Chapter 11

Chi-Square and ANOVA Tests

This chapter presents material on three more hypothesis tests. One is used to determine significant relationship between two qualitative variables, the second is used to determine if the sample data has a particular distribution, and the last is used to determine significant relationships between means of 3 or more samples.

11.1 Chi-Square Test for Independence

Remember, qualitative data is where you collect data on individuals that are categories or names. Then you would count how many of the individuals had particular qualities. An example is that there is a theory that there is a relationship between breastfeeding and autism. To determine if there is a relationship, researchers could collect the time period that a mother breastfed her child and if that child was diagnosed with autism. Then you would have a table containing this information. Now you want to know if each cell is independent of each other cell. Remember, independence says that one event does not affect another event. Here it means that having autism is independent of being breastfed. What you really want is to see if they are not independent. In other words, does one affect the other? If you were to do a hypothesis test, this is your alternative hypothesis and the null hypothesis is that they are independent. There is a hypothesis test for this and it is called the **Chi-Square Test for Independence**. Technically it should be called the Chi-Square Test for Dependence, but for historical reasons it is known as the test for independence. Just as with previous hypothesis tests, all the steps are the same except for the assumptions and the test statistic.

Hypothesis Test for Chi-Square Test

1. State the null and alternative hypotheses and the level of significance

H_o : the two variables are independent (this means that the one variable is not affected by the other)

H_a : the two variables are dependent (this means that the one variable is affected by the other)

Also, state your α level here.

2. State and check the assumptions for the hypothesis test

a. A random sample is taken.

b. Expected frequencies for each cell are greater than or equal to 5 (The expected frequencies, E, will be calculated later, and this assumption means $E \geq 5$).

3. Find the test statistic and p-value

Finding the test statistic involves several steps. First the data is collected and counted, and then it is organized into a table (in a table each entry is called a cell). These values are known as the observed frequencies, which the symbol for an observed frequency is O. Each table is made up of rows and columns. Then each row is totaled to give a row total and each column is totaled to give a column total.

The null hypothesis is that the variables are independent. Using the multiplication rule for independent events you can calculate the probability of being one value of the first variable, A, and one value of the second variable, B (the probability of a particular cell). Remember in a hypothesis test, you assume that is true, the two variables are assumed to be independent.

Now you want to find out how many individuals you expect to be in a certain cell. To find the expected frequencies, you just need to multiply the probability of that cell times the total number of individuals. Do not round the expected frequencies.

If the variables are independent the expected frequencies and the observed frequencies should be the same. The test statistic here will involve looking at the difference between the expected frequency and the observed frequency for each cell. Then you want to find the "total difference" of all of these differences. The larger the total, the smaller the chances that you could find that test statistic given that the assumption of independence is true. That means that the assumption of independence is not true. How do you find the test statistic? First find the differences between the observed and expected frequencies. Because some of these differences will be positive and some will be negative, you need to square these differences. These squares could be large just because the frequencies are large, you need to divide by the expected frequencies to scale them. Then finally add up all of these fractional values. This is the test statistic.

Test Statistic:

Using R: See example 11.1.1 for the process

4. Conclusion

This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value $< \alpha$, then reject H_o. If the p-value$\geq \alpha$, then fail to reject H_o

5. Interpretation

This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to support H_a, or you do not have enough evidence to support H_a.

11.1.1 Example: Hypothesis Test with Chi-Square Test

Is there a relationship between autism and breastfeeding? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what time period they breastfed their children. The data is in table #11.1.1 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to support that breastfeeding and autism are independent? Test at the 1% level.

Table #11.1.1: Autism Versus Breastfeeding

```
##      [,1] [,2] [,3] [,4]
## yes  241  198  164  215
## no    20   25   27   44
```

where [,1] represents number of mothers who did not breast feed, [,2] represents number of mothers who breast feed for less than 2 months, [,3] represents number of mothers who breast feed between 2 and 6 months. and [,4] represents the number of mothers who breast feed more than 6 months. Rows represent whether children has or doesn't have autism.

Solution:

1. State the null and alternative hypotheses and the level of significance

H_o: Breastfeeding and autism are independent

H_a: Breastfeeding and autism are dependent

2. State and check the assumptions for the hypothesis test

a. A random sample of breastfeeding time frames and autism incidence was taken. Check: this was stated in the problem.

b. Expected frequencies for each cell are greater than or equal to 5, $E \geq 5$. See step 3. All expected frequencies are more than 5.

3. Find the test statistic and p-value On R Studio, the command is

```
yes<-c(241, 198, 164, 215) #puts data into row 1
no<-c(20, 25, 27, 44) #puts data into row 2
table<-rbind(yes,no) #combines row 1 and row 2 to make a table
chisq.test(table) #calculates the test statistics and p-value
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 11.217, df = 3, p-value = 0.01061
```

```
chisq.test(table)$expected # shows all the expected frequencies
```

```
##          [,1]      [,2]      [,3]      [,4]
## yes 228.58458 195.30407 167.27837 226.83298
## no   32.41542  27.69593  23.72163  32.16702
```

The test statistic is 11.217 and the p-value is 0.01061.

4. Conclusion

Fail to reject H_o since the p-value is more than 0.01.

5. Interpretation

There is not enough evidence to support that breastfeeding and autism are dependent. This means that you cannot say whether a child is breastfed or not will indicate if that the child will be diagnosed with autism.

11.1.2 Homework

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. The number of people who survived the Titanic based on class and sex is in table #11.1.2 ("Encyclopedia Titanica," 2013). Is there enough evidence to show that the class and the sex of a person who survived the Titanic are independent? Test at the 5% level.

Table #11.1.2: Surviving the Titanic

```
##        [,1] [,2]
## first   134   59
## second   94   25
## third    80   58
```

Where [,1] represents female and [,2] represents male.

2. Researchers watched groups of dolphins off the coast of Ireland in 1998 to determine what activities the dolphins partake in at certain times of

the day ("Activities of dolphin," 2013). The numbers in table #11.1.3 represent the number of groups of dolphins that were partaking in an activity at certain times of days. Is there enough evidence to show that the activity and the time period are independent for dolphins? Test at the 1% level.

Table #11.1.3: Dolphin Activity

Activity	Period			Row Total
	Morning	Noon	Afternoon	Evening	
Travel	6	6	14	13	39
Feed	28	4	0	56	88
Social	38	5	9	10	62
Column Total	72	15	23	79	189

3. Is there a relationship between autism and what an infant is fed? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what they fed their infant. The data is in table #11.1.4 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to show that that what an infant is fed and autism are independent? Breast-feeding (BF), Formula with DHA/ARA (For with), and Formula without DHA/ARA (Form without)Test at the 1% level.

Table #11.1.4: Autism Versus Breastfeeding

Autism	Feeding			Row Total
	BF	Form with	Form out	
Yes	12	39	65	116
No	6	22	10	38
Column Total	18	61	75	154

4. A person's educational attainment and age group was collected by the U.S. Census Bureau in 1984 to see if age group and educational attainment are related. The counts in thousands are in table #11.1.5 ("Education by age," 2013). Do the data show that educational attainment and age are independent? Test at the 5% level.

Table #11.1.5: Educational Attainment and Age Group

Education	Age Group					Row Total
	25-34	35-44	45-54	55-64	>64	
Did not complete HS	5416	5030	5777	7606	13746	37575

Education	Age Group	Row Total				
Competed HS	16431	1855	9435	8795	7558	44074
College 1-3 years	8555	5576	3124	2524	2503	22282
College 4 or more years	9771	7596	3904	3109	2483	26863
Column Total	40173	20057	22240	22034	26290	130794

5. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important good grades were to them (1 very important and 4 least important). The data is in table #11.1.6 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of grades are independent? Test at the 5% level.

Table #11.1.6: Personal Goal and Importance of Grades

Goal	Grades Importance Rating	Row Total			
	1	2	3	4	
Grades	70	66	55	56	247
Popular	14	33	45	49	141
Sports	10	24	33	23	90
Column Total	94	123	133	128	478

6. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important being good at sports were to them (1 very important and 4 least important). The data is in table #11.1.7 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of sports are independent? Test at the 5% level.

Table #11.1.7: Personal Goal and Importance of Sports

Goal	Sports Importance Rating	Row Total			
	1	2	3	4	
Grades	83	81	55	28	247
Popular	32	49	43	17	141
Sports	50	24	14	2	90
Column Total	165	154	112	47	478

7. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important having good looks were to them (1 very important and 4 least important). The data is in table #11.1.8 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of looks are independent? Test at the 5% level.

Table #11.1.8: Personal Goal and Importance of Looks

Goal	Looks Importance Rating	Row Total			
	1	2	3	4	
Grades	80	66	66	35	247
Popular	81	30	18	12	141
Sports	24	30	17	19	90
Column Total	185	126	101	66	478

8. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important having money were to them (1 very important and 4 least important). The data is in table #11.1.9 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of money are independent? Test at the 5% level.

Table #11.1.9: Personal Goal and Importance of Money

Goal	Money Importance Rating	Row Total			
	1	2	3	4	
Grades	14	34	71	128	247
Popular	14	29	35	63	141
Sports	6	12	26	46	90
Column Total	34	75	132	237	478

11.2 Chi-Square Goodness of Fit

In probability, you calculated probabilities using both experimental and theoretical methods. There are times when it is important to determine how well the experimental values match the theoretical values. An example of this is if you wish to verify if a die is fair. To determine if observed values fit the expected values, you want to see if the difference between observed values and expected values is large enough to say that the test statistic is unlikely to happen if you assume that the observed values fit the expected values. The test statistic in this case is also the chi-square. The process is the same as for the chi-square test for independence.

Hypothesis Test for Goodness of Fit Test

1. State the null and alternative hypotheses and the level of significance

H_o : The data are consistent with a specific distribution

H_a : The data are not consistent with a specific distribution

Also, state your α level here.

2. State and check the assumptions for the hypothesis test

a. A random sample is taken.

b. Expected frequencies for each cell are greater than or equal to 5 (The expected frequencies, E, will be calculated later).

3. Find the test statistic and p-value

Using R Studio see example 11.2.1

4. Conclusion

This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value $< \alpha$, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_0

5. Interpretation

This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to support H_a, or you do not have enough evidence to support H_a.

11.2.1 Example: Goodness of Fit Test

Suppose you have a die that you are curious if it is fair or not. If it is fair then the proportion for each value should be the same. You need to find the observed frequencies and to accomplish this you roll the die 500 times and count how often each side comes up. The data is in table #11.2.1. Do the data show that the die is fair? Test at the 5% level.

Table #11.2.1: Observed Frequencies of Die for sides 1 through 6

```
## [1] 78 87 87 76 85 87
```

Solution:

1. State the null and alternative hypotheses and the level of significance

H_o : The observed frequencies are consistent with the distribution for fair die (the die is fair)

H_a : The observed frequencies are not consistent with the distribution for fair die (the die is not fair)

$\alpha = 0.05$

2. State and check the assumptions for the hypothesis test

a. A random sample is taken. check: This is true since each throw of a die is a random event.

b. Expected frequencies for each cell are greater than or equal to 5. See step 3.

3. Find the test statistic and p-value

On R Studio, this would be

```
fair_die<-c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
observed<-c(78, 87, 87, 76, 85, 87)
chisq.test(observed, p=fair_die)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 1.504, df = 5, p-value = 0.9126
```

```
chisq.test(observed, p=fair_die)$expected
```

```
## [1] 83.33333 83.33333 83.33333 83.33333 83.33333 83.33333
```

Test Statistic: The test statistic is 1.504. The p-value is 0.9126.

4. Conclusion

Fail to reject H_o since the p-value is greater than 0.05.

5. Interpretation

There is not enough evidence to support that the die is not consistent with the distribution for a fair die. There is not enough evidence to support that the die is not fair.

11.2.2 Homework

** In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.**

1. According to the M&M candy company, the expected proportion can be found in Table #11.2.2. In addition, the table contains the number of M&M's of each color that were found in a case of candy (Madison, 2013). At the 5% level, do the observed frequencies support the claim of M&M?

Table #11.2.2: M&M Observed and Proportions

	Blue	Brown	Green	Orange	Red	Yellow	Total
Observed Frequencies	481	371	483	544	372	369	2620
Expected Proportion	0.24	0.13	0.16	0.20	0.13	0.14	

2. Eyeglassomatic manufactures eyeglasses for different retailers. They test

to see how many defective lenses they made the time period of January 1 to March 31. Table #11.2.3 gives the defect and the number of defects.

Table #11.2.3: Number of Defective Lenses

Defect type	Number of defects
Scratch	5865
Right shaped – small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape – big	1105
Lost in lab	976
Spots/bubble – intern	976

Do the data support the notion that each defect type occurs in the same proportion? Test at the 5% level.

3. On occasion, medical studies need to model the proportion of the population that has a disease and compare that to observed frequencies of the disease actually occurring. Suppose the end-stage renal failure in southwest Wales was collected for different age groups. Do the data in table 11.2.4 show that the observed frequencies are in agreement with proportion of people in each age group (Boyle, Flowerdew & Williams, 1997)? Test at the 1% level.

   Table #11.2.4: Renal Failure Frequencies

Age Group	16-29	30-44	45-59	60-75	75+	Total
Observed Frequency	32	66	132	218	91	539
Expected Proportion	0.23	0.25	0.22	0.21	0.09	

4. In Africa in 2011, the number of deaths of a female from cardiovascular disease for different age groups are in table #11.2.5 ("Global health observatory," 2013). In addition, the proportion of deaths of females from all causes for the same age groups are also in table #11.2.5. Do the data show that the death from cardiovascular disease are in the same proportion as all deaths for the different age groups? Test at the 5% level.

   Table #11.2.5: Deaths of Females for Different Age Groups

Age	5-14	15-29	30-49	50-69	Total
Cardiovascular Frequency	8	16	56	433	513
All Cause Proportion	0.10	0.12	0.26	0.52	

5. In Australia in 1995, there was a question of whether indigenous people are more likely to die in prison than non-indigenous people. To figure out, the data in table 11.2.6 was collected. ("Aboriginal deaths in," 2013). Do the data show that indigenous people die in the same proportion as non-indigenous people? Test at the 1% level.

   Table #11.2.6: Death of Prisoners

	Pris	oner Dies Pris	oner Did Not Die Tota	l
Indigenous Prisoner Frequency		17	2890	2907
Frequency of Non-Indigenous Prisoner		42	14459	14501

6. A project conducted by the Australian Federal Office of Road Safety asked people many questions about their cars. One question was the reason that a person chooses a given car, and that data is in table #11.2.7 ("Car preferences," 2013).

Table #11.2.7: Reason for Choosing a Car

Safety	Reliability	Cost	Performance	Comfort	Looks
84	62	46	34	47	27

Do the data show that the frequencies observed substantiate the claim that the reasons for choosing a car are equally likely? Test at the 5% level.

11.3 Analysis of Variance (ANOVA)

There are times where you want to compare three or more population means. One idea is to just test different combinations of two means. The problem with that is that your chance for a type I error increases. Instead you need a process for analyzing all of them at the same time. This process is known as **analysis of variance (ANOVA)**. The test statistic for the ANOVA is fairly complicated, you will want to use technology to find the test statistic and p-value. The test statistic is distributed as an F-distribution, which is skewed right and depends on degrees of freedom. Since you will use technology to find these, the distribution and the test statistic will not be presented. Remember, all hypothesis tests are the same process. Note that to obtain a statistically significant result there need

only be a difference between any two of the k means.

Before conducting the hypothesis test, it is helpful to look at the means and standard deviations for each data set. If the sample means with consideration of the sample standard deviations are different, it may mean that some of the population means are different. However, do realize that if they are different, it doesn't provide enough evidence to show the population means are different. Calculating the sample statistics just gives you an idea that conducting the hypothesis test is a good idea.

Hypothesis test using ANOVA to compare *k* means

1. State the random variables and the parameters in words
2. State the null and alternative hypotheses and the level of significance

H_o : all the means are the same

H_a : at least two of the means are different

Also, state your level here.

3. State and check the assumptions for the hypothesis test

a. A random sample of size is taken from each population.

b. All the samples are independent of each other.

c. Each population is normally distributed. The ANOVA test is fairly robust to the assumption especially if the sample sizes are fairly close to each other. Unless the populations are really not normally distributed and the sample sizes are close to each other, then this is a loose assumption.

d. The population variances are all equal. If the sample sizes are close to each other, then this is a loose assumption.

4. Find the test statistic and p-value

The test statistic is F. To find the test statistic, use technology such R Studio.

The test statistic, F, is distributed as an F-distribution, where both degrees of freedom are needed in this distribution. The p-value is also calculated R Studio.

5. Conclusion

This is where you write reject H_O or fail to reject H_O. The rule is: if the p-value $< \alpha$, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o

6. Interpretation

This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to support H_a, or you do not have enough evidence to support H_a.

If you do in fact reject H_o, then you know that at least two of the means are different. The next question you might ask is which are different? You can

look at the sample means, but realize that these only give a preliminary result. To actually determine which means are different, you need to conduct other tests. Some of these tests are the range test, multiple comparison tests, Duncan test, Student-Newman-Keuls test, Tukey test, Scheffé test, Dunnett test, least significant different test, and the Bonferroni test. There is no consensus on which test to use.

11.3.1 Example: Hypothesis Test Involving Several Means

Cancer is a terrible disease. Surviving may depend on the type of cancer the person has. To see if the mean survival time for several types of cancer are different, data was collected on the survival time in days of patients with one of these cancer in advanced stage. The data is in table #11.3.1 ("Cancer survival story," 2013). (Please realize that this data is from 1978. There have been many advances in cancer treatment, so do not use this data as an indication of survival rates from these cancers.) Do the data indicate that at least two of the mean survival time for these types of cancer are not all equal? Test at the 1% level.

Table #11.3.1: Survival Times in Days of Five Cancer Types

```
Cancer <- read.csv(
  "https://krkozak.github.io/MAT160/cancer.csv")
head(Cancer)
```

```
##   survival   organ
## 1      124 Stomach
## 2       42 Stomach
## 3       25 Stomach
## 4       45 Stomach
## 5      412 Stomach
## 6       51 Stomach
```

Code book for data frame Cancer

Description Survival time for several types of cancer was collected.

This data frame contains the following columns:

survival: survival times (months)

organ: the organ that the cancer is in

Source Cancer survival story. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Stories/CancerSurvival.html

References http://lib.stat.cmu.edu/DASL

Solution:

1. State the random variables and the parameters in words

x_1 : survival time of patient with Stomach cancer x_2 : survival time of patient with Bronchus (lung) cancer x_3 : survival time of patient with Colon cancer x_4 : survival time of patient with Ovarian cancer x_5 : survival time of patient with Breast cancer

μ_1 : mean survival time of patient with Stomach cancer μ_2 : mean survival time of patient with Bronchus (lung) cancer μ_3 : mean survival time of patient with Colon cancer μ_4 : mean survival time of patient with Ovarian cancer μ_5 : mean survival time of patient with Brest cancer

Now before conducting the hypothesis test, look at the means and standard deviations. There appears to be a difference between at least two of the means, but realize that the standard deviations are very different. The difference you see may not be significant.

Notice the sample sizes are not the same.

2. State the null and alternative hypotheses and the level of significance

H_o : all the means are equal

H_a : some of the means are different

$\alpha = 0.01$

3. State and check the assumptions for the hypothesis test

a. A random sample of 13 survival times from stomach cancer was taken. A random sample of 17 survival times from bronchus cancer was taken. A random sample of 17 survival times from colon cancer was taken. A random sample of 6 survival times from ovarian cancer was taken. A random sample of 11 survival times from breast cancer was taken. check: These statements may not be true. This information was not shared as to whether the samples were random or not but it may be safe to assume that.

b. The samples are all independent. check: Since the individuals have different cancers, then the samples are independent.

c. Population of all survival times from stomach cancer is normally distributed. Population of all survival times from bronchus cancer is normally distributed. Population of all survival times from colon cancer is normally distributed. Population of all survival times from ovarian cancer is normally distributed. Population of all survival times from breast cancer is normally distributed. check: Looking at the density plots and normal quantile plots for each sample, it appears that none of the populations are normally distributed. The sample sizes are somewhat different for the problem. This assumption may not be true.

d. The population variances are all equal. check: The sample standard deviations are approximately 346.3, 209.9, 427.2, 1098.6, and 1239.0 respectively. This assumption does not appear to be met, since the sample

standard deviations are very different. The sample sizes are somewhat different for the problem. This assumption may not be true.

4. Find the test statistic and p-value

To find the test statistic and p-value on R Studio, the commands would be:

```
results=aov(survival~organ, data=Cancer)
summary(results)
```

```
##              Df   Sum Sq Mean Sq F value   Pr(>F)
## organ         4 11535761 2883940   6.433 0.000229 ***
## Residuals    59 26448144  448274
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test statistic is F = 6.433 and the p-value = 0.000229.

5. Conclusion

Reject H_o : since the p-value is less than 0.01.

6. Interpretation

There is enough evidence to support that at least two of the mean survival times from different cancers are not equal.

By examination of the means, it appears that the mean survival time for breast cancer is different from the mean survival times for both stomach and bronchus cancers. It may also be different for the mean survival time for colon cancer. The others may not be different enough to actually say for sure.

11.3.2 Homework

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. Cuckoo birds are in the habit of laying their eggs in other birds' nest. The other birds adopt and hatch the eggs. The lengths (in cm) of cuckoo birds' eggs in the other species nests were measured and are in table #11.3.2 ("Cuckoo eggs in," 2013). Do the data show that the mean length of cuckoo bird's eggs is not all the same when put into different nests? Test at the 5% level.

Table #11.3.2: Lengths of Cuckoo Bird Eggs in Different Species Nests

```
Eggs <- read.csv(
  "https://krkozak.github.io/MAT160/Birds_eggs.csv")
head(Eggs)
```

```
##   length   bird
## 1  19.65 Meadow
## 2  20.05 Meadow
## 3  20.65 Meadow
## 4  20.85 Meadow
## 5  21.65 Meadow
## 6  21.65 Meadow
```

Code book for data frame Eggs

Description Cuckoo birds are in the habit of laying their eggs in other birds' nest. The other birds adopt and hatch the eggs. The lengths (in cm) of cuckoo birds' eggs in the other species nests were measured

This data frame contains the following columns:

length: length of cuckoo bird's eggs in other species nets (cm)

bird: bids where eggs were found in their nests. The birds are Meadow Pipit, Tree Pipit, Hedge Sparrow, Robin, Pied Wagtail, and Wren

Source Cuckoo eggs in nest of other birds. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Stories/cuckoo.html

References SOCR Home page: http://www.socr.ucla.edu

2. Levi-Strauss Co manufactures clothing. The quality control department measures weekly values of different suppliers for the percentage difference of waste between the layout on the computer and the actual waste when the clothing is made (called run-up). The data is in table #11.3.3, ("Waste run up," 2013). Do the data show that there is a difference between some of the suppliers? Test at the 1% level.

Table #11.3.3: Run-ups for Different Plants Making Levi Strauss Clothing

```
Levi <- read.csv(
  "https://krkozak.github.io/MAT160/Levi_jeans.csv")
head(Levi)
```

```
##   run_up   plant
## 1    1.2 Plant_1
## 2   10.1 Plant_1
## 3   -2.0 Plant_1
## 4    1.5 Plant_1
## 5   -3.0 Plant_1
## 6   -0.7 Plant_1
```

Code book for data frame Levi

Description Levi-Strauss Co manufactures clothing. The quality control department measures weekly values of different suppliers for the percentage differ-

ence of waste between the layout on the computer and the actual waste when the clothing is made (called run-up).

This data frame contains the following columns:

run_up: percentage difference of waste between the layout on the computer and the actual waste when the clothing is made. There are some negative values because sometimes the supplier is able to layout the pattern better than the computer

plant: Which suppliers

Source Waste run up. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Stories/wasterunup.html

References http://lib.stat.cmu.edu/DASL

3. Several magazines were grouped into three categories based on what level of education of their readers the magazines are geared towards: high, medium, or low level. Then random samples of the magazines were selected to determine the number of three-plus-syllable words were in the advertising copy, and the data is in table #11.3.4 ("Magazine ads readability," 2013). Is there enough evidence to show that the mean number of three-plus-syllable words in advertising copy is different for at least two of the education levels? Test at the 5% level.

Table #11.3.4: Number of Three Plus Syllable Words in Advertising Copy

```
Advertising <- read.csv(
  "https://krkozak.github.io/MAT160/three_syllable_words.csv")
head(Advertising)
```

```
##   number education
## 1     34      High
## 2     21      High
## 3     37      High
## 4     31      High
## 5     10      High
## 6     24      High
```

Code book for data frame Advertising

Description Several magazines were grouped into three categories based on what level of education of their readers the magazines are geared towards: high, medium, or low level. Then random samples of the magazines were selected to determine the number of three-plus-syllable words were in the advertising copy

This data frame contains the following columns:

number: number of three=plus-syllable words in advertising copy

education: level of education the magazine is geared towards: high, medium, or low

Source Magazine ads readability. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/magadsdat.html

References http://lib.stat.cmu.edu/DASL

4. A study was undertaken to see how accurate food labeling for calories on food that is considered reduced calorie. The group measured the amount of calories for each item of food and then found the percent difference between measured and labeled food. The group also looked at food that was nationally advertised, regionally distributed, or locally prepared. The data is in table #11.3.5 ("Calories datafile," 2013). Do the data indicate that at least two of the mean percent differences between the three groups are different? Test at the 5% level.

Table #11.3.5: Percent Differences Between Measured and Labeled Food

```
Food <- read.csv(
  "https://krkozak.github.io/MAT160/Food_calories_percent_diff.csv")
head(Food)
```

```
##   percent_diff     food
## 1            2 national
## 2          -28 national
## 3           -6 national
## 4            8 national
## 5            6 national
## 6           -1 national
```

Code book for data frame Food

Description A study was undertaken to see how accurate food labeling for calories on food that is considered reduced calorie. The group measured the amount of calories for each item of food and then found the percent difference between measured and labeled food. The group also looked at food that was nationally advertised, regionally distributed, or locally prepared.

This data frame contains the following columns:

percent_diff: percent difference between the number of calories that are measured in the food and the amount that is labeled on the food.

food: Where the food is created: nationally advertised, regionally distributed, or locally prepared.

Source Calories datafile. (2013, December 07). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/Calories.html

References http://lib.stat.cmu.edu/DASL

5. The amount of sodium (in mg) in different types of hot dogs is in table #11.3.6 ("Hot dogs story," 2013). Is there sufficient evidence to show that the mean amount of sodium in the types of hot dogs are not all equal? Test at the 5% level.

Table #11.3.6: Amount of Sodium (in mg) in Beef, Meat, and Poultry Hotdogs

```
Hotdog_complete <-read.csv(
  "https://krkozak.github.io/MAT160/Hot_Dog_all_Dataset.csv")
head(Hotdog_complete)
```

```
##   type calories sodium
## 1 Beef      186    495
## 2 Beef      181    477
## 3 Beef      176    425
## 4 Beef      149    322
## 5 Beef      184    482
## 6 Beef      190    587
```

Code book for data frame Hotdog

Description Results of a laboratory analysis of calories and sodium content of major hot dog brands. Researchers for Consumer Reports analyzed three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat).

This data frame contains the following columns:

type: Type of hot dog (beef or poultry or meat (mostly pork and beef but up to 15% poultry meat))

calories: Calories per hot dog

sodium: Milligrams of sodium per hot dog

Source SOCR 012708 id data hot dogs. (2013, November 13). Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_012708_ID_Data_HotDogs

References SOCR Home page: http://www.socr.ucla.edu

Data Source:

Aboriginal deaths in custody. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/custody.html

Activities of dolphin groups. (2013, September 26). Retrieved from http://www.statsci.org/data/general/dolpacti.html

Boyle, P., Flowerdew, R., & Williams, A. (1997). Evaluating the goodness of fit in models of sparse medical data: A simulation approach. *International Journal of Epidemiology*, *26*(3), 651-656. Retrieved from http://ije.oxfordjournals.org/content/26/3/651.full.pdf html

Car preferences. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/carprefs.html

Education by age datafile. (2013, December 05). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/Educationbyage.html

Encyclopedia Titanica. (2013, November 09). Retrieved from http://www.encyclopedia-titanica.org/

Global health observatory data respository. (2013, October 09). Retrieved from http://apps.who.int/gho/athena/data/download.xsl?format=xml&target=GHO/MORT/_400&profile=excel&filter=AGEGROUP:YEARS05-14;AGEGROUP:YEARS15-29;AGEGROUP:YEARS30-49;AGEGROUP:YEARS50-69;AGEGROUP:YEARS70 ;MGHEREG:REG6_AFR;GHECAUSES:*;SEX:*

Madison, J. (2013, October 15). *M&M's color distribution analysis*. Retrieved from http://joshmadison.com/2007/12/02/mms-color-distribution-analysis/

Popular kids datafile. (2013, December 05). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html

Schultz, S. T., Klonoff-Cohen, H. S., Wingard, D. L., Askhoomoff, N. A., Macera, C. A., Ji, M., & Bacher, C. (2006). Breastfeeding, infant formula supplementation, and autistic disorder: the results of a parent survey. *International Breastfeeding Journal*, *1*(16), doi: 10.1186/1746-4358-1-16

Cancer survival story. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Stories/CancerSurvival.html

Cuckoo eggs in nest of other birds. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Stories/cuckoo.html

Waste run up. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Stories/wasterunup.html

Magazine ads readability. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/magadsdat.html

Calories datafile. (2013, December 07). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/Calories.html

SOCR 012708 id data hot dogs. (2013, November 13). Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_012708_ID_Data_HotDogs