JOE SEIDEL

STAT 244 HOMEWORK 7

Question 1

Suppose X follows a geometric distribution with parameter p.

1. Derive the likelihood ratio for testing the hypothesis $p = p_0$ versus the alternative $p \neq p_0$.

   THE variable X has PMF given

   $$p(X = x) = p(1-p)^{x-1} \text{ , } x = 1, 2, ...$$

   We have a composite hypothesis a generalized likelihood ratio test is in order.

   $$\Lambda = \frac{\max_{p \in w_0}[L(p)]}{\max_{p \in \Omega}[L(p)]}$$

   Where the rejection region consists of small values for Λ. In this case $w_0 = \{p_0\}$ and $\Omega = \{0 < p < 1\}$

   $$\max_{p \in w_0}[L(p)] = p_0(1-p_0)^{x-1}.$$

   For the denomitor we have to maximize the likelihood for $p \in \Omega$. Which is the mle of $f(x \mid p) = p(1-p)^x$, $\hat{p} = \frac{1}{\overline{X}}$. Which makes

   $$\max_{p \in \Omega}[L(p)] = \frac{1}{\overline{X}}\left(1 - \frac{1}{\overline{X}}\right)^{x-1}.$$

   Therefore

   It's important to note that I've set this up observing only 1 X, otherwise this would look slightly different, in fact I will end up replacing x for $\overline{X}$ soon.

   $$\Lambda = \frac{p_0 x(1-p_0)^{x-1}}{(1-\frac{1}{x})^{x-1}}$$

   Which is the generalized likelihood ratio that will test the hypothesis.

2. For $p_0 = 0.01$, by some combination of numerical experimentation and mathematical analysis, find the set of possible valyes of x for X for which the likelihood ratio is less than 0.1.

   When graphing the the function for Λ with $p_0 = .01$, its clear that there are two values of x that make likelihood ration less than .1.

   I used a computer to find the values. Since $x \in \mathbb{N}$, $x \leq 4$ and $x \geq 488$.

   Hence for $x \leq 4$ we have likelihood ratio less than 0.1.

3. Find the probability of Type 1 error for the test the rejects $p_0 = 0.01$ when the likelihood ratio is less than 0.1. Find the power of this test when $p = 0.5$. Find the power of ths test when $p = 0.001$.

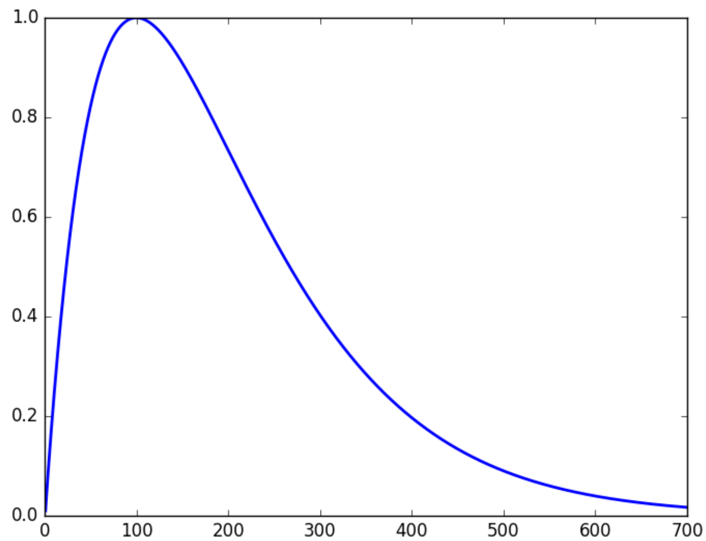


Figure 1: The likelihood function when $p_0 = .01$

To FIND α given $\Lambda \leq .1$ and $H_0 : p = .01$ take

$$\begin{aligned}\alpha &= P(x \leq 4 \mid p = .01) + P(x \geq 488 \mid p = .01)\\ &= \sum_{i=1}^{4} P(x = i \mid p = .01) + 1 - P(x \leq 488 \mid p = .01)\\ &= .04 + .007\\ &= .047\end{aligned}$$

Next I need to find the the power f the test when $p = .5$ and $p = .0001$.

Per Reintiz's office hours, this is to say, type two error given these alternative values for p.

Recall, $\pi = P(H_1 \mid H_1)$.

When $p = .5$

$$\begin{aligned}\pi &= P(H_1 \mid H_1)\\ &= P(x \leq 4 \cup x \geq 488 \mid p = .5)\\ &= \sum_{i=1}^{4} P(x = i \mid p = .5) + 1 - P(x \leq 488 \mid p = .5)\\ &= .9375\end{aligned}$$

When $p = .001$

$$\begin{aligned}\pi &= P(H_1 \mid H_1) \\ &= P(x \leq 4 \cup x \geq 488 \mid p = .001) \\ &= \sum_{i=1}^{4} P(x = i \mid p = .001) + 1 - P(x \leq 488 \mid p = .001) \\ &= .62\end{aligned}$$

Question 2 Rice 9.12

Let $X_1, ..., X_n$ be a random sample from an exponential distribution with the density function $f(x \mid \theta) = \theta e^{-\theta x}$. Derive a likelihood ratio test of H_0: $\theta = \theta_0$ versus H_A: $\theta \neq \theta_0$, and show that the rejection region is of the form $\{\overline{X} e^{-\theta_0 \overline{X}} \leq c\}$.

FIRST, recall that the mle of $L(\theta)$ is $\hat{\theta} = \frac{1}{\overline{X}}$. Set up the likelihood ratio test

I've found this in previous homeworks.

$$\Lambda = \frac{\max_{p \in w_0}[L(p)]}{\max_{p \in \Omega}[L(p)]}.$$

The numerator will be

$$L(\theta_0) = \prod_{i=1}^{n} \theta e^{-\theta_0 x_i}$$

and the denomitaor

$$L(\hat{\theta}) = \prod_{i=1}^{n} \frac{1}{\overline{X}} e^{-\frac{x_i}{\overline{X}}}.$$

Then

$$\begin{aligned}\Lambda &= \frac{\prod_{i=1}^{n} \theta e^{-\theta_0 x_i}}{\prod_{i=1}^{n} \frac{1}{\overline{X}} e^{-\frac{x_i}{\overline{X}}}} \\ &= \frac{\theta_0^n e^{-\theta_0 n \overline{X}}}{\frac{1}{\overline{X}^n} e^{-\frac{\overline{X} n}{\overline{X}}}} \\ &= \frac{\theta_0^n \overline{X}^n e^{-\theta_0 n \overline{X}}}{e^{-n}} \\ &= \left(e \theta_0 \overline{X} e^{(-\theta_0 \overline{X})}\right)^n\end{aligned}$$

Where H_0 is rejected when Λ is small. Since e, n, θ are positive, Λ is small when $\overline{X} e^{(-\theta_0 \overline{X})}$ is small.

$$\left(e\theta_0\overline{X}e^{(-\theta_0\overline{X})}\right)^n < c_1$$
$$e\theta_0\overline{X}e^{(-\theta_0\overline{X})} < c_1^{\frac{1}{n}}$$
$$\overline{X}e^{(-\theta_0\overline{X})} < \frac{c_1^{\frac{1}{n}}}{\theta_0 e}$$

Therefore, we see the rejection region takes the form $\overline{X}e^{(-\theta_0\overline{X})} \leq c = \frac{c_1^{\frac{1}{n}}}{\theta_0 e}$.

Question 3 Rice 9.13

Suppose, to be specific, that in problem 12, $\theta_0 = 1$, $n = 10$, and that $\alpha = .05$. In order to use the test, we must find the appropriate value of c.

1. Show that rejection region is of the form $\{\overline{X} \leq x_0\} \cup \{\overline{X} \geq x_1\}$, where x_0 and x_1 are deterimined by c.

   NOW that we are given a value for θ_0 we have

   $$f(x \mid \theta_0) = xe^{-x}.$$

   But we also found in the previous question that our test rejects when
   $$f(\overline{X} \mid \theta_0) = \overline{X}e^{(-\overline{X})}$$
   is small, specifically when less than c. To see why it takes the form $\{\overline{X} \leq x_0\} \cup \{\overline{X} \geq x_1\}$, consider the graph of the function.

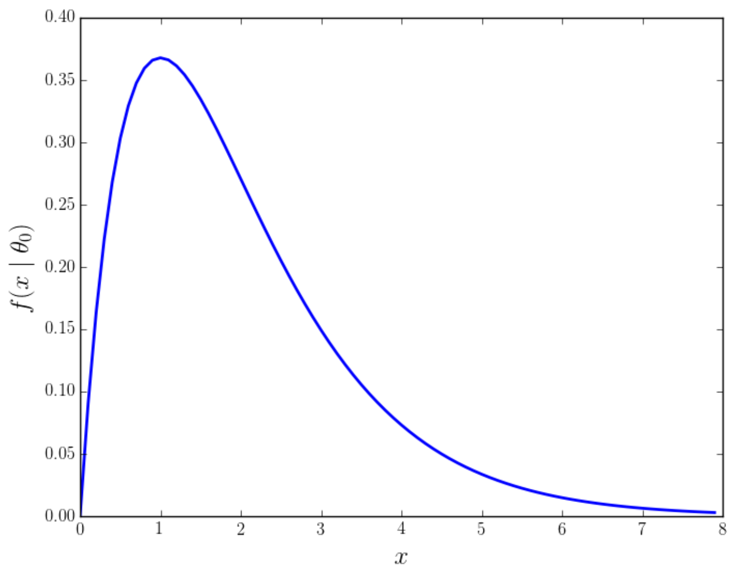
   There would be a c chosen that corresponds with the Y axis of the graph and a horizontal line would interect the function at two values of x.

2. Explain why c should be chosen so that $P(\overline{X}e^{(-\overline{X})} \leq c) = .05$.

   THIS is just a restatement of Type I error under Neyman-Pearson.

   $$Pr(\text{Reject } H_0 \mid H_0) = Pr(x \in [0, c] \mid \theta_0)$$

   Which can be worded to say, the probability that x falls in the rejection zone. Furthermore, we found that $P(\overline{X}e^{(-\overline{X})})$ provides a rejection zone for Λ. What remains is to determine how willing we are to make Type 1 error.



If we want $\alpha = .05$ then we should set

$$Pr(x \in [0, c] \mid \theta_0) = Pr(f(\overline{X} \mid \theta)) < c) = .05$$

to determine what c should be.

3. Explain why $\sum_{i=1}^{1} 0X_i$ and hence $\overline{X}$ follow gamma distributions when $\theta_0 = 1$. How could this knowledge be used to choose c?

   Under H_0: $\theta_0 = 1$ $X_i \sim Exponetial(1)$ which is a special ase of $\Gamma(1, \lambda)$ or in this particular case $\Gamma(1, 1)$. On the last homework, I found that $\sum_{i=1}^{n} X_i \sim \Gamma(n, 1)$ and $\overline{X} \sim \Gamma(n, n)$. Knowing the exact distribution, means I could compute the exact value of c for which 95 coverage for acceptable values of H_0.

   In this day and age, a computer would be the easiest way to do this. The gist would be first solve $f(\overline{X}) = c$ to get $x_0(c)$ and $x_1(c)$. Then solve

   $$\alpha(c) = F(x_0(c)) + 1 - F(x_1(c))$$

   Where $F(x)$ is cdf of $\Gamma(10, 10)$.

4. Suppose you hadn't thought of the preceding fact. Explain how you could determine a good approximation to c by generating random numbers on a computer (simulation).

   GENERATE a bunch of samples from $Exponential(1)$ of size $n = 10$. Then compute $\overline{X}e^{\overline{X}}$ for each sample. Sort the result and set the cuttoff as the value indexed at the 5% of the number of samples generated. For example, if the results are stored in some list of size 10000 then you'd take value indexed at list[500] as the cuttoff value for a close approximation.

Figure 2: James Mill, the economist. Apparently, this guy spent his life in love with a woman who had already been promised to another. After years of her marriage, the husband died. The two finally got together. Then Like 6 months passed and she died. Bummer, right?

Question 4

Call "This, it thus, and" Class I workds; Class II is "everything else". For each of 215 groups of 5 of James Mill's sentences, the number of Class I words was counted.

Test whether a Binomial Distribution $(n = 5, \theta)$ fits these data.

FIRST, knowing the MLE of a binomial will be useful.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \binom{n}{x_i} \theta^{x_i}(1-\theta)^{n-x_i} \\ &= \prod_{i=1}^{n} \binom{n}{x_i} \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

Taking the log and the derivative we have

$$\frac{d}{d\theta} \log L(\theta) = \frac{\overline{X}n}{\theta} + \frac{n - \overline{X}n}{\theta - 1}$$

Setting the above equal to zero and solving for θ give the mle.

$$\hat{\theta} = \overline{X}$$

Using the table we can compute the mle, $\hat{\theta} = 339/215(5) = .31$. With this a few rows can be added to the table.

Expected is calculated

$$E_i = \binom{n}{x_i} \hat{\theta}_i^x (1-\hat{\theta})^{n-x_i} \cdot 225 \text{ for } x_i = 0, 1, 2, .., n = 5$$

Then calculate each component of

$$X^2 = \sum_{i=1}^{m} \frac{[x_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$$

Where $m = 6$ cells.

No. Class I words	0	1	2	3	4	5
No. groups(observed)	87	11	51	42	20	4
No. groups(expected)	34	76	68	30	7	1
Component of Chi-Squared	82.62	55.59	4.25	4.8	24.14	9.0

Table 1: Class I words

Summing up the last row of the table, the chi-square statistic $X^2 = 180.4$ with 4 degrees of freedom (6 cells and one parameter was estimated from the data). Finally, our rejection region is $X^2 > \chi^2_4(\alpha)$. If we specify $\alpha = .005$ (giving our null hypothesis the best chance) our test is

$$180.4 = X^2 > \chi^2_4(.005) = 14.86$$

Which rejects the null hypothesis that $X \sim Bin(n = 5, \theta)$.

Question 5

The members of a community are classified by Blood type:

O	A	B	BA	Total
121	120	79	33	353

Table 2: Community blood types

Theory has is that the probabilities of those types depends on gene frequency parameters r, p, q, where $r + p + q = 1$ and $P(\text{"O"}) = r^2$, $P(\text{"A"}) = p^2 + 2pr$, $P(\text{"B"}) = q^2 + 2qr$, and $P(\text{"AB"}) = 2pq$. Using numerical methods (that is, a method such as that described in Chapter 5 of Stigler's notes) we can find the MLEs of r, p, q; they are

$$\hat{r} = 0.580$$
$$\hat{p} = 0.246$$
$$\hat{q} = 0.173$$

Test if the community fits the theory.

Can't afford to estimate three parameters with only 4 cells!

SINCE there are only 4 cells, I cannot afford to estimate all 3 parameters. Instead, I'll just estimate 2 and the substute the third $r' = 1 - .246 - .173 = .581$

	O	A	B	BA	Total
observed	121	120	79	33	353
expected	119.15	122.26	81.5	30.5	353.41
Comp of Chi-Square	0.02	0.04	0.08	0.2	.342

Table 3: Blood Types with Chi-Square

The table above was completed using similiar methods described in the previous question. The chi-squared statistic is $X^2 = .342$. We have 1 degree of freedom $(4 - 1 - 2 = 1)$. Choose $\alpha = .1$, then

$$.342 = X^2 < \chi^2_1(.1) = 2.71$$

Which supports the null that the community fits the theory.

Question 6

Are finger print patterns genetic, or are the developmental? In 1892, Francis Galton compiled the following table on the relationship between the patterns on the same finger of 105 sibling pairs. Test the hypothesis that the patterns are independent for example, that knowing one sibiling (A) has Whorl on the finger does not help in predicting the pattern of the other (B).

Columbus sailed the ocean blue...or was it 1492?

	A Children			
B Children	Arches	Loops	Whorls	Totals
Arches	5	12	2	19
Loops	4	42	15	61
Whorls	1	14	10	25
Totals	10	68	27	105

Table 4: Galton's Siblings

CALCULATE χ^2 squared statistic

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\left[X_{ij} - \left(\frac{X_{i+}X_{+j}}{n}\right)\right]^2}{\left(\frac{X_{i+}X_{+j}}{n}\right)}$$

Which computes to $\chi^2 = 11.16$. We check this value with the table, given $(r-1)(c-1) = 2 \cdot 2 = 4$ degress of freedom. From the Chi-Squared table, we'd have a p-value $< .025$ which provides fairly strong evidence against the null hypothesis that the patterns are independent.

Small p-values provide strong evidence against Null when chi-square testing. P-values are the chance of false positives, so when they are small the null unlikeyl.

Question 7

For the Bortkiewicz Death by Horsekick Data, test the hypothesis that the data follow a Poisson distribution. You should group the count for "4 or more" a one category.

FIRST, estimate λ. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda}\lambda^{\overline{X}n}\prod_{i=1}^{n}\frac{1}{x_i!} \end{aligned}$$

Then take the log $L(\lambda)$

$$\log L(\lambda) = -n\lambda + \overline{X}n\log(\lambda) + \log(\prod_{i=1}^{n}\frac{1}{x_i!})$$

Next, differentiate

$$\frac{d \log L(\lambda)}{d\lambda} = -n + \frac{\overline{X}n}{\lambda}$$

Then, by setting the above equal to 0, solve for λ to get

$$\hat{\lambda} = \overline{X}$$

If this was a problem about MLE, I'd take second derivative to show this maximizes. I did this in previous homeworks so I'm not doing it here. An important step to remember, none-the-less.

To test the hypothesis, we'll create a X^2 statistic using expected and compaired values. To find expected values, compute $\hat{\theta}$ given the data.

$$\hat{\theta} = \frac{0 \cdot 144 + 1 \cdot 99 + 2 \cdot 32 + 3 \cdot 11 + 4 \cdot 2}{280} = .7$$

No. Deaths	0	1	2	3	≥ 4
Observed	144	91	32	11	2
Expected	139.04	97.33	34.07	7.95	1.61
Comp Chi-Squared	0.18	0.41	0.13	1.17	0.10

Table 5: Horse Kicks

From the table we have chi-squared statistic $X^2 = 1.98$. With $k - 1 - 1 = 5 - 1 - 1 = 3$ degress of freedom, $\chi_3^4(.1) = 6.25$.

$$X^2 = 1.98 < \chi_3^4(.1) = 6.25$$

The value X^2 is even less than the expected value $k = 3$. Additionally, looking at the Chi-Square table we confirm that the p-value is greater than .1 and we accept the null hypothesis that the data follow Poisson distribution.

Question 8

An American roulette whele is spun $n = 3880$ times in order to test if it is fair (i.e. to test if each slot has probability $\frac{1}{38}$). Suppose that each of the 36 number slots (1,2,...,36) comesup exactly 100 timesand each of "0" and "00" comes up 140 times.

1. Test at the 5% level using the χ^2 test if the wheel is fair.

THE NULL hypothesis is that any number appears with probability $\theta = \frac{1}{38}$. We can easily calculate the expected value for each slot, using $n\theta = 102.1$.

Using

$$X^2 = \sum_{i=1}^{m} \frac{[x_i - np_i(\theta)]^2}{np_i(\theta)}$$

Computed, $X^2 = 29.688$, which is even less that our expected value $k - 1 = 37$. Furthemore,

$$X^2 = 29.668 < \chi^2_{37}(.05) = 52.172$$

Which provides no evidence to reject the null hypothesis.

2. Now suppose that before you had looked at the data you had suspected that the numbred slots were less likely that "0" or "00" and you had decided to test the binomial hypothesis H_0: P("0" or "00") $= \frac{2}{38}$ versus H_1: P("0" or "00") $> \frac{2}{38}$. We know that the UMP test of these hypothesis rejects H_0 if Z(=total number of "0" and "00"s) is greater than C, where C is chosen for a level 0.05 test. Use that fact that under H_0 Z has approximately a Normal $N(\frac{2n}{38}, \frac{2n}{38} \cdot \frac{36}{38})$ distribution to find C and perform this test.

The hypothesis to test is H_0: $\theta_0 = \frac{2}{38}$ versus H_1: $\theta_1 > \theta_0$. The goal is to compare $Z = 280$ (the observed number of times that "0" or "00" appear) with some critical value C. In particular, the test well be set up such that if $Z > C$ then reject H_0.

The problem states that Z follows a normal distribution. Using $n = 3880$ this is computed to $Z \sim N(204.21, 193.46)$. With the information about Z's disrtibution and $\alpha = .5$ solve for C.

$$\begin{aligned}
.05 &= P(Z > C \mid H_0) \\
&= P\Big[\Big(\frac{280 - 204.21}{13.9}\Big) > \Big(\frac{C - 204.21}{13.9}\Big) \\
&= P\Big[Z^* > \Big(\frac{C - 204.21}{13.9}\Big)\Big]
\end{aligned}$$

Since Z^* follows standard normal

$$\begin{aligned}
C &= 204 + z_{95}(13.9) \\
&= 204 + 1.645(13.9) \\
&= 220.76
\end{aligned}$$

Which leads us to reject the null, since $Z = 280 > C = 220.76$.

3. Compare the result in (2) with that in (1).

In item (1) we did not reject the H_0, contrary to item 2. The deviation in the data would appear to be large and based on a focused hypothesis test, such as that in (2) enough to reject the hypothesis that the wheel is fair. However, in (1) we see that the Chi-Squared test can help mitigate being mislead by deviations selected as large. I'll conclude from this that, ultimately, one needs to have a good understanding of the world as it is to determine which test best fits the problem.