# Math 181B: Homework 3
## Merrick Qiu

**Exercise 1**

$$P(X_1 = a, X_2 = b)$$

$$= \sum_{\substack{(k_3,\ldots,k_t) \\ \sum_{i-3}^{t} k_i = n-a-b}} P((x_1, x_2, x_3, \ldots, x_t) = (a, b, k_3, \ldots, k_t))$$

$$= \sum_{\substack{(k_3,\ldots,k_t) \\ \sum_{i-3}^{t} k_i = n-a-b}} \frac{n!}{a!b!k_3!\ldots k_t!} p_1^a p_2^b p_3^{k_3} \ldots p_t^{k_t}$$

$$= \frac{n!}{a!b!(n-a-b)!} p_1^a p_2^b (1-p_1-p_2)^{n-a-b} \sum_{\substack{(k_3,\ldots,k_t) \\ \sum_{i-3}^{t} k_i = n-a-b}} \frac{(n-a-b)!}{k_3!\ldots k_t!} \frac{p_3^{k_3} \ldots p_t^{k_t}}{(1-p_1-p_2)^{n-a-b}}$$

$$= \frac{n!}{a!b!(n-a-b)!} p_1^a p_2^b (1-p_1-p_2)^{n-a-b}$$

We have shown that $P(X_1 = a, X_2 = b)$ has the same probability density function as $Multinom(n, p_1, p_2, 1 - p_1 - p_2)$, so we have that $(X_1, X_2, Y) \sim Multinom(n, p_1, p_2, 1-p_1-p_2)$. The last step follows from the fact that the summation is the sum of the probabilities of all possible outcomes of $Multinom(n - a - b, \frac{p_3}{1-p_1-p_2}, \ldots, \frac{p_t}{1-p_1-p_2})$, which is equal to 1.

## Exercise 2

We have the hypotheses that $H_0 : X \sim Poisson(\hat{\lambda})$ and $H_1$ : The model does not fit the data well. The following code gives a p-value of 0.1231678, meaing that we fail to reject the null hypothesis and a Poisson fits the distribution. For the test to work, we assume that $E_i \geq 5$, for the distribution to closely match a chi-squared. This is true for all bins except $E_0 = 4$, but this should not be a problem since 4 is close to 5.s

```
# Import files
setwd("C:/Users/merri/Documents/MATH-31H/MATH 181B/Homework 3")
x = unlist(read.csv("Question.csv"))

#Estimate MLE for Poisson
lambda_hat = mean(x)

# Put data into bins
bins = c(0,1,2,3,4,5,6,7,Inf)
observed = table(cut(x, bins, right=F))

# Expected probabilities
probabilities = dpois(c(0,1,2,3,4,5,6,Inf), lambda_hat)
probabilities[8] = 1 - ppois(6, lambda_hat)
expected = probabilities*length(x)

#Calculate D
D = sum((observed-expected)^2/expected)

#8-1-1 degrees of freedom
pchisq(D, df=6, lower.tail=FALSE)

#P-value is 0.1231678>0.03
```
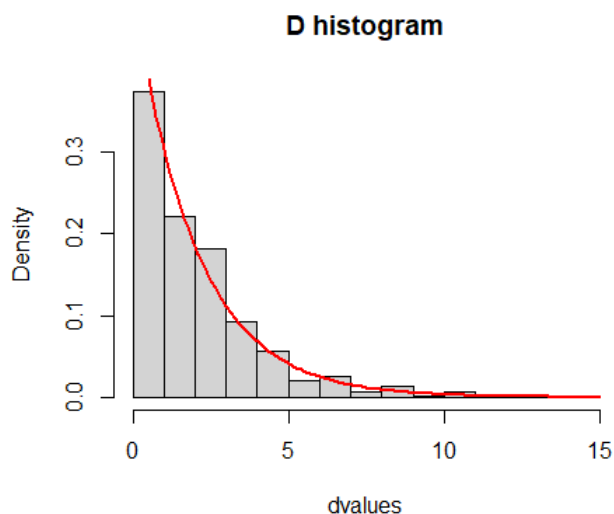
# Exercise 3

1. 
```
#get a random value for d
getd = function() {
    probs = c(0.3,0.5,0.2)
    observed = rmultinom(1, 50, probs)
    expected = 50*probs

    D = sum((observed-expected)^2/expected)
}

# create vector of 1000 samples of D
dvalues = rep(0, 1000)
for (i in 1:1000) {
    dvalues[i] = getd()
}

# Plot histogram
hist(dvalues, breaks = 20, freq=FALSE, main="D histogram")
curve(dchisq(x, df = 2), add = TRUE, col = "red", lwd = 2)
```

**D histogram**



The histogram matches closely with the chi-squared distribution, so my simulation does agree with theory.
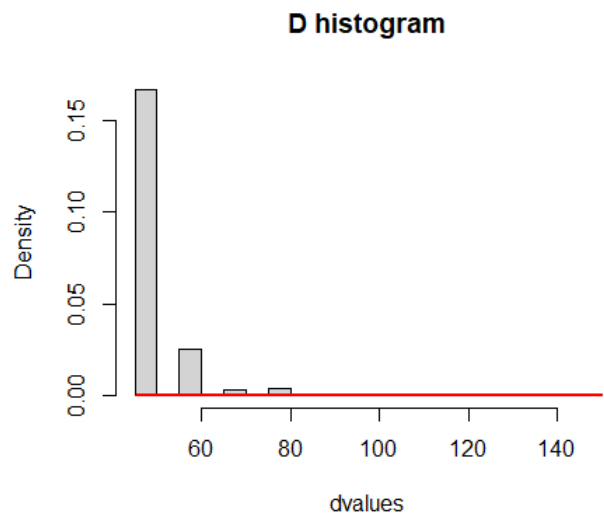
2. 
```
# create vector of rejections
rejections = rep(0, 1000)
for (i in 1:1000) {
  dvalue = getd()
  pvalue = pchisq(dvalue, df=2, lower.tail=FALSE)
  if (pvalue < 0.05) {
    rejections[i] = 1
  }
}

sum(rejections)/1000
#0.057
```

We see that the propotion of rejections is 0.057, which is close to the value of $\alpha$.

**D histogram**



3.

4. Changing part a) to this distribution yields dvalues that are much much larger than before, in a region where the chi-squared distribution has very low density.

   Changing part b) to this distribution yields a rejection rate of 1, which makes sense considering how large the values of d were.