hw2

Mohammad Mottahedi

January 26, 2016

```
set.seed(1234)
```

Problem 5 (30 points). Let X be the number of Chargers (San Diegos NFL team) fans observed in a random sample of n = 30 graduate students at Penn State. It is reasonable to assume that X ~ Bin(30; p). The true proportion p is unknown, but it is likely to be small (especially in comparison to potential number of Pittsburgh Steelers fans). Note: Some helpful R and SAS code is provided on ANGEL; see hw1. R and hw1_ help. sas . For plotting likelihood there is only R code.

(d) Plot the log-likelihood function with respect to the transformed parameter over a range of values from. Comment on this plot in comparison to the one in part (b).

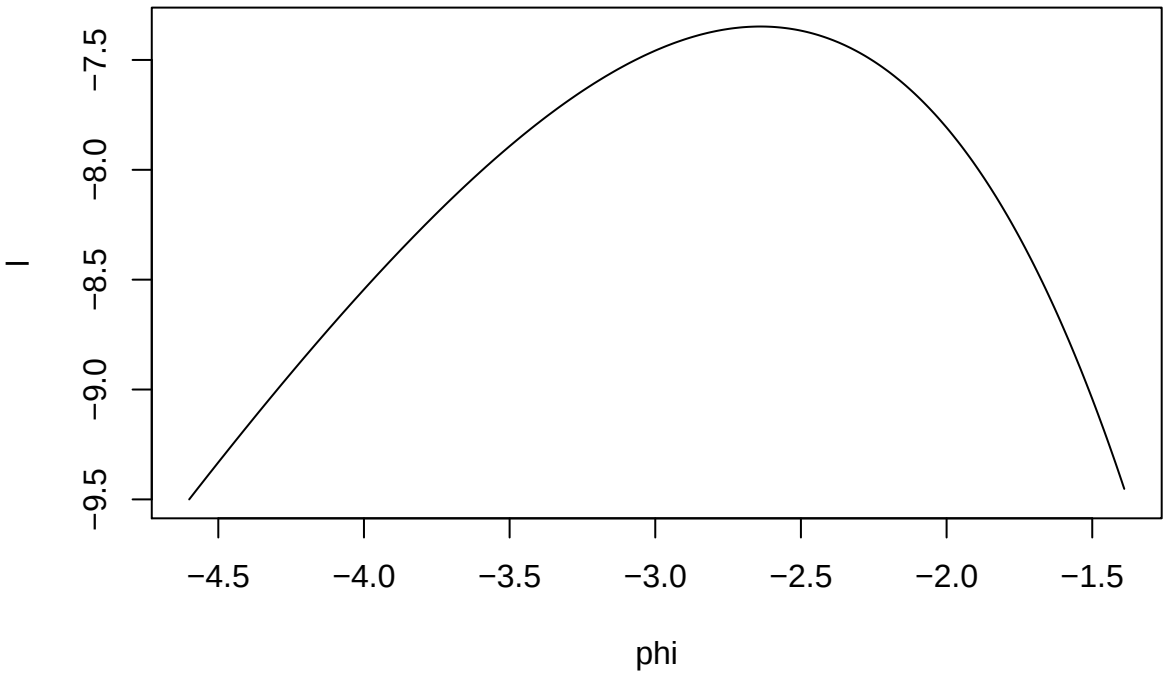
logit transformation: $\phi = log(\frac{p}{1-p})$

$l(p|x) = xlog(p) + (n-x)log(1-p)$

$l(\phi|x) = x\phi + nlog(\frac{1}{1+e^{\phi}})$

$x = 2, n = 30$

```
phi = seq(-4.6,-1.39, .01)
l = 2 * (phi) + 30 * log(1 / (1 + exp(phi)) )
plot(phi, l, type = 'l')
```




```
#par(new = T)
#plot(p, y, xlab="", ylab="log-likelihood", type="l")
```

The transformed scale is less skewed compared to the plot in part C.

(e) What is the 95% confidence interval for based on the expected information, and then transform the end points back to scale of p.

Fisher information for logit ϕ

$I(\hat{\phi}) = \frac{I(\hat{\theta})}{[\phi'(\hat{\theta})]^2}$

$\phi = \phi \pm 1.96\sqrt{\frac{1}{I(\hat{\phi})}}$

$\phi'(\hat{\theta}) = \frac{1}{\theta(1-\theta)}$

$\phi = -2.6390573 \pm 1.960.7319251 = (-4.0736304, -1.2044842)$

transforming the confidence interval to original scale:

$(0.0167308, 0.2306785)$

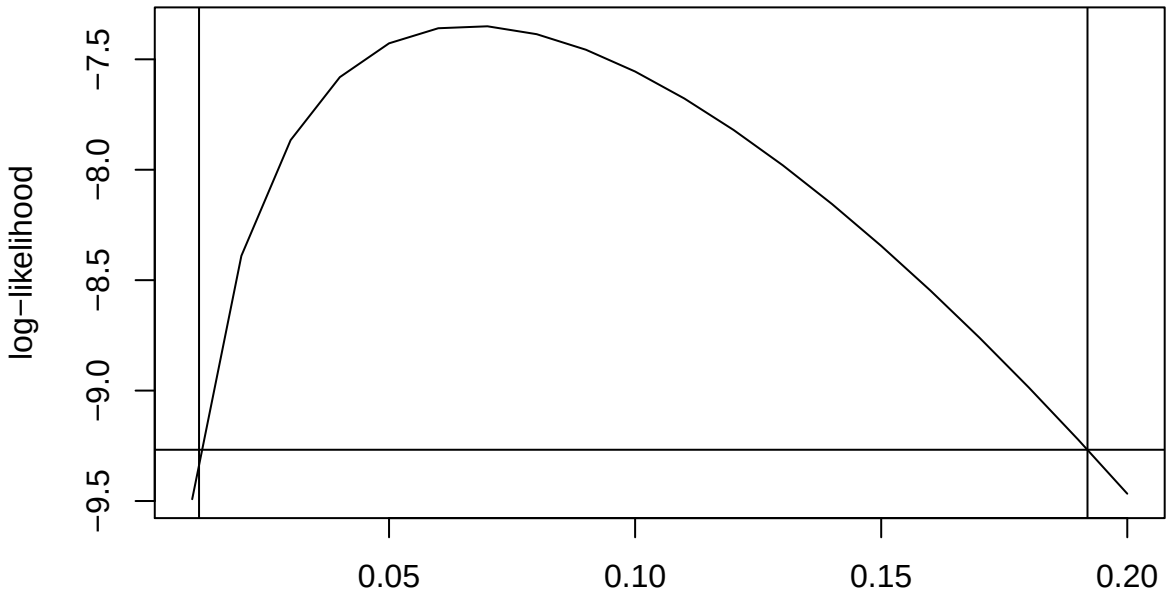
(f) Now find an approximate 95% confidence interval based on the LR method, and compare it to the two other intervals calculated above. Which one is the most appropriate and why?

$l(p|x) = xlog(p) + (n-x)log(1-p)$

$x = 2, n = 30$

$l(p|x) = 2log(p) + (28)log(1-p)$

```
p = seq(from=0.01, to=0.2, by=0.01)
y = 2 * log(p) + 28 * log(1-p)
plot(p, y, xlab="", ylab="log-likelihood", type="l")
abline(-9.2679008,0)
abline(v = 0.1919152)
abline(v = 0.01139688)
```



log-likelihood at $l(\hat{p} = 0.0666667) = -7.3479008$

LR confidence interval: $2[l(\hat{\theta}|x) - l(\theta|x)] \leq 3.85$

$l(\theta|x) \geq l(\hat{\theta}|x) - 1.92 = -9.2679008$

```
l = function(p) {2 * log(p) + 28 * log(1 - p) + 9.2679008}
uniroot(l, c(0.01,0.1))[1]
```

```
## $root
## [1] 0.01139688
```

```
uniroot(l, c(0.1,0.3))[1]
```

```
## $root
## [1] 0.1919152
```

95% confidence interval: $(0.01139688, 0.1919152)$

1. (20 pts) For each of the following scenarios, say whether the binomial distribution is a reasonable probability model for X. If yes, explain why you think the assumptions of binomial are satisifed. If no, explain how you think the assumptions are violated. How could you test data to see if your answer was reasonable?

(a) After warming up with a few throws, you throw a dart at a dartboard 20 times. Let X be the number of times you hit it out of 30 flips.

Yes, the number of trials are fixed ($n = 20$) and each event has two possible outcome (hit, miss) and the outcome of each throw is independent of other throws.

(b) After a warm-up, you throw a dart at a dartboard 20 times at a competition, then 1000 times as practice, then again for 20 times at a competition. Let X be the number of times you hit it out of 40 competition throws.

No, the first and second 20 throw are individually binomial but if we assume that the 1000 practice throws will change the probability of success in each throw the combined 40 throws can be considered as binomial since its violates independent and identical events/trials assumption.

(c) A survey firm draws a random sample of 15 households in State College. Two adults from each household are asked, "Are you generally optimistic about the upcoming years if Donald Trump is elected President?." Let X be the number out of 30 who said "Yes".

No, since the opinion of the first sampled adult in the one households suggest similar answer to the survey by the second adult the trials are not independent of each other.

(d) A survey firm draws a random sample of 30 households in State College containing two or more adults. In each one, one adult member of the household is asked, "Are you generally optimistic about the upcoming years if Donald Trump is elected President?." Let X be the number out of 30 who said "Yes".

Yes, since the adults in the sample have equal chance to be in the sample the assumption for bionomial distribution holds.

2. (20 pts) For each of the following scenarios, say whether a Poisson model for the sample (X 1 , X 2 , . . . , X n) may be appropriate.

(a) Let say that in the Pennsylvania state lottery game Mega Millions, any player may win with probability 1 out of 2 million independently of other players. Also, sup- pose that the number of players each week is always equal to a constant N . Let X 1 , X 2 , . . . , X n be the number of Mega Millions winners in the first n weeks of 2010.

Yes, the Poisson distribution assumptions holds. Events are independent (if someones wins it doesn't affect the chance of wining for others). the homogeniety hold beacause the number of players is constant each weak and the each player has equal probability to win.

(b) Now suppose the scenario as above, but the number of players N varies substantially from week to week.

if N varies and probability of wining is still 1 in 2 Million, the expected number of winners each week cannot be the same anymore and the assumtpion of homogeneity is violated.

(c) Let X 1 , X 2 , . . . , X 12 be the number of cars arriving at a toll booth on a major highway during 12 consecutive hours in a single day.

No, it is expected that the 12 consecutive time interval have different means due to rush hours in the morning and afternoon or other variables.

(d) Let X 1 , X 2 , . . . , X 48 be the number of cars arriving at the toll booth between hours of 1 and 2 pm on 48 Fridays in a single year (excluding holidays).

Yes, since each time interval is defined between 1 and 2 pm every friday we expect to have equal mean during each time interval.

3. (20 pts) Census figures show that the racial/ethnic distribution of persons eligible for jury duty in a large urban county is given by this table:

Black	Hispanic	Other
20%	15%	65%

From this county, 182 persons were summoned to appear for jury duty on a particular day. The observed racial/ethnic distribution of this sample is shown below.

Black	Hispanic	Other
45	35	102

Is it likely that this is a random sample from the eligible population? If not, how does the sample differ from the population? (Be sure to calculate χ^2 , G^2 , residuals, and comment on the appropriateness of the χ^2 approximation, that is of χ^2 sampling distribution.)

$H_0 : (\pi_1, \pi_2, \pi_3) = (0.2, 0.15, 0.65)$

$H_A : (\pi_1, \pi_2, \pi_3) \neq (0.2, 0.15, 0.65)$

$\chi^2 = \sum_{j=1}^{k} \frac{(X_j - n\hat{\pi_j})^2}{n\hat{\pi_j}}$

```
test <- chisq.test(c(45, 35, 102), p= c(0.2, 0.15, 0.65))
```

```
s <- c(45, 35, 102); count <- 182 * c(0.2, 0.15, 0.65)

x2 <- sum( (s - count)^2 / count )
g2 <- 2 * sum( s * log( s / count )  )
```

$\chi^2 = 6.4495633$, $G^2 = 6.2375341$

χ^2 and G^2 are close.

$dof = 2$

$p(\chi_2^2 \geq 6.4495633) = 0.0397645$

$p(G_2^2 \geq 6.2375341) = 0.0442116$

p-values for both G^2 and χ^2 are less than $\alpha = 0.05$.

We can reject the null hypothesis and it's not likely that the sample is from the eligible population.

looking at the residuals we see that the absolute value of the residuals are close to 1.4 and it's slightly higher for hispanic and others.

```
df <- data.frame(residuals=test$residuals, observed = test$observed, expected= test$expected)
df
```

```
##   residuals observed expected
## 1  1.425436       45     36.4
## 2  1.473701       35     27.3
## 3 -1.498633      102    118.3
```

The values of χ^2 and G^2 are large so the model is not tha good and we cannot say that the sample is from the eligible population.

4. (20 pts) R. A. Fisher, in a now-famous analysis of the results of Mendel's breeding experiments, concluded that the data were too good to be true as the data fit Mendel's theory far better than one would expect. In one pea breeding experiment, Mendel hypothesized that four categories of peas: smooth yellow, wrinkled yellow, smooth green, and wrinkled green, should occur with relative frequencies 9:3:3:1. The observed frequencies reported by Mendel from this trial are shown below.

Type of pea	count
Smooth yellow	315
Wrinkled yellow	101
Smooth green	108
Wrinkled green	32

Perform a goodness-of-fit test to show that these data fit the theory too well. Perform your analysis in the software of your choice and make sure to included relevant edited portions of the output.

$H_0 : (\pi_{sy}, \pi_{wy}, \pi_{sg}, \pi_{wg}) = (0.56250.18750.18750.0625)$

$H_A : (\pi_{sy}, \pi_{wy}, \pi_{sg}, \pi_{wg}) \neq (0.56250.18750.18750.0625)$

$\chi^2 = \sum_{j=1}^{k} \frac{(X_j - n\hat{\pi_j})^2}{n\hat{\pi_j}}$

```
s <- c(315, 101, 108, 32); count <- sum(315, 101, 108, 32) * c(0.5625, 0.1875, 0.1875, 0.0625)
```

```
test <- chisq.test(s, p= c(0.5625, 0.1875, 0.1875, 0.0625))
test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  s
## X-squared = 0.47002, df = 3, p-value = 0.9254
```

```
x2 <- sum( (s - count)^2 / count )
g2 <- 2 * sum( s * log( s / count )  )
```

$\chi^2 = 0.470024$, $G^2 = 0.4754452$

χ^2 and G^2 are close.

$dof = 3$

$p(\chi^2_3 \geq 0.470024) = 0.9254259$

$p(G^2_3 \geq 0.4754452) = 0.9242519$

The the pearson and deviance statistics are so small that we can the it seems to good and the data may have been fabricated or altered.

5. (20 pts) Below is a table of 100 values of a random variable X drawn from a discrete distribution. For example, X could be the number of cars passing a point on a road from 1:00pm to 1:10pm on 100 consecutive weekdays, or the number of orders arriving at an exchange from 1:00:00.00pm to 1:00:00.01pm during 100 consecutive weekdays.

x	count
0	22
1	36
2	20
3	16
4	3
≥ 5	3

Perform the χ^2 and G^2 goodness-of-fit tests to see if these data are Poisson. The last category (≥ 5), indicates that outcomes where the count was 6 and higher were rare so we collapsed them to the category (x ≥ 5). Also compute and explain the MLE of the rate and the expected counts.

estimating the mean of Poisson ditribution using MLE:

$\hat{\lambda} = n^{-1} \sum_{i=1}^{n} y_i$

The cell ≥ 5 is broken into two two cells (=5, =6) with two and one observation respectively.

The expected value is $E_i = n\hat{\pi}_i$.

```
ob <- c(22, 36, 20, 16, 3, 3)

lambda_hat <- c( 0 * 22 + 1 * 36 + 2 * 20 + 3 * 16 + 4 * 3 + 2 * 5 + 1 * 6) / 100

p_hat <- dpois(c(0,1,2,3,4), lambda_hat)
p_hat <- c(p_hat, 1-sum(p_hat))
```

```r
ex <- 100 * p_hat


X2 <- sum((ob-ex)^2/ex)
G2 <- 2*sum(ob*log(ob/ex))

#dof k-1 -d = 6 - 1 - 1
p_X2 <- 1-pchisq(X2,4)
p_G2 <- 1-pchisq(G2,4)
```

$\chi^2 = 3.4014129, G^2 = 3.4558682$

$p(\chi^2_2 \geq 3.4014129) = 0.4930261$

$p(G^2_2 \geq 3.4558682) = 0.4846203$

Since both p-values are less 0.05 there's evidence that the data are not Poisson.