Stat 5421 Homework1

Anshuman Narayan

10/2/2018

Q1.2

Q1.2.a

We have 100 multiple-choice questions where each question has 4 possible answers and one of the answers is correct. The distribution of the number of correct answers is given by *Bin(100,1/4)* since the probabilty of an answer being correct is 1/4 and a Binomial distribution is defined by its number of trials(n) and the probablity of success(π). ####Q1.2.b We have n=100 and p=1/4=>0.25. The mean is given by $n * p$ and the standard deviation is given by $\sqrt{(}np(1-p)$

```
n<-100
p<-.25
mu = (n*p)
cat("Mean -> ",mu)
```

```
## Mean ->  25
```

```
sd=sqrt(n*p*(1-p))
cat(" Standard Deviation ->",sd)
```

```
##  Standard Deviation -> 4.330127
```

For a student to get 50 answers correct, he would have answered 25 questions correctly more than the mean. Given our standard deviation of 4.330127, this would be 5.7735027 times more than the standard deviation which is a surprising results as it is a definite outlier to the distribution.

Q1.2.c

The distribution of $\mathsf{n}_1, \mathsf{n}_2, \mathsf{n}_3, \mathsf{n}_4$ where n_i is the number of times a student chooses the answer i is given by the multinomial distribution with the probabilty of choosing option i as 1/4 i.e $\pi_\mathsf{j} = .25$.

Q1.2.d

The Expectation $E(\mathsf{n}_\mathsf{j})$ is given by $n\pi_\mathsf{j}$ where n is the number of trials and π_j is the probabilty of choosing option j. THe variance $var(\mathsf{n}_\mathsf{j})$ is given by $n\pi_\mathsf{j}(1-\pi_\mathsf{j})$. Covariance $Cov(n_j, n_k)$ is given by $-n\pi_j\pi_k$.

```
n<- 100
p<-.25
mu = (n*p)
var_j=n*p*(1-p)
cov_j=-n*p*p
corr_j=cov_j/sqrt(var_j*var_j)
cat("Expectation of nj ",mu)
```

```
## Expectation of nj  25
```

```
cat("\nVariance of nj",var_j)
```

```
##
## Variance of nj 18.75
```

```
cat("\nCovariance between nj and nk",cov_j)
```

```
##
## Covariance between nj and nk -6.25
```

```
cat("\nCorrelation between nh and nk",corr_j)
```

```
##
## Correlation between nh and nk -0.3333333
```

Q1.9

We have an experminent on chlorophyll inheritance in maize for 1103 seedlings, were 854 were green and 249 were yellow. We need to test the hypothesis that the ratio of green:yellow is 3:1 or that probabilty of a seedling being choosen at random is 3/4 or .75. We use the prop.test function in r which performs a t-test to check if the given number of trials and successes match the probability that we expect.

```
prop.test(854,1102,p=0.75)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  854 out of 1102, null probability 0.75
## X-squared = 3.5281, df = 1, p-value = 0.06034
## alternative hypothesis: true p is not equal to 0.75
## 95 percent confidence interval:
##  0.7488977 0.7990662
## sample estimates:
##         p
## 0.7749546
```

From our test we see that the null hypothesis probablity is 0.75 and the 95% confidence interval of p=0.75 includes our sample estimate. Thus we cannot reject the null hypothesis probability. Thus the hypothesis is the true ratio.

We now perform the Wald,Score and Likelihood ratio on the same dataset.

```
n=1103
y=854
p_0=0.75
p_hat=854/1103
zw=(p_hat-p_0)/sqrt((p_hat*(1-p_hat))/n)
cat("Wald test value ",zw)
```

```
## Wald test value  1.926562
```

```
cat("\np-value of whether zw follows chi-squared distribution with df=1 ",1-pchisq(zw,1))
```

```
##
## p-value of whether zw follows chi-squared distribution with df=1  0.1651351
```

We calculate the p-value for whether the calculated wald score follows a chi-squared distribution with degree of freedom as 1 since there is only one parameter in this model. Our p-value tells us that we cannot reject the null hypothesis that the probability is 0.75 as p>0.05 We now compute the Score statistic

```
zs=(p_hat-p_0)/sqrt((p_0*(1-p_0))/n)
cat("Score statistic",zs)
```

```
## Score statistic 1.860096
```

```
cat("\np-value of whether zs follows chi-squared distribution with df=1 ",1-pchisq(zs,1))
```

```
##
## p-value of whether zs follows chi-squared distribution with df=1  0.1726137
```

Here, our score-statistic's p-value for following a chi-squared distribution is greater than 0.05 hence we cannot reject the null hypothesis.

Now, we calculate our Likelihood ratio statistic

```
lr=2*(y*log(p_hat/p_0)+(n-y)*log((1-p_hat)/(1-p_0)))
cat("Likelihood ratio statistic ",lr)
```

```
## Likelihood ratio statistic  3.539017
```

```
cat("\np-value of whether lr follows chi-squared distribution with df=1",1-pchisq(lr,1))
```

```
##
## p-value of whether lr follows chi-squared distribution with df=1 0.05994099
```

Since our p-value is greater than 0.05(since we can approximate our p-value as 0.06) we can say that the null hypothesis cannot be rejected.

Q1.10

We have the data of the number of deaths by mule kicks observed for 10 army corps, over 20 years. This can de considered as 200 trials of observation. We assume Y to connotate the number of deaths observed for all 10 corps in a year. Since the number of death can be independent of on another we can consider Y to be a random variable. The data recorded here tabulates the data in the number of deaths by the years when those sepcific number of deaths occur. We can thus assume that the random variable Y takes on the values from the set {0,1,2,3,4} since 4 and more deaths are treated as the same. These values can be taken by Y with the unkown probabilities $\pi_0*, \pi_1*, \pi_2*, \pi_3*, \pi_4*$ With our null nypothesis, we imply that the probabilities follow a Poisson distribution. We thus can imply that our probabilities would satisfy the following relation, $\pi_j* = (\mu^j/j!)/\sum_{k=0}^{4}(\mu_*^k/k!)$ for j=0,1,2,3,4. Thus our null and alternate hypotheses can be stated as the following $H_0 : \pi_j* = (\mu^j/j!)/\sum_{k=0}^{4}(\mu_*^k/k!)$ for j=0,1,2,3,4 $H_1 : \pi_0*, \pi_1*, \pi_2*, \pi_3*, \pi_4*$ are non-negative parameters with the only condition that $\sum_j \pi_{*j} = 1$ and H_0 is false.

The log likelihood evaluated at $\pi_0*, \pi_1*, \pi_2*, \pi_3*, \pi_4*$ is $log(n!/n_1!....n_c!) + \sum_{j=1}^{5} n_j log\pi_j$. When expressed as a function of μ when H_0 is true, our log likelihood function can be written as: $L(\mu : n_1, ..., n_5) = C_1 + \sum_{j=1}^{5} n_j log((\mu^{j-1}/(j-1)!)/(\sum_{k=0}^{4}\mu^k/k!))$ This function can be simplified to the form $C_2 + log(\mu)\sum_{j=1}^{5} n_j(j-1) - log(\sum_{k=0}^{4}\mu^k/k!)\sum_{j=1}^{5} n_j$ where C_1 and C_2 are constants. We use the optimze function in R to pass our likelihood function as a parameter for which we find the argument(μ) for which our function minimizes. We just need to provide the function over which to optimize and find the range for which to search for our mean. We use the R equivalent of -$-\infty, \infty$. We can use a smaller interval by calculating the mean of all the trials by multiplying the value of y by its instances and dividing by total number of trials and basing an upper bound using that. With our new mu_hat we then go ahead and calculate our probability of y taking the different values. We then calculate the expected counts of each category. Using our observed counts and the expected counts from our calculated mean. We compute the likelihood ratio and Pearson's chi-squared and perform p-tests of whether the values follow the chi-squared distribution.

The completed R code follows.

```
#here we have the minimizing function that represents our log likelihood funciton
minusL=function(mu, n.list)
{
  val=-log(sum(mu^(0:4)/factorial(0:4)))*sum(n.list)
```

```
  val=val+log(mu) * sum(n.list * (0:4))
  return(-val)
}
#n list consists of the observed occurences of y=0,1,2,3 and 4(including >=5).
n.list=c(109,65,22,3,1)
#we run mu.hat on the (-inf,inf) interval.
# we can expect mu to lie around 0.61 as we calculate (0*109+1*65+2*22+3*3+4*1)/200 which is 0.61
mu.hat=optimize(f=minusL, interval=c(1e-10,1e10), n.list=n.list)$min
cat("\nMean over larger interval",mu.hat)
```

```
##
## Mean over larger interval 0.6119384
```

```
mu.hat=optimize(f=minusL, interval=c(1e-10,3),  tol=1e-10, n.list=n.list)$min
cat("\nMean over shorter interval",mu.hat)
```

```
##
## Mean over shorter interval 0.6119398
```

```
#calculating the probability of y taking each of the values using the mu.hat that we calculate.
probs=mu.hat^(0:4) / (factorial(0:4) * sum(mu.hat^(0:4)/factorial(0:4)))
cat("Probablity of each category using the mu we just calculated\n",probs)
```

```
## Probablity of each category using the mu we just calculated
##  0.5425318 0.3319968 0.101581 0.02072049 0.003169923
```

```
#the expected number of counts for each category using the number of
#trials and the probabilities of each category.
expected.counts=200*probs
cat("\nThe expected counts of each category using the probablity we just calculated\n",expected.counts)
```

```
##
## The expected counts of each category using the probablity we just calculated
##  108.5064 66.39936 20.3162 4.144098 0.6339846
```

```
#our likelihood ratio statistic is calculated by summing over our
#categories the result 2*sum of observed occurences*log(observed occurrences/expected occurrences)
g.sq=2*sum(n.list * log(n.list/(200*probs)) )
cat("\nLog Likelihood ratio test statistic",g.sq)
```

```
##
## Log Likelihood ratio test statistic 0.6969812
```

```
cat("\np-value of g.sp following the chi-squared distribution",1-pchisq(g.sq,3))
```

```
##
## p-value of g.sp following the chi-squared distribution 0.8739138
```

```
#the Pearson chi-squared test statistic is calculated as the sum of the difference
#between observed and expected occurences divided by the expected occurences
x.sq=sum((n.list-expected.counts)^2/expected.counts)
cat("\nPearson Chi-squared statistic",x.sq)
```

```
##
## Pearson Chi-squared statistic 0.6984603
```

```
cat("\np=value of x.sq following the chi-squared distribution",1-pchisq(x.sq,3))
```

```
##
```

```
## p=value of x.sq following the chi-squared distribution 0.873566
```

Seeing the p-values of the two tests we can say that both test fail to reject our null hypothesis as both p-values are >0.05. Therefore we can say that our hypothesis that the random variable Y follows the Poisson distribution is true.