Math 440B HW 8

Erick Castillo

3/8/2021

Problem 3: The problem wants us to find the joint distribution of the number of ganes won by each player, given the following information: there are three players, they play ten games, each player has equal probability of winning a game of $\frac{1}{3}$.

Because this is a case of distinct players each having a distinct probability of winning a game, this is a case of the multinomial distribution. Its pmf can be written as follows:

$$\binom{10}{n_1 n_2 n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

Problem 1: The probabilities for the question are generated in the following code:

```
n.val = c(0,1,2,3,4)
lambda = 0.8392
first4 = dpois(n.val, lambda)
first4
```

```
## [1] 0.432056030 0.362581420 0.152139164 0.042558395 0.008928751
```

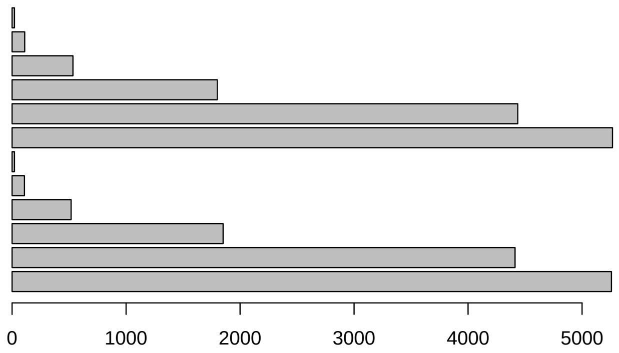
```
fifth = ppois(4, 0.8392, lower.tail = FALSE)
fifth
```

```
## [1] 0.001736239
```

Now notice that the table counts a total of $12,169$ observations. Thus the expected observations given the above probabilities are calculated below, and are plotted on a bar graph where the upper part of the bar graphs represents the actual values and the bottom bar graph represents the expected values. Notice they are very similar in length:

```
expect = c(first4*12169, fifth*12169)
actual = c(5267,4436,1800,534,111,21)

barplot(c(expect, actual), beside = TRUE, horiz = TRUE)
```



```
expect
```

```
## [1] 5257.68983 4412.25330 1851.38149  517.89311  108.65398   21.12829
```

```
actual
```

```
## [1] 5267 4436 1800  534  111   21
```

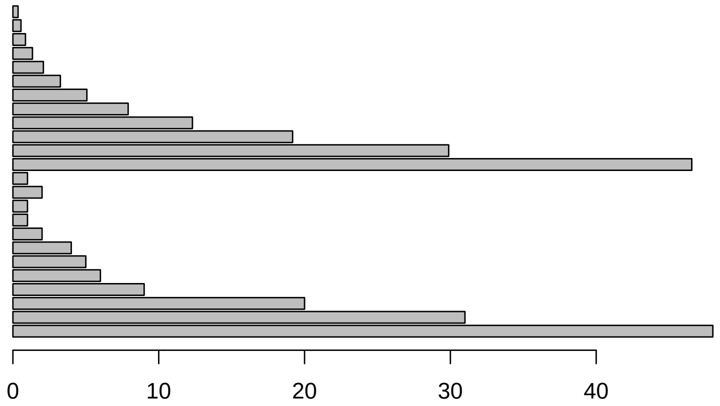
Notice that the actual and expected outputs look similar to each other.

Problem 8A: The question asks to fit a geometric distribution to the provided data. Notice that $\hat{p}_{MLE} = \frac{1}{\bar{x}}$ is the best estimator for p in this problem.

```
numhops = 0:11
actualfreq = c(48,31,20,9,6,5,4,2,1,1,2,1)

#here the mle of p is calculated
xbar = sum((numhops+1)*actualfreq)/sum(actualfreq)
phat = 1/xbar

#calculating the expected frequencies
expprob = dgeom(numhops, phat)
expectedfreq = 130*expprob
barplot(c(actualfreq, expectedfreq), horiz = TRUE, beside = TRUE)
```



Also notice that the expected frequencies, represented by the bar graphs on the top, is noticeably shorter in some parts and noticeably longer in others when compared to the actual frequency.

Problem 8B: The problem now asks to find an approximate 95% CI for p. To begin doing so, we find the $E[X]$ and $Var(X)$ where $X \sim Geom(\hat{p})$. Using the back of the book,

$$E[X] = \frac{1}{\hat{p}} \approx 2.7983$$

and

$$Var(X) = \frac{1-\hat{p}}{\hat{p}^2} \approx 5.0047$$

Now, recall the formula for the confidence interval can be written as follows:

$$E[X] \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{Var(X)}{n}}$$

Plugging in all the information known, I am able to get a confidence interval of $(2.4077, 3.1769)$. Taking the reciprocal of these value will yield the CI for $\hat{p}_{MLE}$ to be $(0.3148, 0.4153)$.

Problem 8C: The question now asks to examine the goodness of fit for the geometric distribution. In class we learned that the expected/observed cells with values less than 5 should be summed with neighboring cells to avoid issues when testing with a chi-square distribution. This is reflected in the following code.

```
actualfreq = c(48,31,20,9,6,5,11)
expectedfreq = c(expectedfreq[1:6],sum(expectedfreq[7:12]))

teststat1 = sum((actualfreq-expectedfreq)^2/expectedfreq)
teststat1
```

```
## [1] 2.237056
```

```r
pchisq(teststat1, df = 5, lower.tail = FALSE)
```

```
## [1] 0.8154657
```

The above p-value indicates that there is strong evidence to suggest that the geometric distribution is a great fit for the data. By convention, I can fail to reject H_0 at a 5% significance level. The model fits the data well.

Problem 8D: A uniform prior is used for p. The likelihood function for p is given by

$$L(p) = p^n(1-p)^{\sum k-n}$$

Now, recall the Bayesian formula $Post \propto Prior \times Likelihood$. So in this case

$$Post \propto L(p) \times 1 \propto p^n(1-p)^{\sum k-n} \sim \beta(n+1,\ \sum k - n + 1) \sim \beta(131,\ 234)$$

So the posterior mean is given by

$$\frac{n+1}{\sum k + 2} = \frac{131}{365}$$

The posterior standard deviation is

$$\sqrt{Var(X)} = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}} \approx 0.0250732$$

As desired.

Problem 25: This problem asks to calculate the likelihood ratio of an earlier example, and to compare it to the Pearson's Chi-Square Statistic.

```r
obs.bac = c(56,104,80,62,42,27,9,20)
exp.bac = c(34.9,85.1,103.8,84.4,51.5,25.1,10.2,5.0)
like.ts = 2*sum(obs.bac*log(obs.bac/exp.bac))
like.ts
```

```
## [1] 54.7727
```

```r
pchisq(like.ts,df=6,lower.tail = FALSE)
```

```
## [1] 5.15213e-10
```

```r
2*obs.bac*log(obs.bac/exp.bac)
```

```
## [1]  52.960864  41.717284 -41.670294 -38.245694 -17.128624   3.940327  -2.252937
## [8]  55.451774
```

The above output includes both the likelihood ratio test statistic, the corresponding p-value, and components of the log likelihood test. Notice that the $\chi^2 = 75.4$ is much greater than the likelihood test statistic; regardless, there is strong evidence in both their p-values that the Poisson distribution is not a good fit for the data. Both reject at 0.5% significance. Also notice that the components of the log-likelihood test have negative values, unlike the Pearson components which are all positive.

Problem 26A: The statement provided is TRUE. The generalized likelihood ratio statistic $\Lambda \leq 1$. This is seen on page 339 in the book where it states $\Lambda = min(\Lambda^*, 1)$. This also makes intuitive sense as the denominator of a LRT is always greater than or equal to the numerator as the denominator considers the maximum likelihood estimator over the entire sample space.

Problem 34: This question ask to test the goodness of fit to Problem 55 in Chapter 8. The $\hat{\theta}_{MLE}$ is given to be 0.0357.

```
theta.hat1 = 0.0357
actual.val = c(1997, 906, 904, 32)
probabs = c(0.5+0.25*theta.hat1, 0.25-0.25*theta.hat1, 0.25-0.25*theta.hat1, 0.25*theta.hat1)
probabs
```

```
## [1] 0.508925 0.241075 0.241075 0.008925
```

```
expect.val = sum(actual.val)*probabs
expect.val
```

```
## [1] 1953.76307  925.48693  925.48693   34.26308
```

```
teststat2 = sum((actual.val-expect.val)^2/expect.val)
teststat2
```

```
## [1] 2.015486
```

```
pchisq(teststat2, df = length(actual.val)-1-1, lower.tail = FALSE)
```

```
## [1] 0.365042
```

The above code generates a very high p-value. This means that there is strong evidence to suggest that the provided probabilities model fits the data well.

Problem 40: This question asks to show that

$$\sum_{i=1}^{2} \frac{(X_i - np_i)^2}{np_i}$$

can be rewritten to be expressed as

$$\frac{(X_1 - np_1)^2}{np_1(1-p_1)}$$

We begin by listing the identities $X_1 + X_2 = n$ and $p_1 + p_2 = 1$.

$$\frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - n - X_1 + np_1)^2}{n(1-p_1)} = (X_1 - np_1)^2 (\frac{1}{np_1} + \frac{1}{n(1-p_1)})$$

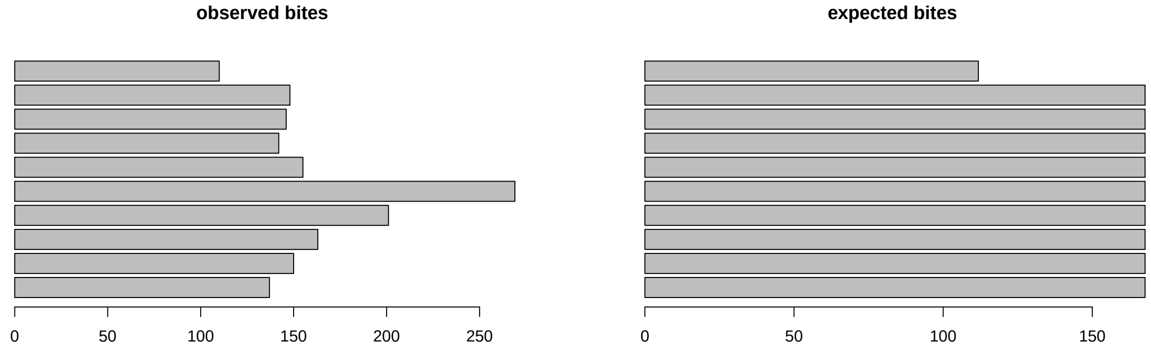
$$(X_1 - np_1)^2 (\frac{n - np_1 + np_1}{np_1(n(1-p_1))}) = \frac{(X_1 - np_1)^2}{np_1(1-p_1)}$$

As desired.

Problem 39: This question asks if there is a temporal trend in the amount of bites recorded at a hospital during a full moon. If there was no trend, then there would be an equal amount of bites on any day. That is, each day would have a uniform probability of there being a bite.

```
#the full moon occurred in the 5th cell
obs.bites = c(137,150,163,201,269,155,142,146,148,110)
probs = c(rep(c(3/29), times = 9),2/29)
exp.bites = probs*sum(obs.bites)
exp.bites
```

```
##  [1] 167.6897 167.6897 167.6897 167.6897 167.6897 167.6897 167.6897 167.6897
##  [9] 167.6897 111.7931
```



I almost don't even need to run a test. I can clearly see that this model does not fit the data, indicating there is some kind of trend. I will perform the Pearson χ^2 test regardless to be certain:

```
pearsonvals = (obs.bites-exp.bites)^2/exp.bites
pearsonvals
```

```
##  [1]  5.61665497  1.86608947  0.13115219  6.61686060 61.20703128  0.96027002
##  [7]  3.93559389  2.80542733  2.31190481  0.02876045
```

```
teststat3 = sum(pearsonvals)
teststat3
```

```
## [1] 85.47975
```

```
pchisq(teststat3, df = length(obs.bites)-1, lower.tail = FALSE)
```

```
## [1] 1.308432e-14
```

The result is as expected. There is strong evidence to suggest that there is a trend in the model as a uniform distribution does not fit the model well.

Problem 43A: This question asks about the data of some coin flips fitting $H_0 : p = 0.5$

```
obs.heads = 9207
obs.tails = 8743
obs.flips = c(obs.heads, obs.tails)
flips = obs.heads+obs.tails
```

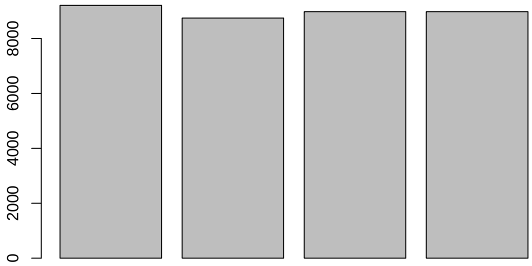
```
p=0.5
exp.heads = flips*p
exp.tails = flips*p
exp.flips = c(exp.heads,exp.tails)

teststat4 = sum((obs.flips-exp.flips)^2/exp.flips)
teststat4
```

```
## [1] 11.99421
```

```
pchisq(teststat4, df=2-1-0, lower.tail = FALSE)
```

```
## [1] 0.000533662
```



The two bar plots on the left pertain to the observed flips, whereas the two bar plots on the right pertain to the expected flips. Notice that the values are very close; however, both the test statistic and the p-value indicate that $H_0 : p = 0.5$ should be rejected. That is, there is strong evidence to suggest that the data does not fit a $bin(n, 0.5)$.

The following code is for a Z-test.

```
teststat5 = (obs.heads - flips*0.5)/sqrt(flips*0.5*0.5)
teststat5
```

```
## [1] 3.463265
```

```
pnorm(teststat5,mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 0.000266831
```

The above output agrees with the results from the χ^2 test performed before.

Problem 43B: Additional information is provided, which comes in the form of a chart with the number of heads that occurred with every 5 flips. The probabilities are binomially distributed. For example, the probability of getting 3 heads occurs $\binom{5}{3}(\frac{1}{2})^5 = \frac{10}{32}$.

```
num.heads = 0:5
obs.heads1 = c(100,524,1080,1126,655,105)
prob.heads = c(1/32,5/32,10/32,10/32,5/32,1/32)

exp.heads1 = sum(obs.heads1)*prob.heads
exp.heads1
```

```
## [1]  112.1875  560.9375 1121.8750 1121.8750  560.9375  112.1875
```

```
teststat6 = sum((obs.heads1-exp.heads1)^2/exp.heads1)
teststat6
```

```
## [1] 21.56813
```

```
pchisq(teststat6, df=length(obs.heads1)-1-0, lower.tail = FALSE)
```

```
## [1] 0.0006323943
```

The above p-value and test statistic suggest that not all the coins were fair when flipping them.

Problem 43C: The question asks if the data are consistent with the hypothesis that all five coins had the same probability of heads, but that the probability was not necessarily 0.5. To test if this is the case, we use $\hat{p}_{MLE} = \frac{\sum k}{n}$ which is the proportion of cases where a head occurred divided by the total number of flips.

```
phat1 = sum(num.heads*obs.heads1)/flips
phat1
```

```
## [1] 0.5129248
```

So $\hat{p}_{MLE}$ is given by the above value. Now the probabilities must be recalculated using this value instead.

```
new.probs = dbinom(num.heads, 5, phat1)
new.probs
```

```
## [1] 0.02741449 0.14434701 0.30401531 0.32014972 0.16857020 0.03550328
```

```
exp.heads2 = new.probs*sum(obs.heads1)
exp.heads2
```

```
## [1]   98.41801  518.20577 1091.41497 1149.33748  605.16700  127.45677
```

The above array shows the new expected amount of heads given $\hat{p}_{MLE}$. Now we calculate the Pearson test statistic and its corresponding p-value.

```
teststat7 = sum((obs.heads1 - exp.heads2)^2/obs.heads1)
teststat7
```

```
## [1] 9.287698
```

```
pchisq(teststat7, df = length(obs.heads1)-1-1, lower.tail = FALSE)
```

```
## [1] 0.05429719
```

The above p-value is somewhat large. There is moderate evidence suggesting that the goodness of fit of the $bin(n, \hat{p}_{MLE})$ distribution is not the great. Reasons why were discussed in office hours.