ASSIGNMENT 1 - PROBABILITY, STATISTICAL MODELS & TESTS
INF2190 - Winter 2022
AUDREY MEDAINO-TARDIF

Note this assignment is to be done individually. Make a copy of this doc and use it as a template for the assignment. Share your google doc with me: tegan.maharaj@utoronto.ca. Don't share it with anyone else. I may ask you for your source R code at any time during the semester. Submit a PDF of your doc via Quercus.

PART I: STATISTICAL MODELS & TESTS

1.1 T-tests and simulation

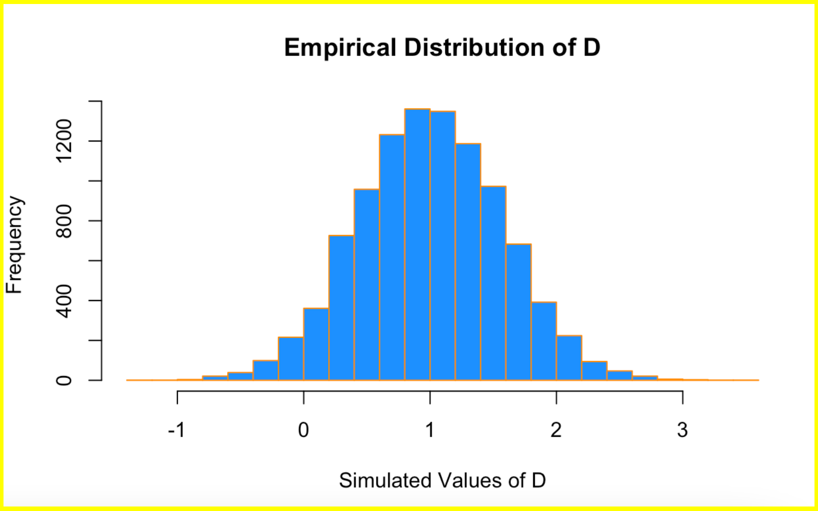
1. What does the value of the pdf of a particular value of x represent?
   The likelihood of a random variable to take on a given value.
2. What does the value of the cdf of a particular value of x represent?
   The probability that the random variable X is less than or equal to the number x
3. What's a quantile?
   Quantiles are cut points dividing the range of a probability distribution or in other words a sample divided into equal subgroups.
4. Paste a screenshot of your r code and the output for generating a random sample of 5 data points from a normal distribution with mean 3 and standard deviation 2.
   ```
   rnorm(n = 5, mean = 3, sd = 2)
   ```
5. Paste a screenshot of your r code and the output for generating a random sample of 6 data points from a binomial distribution with n=4 and p=0.6.
   ```
   rbinom(n= 6, size = 4, prob=0.6 )
   ```
6. Paste a screenshot of your r code to run and store results for a 2-sided t-test for the Captain Crisp data
   ```
   capt_crisp = data.frame(weight = c(15.5, 16.2, 16.1, 15.8, 15.6, 16.0, 15.8, 15.9,
   16.2))
   capt_test_results = t.test(capt_crisp$weight, mu = 16,
               alternative = c("two.sided"), conf.level = 0.95)
   ```
7. Paste a screenshot of your r code and output for accessing the p value from the stored results
   ```
   capt_crisp = data.frame(weight = c(15.5, 16.2, 16.1, 15.8, 15.6, 16.0, 15.8, 15.9,
   16.2))
   capt_test_results = t.test(capt_crisp$weight, mu = 16,
                         alternative = c("two.sided"), conf.level = 0.95)
   names(capt_test_results)
   capt_test_results$p.value
   ```

8. If the p value we calculate is smaller than our chosen significance level, **we reject the null hypothesis**
9. Paste a screenshot of your r code and results for a 2-sided t-test for randomly generated data (using the built-in r function for t test).

```
install.packages('ggplot2')
library(ggplot2)
data(msleep)
print(mean(msleep$awake))

msleep_test_results = t.test(msleep$awake, mu = 13.56,
            alternative = c("two.sided"), conf.level = 0.95)

names(msleep_test_results)
msleep_test_results$conf.int
```

10. Paste your r code and generated plot of a histogram of differences for randomly generated data.

```
print(pnorm(2, mean = 1, sd = sqrt(0.32)) - pnorm(0, mean = 1, sd = sqrt(0.32)))

set.seed(42)

sam = {10000} << NOT IN THE TUTORIAL, ORIGINAL CODE THROWS AN ERROR!

differences = rep(0, sam)

for (s in 1:sam) {
  x1 = rnorm(n = 25, mean = 6, sd = 2)
  x2 = rnorm(n = 25, mean = 5, sd = 2)
  differences[s] = mean(x1) - mean(x2)
}

mean(0 < differences & differences < 2)

hist(differences, breaks = 20,
    main   = "Empirical Distribution of D",
    xlab   = "Simulated Values of D",
    col    = "dodgerblue",
    border = "darkorange")
```



11. What do we learn from a histogram of differences?
    It looks like a normal distribution.

1.2 Chi-square goodness of fit test
http://www.sthda.com/english/wiki/chi-square-goodness-of-fit-test-in-r

1. The chi-square goodness of fit test is used to compare an **observed** distribution to an **expected** distribution. We use it **when we have two or more categories, for discrete data.**
2. Give an example (not the ones in the tutorial) of a question we could answer with a chi-square goodness of fit test.
   The colours of m&ms in an m&m bag. If we calculate the colour ratio or colour distribution in one bag as an observed distribution, then we could test with the chi-square goodness of fit test to see if there's a difference between the observed and expected distribution.

3. Paste a screenshot of your r code and output for running a chi-square goodness of fit test on randomly generated data of length 4 and expected probabilities of 0.25 for each category.

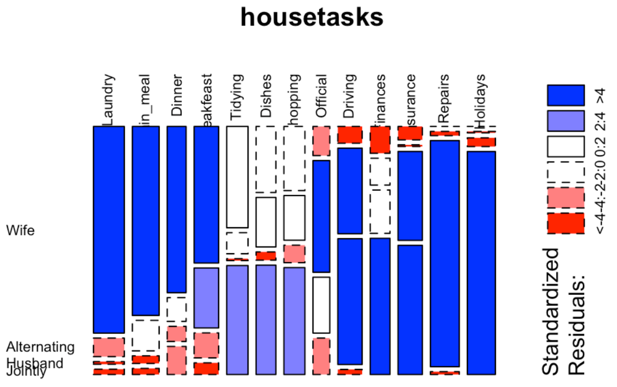
```
> mnms <- c(56, 40, 24, 56)
> res <- chisq.test(mnms, p = c(0.25, 0.25, 0.25, 0.25))
> res

      Chi-squared test for given probabilities

data:  mnms
X-squared = 16, df = 3, p-value = 0.001134
```

4. What do you results tell you about the data (reference the p value)?
   If the p value is more than the significance value, we can determine if our hypothesis or observation vs expectation is correct or incorrect.

1.3 Chi-Square Test of Independence
http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r

1. Paste a screenshot of your r code and output for a mosaic plot of the household tasks dataset.

```
file_path <- "/Users/audreymedaino-tardif/INF2190-DataVis/housetasks.txt"
housetasks <- read.delim(file_path, row.names = 1)


head(housetasks)
library("gplots")
dt <- as.table(as.matrix(housetasks))
# balloonplot(t(dt), main ="housetasks", xlab ="", ylab="",
         #label = FALSE, show.margins = FALSE)

library("graphics")
mosaicplot(dt, shade = TRUE, las=2,
         main = "housetasks")
```



2. Paste a screenshot of your r code and output for performing a chi-square test of independence on the same data.

```
chisq <- chisq.test(housetasks)
```

chisq

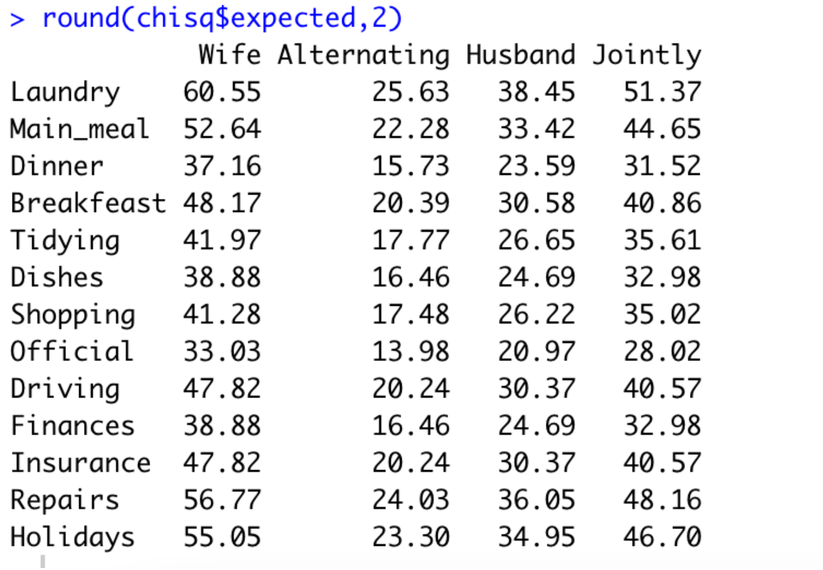
Output: Pearson's Chi-squared test

data: housetasks
X-squared = 1944.5, df = 36, p-value < 2.2e-16

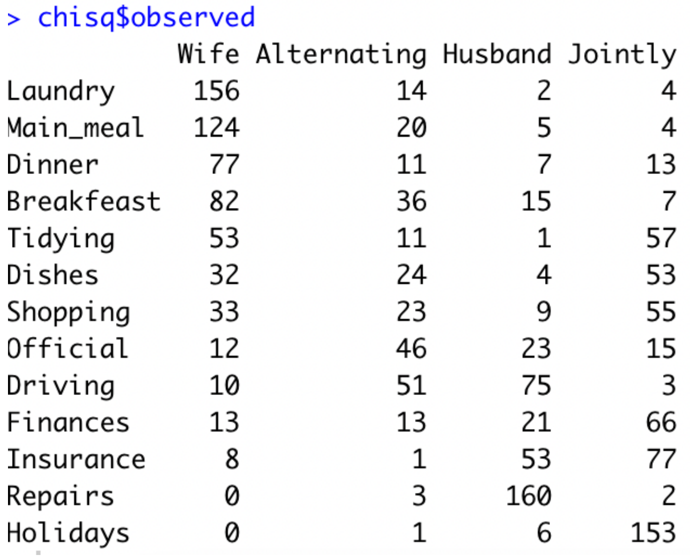
3. What question are we asking by doing the above test (reference specifics of the dataset, not just abstractly about the independence)?

   We're testing whether what we observed as housetasks associated to the wife, the husband, alternating or joint will be what we expected with a larger set of data.

4. Paste a screenshot of your r code and output for the observed and expected tables.



```
> round(chisq$expected,2)
            Wife Alternating Husband Jointly
Laundry    60.55       25.63   38.45   51.37
Main_meal  52.64       22.28   33.42   44.65
Dinner     37.16       15.73   23.59   31.52
Breakfeast 48.17       20.39   30.58   40.86
Tidying    41.97       17.77   26.65   35.61
Dishes     38.88       16.46   24.69   32.98
Shopping   41.28       17.48   26.22   35.02
Official   33.03       13.98   20.97   28.02
Driving    47.82       20.24   30.37   40.57
Finances   38.88       16.46   24.69   32.98
Insurance  47.82       20.24   30.37   40.57
Repairs    56.77       24.03   36.05   48.16
Holidays   55.05       23.30   34.95   46.70
```

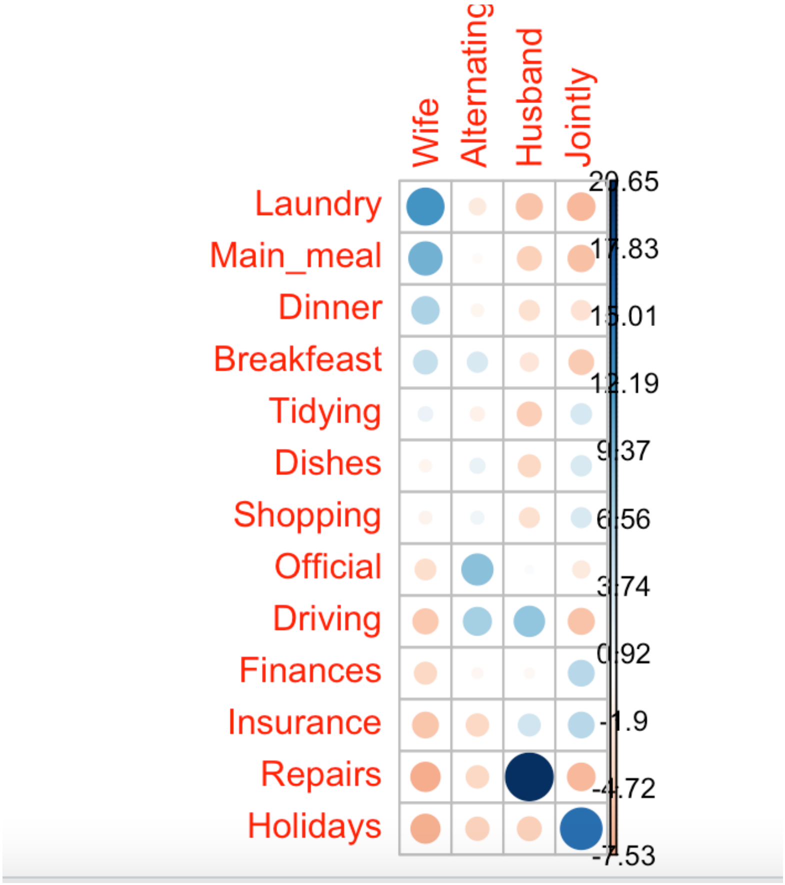


```
> chisq$observed
           Wife Alternating Husband Jointly
Laundry     156          14       2       4
Main_meal   124          20       5       4
Dinner       77          11       7      13
Breakfeast   82          36      15       7
Tidying      53          11       1      57
Dishes       32          24       4      53
Shopping     33          23       9      55
Official     12          46      23      15
Driving      10          51      75       3
Finances     13          13      21      66
Insurance     8           1      53      77
Repairs       0           3     160       2
Holidays      0           1       6     153
```

5. Paste a screenshot of your r code and output for a plot of the residuals.

   > library("corrplot")

```
corrplot 0.92 loaded
> corrplot(chisq$residuals, is.cor = FALSE)
```



6. What do the residuals show us?

It allows us to interpret the relationship or correlation between rows and columns;
Positive residuals: positive relationship between rows and columns
Negative residuals: negative or no relationship between rows and columns.

In concrete terms; husband & repairs has a strong positive residual versus wife and repairs.

7. Give an example of a decision these plots might be useful for making.

If I am a company trying to sell laundry detergent, I'd likely market it towards women and not men. Vice versa, if I want to sell tools, I'd market to men.