Final Exam_Spring_2021

Sunny Lee

2021-04-24

You are required to show your work on each problem on this exam. The following rules apply:

a) The majority of the credit you receive will be based on the completeness and clarity of your response.

b) Organize your work, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.

c) Mysterious or unsupported answers will not receive full credit. A correct answer, unsupported by calculations, explanation, or algebraic work will receive no credit; an incorrect answer supported by substantially correct calculations and explanations might still receive partial credit.

d) Its due on or before Wednesday 04/27/2021 at 11:59am via grade-scope.

1

) I will not share any of the information or collaborate with any one. I can only use the materials discussed in class and lecture videos to work on this exam. Please write your name as if you agree with the above statement.

Sunny Lee

2) A study was conducted by introductory statistics students to see whether majors in the sciences are as likely to take a statistics course as majors in the humanities. They sampled 50 seniors from among humanities majors and another 50 seniors from among science majors and asked if they had taken a statistics course or were scheduled to take one before they graduated. Their data are provided below.

```
Humanities<-cbind(c(23,32,55),c(27,18,45),c(50,50,100))
rownames(Humanities)<-c("Humanities Major", "Science Major","Total")
colnames(Humanities)<-c("Yes", "No", "Total")
Humanities
```

```
##                  Yes No Total
## Humanities Major  23 27    50
## Science Major     32 18    50
## Total             55 45   100
```

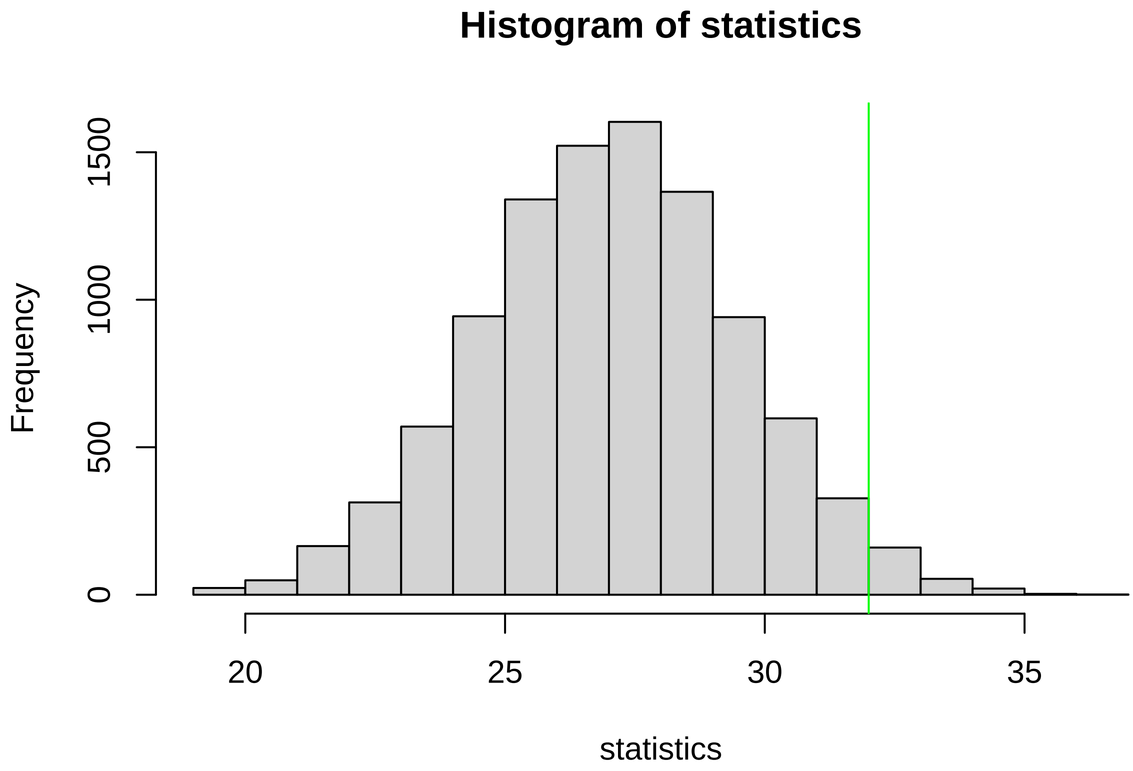
a) Create a simulation study to test the one-sided hypothesis that science majors are more likely to take a statistics course. State the null and alternative hypothesis. Provide a p-value and write your conclusion.

Since we are checking to see if science majors are more likely to take a statistics class, our null hypothesis should be that the proportion of science majors who take a statistics class is equal to the proportion of humanities majors who take a statistics class: $H_o : P_{science} = P_{humanities}$ and our alternative hypothesis is the proportion of science majors who take a statistics class is higher than the proportion of humanities majors who take a statistics class: $H_a : P_{science} > P_{humanities}$.

```
stats<-c(rep("yes",55),rep("no",45))
exp1<-sample(stats,50,replace=FALSE)
sum(exp1=="yes")
```

```
## [1] 23
```

```
p=10000
statistics<-rep(0,p)
for(i in 1:p)
{
  temp<-sample(stats,50,replace=FALSE)
  statistics[i]<-sum(temp=="yes")
}
hist(statistics)
abline(v=32,col="green")
```



```
p_value<-mean(statistics>=32)
p_value
```

```
## [1] 0.0566
```

From our simulated study, we obtain the p-value = 0.0572. Since our simulated p-value is greater than .05, we fail to reject the null hypothesis and conclude that the proportion of science majors who take a statistics course is equal to the proportion of humanities majors who take a statistics course.

b) Use Fisher's exact test to test the one-sided hypothesis that science majors are more likely to take statistics course. Provide a p-value and include your conclusion.

```
sum(dhyper(32:50, 55, 45, 50))
```

```
## [1] 0.05367785
```

Using Fisher's exact test, we obtain a p-value of 0.0537 which is quite close to the simulated p-value we obtained above. Since the p-value obtained from Fisher's exact test is still greater than 0.05, we fail to reject the null hypothesis and conclude that the proportion of science majors who take a statistics course is equal to the proportion of humanities majors who take a statistics course.

c) Conduct a chi-square test to determine if there is a significant difference between what was observed and what was expected.

To conduct a chi-square test, we must first calculate the expected values for each group:

```
expectedHumanities<-cbind(c(27.5,27.5,55),c(22.5,22.5,45),c(50,50,100))
rownames(expectedHumanities)<-c("Humanities Major", "Science Major","Total")
colnames(expectedHumanities)<-c("Yes", "No", "Total")
expectedHumanities
```

```
##                     Yes   No Total
## Humanities Major 27.5 22.5    50
## Science Major    27.5 22.5    50
## Total            55.0 45.0   100
```

To calculate the chi-square statistic, we sum the square difference of each of the observed and expected values and divide by the expected value:

```
chisq <- 0
for(i in 1:2){
  for (j in 1:2){
    chisq <- chisq + ((Humanities[i, j] - expectedHumanities[i, j])^2 / expectedHumanities[i, j])
  }
}

chisq
```

```
## [1] 3.272727
```

Thus, our chi-square statistic is 3.27. To calculate the p-value from this statistic we can use the pchisq with degree of freedom 1 and lower.tail set to false:

```
pchisq(3.272727, df = 1, lower.tail = FALSE)
```

```
## [1] 0.07044044
```

Since we obtain a p-value of 0.0704, which is greater than 0.05, we fail to reject the null hypothesis and conclude there is no significant difference between what was observed abnd what was expected.

3

) TRUE OR FALSE (a) The Kruskal Wallis test is a non-parametric test.

TRUE

(b) In a logistic regression model, if you add more variables the residual deviance either decreases or stays the same.

TRUE

(c) If a linear model has two variables independent variables, the one with the larger slope is more significant.

FALSE

(d) You can use Fisher's exact test to compare two continous numeric variables.

FALSE

(e) An assumption of linear regression is that the error terms follow a normal distribution centered at zero with a fixed variance.

TRUE

4

) Matched Pairs Let T_1 represent the time it takes you to complete a task the first time. Let T_2 represent the time it takes you to complete a task the second time. Create a dataset with two variables T_1 and T_2 each with 50 observations. Let T_1 be 50 randomly generated numbers from a normal distribution with mean 10 and standard deviation 3. Let T_2 be 50 randomly generated numbers from a normal distribution with mean 12 and standard deviation 3.

```
T_1<-rnorm(50, mean=10,sd=3)
T_2<-rnorm(50,mean=12,sd=3)
```

Create a simulation that builds a distribution of the difference of the mean times if the null hypothesis, that on average it always takes people the same amount of time to complete the task, is true. Report a histogram and p-value associated with our data and this null hypothesis?

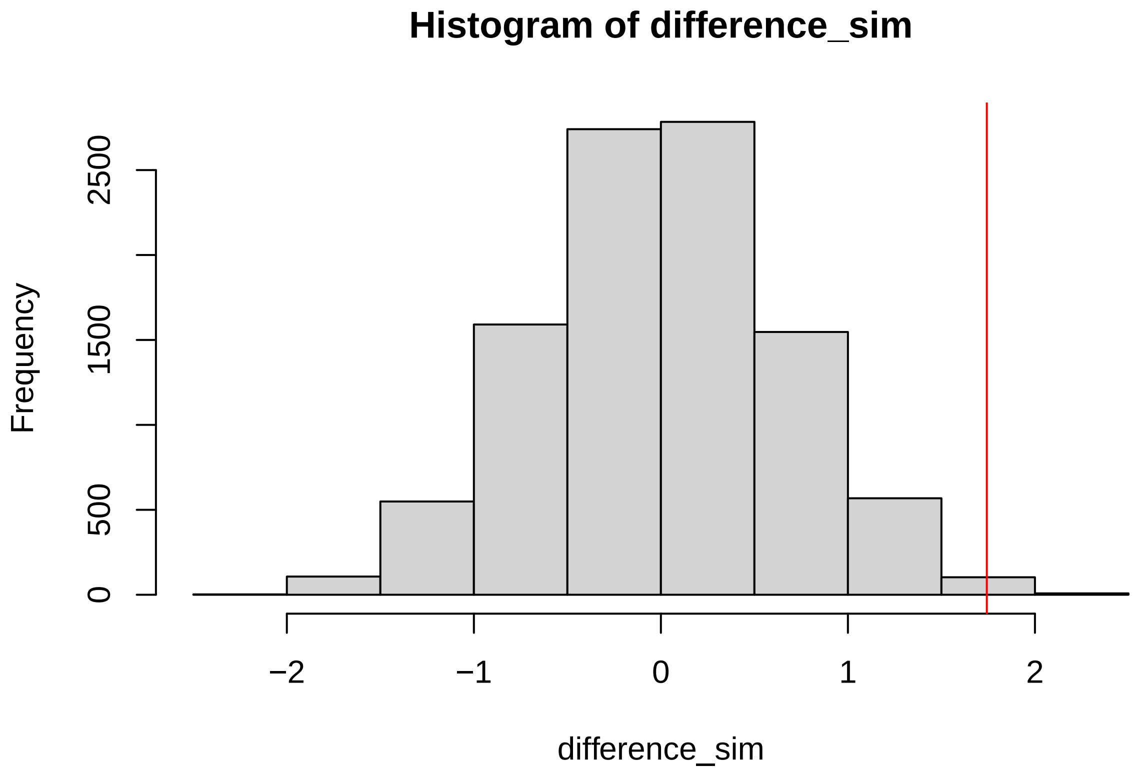
```
difference <- T_2 - T_1
obs <- mean(difference)
obs
```

```
## [1] 1.742679
```

```
multiplier <- multiplier <- sample(c(1, -1), length(difference), replace = TRUE)

mean(difference * multiplier)
```

```
## [1] -0.921172
```

```
p <- 10000
difference_sim <- rep(0, p)

for (i in 1:p){
  tempMult <- sample(c(1, -1), length(difference), replace = TRUE)
  difference_sim[i] <- mean(tempMult * difference)
}

hist(difference_sim)
abline(v = obs, col = "red")
```



```
p_value <- mean(difference_sim > obs)
p_value
```

```
## [1] 0.0028
```

Since we obtain a simulated p-value of 0, we reject the null hypothesis which was that it takes the same time to complete the task the first and second times and accept the alternative hypothesis that on average, the time it takes to complete the task the second time is higher than the first.

5) Compute an F test for the one-way analysis of variance data below. Show your work using R commands and the aov() function. Include the null and alternative hypothesis and interpret your results using the anova() function to output the analysis of variance summary table. Do the average air temperatures differ by season?

Since we are applying ANOVA to our null hypothesis is $H_o : \alpha_1 = \alpha_2 = \alpha_3$ and the alternative hypothesis is at least one of the effects are different. Using the aov() function we can find the p-value of the F-statistic:

```
Temperature<-cbind(c(60,65,62,64,63),c(80,85,82,84,83),c(90,95,101,99,100))
colnames(Temperature)<-c("Fall", "Spring", "Summer")
Temperature
```

```
##      Fall Spring Summer
## [1,]   60     80     90
## [2,]   65     85     95
## [3,]   62     82    101
## [4,]   64     84     99
## [5,]   63     83    100
```

```
temp <- c(Temperature[, 1], Temperature[, 2], Temperature[, 3])
season <- c(rep("Fall", 5), rep("Spring", 5), rep("Summer", 5))
anova <- aov(temp~season)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## season       2 2952.1  1476.1   158.7 2.34e-09 ***
## Residuals   12  111.6     9.3                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since our p-value is less than 0.05, we reject the null hypothesis and conclude that at least one of the effects is different from the others. Since we reject the null hypothesis, the air temperatures differ by season.

6

) Given a logistic regression with a dependent variable saver (0=does not save money on a regular basis, 1= saves money on a regular basis) and three predictor variables, age (0=under 30, 1= over 30), education (0=high school, 1= college), and income (0=less than \$50,000, 1=more than \$50,000), the following logistic regression coefficients were reported.

```
Coefficients<-cbind(c(-1.00,1.5,1.15,2.5),c(2.34,3.67,1.67,2.89))
rownames(Coefficients)<-c("Intercept","Age","Education","Income")
colnames(Coefficients)<-c("Estimate","Standard Error")
Coefficients
```

```
##           Estimate Standard Error
## Intercept    -1.00           2.34
## Age           1.50           3.67
## Education     1.15           1.67
## Income        2.50           2.89
```

```
Null_Deviance= 107.97
Residual_Deviance= 102.34
Null_Deviance
```

```
## [1] 107.97
```

```
Residual_Deviance
```

```
## [1] 102.34
```

a) State the form of the logistic regression equation using these coefficients.

$$\frac{e^{-1+1.5(Age)+1.15(Education)+2.5(Income)}}{1+e^{-1+1.5(Age)+1.15(Education)+2.5(Income)}}$$

b) Predict the probability value for an individual with the following characteristics: Over 30, college educated, with income more than \$50,000.

```
exp(-1 + 1.5 + 1.15 + 2.5) / (1 + exp(-1 + 1.5 + 1.15 + 2.5))
```

```
## [1] 0.9844802
```

The probability value for an individual who is over 30, college educated and has an income of more than \$50,000 is 98.45%.

```
c) Predict the probability value for an individual with the following characteristics: over 30, high sc
```

```
exp(-1 + 1.5) / (1 + exp(-1 + 1.5))
```

```
## [1] 0.6224593
```

The probability value for an individual who is over 30, high school education and an income less than \$50,000 is 62.25%

```
d) Compare (b) and (c), which one is more likely to be a saver?
```

(b) or is more likely to be a saver since the probability is much higher.

e) Test $H_o : \beta_i = 0$ versus $H_a : \beta_i \neq 0$ using Wald's test. State your conclusion based on these tests.

```
exp(1.5 - (1.96*3.67))
```

```
## [1] 0.003368796
```

```
exp(1.5 + (1.96*3.67))
```

```
## [1] 5962.231
```

```
exp(1.15 - (1.96*1.67))
```

```
## [1] 0.1196481
```

```
exp(1.15 + (1.96*1.67))
```

```
## [1] 83.36262
```

```
exp(2.5 - (1.96*2.89))
```

```
## [1] 0.04223948
```

```
exp(2.5 + (1.96*2.89))
```

```
## [1] 3513.613
```

Let β_1 be the coefficient of the Age variable, β_2 be the coefficient of the Edcuation variable and β_3 be the coefficient of the Income variable. Using the Wald's test for β_1, we obtain a confidence interval of $(0.003368796, 5962.231)$. Since zero is not included in the confidence interval, we can conclude that β_1 is significant in our model. For β_2, we obtain a confidence interval of $(0.1196481, 83.36262)$ and since zero is not part of our confidence interval, we can also conclude that β_2 is significant in our model. For β_3, we obtain a confidence interval of $(0.04223948, 3513.613)$ and since zero is not part of our confidence interval, we conclude that β_3 is significant to our model.

```
f) Check if the model is significant
```

To check if our model is significant, our null hypothesis must be that the β_i are zero: $H_o : \beta_1 = \beta_2 = \beta_3 = 0$ and our alternative hypothesis must be that at least one of the β_i are not zero: H_a :at least one of the β_i is not zero. To calculate the test statistic, we subtract the residual deviance from the null deviance. To calculate our p-value we check our test statistic against a chi-square distribution with degrees of freedom 3 and lower.tail set to FALSE.

```
pchisq(Null_Deviance - Residual_Deviance, df = 3, lower.tail = F)
```

```
## [1] 0.1310667
```

Since our p-value is 0.1311 which is greater than 0.05, we reject the null hypothesis and conclude that our at least one of our β_i is not zero and thus our model is significant.

```
\newpage
```

7) Explain clearly the difference between parametric and non-parametric test. Give an example of each tests.

A parametric test assumes that data comes from a population that follows a certain distribution. One such parametric test would be the t-test. A non-parametric test does not require assumption about the distribution of a population or a large sample size. One such example would be the randomization test.

8) Distribution test

```
Temperature<-cbind(c(60,65,62,64,63),c(80,85,82,84,83),c(90,95,101,99,100))
colnames(Temperature)<-c("Fall", "Spring", "Summer")
Temperature
```

```
##      Fall Spring Summer
## [1,]   60     80     90
## [2,]   65     85     95
## [3,]   62     82    101
## [4,]   64     84     99
## [5,]   63     83    100
```

a) What test would you use if you want to compare the distribution of Fall and Spring temperature only. State null and alternative hypothesis and calculate the test statistic.

To compare the distribution of Fall and Spring, we would use the Wilcoxon Rank Sum test. Our null hypothesis would be that the two groups have the same distribution $H_o : G_1, G_2$ same distribution and the alternative hypothesis is that the two groups have different distributions $H_a : G_1, G_2$ different distributions.

```
wilcox.test(Temperature[, "Fall"], Temperature[, "Spring"])
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Temperature[, "Fall"] and Temperature[, "Spring"]
## W = 0, p-value = 0.007937
## alternative hypothesis: true location shift is not equal to 0
```

From the wilcox.test function, we find that the test statistic is $W = 0$.

b) Use a software to determine if the distribution of temperature is the same for both Fall and Spring.

Using the wilcox.test(), we find that our p-value is 0.0079 which is less than 0.05 and thus we reject the null hypothesis. Since we reject the null hypothesis, we conclude that Fall and Spring have different distribution.

c) State the null and alternative hypothesis test of Kruskal-Wallis test.

The null hypothesis is H_o : distribution for all response variables is the same for all groups. The alternative hypothesis H_a : at least one of the distributions are different.

d) Calculate Kruskal-Wallis test statistic by hand.

Ranking the values given:

```
rank<-cbind(c(1, 5, 2, 4, 3),c(6, 10, 7, 9, 8),c(11, 12, 15, 13, 14))
colnames(rank)<-c("Fall", "Spring", "Summer")
rank
```

```
##      Fall Spring Summer
## [1,]    1      6     11
## [2,]    5     10     12
## [3,]    2      7     15
## [4,]    4      9     13
## [5,]    3      8     14
```

Since $N = 15$ we can calculate the test statistic:

```
sum <- (15^2 + 40^2 + 65^2) / 5
(12/(15*16))*sum - (3*16)
```

```
## [1] 12.5
```

which yields $H = 12.5$

e) Use a soft ware to run Kruskal-Wallis test to determine if the distribution of temperature for the three difference seasons is the same. State the null and alternative hypothesis, p-value,

The null hypothesis is H_o : distribution for Fall, Spring, and Summer is the same for all groups. The alternative hypothesis H_a : at least one of the distributions are different.

```
kruskal.test(list(Temperature[, "Fall"], Temperature[, "Spring"], Temperature[, "Summer"]))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(Temperature[, "Fall"], Temperature[, "Spring"], Temperature[, "Summer"])
## Kruskal-Wallis chi-squared = 12.5, df = 2, p-value = 0.00193
```

From the Kruskal Wallis test, we obtain a p-value of 0.00193. Since our p-value is less than 0.05 we reject the null hypothesis and accept the alternative hypothesis. Thus, we conclude that at least one of the distributions are different.

9

) Logistic Regression: Imagine you played a game where you try and throw a play pen ball into a bucket and find the following results. The first 3 shots are taken from a distance of 8 feet and 2 of them go into the bucket. The next 4 shots are taken from 10 feet and 2 of them go into the bucket. Finally, the last 4 shots are taken from 12 feet, and only 1 goes into the bucket.

a) Use the maximum likelihood estimate to estimate the probability of successful shooting into the bucket.

```
feet <- c(rep(8, 3), rep(10, 4), rep(12, 4))
success <- c(1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0)
logit <- glm(success~feet, family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = success ~ feet, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5230  -0.9559  -0.7844   1.0478   1.6301
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.3926     4.2946   1.023    0.306
## feet         -0.4511     0.4200  -1.074    0.283
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.158  on 10  degrees of freedom
## Residual deviance: 13.887  on  9  degrees of freedom
## AIC: 17.887
##
## Number of Fisher Scoring iterations: 4
```

```
MLE <- exp(logit$coefficients[1] + logit$coefficients[2]*feet) / (1 + exp(logit$coefficients[1] + logit$
MLE
```

```
##  [1] 0.6864364 0.6864364 0.6864364 0.4703454 0.4703454 0.4703454 0.4703454
##  [8] 0.2648273 0.2648273 0.2648273 0.2648273
```

Using the maximum likelihood estimate, we estimate the probability of successful shooting into the bucket as 68.64% for 8 feet 47.03% for 10 feet and 26.48% for 12 feet.

b) Calculate and interpret the odds of success.

To calculate the odds of success, we take the probability of success and divide it but the probability of failure for 8, 10 and 12 feet.

```
0.6864364 / (1 - 0.6864364)
```

```
## [1] 2.189146
```

```
0.4703454 / (1 - 0.4703454)
```

```
## [1] 0.8880229
```

```
0.2648273 / (1 - 0.2648273)
```

```
## [1] 0.3602246
```

The probability of success for 8 feet is 2.189 times more likely than the probability of failure. The probability of succes for 10 feet is .888 times more likely than the probability of missing the bucket. The probability of success for 12 feet is .36 times more likely than the probability of missing the bucket.

c) Calculate the odds ratio of successful shooting between 8 feet and 12 feet. Provide a confidence interval for this odds ratio and interpret your result.

```
2.189146/0.3602246
```

```
## [1] 6.07717
```

To calculate the odds ratio, we take the odds given feet is 8 and the odds given feet is 12 and divide one from the other. Thus, our odds ratio is 6.08 which means the odds of success at 8 feet is 6.08 times more likely than the odds of success at 12 feet. To calculate the confidence interval for the odds ratio we take β_1 add and subtract the standard error times 1.96.

```
exp(logit$coefficients[2] - (1.96*-1.074))
```

```
##     feet
## 5.227354
```

```
exp(logit$coefficients[2] + (1.96*-1.074))
```

```
##       feet
## 0.07760102
```

Using these values, we obtain the confidence interval $(0.07760102, 5.227354)$.

d) Check overall adequacy of the model. Calculate the test statistics, p-value, and a concluding statement.

10

) In a sample of 1000 animals 400 are dogs, 500 are cats, and 100 are rabbits. Of the animals 250 have blue eyes and 750 have brown eyes.

a) Assuming these two variables (type of animal and color of eye) are independent, fill in the table of how many of each type of animal you would expect to see in a sample of size 1000.

```
expectedAnimal<-cbind(c(100,125, 25),c(300, 375, 75))
rownames(expectedAnimal)<-c("Dog","Cat","Rabit")
colnames(expectedAnimal)<-c("Blue","Brown")
expectedAnimal
```

```
##       Blue Brown
## Dog    100   300
## Cat    125   375
## Rabit   25    75
```

b) If you were using a χ^2 goodness of fit test, what would be the χ^2 statistic to measure how far the following table is from what you expect, if the null hypothesis is true? (Do not simplify your answer!)

```
Animal<-cbind(c(100,150,0),c(300,350,100))
rownames(Animal)<-c("Dog","Cat","Rabit")
colnames(Animal)<-c("Blue","Brown")
Animal
```

```
##       Blue Brown
## Dog    100   300
## Cat    150   350
## Rabit    0   100
```

The χ^2 statistic would be $\frac{(100-100)^2}{100}+\frac{(300-300)^2}{300}+\frac{(125-150)}{125}+\frac{(375-350)^2}{375}+\frac{(25-0)^2}{25}+\frac{(75-100)^2}{75}$

c) How many degrees of freedom are there for the above goodness of fit test?

Since our degrees of freedom is number of rows minus 1 times number of columns minus 1, we have 2 degrees of freedom.