | **Left-sided Test** | **Two-sided Test** |
| ***Step 3: Compute the value of the test statistic.*** | | |
|  | | |
| ***Step 4: Determine the p-value.*** | | |
| |  |  | | --- | --- | | **Test** | **p-value** | | **Right-Sided** |  | | **Left-Sided** |  | | **Two-Sided** |  |   *p-value:*  *Calculate the p-value from the following below:* | | |
| ***Step 5: Based on p-value, make your conclusion.*** | | |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* | | |

Using SAS to Perform Large-Sample Inference on One Proportion

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the BINOMIAL option as well as the level we wish to test.

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VARIABLE\_DESIRED** / BINOMIAL(level="CATEGORY\_TO\_TEST" p=**P\_0**);

**RUN**;

* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).

Using R to Perform Large-Sample Inference on One Proportion

We will use the **prop.test** procedure, selecting **FALSE** for the correction factor.

* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.

**prop.test(x, n, p=P\_0, alternative= “two.sided”/ “less”/“greater” , conf.level = CONF\_LEVEL, correct=F)**

* As mentioned earlier, R does not *directly* provide . To obtain , use the process below.

**ztest <- prop.test(x, n, p=P\_0, alternative= “two.sided”/ “less”/“greater” ,   
conf.level = CONF\_LEVEL, correct=F)**

**sqrt(ztest$statistic)\*sign(x/n-P\_0)**

**Example 5.5:**

Central State University is considering constructing a new student parking lot. The administration believes that 60% of the students drive cars to school. If in a random sample of 350 students, 231 drive cars to school (so 119 do not), is there evidence that the administration is incorrect? Let .

|  |
| --- |
| SAS Code: (Create Data Set CSU)  **PROC** **FREQ** DATA=CSU;  WEIGHT number;  TABLES car / BINOMIAL(level="Car" AC p=**0.6**);  **RUN**; |
| R Code:  **ztest <- prop.test(231,350, p=0.6, alternative= “two.sided”/ , conf.level = 0.90, correct=F)**  **ztest**  **sqrt(ztest$statistic)\*sign(231/350-0.6)** |

**Example 5.6:**

Consider the data set ***carseats.csv***. The variable ***ShelveLoc*** indicates whether the carseats were shelved in either a “Bad”, “Good”, or “Medium” location in the store. Is there evidence to conclude that more than half of the stores received “Medium” ratings on their shelving location?   
Provide all steps to a hypothesis test.

|  |
| --- |
| SAS Code:  **PROC** **FREQ** DATA=carseats;  TABLES Shelveloc / BINOMIAL(level="Medium" p=**0.5**);  **RUN**; |
| R Code:  **carseats <- read.csv("carseats.csv")**  **SL.table <-table(carseats$ShelveLoc)**  **SL.table**  **sum(SL.table)**  **prop.test(219, 400, p=0.50, alternative= "greater", conf.level = 0.95, correct=F)** |

**5.3 The Exact Binomial Test**

Most surveys and studies employ large samples, but in the case of small samples, the large-sample test in Section 5.2 we’ve been using may not be valid, as the sampling distribution may be skewed. However, with small samples, we can use yet another *nonparametric procedure*, called an ***Exact Binomial Test***.

[Note: The WAC confidence interval has been found still be useful when the assumptions are violated.]

The Exact Binomial Test is actually an extension of the sign test we performed back in Chapter 2. The idea is that we *expect* that a specific proportion of the population to feature a specific category (typically under ), and thus sampling a single individual that possesses that category is like ***flipping a weighted coin***, one with probability .

Since each of these observations are like a coin flip, that suggests that we have a ***binomial distribution****,* with observations and probability . []

We will let  ***be our test statistic***. Then, for an alternative hypothesis our p-value comes from a binomial distribution like the sign test, but now we generalize our proportion to be instead of just .

**The Exact Binomial Test for a Population Proportion**

|  |  |  |
| --- | --- | --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | | |
| (1) A simple random sample  (2) The distribution is binomial. | | |
| ***Step 2: State the null and alternative hypotheses.*** | | |
| **Right-sided Test** | **Left-sided Test** | **Two-sided Test** |
| ***Step 3: Compute the value of the test statistic.*** | | |
|  | | |
| ***Step 4: Determine the p-value.*** | | |
| *p-value:*   |  |  | | --- | --- | | **Test** | **P-Value** | | **Left-Sided** |  | | **Two-Sided** |  | | **Right-Sided** |  |   *Calculate the p-value from the following, with* | | |
| ***Step 5: Based on Rejection Region / p-value, make your conclusion.*** | | |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* | | |

Using SAS to Perform Exact Binomial Test on One Proportion

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the BINOMIAL option as well as the level we wish to test.

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VARIABLE\_DESIRED** / BINOMIAL(level="CATEGORY\_TO\_TEST" p=**P\_0**);

EXACT Binomial;

**RUN**;

* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).

Using R to Perform Exact Binomial Test on One Proportion

We will use the **binom.test** procedure.

* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.

**binom.test (x, n, p=P\_0, alternative= “two.sided”/ “less”/“greater” , conf.level = CONF\_LEVEL)**

**Example 5.7:** (*Page 490 of Ott*)

The public health department in a county with a large number of oil wells has been assigned the task of evaluating whether the wastewater from the oil wells has polluted the water from private water wells near the drilling sites. In a preliminary study, a random sample of 15 oil wells was selected in the county. For each of the selected oil wells, a water well within 0.25 kilometers of the oil well is examined. In 4 of the 15 wells, the level of endocrine-disrupting chemicals was above the level that can cause interferences with the body’s normal hormonal function. These chemicals are known to occur naturally in approximately 20% of water wells. Is there significant evidence that more than 20% of the water wells near an oil well are contaminated with endocrine-disrupting chemicals? Use .

|  |
| --- |
| SAS Code: (Create Data Set Wells with variable OilWells)  **PROC** **FREQ** DATA=Wells;  WEIGHT Number;  TABLES OilWells / BINOMIAL(level="Oil" p=**0.2**);  EXACT Binomial;  **RUN**; |
| R Code:  **binom.test(4,15, p=0.2, alternative="greater", conf.level = 0.95)** |

**5.4 Goodness-of-Fit Test and the Multinomial Experiment**

Sometimes experiments will result in measurements that are qualitative or categorical, such as:

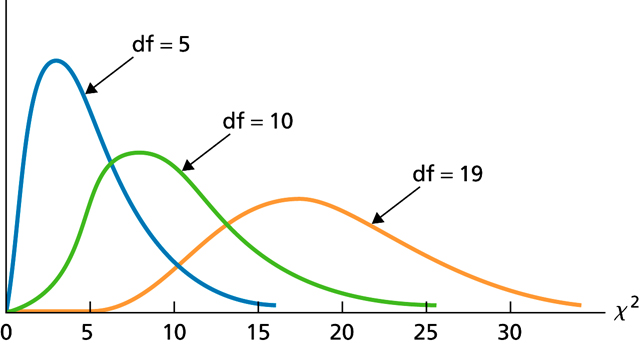
* An M&M can be one of 6 colors
* Airline ticket: coach, business class, or first class
* Survey question: Strongly disagree, disagree, no opinion, agree, strongly agree

We can summarize this type of data by reporting the count of the number of measurements in each category, as it is considered to be a **multinomial experiment** (like binomial, but with categories).

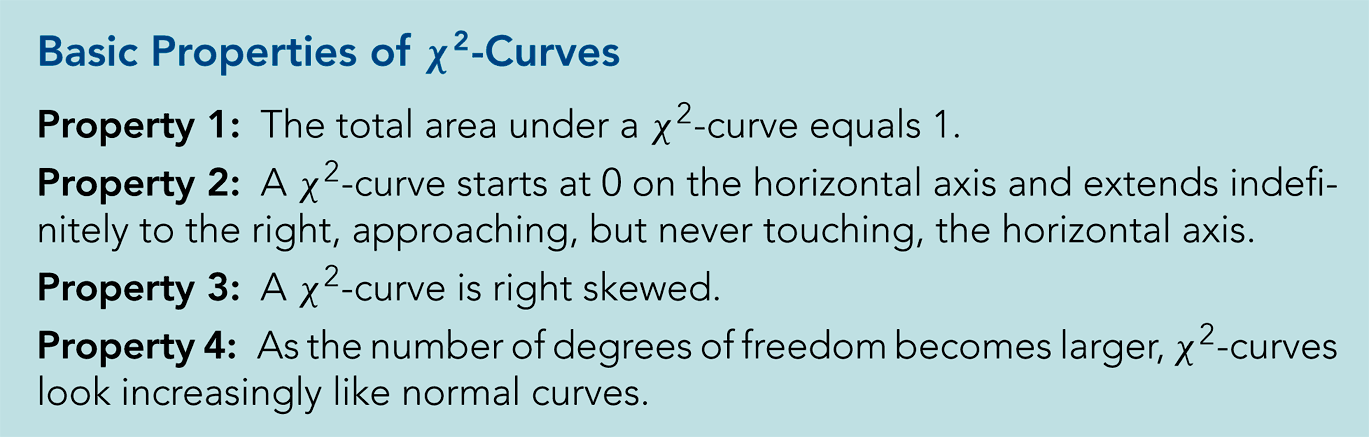
|  |
| --- |
| **Properties of the Multinomial Experiment**  (1) The experiment consists of identical trials. (2) There are possible outcomes to each trial. These outcomes are sometimes called   **classes**, **categories**, or **cells**.  (3) The probabilities of the outcomes, (), where ,   remain the same among trials.  (4) The trials are independent.  (5) The random variables of interest are the counts the # of observations   in each of the categories. |

Our previous procedures (using z-procedures and t-procedures) don’t fit the conditions we wish to consider, as we typically have more than two categories and proportions for each variable. So, we need to derive something new:

**The Chi-Square Distribution**

A variable has a chi-square distribution if its distribution has the shape of a special type of right-skewed curve, called a chi-square curve. Like the *t*-distribution, a chi-square distribution has an associated degrees of freedom (df).

**Notation:** denotes the -value with (df) degrees of freedom that has area α (our significance level) to its right under a chi-square curve.



The Chi-Square Tests have a multitude of uses. In this section, we consider a multinomial experiment with outcomes that correspond to categories in a *single* *qualitative variable*.

Example:   
According to a past post on the M&M’s website, M&M’s are manufactured in the percentages listed in the following table. A 14-ounce bag of M&M’s is randomly selected and contains the following number of each color. Do the data contradict the claims made by the Mars Company at the 5% significance level?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Color** | **Brown** | **Yellow** | **Red** | **Blue** | **Orange** | **Green** |
| **Percentage** | 13% | 14% | 13% | 24% | 20% | 16% |
| **Observed #** | 70 | 72 | 61 | 118 | 108 | 85 |

The table above is what we call a **one-way table**. The term *one-way* is used because only one variable is classified. Typically, we will wish to perform hypothesis tests on the true proportions of the categories.

For this problem, we are testing:

: The color distribution of M&M’s is the same as the distribution given on their website.

: The color distribution of M&M’s differs from the given distribution.

If we let represent the assumed proportion for the first category for the second category, and so on, we also can write our hypotheses as:

We will compare the **observed** number of M&M’s of each color to the **expected** number of M&M’s that we get of each color under the assumption that the null hypothesis is true.

**Note:**If *n* is the sample size, then the ***expected frequencies*** are given by where is the hypothesized proportion.

Do the observed and expected frequencies differ significantly from each other?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Observed ()** | **Expected ()** | **Difference ()** |  |  |
| Brown | 70 | 66.82 | 3.18 | 10.1124 | 0.1513 |
| Yellow | 72 | 71.96 | 0.04 | 0.0016 | 0.0000 |
| Red | 61 | 66.82 | -5.82 | 33.8724 | 0.5069 |
| Blue | 118 | 123.36 | -5.36 | 28.7296 | 0.2329 |
| Orange | 108 | 102.8 | 5.2 | 27.0400 | 0.2630 |
| Green | 85 | 82.24 | 2.76 | 7.6176 | 0.0926 |
| **Total** | 514 | 514 |  |  | = 1.2468 |

|  |  |
| --- | --- |
| The chi-square test statistic is the following,  summed over all the cells, where : |  |

When the null hypothesis is true, the test statistic has an approximate chi-square distribution in repeated sampling if the following Assumptions are met:

1. Simple random sample from a multinomial distribution.
2. The sample size is large. This condition can be satisfied if for every cell, the expected count, ,   
   is 5 or greater (.

|  |
| --- |
| ***Notice*:**  Sometimes, other researchers relax the second assumption, trading that assumption for those below:   1. All expected frequencies are 1 or greater. 2. At most 20% of the expected frequencies are less than 5. |

**Notes:**

1. If the expected cell counts are correct, the differences are likely to be small in absolute value and .
2. If the expected cell counts are wrong, the values are likely to be large in absolute value and is large.
3. The number of degrees of freedom is , where
   * Note that comes from the fact that, given a total value, you can freely assign the values of all but one of the categories, as the last category must equal the remaining unassigned observations.

**Chi-Square Goodness-of-Fit Test Procedure (Multinomial Probabilities: One-Way Table)**

|  |  |
| --- | --- |
| Purpose: To perform a hypothesis test for the distribution of a single variable and   whether it fits the suggested proportions previously stated. | |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | |
| 1. Simple random sample from a multinomial distribution. 2. The sample size is large enough so that for all cells. | |
| ***Step 2: State the null and alternative hypotheses.*** | |
| (*The variable has the specified distribution of   probabilities.*) | (*The variable does not have the specified   distribution given.*) |
| ***Step 3: Compute the value of the test statistic.*** | |
| , where . Use | |
| ***Step 4: Determine the p-value.*** | |
| ***P*-value** | |
| ***Step 5: Based on p-value, make your conclusion.*** | |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* | |

Note:

The Chi-Square test works for categorical variables with categories. When , then we are back to testing binomial proportions as in Sections 5.2 and 5.3. Because of that special case, our chi-square test statistic will be equal to the *square* of the z-score test statistic we obtain if we ran a large-sample proportion test.

Using SAS to Perform Goodness-of-Fit Test

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the TESTP option as well as the level we wish to test.
* When creating your data, be sure to format your qualitative variable names to correspond with the order your provide in the TESTP command.

**PROC** **FREQ** DATA=MandM;

TABLES Color / TESTP=(**p1\_0**, **p2\_0**,…,**pk\_0**);

**RUN**;

* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).

Using R to Perform Goodness-of-Fit Test

We will use the **chisq.test** procedure, selecting **FALSE** for the correction factor.

* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.

**chisq.test(x=c(x1, x2, …, xk) , p=p0=c(p1\_0, p2\_0,…,pk\_0), correct=FALSE)**

Thus, let us finish the M&M’s example:

**Example 5.8:**   
According to a past post on the M&M’s website, M&M’s are manufactured in the percentages listed in the following table. A 14-ounce bag of M&M’s is randomly selected and contains the following number of each color. Do the data contradict the claims made by the Mars Company at the 5% significance level?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Color** | **Brown** | **Yellow** | **Red** | **Blue** | **Orange** | **Green** |
| **Percentage** | 13% | 14% | 13% | 24% | 20% | 16% |
| **Observed #** | 70 | 72 | 61 | 118 | 108 | 85 |

|  |
| --- |
| SAS Code: (Create Dataset MandM with column order the same as above.)  **PROC** **FREQ** DATA=MandM;  WEIGHT Number;  TABLES Color / TESTP=(**0.13**,**0.14**,**0.13**,**0.24**,**0.20**,**0.16**);  **RUN**; |
| R Code: (Create Dataset MandM with column order the same as above.)  **p0 <- c(0.13,0.14,0.13,0.24,0.20,0.16)**  **chisq.test(ObservedMM, p=p0, correct=FALSE)** |

**5.5 Testing Categorical Probabilities: Two-Way Tables & The Chi-Square Homogeneity Test**

We have been constructing confidence intervals and conducting tests of hypotheses in different situations. We have considered making inferences about a mean & median, making inferences about a proportion, and comparing two means & even two medians. Now we are going to compare two or more distributions with qualitative data.

To compare the distributions of a qualitative variable for two or more populations, we will use the   
***Chi-Square Test for Homogeneity*** to determine whether the distributions of the two or more populations are the same with respect to the categorical variable.

The idea is that, if the populations are ***homogenous*** (the same) with respect to a qualitative variable, then the marginal percentages are the same with respect to the population totals.

We will enhance the chi-square test statistic to be the following: , where: such that  **= {row sum that is in}** and  **= {column sum that is in}**.

[*Note: The text uses instead of , and instead of . We will ignore this notation.*]

When the null hypothesis is true, in repeated sampling, the test statistic has an approximate chi-square distribution with , where  **= {# of rows}** and **= {# of columns}**.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Why***   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **C1** | **C2** | **C3** | **Total** | | **R1** |  |  |  |  | | **R2** |  |  |  |  | | **R3** |  |  |  |  | | **Total** |  |  |  |  |   The reasoning comes from a mix of math and puzzles.   If we are to construct a table of cells, then, given our marginal totals, once you fill in any set of cells, the rest of the table can be filled in through logic. |

We can perform this test if the following assumptions are made:

1. The Simple random samples each come from a multinomial distribution.
2. The sample size is large enough such that for all cells.  
   [*The alternative assumptions for Assumption 2 are analogous to the Goodness-of-Fit Test*]
3. The two samples / populations are Independent.

*What can we do if the second assumption is not met?*

1. Combine rows or columns to increase the frequencies in those cells in which they are too small.
2. Eliminate certain rows or columns in which the small expected frequencies occur.
3. Increase the sample size.

Caution:

When examining effects on population, combining certain variables can change the conclusion made by your inference test. More on such a phenomenon in Section 5.6.

**Chi-Square Homogeneity Test**

|  |  |
| --- | --- |
| Purpose: To perform a hypothesis test to compare the distributions of a qualitative variable for two or   more populations and see if they are similar. | |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | |
| 1. Simple random samples from multinomial distributions. 2. The sample size is large enough such that for all cells.\* 3. Independent samples. | |
| ***Step 2: State the null and alternative hypotheses.*** | |
| The populations are homogenous with respect   to the variable (same). | The populations are non-homogenous with   respect to the variable (different). |
| ***Step 3: Compute the value of the test statistic.*** | |
| , where . Use | |
| ***Step 4: Determine the p-value.*** | |
| ***P*-value** | |
| ***Step 5: Based on p-value, make your conclusion.*** | |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* | |

Using SAS to Perform Chi-Square Test of Homogeneity

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the CHISQ option as well as the level we wish to test.
* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).
* If you would like the expected count for each cell shown, then enter the **EXPECTED** command in the **TABLES** line. If you would like the contribution to the chi-square statistic for each cell, enter **CELLCHI2**. If you would not like the column, row and cell percentages to be shown, use the commands **NOCOL NOROW NOPERCENT**, respectively.

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VAR1\*VAR2** / CHISQ;

**RUN**;

Using R to Perform Chi-Square Test of Homogeneity

We will use the **CrossTable** procedure, selecting **expected=T,** **chisq = T**.

* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.

**CrossTable(TABLE, prop.r=T/F, prop.c=T/F, prop.t=T/F, prop.chisq=F,** **expected=T,** **chisq = T)**

|  |
| --- |
| **Question:** *“What can be done if our sample size is too small and we wish to keep all categories?”* If the sample is small, it was previously mentioned that we could remove or combine categories in order to get our expected frequencies to meet out requirements. However, sometimes, we desire to keep all categories intact, and so the -values attained most likely would not be accurate.  We must use the ***Fisher’s Exact Test*** a *nonparametric test* that uses the *hypergeometric distribution*. |

**Example 5.9:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Year | | |
| *1996* | *2002* | **Total** |
| Age (yr) | *17 or younger* | 5 | 6 | **11** |
| *18-24* | 77 | 78 | **155** |
| *25-34* | 88 | 103 | **191** |
| *35-44* | 72 | 66 | **138** |
| *45-54* | 28 | 17 | **45** |
| *55 or older* | 6 | 4 | **10** |
| **Total** | **276** | **274** | **550** |

The Bureau of Justice Statistics surveys jail inmates on various issues and reports its findings in *Profile of Jail Inmates*. Independent simple random samples of jail inmates in two different years gave the following information on age. At the 5% significance level, do the data provide sufficient evidence to conclude that, in the two years, jail inmates are nonhomogeneous with respect to age?

|  |
| --- |
| SAS Code: (Create Data Set inmates.)  **PROC** **FREQ** DATA=inmates;  WEIGHT number;  TABLES age\*year / CHISQ;  FORMAT age afmt. year yfmt.;  **RUN**; |
| R Code: (Create Data Set inmates. Turn off total proportions)  **CrossTable(inmates,prop.r=T, prop.c=T, prop.t=F, prop.chisq=F,** **expected=T,** **chisq = T)** |

Special Case – Comparing Two Population Proportions: Independent Sampling with a 2x2 Table

In introductory statistics, you may have also tested to test whether the probability of success is the same for populations. For most of y’all, these inference problems were performed using a ***two-sample Z-test on difference in proportions***.

However, note that the Chi-square Test for Homogeneity extends this test to compare a whole bunch of proportions all at once. If there are more than 2 categories (i.e., success and failure), then the test for homogeneity is comparing several multinomial populations. Thus, the two-proportion test can also be done with the Chi-Square test.

One point of importance is obtaining a confidence interval for the difference between two population proportions, . We will estimate this difference with . Note the following:

**The Sampling Distribution of :**

|  |
| --- |
| 1. The **mean** of the sampling distribution is: . 2. The **standard** **deviation** of the sampling distribution is: . 3. When the samples are large, the distribution of is approximately Normal.  That is, when , , and are all at least 5. |

**Note:**

Like one-sample procedures, we can approximate and in the confidence interval below.

|  |  |
| --- | --- |
| **Large-Sample Confidence Interval (CI) for** | |
| **Assumptions:**   1. The samples are independently chosen random samples from the two populations 2. The sample sizes and are large enough such that the number of successes and failures from each sample is at least 5. |  |

Unless are both extremely large, the large-sample procedure presented in this section performs poorly whenor(or or )are near or near 1. Again, it is recommended using the “Plus Four” version of the confidence interval for two samples in order to improve accuracy.

|  |
| --- |
| **Adjusted (“Plus Four”) Confidence Interval for a Population Proportion,**  where  (1) and  (2) is the number of successes in sample 1, is the number of successes in sample 2, and  (3) is the size of sample 1, and is the size of sample 2.   * Use this interval when the sample size is at least 5 in each group, with any counts of successes and failures. * Thus, you add four imaginary observations, one success and one failure to each sample. |

If you do desire to run a two-proportion test, note that it *pools* your proportion (under the belief that the two proportions are equal). From that, we can test with the test statistic

where

Using SAS to Perform Large Two-Sample Proportion Test

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the RISKDIFF option with EQUAL as well as the level we wish to test.
* We will also set CL=AC to get a WAC confidence interval.

**PROC** **FREQ** DATA=**DATASET**;

WEIGHT number;

TABLES **VAR1\*VAR2**/RISKDIFF(EQUAL CL=AC) ALPHA=**ALPHA\_VALUE**;

**RUN**;

* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).

Using R to Perform Large Two-Sample Proportion Test

We will use the **prop.test** procedure, selecting **FALSE** for the correction factor.

* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.
* We will also use the **wald2ci** command from the package **PropCIs** to get our point estimate and WAC confidence interval.

**prop.test(x=c(x1, x2), n=(n1, n2), alternative= "greater/less/two.sided", conf.level =CONF\_LEVEL, correct=F)**

**library(PropCIs)**

**wald2ci(x1, n1, x2, n2, conf.level=CON\_LEVEL, adjust="AC")**

**Example 5.10:**   
Olestra is a fat substitute approved by the FDA for use in snack foods. To investigate reports of gastrointestinal problems associated with olestra consumption, an experiment was carried out to compare olestra potato chips with regular potato chips (“Gastrointestinal Symptoms Following Consumption of Olestra or Regular Triglyceride Potato Chips,” Journal of the American Medical Association, [1998]:   
150-152). Subjects were assigned at random to either the olestra chip group or the regular chip group. Of the 529 individuals in the regular chip group, 93 experienced an adverse gastrointestinal symptoms, whereas 89 of the 563 individuals in the olestra chip group experienced symptoms. Does it appear that the rate of gastrointestinal problems for those who consume olestra chips differs from the rate of those who experience symptoms after consuming regular chips?

Test the claim at the 5% significance level and provide the associated 95% WAC confidence interval.

|  |
| --- |
| SAS Code: (Create Data Set for Potato)  **PROC** **FREQ** DATA=potato;  WEIGHT number;  TABLES chip\*GIProb/RISKDIFF(EQUAL CL=AC) ALPHA=**0.05**;  FORMAT chip $chipfmt. GIProb $GIfmt.;  **RUN**; |
| R Code: (Create Data Set for Potato)  **ztest<- prop.test(potato, alternative ="two.sided", correct=F)**  **ztest**  **sqrt(ztest$statistic)**  **library(PropCIs)**  **wald2ci(89, 89+474,93, 93+436, conf.level=0.95, adjust="AC")** |

**5.6 Two-Way Tables & The Chi-Square Test for Independence**

In the last section, we explored differences in distributions between two independent populations. However, we can use similar principles to test where *independence* *even occurs* within a single population*.*

Association between Qualitative Variables

Are two variables of a population associated?

* They are associated if knowing the value of one of the variables gives us information about the value of the other variable. *(We typically check this statistically instead of common sense, since that gives us an objective viewpoint.)*
* If we have bivariate data for only ***a sample*** of the population, then we must apply **inferential methods** to decide whether the two variables are associated.

**Note:**   
We must keep in mind that association does not imply causation. However, if two variables are not associated, there is no sense in looking for a causal relationship. In other words, association is necessary for causation.  
  
Example:  
*What determines your overall happiness? Does it depend on whether you have a partner, how many friends you have, whether or not you like your job, how much money you make, how often you attend religious services, how good your health is, how attractive you are, or which country you live in? To investigate which variables are associated with happiness, we can use the data taken from the General Social Survey (GSS) in 2010.*

*How can we determine if there is an association between happiness and family income in the population of all adult Americans? In each survey, the GSS asks, “Taken all together, would you say that you are very happy, pretty happy, or not too happy? They also asked “Compared with American families in general, would you say that your family income is below average, average, or above average?”*

The results are given in the table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Happiness** | | |  |
| **Income** | **Not too Happy** | **Pretty Happy** | **Very Happy** | **Total** |
| **Above Average** | 31 | 241 | 140 | **412** |
| **Average** | 96 | 506 | 248 | **850** |
| **Below Average** | 187 | 425 | 142 | **754** |
| **Total** | **314** | **1172** | **530** | **2016** |

We will calculate the expected cell frequencies, assuming the null hypothesis is true.

Recall from introductory probability, where we looked for an association or a relationship between two variables.

* If there was a relationship [], we considered the two variables **dependent**.
* If two variables are **independent** [], then the variables are not associated.

Consider Event *A* = “not too happy” and Event *B* = “above average income.”

If the events are independent, then this means that knowing if someone has “above average income” would provide us with no information on whether that person would respond with being “not too happy”.  
  
Once again, the chi-square test statistic is the following: , where: such that  **= {row sum that is in}** and  **= {column sum that is in}**.

[*Note: The text uses instead of , and instead of . We will ignore this notation.*]

When the null hypothesis is true, in repeated sampling, the test statistic has an approximate chi-square distribution with , where  **= {# of rows}** and **= {# of columns}**.

We can perform this test if the following assumptions are made:

1. Simple random sample from a multinomial distribution.
2. The sample size is large enough such that for all cells.\*  
   [\**The alternative assumptions for Assumption 2 are analogous to the Goodness-of-Fit Test*]

Notes:

* The expected frequencies and test statistic are computed the **same** as the Test for Homogeneity.
* The assumptions and null and alternative hypotheses are **different**.
  + We now will test that the variables are *independent* of each other.

*What can we do if the second assumption is not met?*

1. Combine rows or columns to increase the frequencies in those cells in which they are too small.
2. Eliminate certain rows or columns in which the small expected frequencies occur.
3. Increase the sample size.

If the test statistic is sufficiently large, the null hypothesis claim of independence is rejected and the variables are declared to be **dependent**.  
 **Chi-Square Test for Independence**

|  |  |
| --- | --- |
| Purpose: To perform a hypothesis test to decide whether two variables are associated. | |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | |
| 1. Simple random sample from a multinomial distribution. 2. The sample size is large enough such that for all cells. | |
| ***Step 2: State the null and alternative hypotheses.*** | |
| The two variables are independent. | The two variables are dependent. |
| ***Step 3: Compute the value of the test statistic.*** | |
| , where . Use | |
| ***Step 4: Determine the p-value.*** | |
| ***P*-value** | |
| ***Step 5: Based on p-value, make your conclusion.*** | |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* | |

Using SAS to Perform Chi-Square Test of Independence

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the CHISQ option as well as the level we wish to test.
* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).
* If you would like the expected count for each cell shown, then enter the **EXPECTED** command in the **TABLES** line. If you would like the contribution to the chi-square statistic for each cell, enter **CELLCHI2**. If you would not like the column, row and cell percentages to be shown, use the commands **NOCOL NOROW NOPERCENT**, respectively.

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VAR1\*VAR2** / CHISQ;

**RUN**;

Using R to Perform Goodness-of-Fit Test

We will use the **CrossTable** procedure, selecting **chisq = T**.

* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.

**CrossTable(TABLE, expected=T, prop.r=T/F, prop.c=T/F, prop.t=T/F, prop.chisq=F,** **chisq = T)**

**Example 5.11:**  
Consider the GSS data above.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Happiness** | | |  |
| **Income** | **Not too Happy** | **Pretty Happy** | **Very Happy** | **Total** |
| **Above Average** | 31 | 241 | 140 | **412** |
| **Average** | 96 | 506 | 248 | **850** |
| **Below Average** | 187 | 425 | 142 | **754** |
| **Total** | **314** | **1172** | **530** | **2016** |

*Now*, while we may have inclinations on whether happiness is dependent on income, we need more than perception: we need the data to provide sufficient evidence for us to make this claim.   
Test this hypothesis at the 1% significance level.

|  |
| --- |
| SAS Code: (Create Data Set for GSS)  **PROC** **FREQ** DATA=GSS;  FORMAT Happiness $Happ. Income $Inc.;  WEIGHT number;  TABLES Income\*Happiness / CHISQ;  **RUN**; |
| R Code: (Create Data Set for GSS)  **library(gmodels)**  **CrossTable(GSS,expected=T, prop.r=T, prop.c=T, prop.t=T, prop.chisq=F, chisq = T)** |

|  |
| --- |
| **Question:** *“What can be done if our sample size is too small and we wish to keep all categories?”* If the sample is small, it was previously mentioned that we could remove or combine categories in order to get our expected frequencies to meet out requirements. However, sometimes, we desire to keep all categories intact, and so the -values attained most likely would not be accurate. Thus, we will need to use the ***Fisher’s Exact Test***. It uses the *hypergeometric distribution*. |

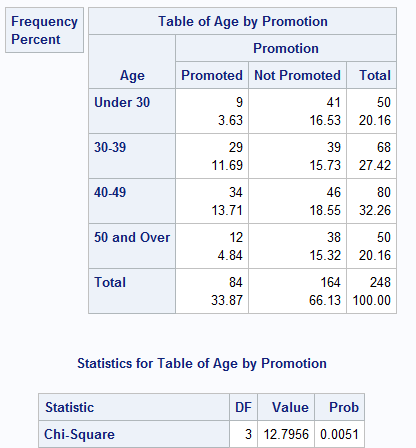
Point of Caution: Independence May Not Be Immediately Evident

Note that in the previous situations provided above, the test for independence or the test for homogeneity is merely a procedure. However, there are cases that arise where effects must depend on the skills of the researcher.

**Example 5.12:** (Exercises 10.42 and 10.43 in Text)

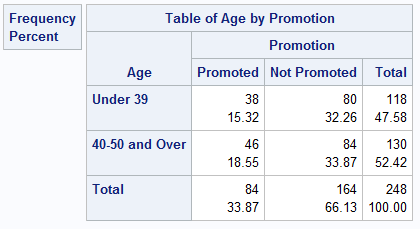
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Age** | | | |
| **Promotion** | **Under 30** | **30-39** | **40-49** | **50 & Over** |
| **Promoted** | 9 | 29 | 34 | 12 |
| **Not Promoted** | 41 | 39 | 46 | 38 |

 *The Fire Department in a large city is examining its promotion policy to assess if there is the potential for an age discrimination lawsuit. A random sample of 248 promotion decisions over the past 5 years yield the following information:*

When constructing a test, one finds that with the structure of the table provided, there is significant evidence that age does have an effect on the hiring practices of the fire department.

|  |  |  |
| --- | --- | --- |
|  | **Age** | |
| **Income** | **39 and Under** | **40 and Older** |
| **Promoted** | 38 | 46 |
| **Not Promoted** | 80 | 84 |

Now, when provided the following information to the fire department, their legal defense team provides their own statistical analysis, claiming there is no age discrimination. When asked to provide proof, they submit the following table and analysis provided below:



**Question:**

Which argument is correct?   
How would you refute the defense?

Notes:

* Suitable visuals can help with explaining effects from variables.
* When combining categories, be careful which categories you are collapsing. Make sure you have a reasonable explanation when you perform unnecessary collapsing, as you may mitigate effects.

**5.7 Nonparametric Procedures: Multinomial Test & Fisher’s Exact Test**

We will now explore the nonparametric procedures that can be used with qualitative data beyond a single proportion, focusing on the ***Exact Multinomial Test*** and the ***Exact Fisher Test***.

* Note that these tests are useful when cell counts are less than 5 and we cannot / do not wish to collapse categories. If the cell counts fit the criteria, it is more computationally sound to use Chi-Square Tests.
* Sometimes, other researchers relax the assumption, trading that assumption for those below:

1. All expected frequencies are 1 or greater.
2. At most 20% of the expected frequencies are less than 5.

The Exact Multinomial Test

We can summarize this type of data by reporting the count of the number of measurements in each category, as it is considered to be a **multinomial experiment** (like binomial, but with categories).   
(This test is our nonparametric version of the Chi Square Goodness of Fit Test.)

Unlike most inference tests, *there is no closed test statistic*. The computation of the p-value is unique in that ***we take the probability of this combination*** and then ***find all combinations with probabilities at most equal to our case &*** ***add them all up***.For small samples, the computation works well, but larger samples appear to draw in a lot of needed computing power, and so it is wise to simply use the Chi-Square tests when sample sizes requirements are met (or use a *Monte Carlo method* that can be seen in Bayesian coursework).

**Exact Multinomial Test Procedure (Multinomial Probabilities: One-Way Table)**

|  |  |
| --- | --- |
| Purpose: To perform a hypothesis test for the distribution of a single variable and   whether it fits the suggested proportions previously stated. | |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | |
| 1. Simple random sample from a multinomial distribution. | |
| ***Step 2: State the null and alternative hypotheses.*** | |
| (*The variable has the specified distribution of   probabilities.*) | (*The variable does not have the specified   distribution given.*) |
| ***Step 3: Compute the value of the test statistic.*** | |
| * In a rare case, there is no major test statistic to report, due to the construction of the p-value. | |
| ***Step 4: Determine the p-value.*** | |
| ***P*-value** | |
| ***Step 5: Based on p-value, make your conclusion.*** | |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* | |

Using R to Perform Exact Multinomial Test

* SAS unfortunately does not have an easy procedure for the exact multinomial test, but R does.
* We will use the **xmulti** procedure, using the **XNomial** package.

**library(XNomial)**

**xmulti(TABLE, p0, detail=3, safety = 1E9)$observedProb**

Fisher’s Exact Test

The second of the two nonparametric tests we will cover is what is known as the ***Fisher’s Exact Test***.

Originally developed for a 2x2 table, the Fisher’s Exact Test is constructed by taking the values found in our cells, and viewing it through the lens of what is called a ***multivariate hypergeometric distribution***.

No fear, like the exact multinomial test, *there is no closed test statistic* and we will rely on computers to do the dirty work. The computation of the p-value is unique in that ***we take the probability of this combination*** and then ***find all combinations with probabilities at most equal to our case &*** ***add them all up***.For small samples, the computation works well, but larger samples appear to draw in a lot of needed computing power, and so it is wise to simply use the Chi-Square tests when sample sizes requirements are met (or use a *Monte Carlo method* that can be seen in Bayesian coursework).

**Fisher’s Exact Test**

|  |  |
| --- | --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | |
| 1. Simple random samples from a multinomial distribution. 2. (*If testing for homogeneity*) The samples are independent. | |
| ***Step 2: State the null and alternative hypotheses.*** | |
|  |  |
| The two variables are independent.  The two variables are dependent. | The populations are homogenous.  The populations differ with respect to a given variable. |
| ***Step 3: Compute the value of the test statistic.*** | |
| * In a rare case, there is no major test statistic to report, due to the construction of the p-value. | |
| ***Step 4: Determine the p-value.*** | |
| ***P*-value** | |
| ***Step 5: Based on p-value, make your conclusion.*** | |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* | |

Using SAS to Perform Fisher’s Exact Test

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the CHISQ option as well as the EXACT CHISQ command.
* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VAR1\*VAR2** / CHISQ;

EXACT CHISQ;

**RUN**;

Using R to Perform Fisher’s Exact Test

We will use the **CrossTable** procedure, selecting **fisher = T**.

* We will use the numbers attained from our table, either provided in the problem or calculated by the **table** command in R.

**CrossTable(TABLE, expected=T, prop.r=T/F, prop.c=T/F, prop.t=T/F, prop.chisq=F,** **fisher = T)**

**Example 5.13:**   
At an local pharmacy in town, the store owner sells four types of ice cream: strawberry, chocolate, vanilla, and butterscotch. The owner wants to place his supply orders proportionate to the sales of each flavor. Five years ago, it was believed that roughly 25% of customers chose strawberry, 40% chose chocolate, 20% chose vanilla, and 15% chose butterscotch. To be certain flavor preferences hadn’t changed, the owner records the flavor selections of his customers during a day. The following data was collected below. Based on the information provided, is there evidence to suggest there is a difference in preference? Evaluate at the 10% significance level.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Color** | **Strawberry** | **Chocolate** | **Vanilla** | **Butterscotch** |
| **Percentage** | 25% | 40% | 20% | 15% |
| **Observed #** | 7 | 15 | 5 | 2 |

|  |
| --- |
| SAS Code: (Create Data Set for GSS)  **PROC** **FREQ** DATA=icecream;  WEIGHT number;  TABLES flavor / TestP=(**0.25**,**0.40**,**0.20**,**0.15**);  EXACT CHISQ;  **RUN**; |
| R Code: (Create Data Set for ice.cream)  **p0 <- c(0.25,0.40,0.20,0.15)**  **library(XNomial)**  **xmulti(ice.cream, p0, detail=3, safety = 1E9)$observedProb**  **xmulti(ice.cream, p0, detail=3, safety = 1E9)$observedChi** |

**Example 5.14:**

A clinical trial is conducted to compare two drug therapies for leukemia: P & PV. Twenty-one patients were assigned to drug P and forty-two patients to drug PV. The data is provided below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Success** | **Failure** | **Total** |
| **PV** | 38 | 4 | 42 |
| **P** | 14 | 7 | 21 |
| **Total** | 52 | 11 | 63 |

Test the null hypothesis that the proportion success for the PV therapy differs from that of the P therapy, using .

|  |
| --- |
| SAS Code: (Create Data Set for leukemia)  **PROC** **FREQ** DATA=icecream;  WEIGHT number;  TABLES flavor / TestP=(**0.25**,**0.40**,**0.20**,**0.15**);  EXACT CHISQ;  **RUN**; |
| R Code: (Create Data Set for leukemia)  **library(gmodels)**  **CrossTable(leukemia, prop.r=T, prop.c=T, prop.t=F, expected=T, prop.chisq=F, fisher=T)** |

**5.8 The Odds Ratio**

For the last lesson of this chapter, we will return to the concept of the 2×2 table and look at another important metric in qualitative analysis: ***The Odds Ratio***.

**Definition:**

The **odds** for an Event is defined as .

The ***odds*** of an event are useful to help determine the advantage (or *disadvantage*) of obtaining an event. For example, if , then , and so the odds of getting such a result is 1:1, or *equally likely*. If , then . And so it is 4 times more likely for the Event to occur than for it to not.

However, if an event had , then . If the odds of an event ends up being less than 1, we typically do not report than an event is “0.5 times more likely”, but instead *take the reciprocal* and report it as either “2 times less likely” or “the opposite is 2 times more likely”.

The nuances of talking about numbers, am I right?

This relationship between the *odds* and the *probability* can also be reversed. That is, .

The concept of the odds of an event is useful for a single proportion, but like much of statistics, we wish to use these metrics as a way of *comparison* to another group. In Section 5.5, we discussed how *sampling two proportions* is equivalent to a 2×2 table, and we can use both Chi-Square Tests as well as 2-sample   
Z-tests to perform inference.

Another closely-related inferential measure, one that is widely used in biomedical studies, is the **odds ratio.** As the name indicates, it is the ratio of the odds of an event (for example, contracting

a certain form of a disease) for one group (say, one with a family history) to the odds of the same event for another group (one that doesn’t). The odds ratio is usually defined using *conditional* *probabilities* but can be stated equally well in terms of joint probabilities (which we’ll do).

**Definition:**

The **odds ratio (OR)** is the ratio of the odds between two events/populations.

* Let . Then, it is defined as .

**Interpretation of the Odds Ratio:**

* If , this suggests that there is little difference between the odds of an event happening to Group 1 and the odds of an event happening to Group 2.
* If , then an event is times more likely to happen to Group 1   
  as it is to happen to Group 2.
* If , then an event is times less likely to happen to Group 1   
  as it is to happen to Group 2.

**Recall:**

|  |  |  |
| --- | --- | --- |
|  | **Group 1** | **Group 2** |
|  |  |  |
|  |  |  |

For a 2×2 table, sampling individuals total, the *observed counts* can be viewed as the table to our right, where is the observed count for row , column .

* It is through this construction that we can estimate the odds ratio, given our data.

**Definition:**

The **estimated** **odds ratio** is defined as .

Now, in more-advanced statistics courses, we would spend the time to perform inferential methods using the odds ratio (we will see a little of this in Section 9.6). One neat aspect of the estimated odds ratio is that *for large sample sizes ,* the ***natural logarithm of the estimated odds ratio*** is *approximately normal.*

Due to time restraints, we will just state the sampling distribution of and how to compute /interpret the confidence interval for the , but rely on our previous tests for inference.

**The Sampling Distribution of**

|  |  |
| --- | --- |
| For samples of size ***n***, | |
|  |  |
| 1. The sampling distribution of is approximately Normal for large *n*. | |

To obtain a confidence interval for the , we use the to get a confidence interval (using   
Z-scores), getting a confidence interval of the form , our lower and upper bounds, respectively.

|  |
| --- |
| **Asymmetric Confidence Interval for :** |

Since the exponential function is one-to-one (only one value for every value), we can get rid of the natural logarithm to get what we call an ***asymmetric confidence interval*** (one that isn’t symmetrical around the point estimate) :

No fear, SAS and R does all that computational stuff for you 😊.

**Notes:**

* An estimated odds ratio near 1 (or one with a confidence interval that contains 1) suggests that there is little difference between the odds of an event happening to Group 1 and the odds of an event happening to Group 2.
* If the hypothesis test for two proportions is rejected, that suggests that and vice-versa.
* If the hypothesis test for two proportions is not rejected, that suggests that the confidence interval for contains 1 and vice-versa.
* **BE CERTAIN** you have your variables in the correct order. While it won’t matter which variable is horizontal and which is vertical in your table, switching the order of the groups will give you the incorrect OR estimates and CIs.

Using SAS to Obtain Estimates and (Wald) Confidence Intervals for the Odds Ratio

* We will use the **PROC** **FREQ** procedure, and use the TABLES command. We will specify the CHISQ option as well as the RELRISK command.
* We will use EXACT CHISQ when necessary.
* When creating a dataset that uses tables instead of raw data, you will need to add the   
  WEIGHT command (between the DATA and TABLES commands).

**PROC** **FREQ** DATA=**DATASET**;

TABLES **VAR1\*VAR2** / CHISQ RELRISK ALPHA**=ALPHA\_LEVEL**;

**RUN**;

Using R to Estimates and (Wald) Confidence Intervals for the Odds Ratio

We will use the **oddsratio** procedure from the package **epitools**.

* We will use the table, either constructed manually or calculated by the **table** command in R.

**library(epitools)**

**oddsratio(TABLE, conf.level = CONFIDENCE\_LEVEL, method="wald")$measure**

**Example 5.15:**

Consider the fire department discrimination case at the end of Section 5.6. Recall that when the original data was provided to the fire department, their legal defense team provided their own statistical analysis, using a 22 table. Use the following table to provide an estimate for the odds ratio, along with a corresponding 95% confidence interval.

|  |  |  |
| --- | --- | --- |
|  | **Age** | |
| **Income** | **39 and Under** | **40 and Older** |
| **Promoted** | 38 | 46 |
| **Not Promoted** | 80 | 84 |

|  |
| --- |
| SAS Code: (Create Data Set for discrimination and format correctly.)  **PROC** **FREQ** DATA=Discrimination;  WEIGHT number;  TABLE age \* promotion /CHISQ RELRISK ALPHA=**0.05**;  FORMAT age age2fmt. promotion promotion2fmt.;  **RUN**; |
| R Code: (Create Data Set for fire2)  **library(epitools)**  **oddsratio(fire2, conf.level=0.95)$measure** |