MichaelY_HW6_Inference_Categorical

Michael Y.

March 31st, 2019

```
###setwd("c:/Users/Michael/DROPBOX/priv/CUNY/MSDS/201902-Spring/DATA606-Jason/Homework")
```

Homework - Chapter 6 - Inference for categorical data (pp.274-330)

Exercises: 6.6, 6.12, 6.20, 6.28, 6.44, 6.48 (pp.312-330)

Datasets:

6.6 - scotus_healthcare

6.12 - gss2010

6.20 - leg_mari

#.

Exercise 6.6 2010 Healthcare Law.

On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional.

A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.

```r
# look at the exact data set referenced
data("scotus_healthcare")
summary(scotus_healthcare)
```

```
##   response
##  agree:466
##  other:546
```

```r
# extract the counts from the data
p66_agree <- table(scotus_healthcare)[1]
p66_disagree <- table(scotus_healthcare)[2]
p66_total <- dim(scotus_healthcare)[1]

# What is the exact percentage of subjects who do agree? (text rounds to 46 percent)
p66_percent_agree <- as.numeric(p66_agree / p66_total)
p66_percent_agree
```

```
## [1] 0.46047431
```

```
# what is the exact percent of subjects who do not agree? (text rounds to 54 percent)
p66_percent_disagree <- as.numeric(p66_disagree / p66_total)
p66_percent_disagree
```

```
## [1] 0.53952569
```

```
# What is the standard deviation on the actual data?
p66_stdev <- as.numeric(sqrt( p66_percent_agree * p66_percent_disagree / p66_total))
p66_stdev
```

```
## [1] 0.015668179
```

```
# what is the Z-score at 95 percent confidence?
p66_Z_95 <- qnorm(0.975) # 2-tailed

# What is the margin of error at 95 percent confidence (text says 3 percent)
p66_ME_95 <- p66_Z_95 * p66_stdev
p66_ME_95
```

```
## [1] 0.030709066
```

(a) We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

No, we are ***100% confident*** that 46% of Americans ***in this sample*** support the decision of the US Supreme Court on the 2010 healthcare law.

(b) We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

Yes, this is correct.

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

Yes, this is correct.

(d) The margin of error at a 90% confidence level would be higher than 3%.

No – the margin of error would be **smaller** than 3% at the reduced confidence level:

$$MarginOfError = StandardError * Zscore$$

```
# what is the Z-score for 90 percent confidence
p66d_Z_90 <- qnorm(0.95) # 2-tailed
p66d_Z_90
```

```
## [1] 1.6448536
```

```
# percentage reduction in ME at lower confidence
p66d_ME_reduction <- p66d_Z_90 / p66_Z_95
p66d_ME_reduction
```

```
## [1] 0.83922646
```

```
# compute the margin of error at 90 percent confidence
p66_ME_90 <- p66d_Z_90 * p66_stdev
p66_ME_90
```

```
## [1] 0.02577186
```

At a 95 percent confidence level, we are using a Zscore of 1.95996398.

At a 90 percent confidence level, we would be using a Zscore of 1.64485363 , which is **smaller**.

Thus, the margin of error at 90% confidence would be **reduced** to 0.83922646 percent of the margin of error at 95% confidence, i.e., the margin of error would become 0.02577186 .

#·#

Exercise 6.12 Legalization of marijuana, Part I.

The 2010 General Social Survey asked 1,259 US residents:

“Do you think the use of marijuana should be made legal, or not?”

48% of the respondents said it should be made legal.

```
# load up the data file
p612_gss2010      <- read.csv("gss2010.csv")

# select the column which contains the responses for "should marijuana ever be legalized"
p612_grass        <- p612_gss2010$grass

# summarize the responses
p612_grass_survey <- summary(p612_grass)
p612_grass_survey
```

```
##     LEGAL NOT LEGAL      NA's
##       603       656       785
```

```
# capture the affirmative responses
p612_legal        <- p612_grass_survey[1]
p612_legal
```

```
## LEGAL
##   603
```

```
# capture the negative responses
p612_notlegal     <- p612_grass_survey[2]
p612_notlegal
```

```
## NOT LEGAL
##       656
```

```r
# count the NAs
p612_nonresponse  <- p612_grass_survey[3]
p612_nonresponse
```

```
## NA's
##  785
```

```r
# count the total survey size (including NAs)
p612_surveysize   <- length(p612_grass)
p612_surveysize
```

```
## [1] 2044
```

```r
# count the number of responses (excluding NAs)
p612_responses    <- sum(!is.na(p612_grass))
p612_responses
```

```
## [1] 1259
```

```r
# get the percentage of affirmative responses
p612_pct_yes      <- p612_legal     / p612_responses
p612_pct_yes
```

```
##      LEGAL
## 0.47895155
```

```r
# get the percentage of negative responses
p612_pct_no       <- p612_notlegal  / p612_responses
p612_pct_no
```

```
##  NOT LEGAL
## 0.52104845
```

(a) Is 48% a sample statistic or a population parameter? Explain.

It is a sample statistic, based on 603 responses supporting legalization out of a total of 1259 responses (excluding non-respondents). (It is noteworthy that there were also 785 "NAs"" out of 2044 total subjects in the overall survey. One wonders whether there is any pattern among those participants in the overall study who did not respond to this question, and whether their reponses, if made, may have differed from the results.)

(b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

```r
# calculate the standard error
p612_stderror <- as.numeric( sqrt(p612_pct_yes * p612_pct_no
                                  /
                                  p612_responses)
                             )
p612_stderror
```

```
## [1] 0.014079006
```

```r
# get the Z-score for a 95% confidence interval (2-tailed):
p612_Z95 <- qnorm(p = 0.975)
p612_Z95
```

```
## [1] 1.959964
```

```
# get the margin of error
p612_margin_of_error <- p612_Z95 * p612_stderror
p612_margin_of_error
```

```
## [1] 0.027594344
```

```
# get the lower_CI
p612_lower_CI <- p612_pct_yes - p612_margin_of_error
p612_lower_CI
```

```
##      LEGAL
## 0.45135721
```

```
# get the upper_CI
p612_upper_CI <- p612_pct_yes + p612_margin_of_error
p612_upper_CI
```

```
##      LEGAL
## 0.50654589
```

```
p612_CI <- c(p612_lower_CI,p612_upper_CI)
p612_CI
```

```
##      LEGAL      LEGAL
## 0.45135721 0.50654589
```

$$StandardError = \sqrt{\frac{p(1-p)}{SampleSize}}$$

$$MarginOfError = StandardError * Zscore$$

The 95% Confidence Interval for the proportion of US residents who think Marijuana should be made legal is

(0.45135721, 0.50654589) .

(c) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

By the central limit theorem, we can assume that the normal distribution is valid for the standard error of the proportion because we have more than 10 "successes" and 10 "failures", and the survey is drawn from a random cross-section of the population, as it is less than 10 percent of all Americans. Therefore the conditions for inference are satisfied as the sample is "nearly normal."

(d) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

Because the confidence interval (narrowly) includes 50 percent, there is a chance that the true population parameter may be (very) slightly above half. However, to accept the above claim,

we would need for the entire confidence interval to be above 50%. Therefore, the news piece's statement is not justified.

However, based upon the above confidence interval, it would be appropriate to assert that "Half of Americans think Marijuana could be legalized."

If we are trying to perform a hypothesis test to determine the veracity of the above statement, we would actually do a slightly different calculation.

Our Null Hypothesis would be $H_0: \quad p = 0.5$ and our Alternative Hypothesis would be $H_A: \quad p > 0.5$. Checking this would require performing a one-tailed test rather than generating a confidence interval using the above 2-tailed methodology.

We would have a slightly different standard error:

$$StandardError = \sqrt{\frac{p_0(1-p_0)}{SampleSize}} = \sqrt{\frac{0.5(1-0.5)}{1259}} = 0.014091497$$

```
# under the null hypothesis, p0 = 0.5
p612d_p0 <- 0.5
# calculate the standard error
p612d_stderror <- as.numeric( sqrt(p612d_p0 * (1 - p612d_p0)
                                   /
                                   p612_responses)
                              )
p612d_stderror
```

```
## [1] 0.014091497
```

```
# compute the Z-score
p612d_Z <- as.numeric( (p612_pct_yes - p612d_p0) / p612d_stderror)
p612d_Z
```

```
## [1] -1.4936987
```

```
# compute the critical Z-value for a one-tailed test
p612d_critZ <- qnorm(p=0.95)
p612d_critZ
```

```
## [1] 1.6448536
```

```
# compute the corresponding p-value for the Z-score:
p612d_pval <- pnorm(-p612d_Z)   # note the negative sign
p612d_pval
```

```
## [1] 0.93237281
```

```
# compute the required value for the point estimate to pass the one-sided test
p612d_required_proportion <- p612d_p0 + p612d_stderror * p612d_critZ
p612d_required_proportion
```

```
## [1] 0.52317845
```

```
p612d_required_affirmatives <- ceiling(p612d_required_proportion * p612_responses)
p612d_required_affirmatives
```

```
## [1] 659
```

and we would compute the Z-score as follows:

$$Z = \frac{pointestimate - nullvalue}{SE} = \frac{0.47895155 - 0.5}{1259} = -1.4936987$$

Based upon the critical value, we would need to have a point estimate which would translate to a Z-score **greater than** 1.64485363 in order to conclude reject the null hypothesis and accept the alternative. The point estimate that we have yields z Z-score which is **not even the correct sign** – it is **negative** which means that the corresponding p-value is 0.93237281 (rather than 0.06762719 .) This means we are quite far from rejecting the null hypothesis in favor of the alternative.

Indeed, in order to do so, we would have needed to have 659 affirmative responses in favor of legalization rather than the 603 that we actually have.

#·#

Exercise 6.20 Legalize Marijuana, Part II.

As discussed in Exercise 6.12, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal.

If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

$$Z\sqrt{\frac{p(1-p)}{n}} \leq \quad 0.02 \quad \Rightarrow \quad Z^2\left(\frac{p(1-p)}{n}\right) \quad \leq \quad (0.02)^2 \quad \Rightarrow \quad n \quad \geq \quad Z^2\left(\frac{p(1-p)}{(0.02)^2}\right)$$

```
p620_required_ME <- 0.02
p620_p_estimate <- p612_pct_yes
p620_Z <- qnorm(p = 0.975)

p620_required_n <- as.numeric(p620_p_estimate * (1 - p620_p_estimate) * (p620_Z / p620_required_ME )^2)
p620_required_n
```

```
## [1] 2396.657
```

```
p620_required_n <- ceiling(p620_required_n)
p620_required_n
```

```
## [1] 2397
```

The required sample size in order to obtain a margin of error within 2 percent at 95% confidence is *2397* .

#·#

Exercise 6.28 Sleep deprivation, CA vs. OR, Part I.

According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents.

These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.

$$SE_{\hat{p_1}-\hat{p_1}} = \sqrt{SE^2_{\hat{p_1}} + SE^2_{\hat{p_2}}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

```
## get CA SE
p628_CA_prop <- 0.080
p628_CA_n    <- 11545
p628_CA_SE2  <- p628_CA_prop * (1-p628_CA_prop) / p628_CA_n
p628_CA_SE   <- sqrt(p628_CA_SE2)
p628_CA_SE
```

```
## [1] 0.002524887
```

```
## get OR SE
p628_OR_prop <- 0.088
p628_OR_n    <-  4691
p628_OR_SE2  <- p628_OR_prop * (1-p628_OR_prop) / p628_OR_n
p628_OR_SE   <- sqrt(p628_OR_SE2)
p628_OR_SE
```

```
## [1] 0.0041362429
```

```
## get SE diff
p628_SE2_diff <- p628_CA_SE2 + p628_OR_SE2
p628_SE_diff <- sqrt(p628_SE2_diff)
p628_SE_diff
```

```
## [1] 0.0048459839
```

```
## get prop diff
p628_prop_diff <- p628_CA_prop - p628_OR_prop
p628_prop_diff
```

```
## [1] -0.008
```

```
## get Z 95
p628_Z_95 <- qnorm(0.975)
p628_Z_95
```

```
## [1] 1.959964
```

```
## get ME diff
p628_ME_diff <- p628_Z_95 * p628_SE_diff
p628_ME_diff
```

```
## [1] 0.0094979539
```

```
## get CI 95
p628_lower_CI_95 <- p628_prop_diff - p628_ME_diff
p628_upper_CI_95 <- p628_prop_diff + p628_ME_diff
p628_CI_95 <- c(round(p628_lower_CI_95, 4), round(p628_upper_CI_95, 4))
p628_CI_95
```

```
## [1] -0.0175  0.0015
```

The difference in sample proportions (California minus Oregon) is **-0.008** .

The standard error of the proportion for California is **0.00252489** and for Oregon is **0.00413624** .

The standard error of the difference in proportions is **0.00484598** and the margin of error at 95% confidence is **0.0095** . [Not a typo – the similarity in values here is a coincidence.]

The 95% Confidence Interval for the difference in proportions (California minus Oregon) is (-0.0175, 0.0015) .

The interpretation is that although these samples suggest that a larger proportion (.088 - .080 = .008) of Oregonians than Californians are sleep-deprived, **we do not have sufficient data** to declare (with 95% confidence) that the ***actual population proportion*** of all Oregonians who are sleep-deprived exceeds that of all Californians, because *the above confidence interval includes zero* . **For the same difference in proportions, we would need larger samples to reach such a conclusion.** For the given data, we would have to conclude that the proportions of sleep-deprived people in California vs. Oregon are ***not statistically different*** .

(Indeed, if each sample were increased in size by 41 percent, with the proportions remaining the same, then the margin of error would decrease below 0.008. This would enable us to declare that Oregonians are more sleep-deprived than Californians.)

#·#

Exercise 6.44 - Barking deer

Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002.

In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7% and deciduous forests makes up 39.6%.

Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and *61* as deciduous forests.

```
p644_text_names <- c("Woods" , "Cultivated grassplot" , "Deciduous forests" , "Other" , "Total")
p644_text_avail <- c(.048, .147, .396, 1-(.048+.147+.396), 1)
p644_text_used <- c(4, 16, 61, 345, 426)

p644_textbook_data <- array(data=c(p644_text_avail,p644_text_used),
                            dim=c(5,2),
                            dimnames=list(p644_text_names,
                                          c("avail","used")))
p644_textbook_data
```

```
##                        avail used
## Woods                  0.048    4
## Cultivated grassplot 0.147   16
## Deciduous forests    0.396   61
## Other                0.409  345
## Total                1.000  426
```

```
#names(p644_used) = p644_names
#names(p644_avail) = p644_names
```

Note that there is an inconsistency in the data shown in the textbook:

In the problem discussion, the text indicates that the count of deciduous forest sites was *61*,

but the *table* within the text indicates that the number of such sites was *67*.

Both cannot be correct. (As will be discussed below, *NEITHER* is correct!!!!)

However, because the textbook indicates that the "Total" is 426 and "Other" is 345, only the *61* figure can be consistent with the data given in this textbook. So, for the moment, I'll disregard the 67 figure.

The table below summarizes these data.

```
p644_textbook_data
```

```
##                        avail used
## Woods                  0.048    4
## Cultivated grassplot 0.147   16
## Deciduous forests    0.396   61
## Other                0.409  345
## Total                1.000  426
```

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

H_0 : The respective usage of each of the foraging habitats ***matches*** its availability

(thus indicating that ***no habitat is "preferred"*** over others)

For this hypothesis to be accepted, the equivalent multiple-hypothesis test states that for each of the specified habitats, the percentage of **actual usage** (obtained by dividing the respective count by the total across all such counts, 426) is equal to the **expectation** which corresponds to the respective **availability** of such habitat. (Equivalently, the **percentage availability** can be multiplied by the total of the observed counts to establish the expected usage of each habitat.)

H_A : The respective usage of each of the foraging habitats ***does not match*** its respective availabilty

(thus indicating that ***certain habitats are "preferred"*** – or ***avoided*** – depending on whether actual is greater than or less than expected).

(b) What type of test can we use to answer this research question?

This can be answered using a Chi-Squared test, which will indicate whether the Null Hypothesis should be rejected in favor of the alternative, or whether the data is not sufficient to reject the null hypothesis. Based on the 4 classes specified, we would check the Chi-squared statistic against the critical value with 3 degree of freedom.

Note that this only indicates a "yes" or "no" to the overall question. If the null hypothesis is rejected, the Chi-Squared test does not specify which habitat(s) have a differential significant enough to fall into the appropriate category (preferred/avoided) vs. which are "close enough" to be within a margin of error. Additional analysis (e.g., Bonferroni-adjusted confidence intervals) are needed to make specific judgments on a habitat-by-habitat basis.

(c) Check if the assumptions and conditions required for this test are satisfied.

The conditions require firstly that each case that contributes to a count must be independent of the other cases in the table. While I am not an expert on the behavior of deer and the various types of vegetation, I will presume that the experimental design incorporated this requirement.

Secondly, the conditions require that each cell-count must have at least 5 expected cases. Multiplying the given availabilities by the total number of actual observatious, we observe that this is indeed true, as the smallest expected count is 20:

```
p644_text_expected <- round(p644_text_avail * 426,0)
p644_textbook_data_expanded <- cbind(p644_textbook_data,expected=p644_text_expected)
p644_textbook_data_expanded
```

```
##                        avail used expected
## Woods                  0.048    4       20
## Cultivated grassplot 0.147   16       63
## Deciduous forests      0.396   61      169
## Other                  0.409  345      174
## Total                  1.000  426      426
```

(d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

```

##p644_text_chisqtest_noround <- chisq.test(x=p644_text_used[1:4] , p=p644_text_avail[1:4])

###We can compute the Chi-Squared statistic manually:
sum((p644_text_used[1:4] - p644_text_expected[1:4])^2 / p644_text_expected[1:4])
```

```
## [1] 284.93297
```

```
### We can compute the Chi-Squared Statistic using the R function:
p644_text_chisqtest_rounded <- chisq.test(x=p644_text_used[1:4] , p=p644_text_expected[1:4]/426)
p644_text_chisqtest_rounded
```

```
##
##  Chi-squared test for given probabilities
##
## data:  p644_text_used[1:4]
## X-squared = 284.933, df = 3, p-value < 0.000000000000000222
```

```
### The Chi-Squared Statistic matches the above calculation:
p644_text_chisqtest_rounded$statistic
```

```
## X-squared
## 284.93297
```

```
### The Chi-Squared p-value contains many zeroes:
p644_text_chisqtest_rounded$p.value
```

```
## [1] 0.00000000000000000000000000000000000000000000000000000000000018130922
```

```
### Computing the above quantities in Excel indicate that the p-value is 1.8130922*10^-61

###Multiplying the above by 10^+61 does make the zeroes go away:
p644_text_chisqtest_rounded$p.value*10^61
```

```
## [1] 1.8130922
```

The result indicates that the Chi-Squared value is 284.93296768 and the corresponding p-value is close to zero (it is 1.81309216 * 10ˆ-61)

This means that the null hypothesis should be rejected in favor of the alternative.

These results suggest that the deer ***avoid*** the named regions (woods, cultivated grassplot, and deciduous forests) and prefer those regions lumped into "***other***" .

HOWEVER. . .

It is important to note that the data in the textbook does *not* accurately reflect that in the original paper!!!

While the above figures for "Woods" and "Cutivated Grassplot" are correct, ***those listed alongside "Deciduous Forests" are not associated with that environment.***

The *actual* data in the paper are as follows:

```
p644_paper_names <- c("Woods" ,
                     "Cultivated grassplot" ,
                     "Deciduous forests" ,
                     "Thorny Shrubland",
                     "Shrub Grassland",
                     "Dry Savannah",
                     "Total")
p644_paper_avail <- c(.048, .147, .189, .055, .396, .165,   1)
p644_paper_used  <- c(   4,   16,  149,  129,   67,   61, 426)

p644_paper_data <- array(data=c(p644_paper_avail,p644_paper_used),
                             dim=c(7,2),
                             dimnames=list(p644_paper_names,
                                           c("avail","used")))
p644_paper_data
```

```
##                      avail used
## Woods                0.048    4
## Cultivated grassplot 0.147   16
## Deciduous forests    0.189  149
## Thorny Shrubland     0.055  129
## Shrub Grassland      0.396   67
## Dry Savannah         0.165   61
## Total                1.000  426
```

The percentage which the textbook gives for the ***availability*** of ***Deciduous Forests*** (0.396) is actually associated with ***Shrub Grassland***; the correct percentage for availability of Deciduous Forests should be 0.189 .

The count which the textbook gives for the ***actual use*** of ***Deciduous Forests*** (61) is actually associated with ***Dry Savannah***; the correct count for actual use of Deciduous Forests should be 149 .

This is a serious blunder because the textbook authors are incorrectly indicating that the availability of deciduous forests is high (39.6%) while the deer's usage of such areas is low (61/426 = 14.3%).

In reality, the availability of deciduous forests is ***much lower*** (18.9%) while the deer's actual usage of such habitat is ***much greater*** (149/426 = 35%) !!!!!!!!!

The results from the chi-squared test indicate that there is a significant difference between availabilty vs. usage which reflects that the deer ***avoid*** certain areas while ***prefering*** others. However, further analysis is needed to determine whether the difference in actual vs. expected usage of each habitats is large enough to declare that such area is "avoided" or "preferred" vs. whether the differential is small enough to fall within a margin of error.

The paper goes on to calculate a **Bonferroni-adjusted confidence interval** for each habitat, centered about the respective percentage of ***actual usage***.

In the case where the percentage ***availability*** of such region is ***higher*** than the ***maximum*** of the Bonferroni Confidence Interval for such region, this gives evidence that the deer ***avoid** such region. This is the case for two of the regions detailed in the textbook (**Woods** and **Cultivated Grassplot**) as well as a third region which the textbook lumps into "other" ("**Shrub Grassland**") but for the fact that, as indicated above, the textbook misattributes the high availability of this habitat to **Deciduous Forests.**.

In the case where the percentage ***availability*** of such region is ***lower*** than the ***minimum*** of the Bonferroni Confidence Interval for such region, this gives evidence that the deer ***prefer** such region. This is the case for **Deciduous Forests** (which is mischaracterized in the textbook) as well as another region which the testbook also lumps into "other" ("**Thorny Shrubland**").

There is only a single habitat for which the Bonferroni Confidence Interval includes the percentage of availability; that region is "**Dry Savannah**" (which the textbook also attempts to lump into "other", except, as indicated above, it incorrectly misattributes the Actual Use count for this habitat to **Deciduous Forests**).

The paper explains that the "barking deer" is an unusually miniature deer, corresponding in size to a medium size dog. Because of its vulnerability to predators, the animal prefers to hide in areas where it is less likely to be found. Thus, its preference for covered habitats ("Deciduous Forests" and "Thorny Shrubland") where predators are less likely to find (and eat) it.

Unfortunately, the errors in the textbook prevent the reader from understanding this key distinction.

Below is the chi-squared analysis using the *CORRECT* data, as presented in the original paper:

```
### The result for the chisq value is close to the figure in the paper (656.191), but not exact.
p644_paper_chisq_1 <- chisq.test(x=p644_paper_used[1:6] , p=p644_paper_avail[1:6])
p644_paper_chisq_1
```

```
##
```

```
##  Chi-squared test for given probabilities
##
## data:  p644_paper_used[1:6]
## X-squared = 644.402, df = 5, p-value < 0.000000000000000222
```

```
### Adjust the rounded probabilities by computing the expected number of sites of each type
p644_paper_expected <- p644_paper_avail * 426
p644_paper_expected
```

```
## [1]  20.448  62.622  80.514  23.430 168.696  70.290 426.000
```

```
### The above figures are close to the values in the paper, up to rounding.
p644_paper_expected <- round(p644_paper_expected)
p644_paper_expected
```

```
## [1]  20  63  81  23 169  70 426
```

```
### These figures match those given in the original paper.

###We can compute the Chi-Squared statistic manually:
sum((p644_paper_used[1:6] - p644_paper_expected[1:6])^2 / p644_paper_expected[1:6])
```

```
## [1] 656.19092
```

```
### Compute the chisq value by dividing the integer expected figures by their total (426) to obtain pro
p644_paper_chisq_2 <- chisq.test(x=p644_paper_used[1:6] , p=p644_paper_expected[1:6]/426)
p644_paper_chisq_2$statistic
```

```
## X-squared
## 656.19092
```

```
### This figure exactly matches that in the paper.

p644_paper_chisq_2$p.value
```

```
## [1] 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
```

Based on the 6 classes specified in the paper, we would now check the Chi-squared statistic against the critical value with 5 degrees of freedom.

```
### Testing the upper tail on the chisq value to obtain the p-value results in a figure with many zeroe
pchisq(p644_paper_chisq_2$statistic, 5, lower.tail=FALSE)
```

```
##
## 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
```

```
### performing manual calculations (in Excel) using the old =CHITEST() and the new =CHISQ.DIST.RT() eac

### Multiply the above by 10^+139 to check whether all of the zeroes are eliminated:
pchisq(p644_paper_chisq_2$statistic, 5, lower.tail=FALSE)*10^139
```

```
## X-squared
## 1.4531133
```

The numbers match.

#.#

Exercise 6.48 Coffee and Depression.

Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women.

They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006.

The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants.

The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.

```
### construct an array with the table from the textbook
p648_Coffee_Quantities <- c("<=1 cup/week" ,
                            "2-6 cups/week" ,
                            "1 cup/day" ,
                            "2-3 cups/day",
                            ">=4 cups/day",
                            "TOTAL")
p648_depression_yes   <- c(  670,  373,   905,   564,   95,  2607)
p648_depression_no    <- c(11545, 6244, 16329, 11726, 2288, 48132)
p648_depression_total <- p648_depression_yes + p648_depression_no
p648_coffee_grid      <- array(data=c(p648_depression_yes , p648_depression_no , p648_depression_total)
                             dim=c(6,3),
                             dimnames=list(p648_Coffee_Quantities,
                                           c("depression_YES","depression_NO", "TOTAL")))
p648_depression_grid  <- t(p648_coffee_grid)
p648_depression_grid
```

```
##                <=1 cup/week 2-6 cups/week 1 cup/day 2-3 cups/day
## depression_YES          670           373       905          564
## depression_NO         11545          6244     16329        11726
## TOTAL                 12215          6617     17234        12290
##                >=4 cups/day TOTAL
## depression_YES           95  2607
## depression_NO          2288 48132
## TOTAL                  2383 50739
```

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

A chi-squared test with (5-1) * (2-1) = 4*1 = 4 degrees of freedom.

(b) Write the hypotheses for the test you identified in part (a).

H_0 : There is **no significant association** between amount of coffee consumption and incidence of clinical depression among women.

H_A : Incidence of clinical depression among women **does have a significant association** with amount of coffee consumption.

(c) Calculate the overall proportion of women who do and do not suffer from depression.

```
### compute the overall proportion of women who do suffer from depression
p648_proportion_depression_yes <-  p648_depression_yes[6]/p648_depression_total[6]
p648_proportion_depression_yes
```

```
## [1] 0.051380595
```

```
### compute the overall proportion of women who do not suffer from depression
p648_proportion_depression_no  <-  p648_depression_no[6] /p648_depression_total[6]
p648_proportion_depression_no
```

```
## [1] 0.94861941
```

```
## assemble the column of proportions (one for depression=yes ; one for depression=no)
p648_column_proportions <- array(c(p648_proportion_depression_yes,p648_proportion_depression_no),dim=c(
p648_column_proportions
```

```
##             [,1]
## [1,] 0.051380595
## [2,] 0.948619405
```

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. (Observed - Expected)2/Expected.

```
### Compute the row of proportions (one for each coffee column) by dividing each column total by the gra
p648_row_proportions   <- p648_depression_total[1:5] / p648_depression_total[6]
p648_row_proportions
```

```
## [1] 0.240741836 0.130412503 0.339659828 0.242219989 0.046965845
```

```
### Compute the table of proportions by computing the outer product of column proportions (2x1) and row
p648_proportion_table  <- p648_column_proportions %*% t(p648_row_proportions)
p648_proportion_table
```

```
##             [,1]        [,2]        [,3]        [,4]        [,5]
## [1,] 0.012369459 0.006700672 0.017451924 0.012445407 0.002413133
## [2,] 0.228372377 0.123711831 0.322207904 0.229774581 0.044552712
```

```
### Compute the table of expectations by multiplying the proportion_table by the grand total
p648_expectation_table <- p648_proportion_table    *    p648_depression_total[6]
p648_expectation_table
```

```
##             [,1]      [,2]        [,3]        [,4]       [,5]
## [1,]   627.61397  339.9854   885.49317   631.46751  122.43996
## [2,] 11587.38603 6277.0146 16348.50683 11658.53249 2260.56004
```

```r
### extract the count for the requested cell
expected_count_cell12 <- p648_expectation_table[1,2]
expected_count_cell12
```

```
## [1] 339.9854
```

```r
### assemble a similar table of the observed values, to enable chi-squared calculations
p648_observed_table  <- t(array(c(p648_depression_yes[1:5],p648_depression_no[1:5]),dim = c(5,2)))
p648_observed_table
```

```
##       [,1] [,2]  [,3]  [,4] [,5]
## [1,]   670  373   905   564   95
## [2,] 11545 6244 16329 11726 2288
```

```r
### compute the chi-squared table
p648_chisq_table <- (p648_observed_table - p648_expectation_table)^2 / p648_expectation_table
p648_chisq_table
```

```
##            [,1]       [,2]        [,3]       [,4]       [,5]
## [1,] 2.86254929 3.20591443 0.429722546 7.20839134 6.14955509
## [2,] 0.15504583 0.17364371 0.023275299 0.39043207 0.33308174
```

```r
### extract the chi-squared contribution for the requested cell
chi_squared_contribution_cell12 <- p648_chisq_table[1,2]
chi_squared_contribution_cell12
```

```
## [1] 3.2059144
```

```r
### compute the chi-squared statistic
p648_chi_squared_statistic <- sum(p648_chisq_table)
p648_chi_squared_statistic
```

```
## [1] 20.931611
```

The expected count for the highlighted cell [1,2] is 339.98539585 . The contribution of this cell to the test statistic is 3.20591443 .

(e) The test statistic is #2 = 20.93. What is the p-value?

```r
p648_chisq_results <- chisq.test(x=p648_observed_table,p=p648_proportion_table)
p648_chisq_results
```

```
##
##  Pearson's Chi-squared test
##
## data:  p648_observed_table
## X-squared = 20.9316, df = 4, p-value = 0.00032671
```

```r
### check the chi-squared statistic from the above calculation
p648_chisq_results$statistic
```

```
## X-squared
## 20.931611
```

```r
### check the chi-squared p-value from the above calculation
p648_chisq_results$p.value
```

```
## [1] 0.00032671037
```

The Chi-Squared p-value is 0.00032671 .

(f) What is the conclusion of the hypothesis test?

Because the p-value is so low, reject the Null Hypothesis, which stated :

H_0 : There is **no significant association** between amount of coffee consumption and incidence of clinical depression among women.

Instead, accept the alternative, which asserts:

H_A : Incidence of clinical depression among women **does have a significant association** with amount of coffee consumption.

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.

https://well.blogs.nytimes.com/2011/09/26/coffee-drinking-linked-to-less-depression-in-women

Do you agree with this statement? Explain your reasoning.

The comment from Dr. Albert Ascherio, an author of the study and professor of epidemiology and nutrition at the Harvard School of Public Health, stated that it was too early to recommend that women load up on extra lattes. More research is needed, he said, and "a very high level of caffeine can increase anxiety" and insomnia, potentially reversing any mood-lifting effects.

One key item which is not mentioned at all in the textbook, and which is only vaguely referenced in the New York Times article, is that ***all participants*** in this study were employed as ***nurses.***

https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/1105943

It is unclear whether the stresses on women who work as nurses are representative of women in general. Likewise, it is not clear whether the incidence of depression, and the consumption of coffee, among nurses are representative of women in general. Additionally, women who work as nurses may be more likely to be aware of various medical issues, and may have more access to various medical treatments/therapies vs. women who are not nurses. Furthermore, the average age of participants in the study was 63 years as of the ***start*** of the study (in 1996). This would suggest that the participants skew toward the elderly vs. the general population.

Accordingly, despite the observed 20% reduction in depression amongst participant consuming 4 or more cups of coffee per day, I would concur that additional research is needed to ensure that consumption of such a large amount of coffee is indeed safe, given other possible confounding health effects other than reduction in depression.

#·#