Homework 2

Submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have you 1) familiarized with sample size calculation procedures; 2) concepts of power & Type II error for significance tests; 3) practice your R coding.

Problem #1

1. Write your own `my.chisq.text()` function that will perform a X^2-test. As a single argument, it should just take a contingency table of arbitrary size. As output, it should provide

    - Calculated X^2 statistic
    - p-value,

Calculating the expected cell counts under H0 hypothesis should constitute a critical part of your function definition. Don't use neither `chisq.test()`, nor `prop.test()`, nor any other "fancy cheat" R's built-in functions inside your function's definition.

```
my.chisq.test <- function(table){
  table.expected <- (rowSums(table) %o% colSums(table)) / sum(table)
  statistic <- sum( (table.expected - table)^2/table.expected )
  c(statistic = statistic,
       p.value = pchisq(statistic,
                        df = (nrow(table)-1)*(ncol(table)-1),
                        lower.tail = F)
  )
}
```

2. For NYC Airbnb data set (listings.csv on Canvas), you would like to know whether there are differences in Airbnb room types offered in different NYC burrows. Proceed to formulate this in the form of a hypothesis test, as in

```
df <- readr::read_csv(sprintf("https://docs.google.com/uc?id=%s&export=download", "1au2tYZUtxb1AHckSTkd
```

a) *What variables are we interested in?*
We are interested in the `room_type` and `neighbourhood_group` variables.

b) *What are the hypotheses?*
H_0 :Airbnb room type offering and NYC burrow are independent
H_a :Airbnb room type offering and NYC burrow are dependent

c) *Print the contingency table. Under H0 hypothesis, proceed to calculate expected counts for two arbitrary cells of the contingency table (simply for practice).*

```
temp.df <- table(df$room_type, df$neighbourhood_group)
knitr::kable(stats::addmargins(temp.df), caption = "Contingency Table")
```

Table 1: Contingency Table

	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Sum
Entire home/apt	378	9565	13054	2118	181	25296
Private room	659	10131	7931	3489	187	22397
Shared room	68	418	471	204	10	1171
Sum	1105	20114	21456	5811	378	48864

Staten Island and Shared Room: $\frac{1171 \times 378}{48864} \approx 9.0585707$
Manhattan and Entire home/apt: $\frac{25296 \times 21456}{48864} \approx 1.1107379 \times 10^4$

d) *Proceed to apply your my.chisq.test() and interpret the results. As a sanity check, also run R's built-in chisq.test() function on that same data, make sure the outputted X2 and p-values match with those provided by my.chisq.test().*

```
my.chisq.test(temp.df)
```

```
## statistic   p.value
##  1609.391     0.000
```

```
chisq.test(temp.df)
```

```
##
##  Pearson's Chi-squared test
##
## data:  temp.df
## X-squared = 1609.4, df = 8, p-value < 2.2e-16
```

The p-value (≈ 0) is significantly less than the 0.05 significance level, which leads us to reject the null hypothesis and conclude that Airbnb room offerings and NYC burrow are dependent.

e) *In case you end up claiming that variables are not independent, proceed to make a few comments on strength of the relationship (as was done for Income & Happiness example in class).*

Table 2: Conditional Percentages

	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Entire home/apt	1.49	37.81	51.60	8.37	0.72
Private room	2.94	45.23	35.41	15.58	0.83
Shared room	5.81	35.70	40.22	17.42	0.85

- Shared room Airbnb offerings are 3.8993 times more likely than Entire home/apt Airbnb offerings in Bronx whereas they are only 2.0812 times more likely in Queens.
- The percentage of private rooms Airbnb offerings is 9.53 percentage-points higher than Shared room Airbnb offerings in Brooklyn.

Table 3: Vote for Femail President

Political.Views	Yes	No	Total
Extremely Liberal	56		58
Moderate	490		509
Extremely Conservative			61
Total	604	24	628

Political.Views	Yes	No	Total
Extremely Liberal			58
Moderate		19	509
Extremely Conservative		3	61
Total	604	24	628

Problem #2

1. **exercise** 11.84 from Agresti book.

a) *Fill in the cell counts that must appear in the blank cells for the left table*

Table 4: Vote for Female President

Political.Views	Yes	No	Total
Extremely Liberal	56	(58-56)	58
Moderate	490	(509-490)	509
Extremely Conservative	(604-56-490)	(61-58)	61
Total	604	24	628

b) *Fill in the cell counts that must appear in the blank cells for the left table*

Table 5: Vote for Female President

Political.Views	Yes	No	Total
Extremely Liberal	(58-2)	(24-19-3)	58
Moderate	(509-19)	19	509
Extremely Conservative	(61-3)	3	61
Total	604	24	628

2. **exercise** 11.9 from Agresti book.

11.9 Happiness and gender For the 2×3 table on gender and happiness in Exercise 11.4 (shown below), software tells us that $X^2 = 1.04$ and the P-value $= 0.59$.

	Happiness		
Gender	**Not**	**Pretty**	**Very**
Female	154	592	336
Male	123	502	257

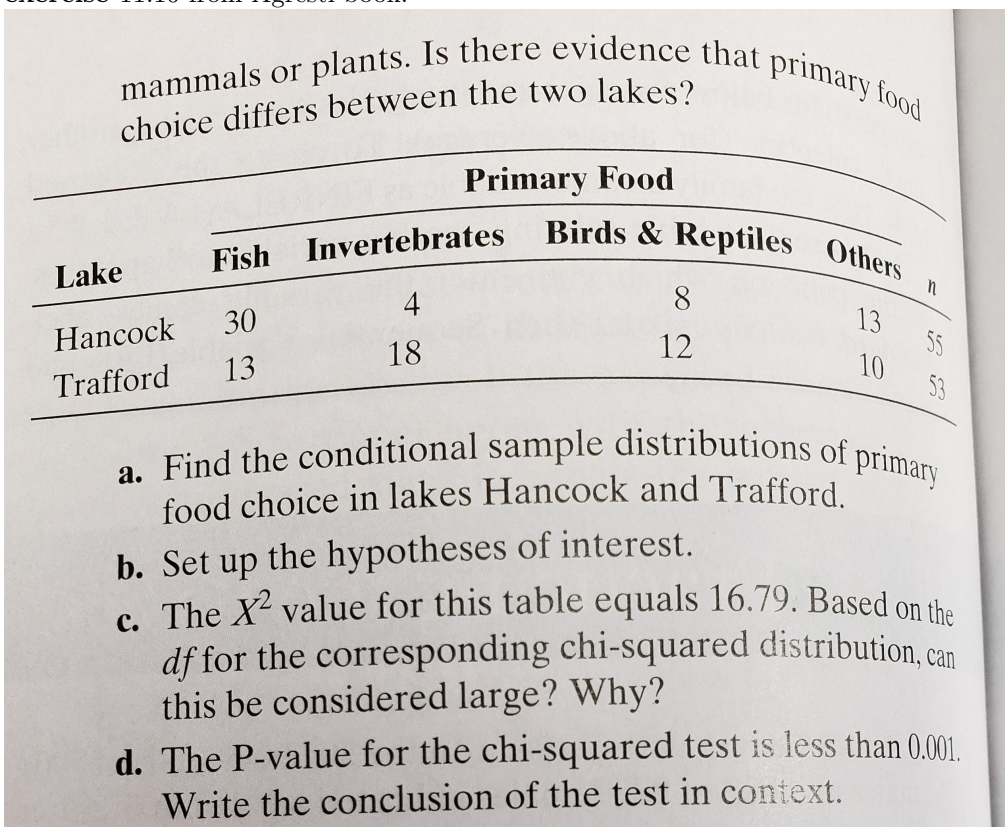
a. State the null and alternative hypothesis, in context, to which these results apply.

b. Interpret the P-value.

a) H_0 :The gender of an adult does not affect their happiness levels (Gender and Happiness are independent)
H_a :The gender of an adult affects their happiness levels (Gender and Happiness are dependent)

b) It is plausible that gender of an adult does not affect their happiness levels (Gender and Happiness are independent)

3. **exercise** 11.16 from Agresti book.



mammals or plants. Is there evidence that primary food choice differs between the two lakes?

	Primary Food				
Lake	Fish	Invertebrates	Birds & Reptiles	Others	n
Hancock	30	4	8	13	55
Trafford	13	18	12	10	53

a. Find the conditional sample distributions of primary food choice in lakes Hancock and Trafford.

b. Set up the hypotheses of interest.

c. The X^2 value for this table equals 16.79. Based on the *df* for the corresponding chi-squared distribution, can this be considered large? Why?

d. The P-value for the chi-squared test is less than 0.001. Write the conclusion of the test in context.

a)

Table 6: Conditional Percentages

	Fish	Invertebrates	Birds & Reptiles	Others
Hancock	54.54545	7.272727	14.54545	23.63636
Trafford	24.52830	33.962264	22.64151	18.86792

b) H_0 :The lake in which an alligator resides does not affect the proportion of primary foods eaten (Lake and Primary food are independent).
H_a : The lake in which an alligator resides affects the proportion of primary foods eaten (Lake and Primary food are dependent)

c) The X^2 value can be considered large because the chi-squared distribution with 3 degrees of freedom has an equivalent mean ($\mu = 3$) and is heavily right skewed. A value of 16.79 is more than five times the mean.

d) A p-value less than 0.001 indicates that the primary foods an alligator eats is dependent on the lake in which they live.

Problem #3

1. For all three examples on the "X2 Does NOT Measure Strength of Association" slide, proceed to **a)** *Use the my.chisq.test() function you've defined previously in order to confirm the X2 and p-values. Hint: Make sure to convert the %'es into counts first.*

```
df <- matrix(c(.51,.49,.49,.51), nrow = 2)
data.frame(A=my.chisq.test(df*100), B=my.chisq.test(df*200), C=my.chisq.test(df*10000)) %>%
  knitr::kable(digits = 3)
```

	A	B	C
statistic	0.080	0.160	8.000
p.value	0.777	0.689	0.005

b) *Calculate the difference in proportion between males & females that attend religious services weekly. Calculate the risk ratio between males & females that attend religious services weekly.*
Proportion Difference: $0.51 - 0.49 = 0.02$
Risk Ratio: $\frac{0.51}{0.49} \approx 1.04$

c) *Based on your answers to parts (a)-(b), as n increases, what do you notice with respect to statistical significance? Practical significance?*
As the sample size increases, the statistical significance increases. Both the proportion difference and the risk ratio indicate a weak practical significance since the resulting value is closer to 0 and close to 1, respectively.

2. **exercise** 11.32 from Agresti book.

11.32 Marital happiness The table shows 2012 GSS data on marital and general happiness for married respondents.

	General Happiness		
Marital Happiness	**Not Too Happy**	**Happy**	**Very Happy**
Not Too Happy	11	11	4
Happy	31	215	34
Very Happy	20	231	337

a. The chi-squared test of independence has $X^2 = 214$. What conclusion would you make using a significance level of 0.05? Interpret.

b. Does this large chi-squared value imply there is a strong association between marital and general happiness? Explain.

c. Find the difference in the proportion of being not too happy between those that are not too happy in their marriage and those that are very happy in their marriage. Interpret that difference.

d. Find and interpret the relative risk of being not too happy, comparing the lowest and highest marital happiness group. Interpret.

a) The p-value (≈ 0) indicates that general happiness of an adult is dependent on their marital happiness.

b) X^2 value cannot be used to determine the practical strength of an association.

c)

$p_{not\ happy,not\ happy} \approx 0.4231$
$p_{very\ happy,not\ happy} \approx 0.034$
Adults who are not happy in their marriage are 38.91 percentage points more likely to be unhappy in general than adults who are very happy in their marriage.

d)

$p_{not\ happy,not\ happy} \approx 0.4231$
$p_{very\ happy,not\ happy} \approx 0.034$
Adults who are not happy in their marriage are 12.4441176 times more likely to be unhappy in general than adults who are very happy in their marriage.