Homework #7

Due: Tuesday, November 16 @ 6pm

Please remember to give R code, as well as answers, for any problems where you used R

Problem 1:

A lake contains 600 fish, 80 of which have been tagged by scientists. Suppose a researcher randomly catches 15 fish from the lake: 3 are tagged and 12 are not tagged.

a. Is this a binomial random variable? Why or why not?

No this is not a binomial random variable because this is sampling without replacement, thus each sample is not independent from the other and the probability of sampling a tagged fish changes each time.

b. What is the probability of catching at least 3 tagged fish out of a sample of 15? (*In other words, is there an enrichment of tagged fish in the researcher's sample?*). **Solve this problem by hand (not using R functions like `dbinom()` or `dhyper()`). You may use R to help with the calculation in other ways**

This is a hypergeometric distribution.

$P(3+) = P(3) + P(4) + ... + P(15)$

$P(3+) = 1 - [P(0) + P(1) + P(2)]$

$P(0) = \frac{({}_{80}C_0)({}_{520}C_{15})}{{}_{600}C_{15}}$

$P(1) = \frac{({}_{80}C_1)({}_{520}C_{14})}{{}_{600}C_{15}}$

$P(2) = \frac{({}_{80}C_2)({}_{520}C_{13})}{{}_{600}C_{15}}$

Using the 'choose' function in R to calculate combinations:

```
p0 = choose(80,0)*choose(520,15)/choose(600,15)
p1 = choose(80,1)*choose(520,14)/choose(600,15)
p2 = choose(80,2)*choose(520,13)/choose(600,15)

1 - (p0 + p1 + p2)
```

```
## [1] 0.3223572
```

c. Now, using the `*binom()` and/or `*hyper()` functions in R, repeat (b). How does your answer compare?

```
# one option:
p0 = dhyper(0, 80, 520, 15)
p1 = dhyper(1, 80, 520, 15)
p2 = dhyper(2, 80, 520, 15)

1 - (p0 + p1 + p2)
```

```
## [1] 0.3223572
```

```r
# another option:
# remember phyper is only greater than, not equal to, so choose 2 not 3
phyper(2, 80, 520, 15, lower.tail = F)
```

```
## [1] 0.3223572
```

```r
# note: we could also use fisher.test for enrichment!
fisher.test(data.frame(tagged = c(3, 77), not = c(12, 520-12)), alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data.frame(tagged = c(3, 77), not = c(12, 520 - 12))
## p-value = 0.3224
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3891006       Inf
## sample estimates:
## odds ratio
##   1.647717
```

d. If this problem is binomial, how could you re-write the problem to be a hypergeometric distribution? Alternatively, if this problem is hypergeometric, how could you re-write the problem to be binomial?

To make this a binomial question, the researcher could select a fish, note if it was tagged or not, then replace the fish into the lake before preceeding to catch another fish. This way each catch is independent and the probability of selecting a tagged fish remains constant at $80/600 = 0.1333$

e. Solve this new problem from (d) using any method. How does your answer compare to (b) and (c)?

```r
# prob of at least 3 tagged fish:
p0 = dbinom(0, 15, 80/600)
p1 = dbinom(1, 15, 80/600)
p2 = dbinom(2, 15, 80/600)

1 - (p0 + p1 + p2)
```

```
## [1] 0.3228617
```

```r
# another option:
# pbinom is still greater than only
pbinom(2, 15, 80/600, lower.tail = F)
```

```
## [1] 0.3228617
```

In this case, the probabilities are very similar! This is likely because the probability of getting a tagged fish is so low and the number of fish is so high there is not a huge difference if your replace or not.

Problem 2:

Consider a fictitious population of mice in which each animal's coat is either black (B) or gray (G) in color and is either wavy (W) or smooth (S) is texture. Suppose a random sample of mice is selected and the coat color and texture are observed. Consider the accompanying contingency table for the data.

		Color		
		B	G	Total
Texture	W	40	50	**90**
	S	20	100	**120**
	Total	**60**	**150**	**210**

a. Express the following conditional probabilities as numbers:

- P(B | W) $40/90 = 0.444$
- P(S | G) $100/150 = 0.667$
- Smooth coats that are black $20/60 = 0.333$

b. Using a hand-held calculator or computer (not R functions), calculate the expected counts for each cell under the null hypothesis that coat color is independent from texture.

W|B = 90*60/210 = 25.7

W|G = 90*150/210 = 64.3

S|B = 120*60/210 = 34.3

S|G = 120*150/210 = 85.7

Can check that the row, columns, and grand total all add up: 25.7 + 64.3 = 90; 34.3 + 85.7 = 120; 25.7 + 64.3 + 34.3 + 85.7 = 210

c. Using a hand-held calculator or computer (not R functions), calculate the χ^2 statistic to test that hypothesis.

$\chi^2 = \sum \frac{(o-e)^2}{e}$

$\chi^2 = \frac{(40-25.7)^2}{25.7} + \frac{(50-64.3)^2}{64.3} + \frac{(20-34.3)^2}{34.3} + \frac{(100-85.7)^2}{85.7}$

$\chi^2 = 19.48$

d. Using your answer to (c) and the `*chisq()` suite of functions (*i.e.* `pchisq()`, `dchisq()`, `rchisq()`, or `qchisq()`), test your hypothesis that coat color is independent from texture. *Give your code, the p-value, and a written interpretation of your results*

```
# we want to use pchisq to calculate the p-value.
# Since we are testing for independence we can use a non-directional alternative
# df for 2x2 table is always 1
pchisq(19.48, 1, lower.tail = F)
```

```
## [1] 1.016585e-05
```

Coat color is not independent from texture (p = 1.01e-05)

e. What assumptions did you make to perform the chi-square test for independence?

We assumed that the data is one random sample of mice observed with resoect to two categorical variables, hair color and hair texture. Sample size must also be considered, but all counts are greater than 5 so we don't have to worry about this assumption.

f. Read the help page for the `chisq.test()` function in R and apply it to the data above. How does it compare to your answer in (d)?

```
# create a data frame for mice. Notice the correct = F must be used to see the same
# value we calculated by hand
chisq.test(data.frame(B = c(40,20), G = c(50, 100)), correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.frame(B = c(40, 20), G = c(50, 100))
## X-squared = 19.444, df = 1, p-value = 1.036e-05
```

```
# look at correct = T (default)
chisq.test(data.frame(B = c(40,20), G = c(50, 100)), correct = T)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data.frame(B = c(40, 20), G = c(50, 100))
## X-squared = 18.107, df = 1, p-value = 2.088e-05
```

```
# for grading purposes, either answer is okay.
```

Without the continuity correction, the chi-squared test statistic and p-value were identical to what I calculated!

g. We learned in class that the Fisher's Exact Test is more accurate (and more computationally expensive) than the χ^2-test for count data. Read the help page for the `fisher.test()` function and apply it here. How does it compare to your answers from (c) and (d)?

```
# we can use the same data frame for fisher test
fisher.test(data.frame(B = c(40,20), G = c(50, 100)))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data.frame(B = c(40, 20), G = c(50, 100))
## p-value = 1.333e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.028624 7.981731
## sample estimates:
## odds ratio
##   3.971463
```

The Fisher's Exact Test does not provide a test statistic, but rather calculates the p-value "exactly" by finding the proportion of possible data combinations that are as extreme or more extreme than our data. It calculated a p-value of 1.33e-05 compared to the p-value of 1.03e-05 from the chi-square test. These two values are very similar (same order of magintude)

Problem 3:

Prior to an influenza season subjects were randomly assigned to receive either a flu vaccine or a placebo. During that season there were 28 cases of the flu among 813 vaccine recipients and 35 cases of the flu among the 325 subjects who were given the placebo.

a. Calculate the relative risk (conditional probability) of getting the flu for individuals who received the placebo versus those who received the vaccine. Write one sentence explaining this value.

```
# relative risk = p1 / p2
# p1 = P(flu | placebo)
# p2 = P(flu | vaccine)

p1 <- 35/325
p1
```

```
## [1] 0.1076923
```

```
p2 <- 28/813
p2
```

```
## [1] 0.03444034
```

```
# relative risk:
p1/p2
```

```
## [1] 3.126923
```

Individuals who received the placebo were 3.12 times more likely to get the flu than individuals who received the flu vaccine.

b. Calculate the odds ratio for comparing flu cases among individuals who received the placebo to flue cases among individuals who received the vaccine.

```
# odds ratio = odds1 / odds2
# odds = P(E) / (1 - P(E))
# odds1 = P(flu | placebo)
# odds2 = P(flu | vaccine)

# we can use p1 and p2 calculated above ^

odds_placebo <- p1 / (1 - p1)
odds_placebo
```

```
## [1] 0.1206897
```

```
odds_vaccine <- p2 / (1 - p2)
odds_vaccine
```

```
## [1] 0.03566879
```

```r
# odds ratio
odds_placebo / odds_vaccine
```

```
## [1] 3.383621
```

The odds of getting the flu are about 3.38 times as great for individuals who got the placebo as for individuals who did get the vaccine.

c. The output from the `fisher.test()` in R also gives an odds ratio. Perform this test for this data. How does it compare to the answer from (b)?

```r
# need to create a data frame for fisher test
fisher.test(data.frame(flu = c(35, 28), no_flu = c(325-35, 813-28)))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data.frame(flu = c(35, 28), no_flu = c(325 - 35, 813 - 28))
## p-value = 4.509e-06
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.958105 5.880938
## sample estimates:
## odds ratio
##   3.379365
```

```r
# notice it doesn't matter if you do rows or columns
fisher.test(data.frame(placebo = c(35, 325-35), vaccine = c(28, 813 - 28)))
```

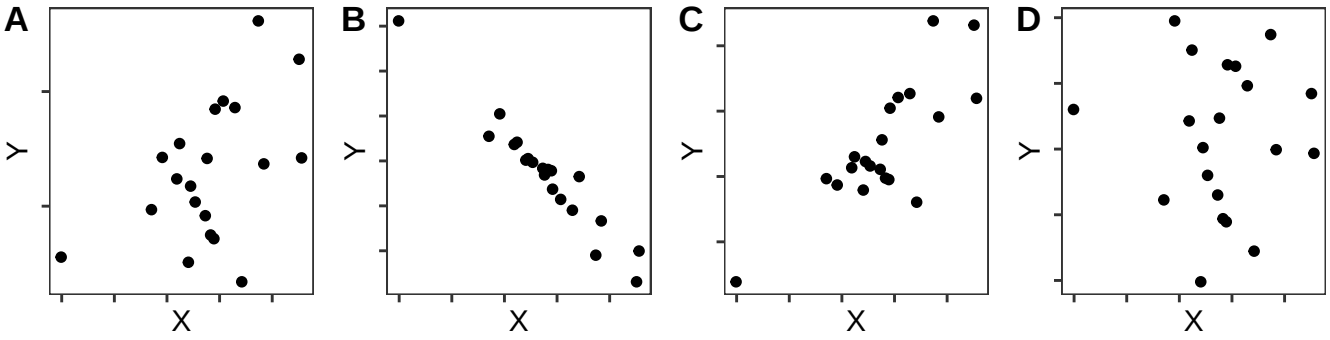
```
##
##  Fisher's Exact Test for Count Data
##
## data:  data.frame(placebo = c(35, 325 - 35), vaccine = c(28, 813 - 28))
## p-value = 4.509e-06
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.958105 5.880938
## sample estimates:
## odds ratio
##   3.379365
```

Yes! The odds ratio calculated with the fisher.test is very close to the ratio I calculated by hand.

d. Does the odds ratio give a good approximation to the relative risk for these data? Why or why not?

Yes, the relative risk and odds ratio were very similar. This makes sense because odds ratio = relative risk * (1-p2)/(1-p1), so if p1 and p2 are very small, the odds ratio will be about equal to the relative risk. And in this case p1 = 0.1 and p2 = 0.03 (very small).

Problem 4:



a. Arrange the plots in order of their correlations (from closest to -1 to closest to +1)

B, D, A, C

b. Arrange the plots in order of their corresponding P-values (smallest to largest) for the test $H_0 : \rho = 0$. *Note: all of the plots display the same number of observations*

B, C, A, D

Problem 5:

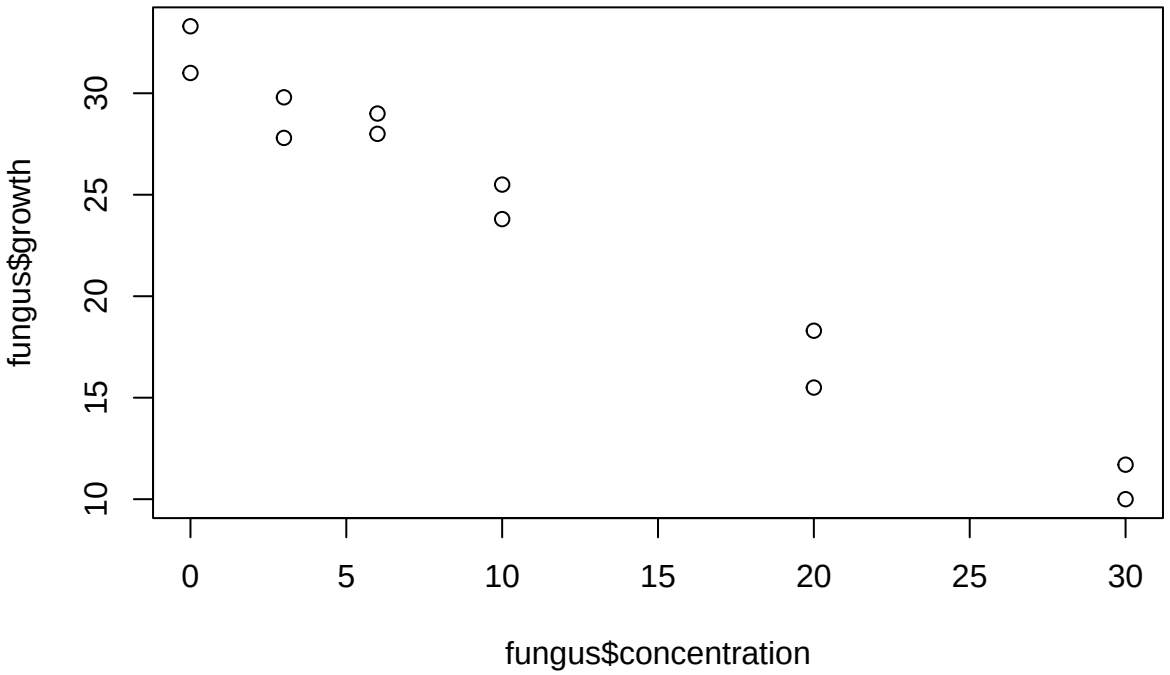
Laetisaric acid is a compound that holds promise for control of fungus diseases in crop plants. The accompanying data show the results of growing the fungus *Pythium ultimum* in various concentrations of laetisaric acid. Each growth value is the average of four radial measurements of a *P. ultimum* colony grown in a petri dish for 24 hours; there were two petri dishes at each concentration.

Laetisaric acid concentration (uG/mL)	0	0	3	3	6	6	10	10	20	20	30	30
Fungus growth (mm)	33.3	31	29.8	27.8	28	29	25.5	23.8	18.3	15.5	11.7	10

a. Plot the data in a way that you can visualize the relationship between laetisaric acid concentration and fungus growth. By eye, does there seem to be a linear relationship?

```
fungus <- data.frame(concentration = c(0,0,3,3,6,6,10,10,20,20,30,30),
                     growth = c(33.3,31,29.8,27.8,28,29,25.5,23.8,18.3,15.5,11.7,10))

plot(fungus$concentration, fungus$growth)
```



Yes, there appears to be a very strong linear relationship between concentration of laetisaric acid and fungal growth.

b. Calculate the correlation coefficient using the R function `cor()`. Give a one sentence interpretation of this value.

```
cor(fungus$concentration, fungus$growth)
```

```
## [1] -0.9875349
```

There is a strong negative linear correlation (r = -0.987) between laetisaric acid and fungal growth.

c. Consider testing whether the population correlation is zero. Compute the value of the test statistic. You may choose to use the `cor.test()` function in R. *Give your code, the p-value, and a written interpretation of your results*

```
cor.test(fungus$concentration, fungus$growth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fungus$concentration and fungus$growth
## t = -19.84, df = 10, p-value = 2.321e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9966099 -0.9547173
## sample estimates:
##        cor
## -0.9875349
```

There is a strong negative linear correlation (r = -0.987) between laetisaric acid and fungal growth. We reject the null hypothesis that there is no linear correlation (p = 2.32e-09)

e. Is this study an observational study or an experiment?

This is an experiment because the fungus was grown at different chosen concentrations - researchers were in control of one of the variables.

e. It is suggested that acid could be used to retard fungus growth. Could these data be used to verify this claim? If not, what could be said? Briefly explain.

Although correlation does not equal causation, in this case a controlled experiment where only one factor (acid concentration) was varied and one factor (fungus growth) was measured. We could use this information to infer causality that acid can retard fungus growth (within the realm of the experiment - i.e. these concentrations tested)