Homework 6 Solutions
Statistics 200B
Due Mar 14, 2019

1. Let $X_1, \ldots, X_n$ be *iid* with density $f(x;\beta) = \beta e^{-\beta x}$ for $x > 0$ and $\beta > 0$. Find the asymptotic (large sample) likelihood ratio test of size α for $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$.

   Solutions:

   The MLE of β can be calculated as follows.

$$L_n(\beta) = \prod_{i=1}^{n} f(X_i;\beta) = \beta^n e^{-\beta \sum_{i=1}^n X_i}.$$

$$\Rightarrow \ell_n(\beta) = \log L_n(\beta) = n \log \beta - \beta \sum_{i=1}^{n} X_i.$$

$$\Rightarrow \frac{d\ell_n}{d\beta}(\beta) = \frac{n}{\beta} - \sum_{i=1}^{n} X_i.$$

$$\frac{d\ell_n}{d\beta}(\beta) = 0 \Rightarrow \hat{\beta} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}_n}.$$

$$\frac{d^2\ell_n}{d\beta^2}(\beta) = -\frac{n}{\beta^2} < 0.$$

   So the MLE is $\hat{\beta} = \frac{n}{\sum_{i=1}^n X_i}$.

   The likelihood ratio statistic is

$$\lambda = 2 \log \frac{L_n(\hat{\beta})}{L_n(\beta_0)} = 2(\ell_n(\hat{\beta}) - \ell_n(\beta_0)) = 2\left[-n \log \bar{X}_n - n - \left(n \log \beta_0 - n\beta_0 \bar{X}_n\right)\right]$$
$$= 2n\left[\beta_0 \bar{X}_n - \log(\beta_0 \bar{X}_n) - 1\right].$$

   The asymptotic distribution of λ is χ^2_1. So the rejection region is

$$\lambda > \chi^2_{1,\alpha},$$

   where $\chi^2_{1,\alpha}$ is the $(1-\alpha)$th quantile of χ^2_1 distribution.

2. Let X and Y be two random variables with joint distribution F. Suppose we observe pairs $(x_1, y_1), \ldots, (x_n, y_n)$, a random sample from F. Without making any assumptions about F, form a statistic for testing $H_0 : P(X > Y) = 0.5$. How would you calculate the p-value?

Solution:

Define $Z = I(X > Y)$, and $p = P(X > Y) = P(Z = 1)$. We have observed $z_1, \ldots, z_n$, where $z_i = I(x_i > y_i)$, $i = 1, \ldots, n$.

Define $S = \sum_{i=1}^{n} Z_i$. Then $S \sim Binomial(n, p)$. The MLE of p is $\hat{p} = S/n = \sum_{i=1}^{n} Z_i/n = \bar{Z}_n$, and $se(\hat{p}) = \sqrt{p(1-p)/n}$.

Then the test becomes tesing $H_0 : p = 0.5$.

We can use S as the test statistic. Under H_0, $S \sim Binomial(0, 0.5)$. The Wald statistic is

$$W = \frac{\hat{p} - 0.5}{\sqrt{\hat{p}(1-\hat{p})/n}} = \frac{\bar{Z}_n - 0.5}{\sqrt{\bar{Z}_n(1-\bar{Z}_n)/n}}.$$

The asymptotic distribution of W under the null hypothesis is $N(0, 1)$.

- If $H_1 : P(X > Y) \neq 0.5$,

$$\text{p-value } = P(|W| > |w|) = 2\Phi(-|w|) = 2\phi\left(\left|-\frac{\bar{z}_n - 0.5}{\sqrt{\bar{z}_n(1-\bar{z}_n)/n}}\right|\right).$$

- If $H_1 : P(X > Y) > 0.5$,

$$\text{p-value } = P(W > w) = 1 - \Phi(w) = 1 - \phi\left(\frac{\bar{z}_n - 0.5}{\sqrt{\bar{z}_n(1-\bar{z}_n)/n}}\right).$$

- If $H_1 : P(X > Y) < 0.5$,

$$\text{p-value } = P(W < w) = \Phi(w) = \phi\left(\frac{\bar{z}_n - 0.5}{\sqrt{\bar{z}_n(1-\bar{z}_n)/n}}\right).$$

Alternatively, we can calculate the p-value by doing the exact binomial test and use S as the test statistic. Under the null hypothesis, $S \sim Binomial(n, 0.5)$.

- If $H_1 : P(X > Y) \neq 0.5$,

– If $s \geq 0.5n$,

$$\text{p-value } = 2P(S > s) = 2\sum_{i=s+1}^{n} \binom{n}{i} 0.5^i (1-0.5)^{n-i} = 2^{-n+1} \sum_{i=s+1}^{n} \binom{n}{i}.$$

– If $s < 0.5n$,

$$\text{p-value } = 2P(S < s) = 2\sum_{i=1}^{s-1} \binom{n}{i} 0.5^i (1-0.5)^{n-i} = 2^{-n+1} \sum_{i=1}^{s-1} \binom{n}{i}.$$

- If $H_1 : P(X > Y) > 0.5$,

$$\text{p-value } = P(S > s) = \sum_{i=s+1}^{n} \binom{n}{i} 0.5^i (1-0.5)^{n-i} = 2^{-n} \sum_{i=s+1}^{n} \binom{n}{i}.$$

- If $H_1 : P(X > Y) < 0.5$,

$$\text{p-value } = P(S < s) = \sum_{i=1}^{s-1} \binom{n}{i} 0.5^i (1-0.5)^{n-i} = 2^{-n} \sum_{i=1}^{s-1} \binom{n}{i}.$$

3. The file `pvals.R` contains `R` code for a simulation examining the proportion of times H_0 is actually true for tests within a specified range of p-values. The set-up (due to Sellke, Bayarri, and Berger, 2001) is as follows:

   Suppose for each dataset we observe $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. We form the Wald test statistic and calculate the p-value for testing $H_0 : \mu = 0$.

   The variable `nullprop` specifies the proportion of times in the simulation that H_0 is actually true. The rest of the time, the true value of μ is drawn from a pre-specified distribution. This is currently set to be $Unif(-1, 1)$. The last two lines of the file calculate and print the proportion of times, among tests for which $0.01 <$ p-value < 0.05 (what is usually called "strong evidence against H_0"), that the null hypothesis was actually true.

   (a) Experiment with the code to see what the effect is of changing the distribution of μ when it is not zero. For example, try changing the parameters of the uniform distribution, using a different distribution, or using a constant. Summarize your findings. (*Hint: If you change the file* `pvals.R`*, you can rerun everything in it using* `source("pvals.R")`*. You may need to add the directory to the beginning of the file if it is not in your working directory.*)

(b) Now experiment with changing `nullprop`, the proportion of times in the simulation that H_0 is actually true. Again, summarize your findings.

Solution:

(a) When $\sigma^2 = 1$,

$$0.01 < \text{p-value} < 0.05 \Leftrightarrow 1.960 < \sqrt{n}|\bar{X}_n| < 2.576 \overset{n=100}{\Leftrightarrow} 0.196 < |\bar{X}_n| < 0.258.$$

When we set `nullprop`= 0.5 and change the distribution of μ when its not zero. The results are summarized in Table 1.

Distribution	% true H_0 when $0.01 <$ p-value < 0.05
$Unif(-1, 1)$	39.9%
$Unif(-2, 2)$	58.0%
$Unif(-3, 3)$	66.5%
$N(0, 1)$	46.1%
$N(0, 4)$	64.2%
$N(0, 9)$	71.6%
$Gamma(\alpha = 1, \beta = 1)$	46.2%
$Gamma(\alpha = 1, \beta = 2)$	58.9%
$Gamma(\alpha = 2, \beta = 1)$	78.6%

Table 1: Effect of changing the distribution of μ when it is not zero, with nullprop=0.5.

From Table 1, we can see that as the distribution becomes more diverged from $\mu = 0$, the proportion of times in which the null hypothesis is true among tests with p-value $\in$ (0.01, 0.05) increases. The reason for that is, as the distribution of μ diverges from 0, there will be more μ having larger absolute values, and their generated sample mean are less likely to fall in the rejection region $|\bar{X}_n| \in$ (0.196, 0.258). On the other hand,the sample mean coming from the null distribution falls into the rejection region with probability 0.04. Hence, out of the cases where the sample mean falls into that rejection region, the proportion of the true null hypothesis becomes higher.

(b) Now we change `nullprop`, and fixed the distribution of μ as $Unif(-1, 1)$, which is mixed with $\mu = 0$ under the null. The results are plotted in Figure 1.

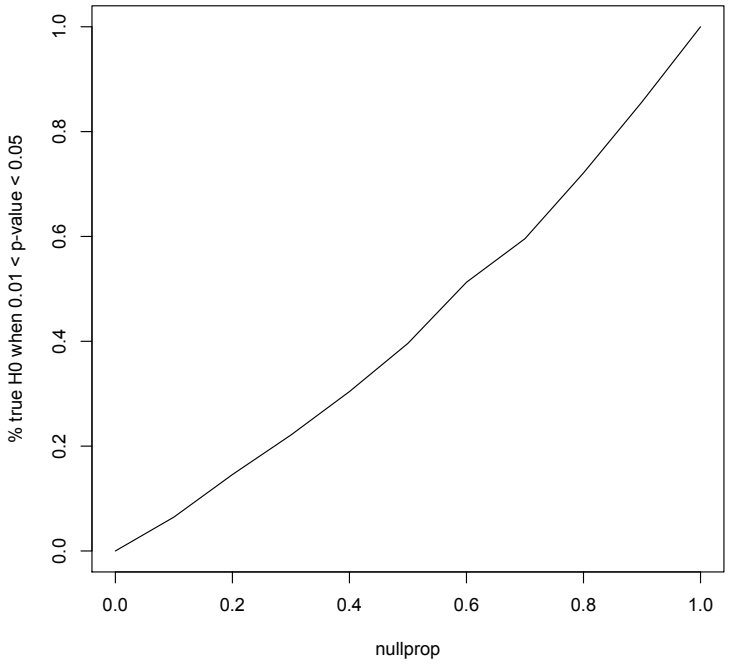

Figure 1: Change in the proportion of true null hypothesis when p-value $\in$ (0.01, 0.05).

From Figure 1, we can see that as `nullprop` increases, the proportion of the true H_0 when p-value $\in$ (0.01, 0.05) is increasing. This is reasonable, as when `nullprop` increases, the proportion of samples simulated from the null $N(0,1)$ increases, and thus the proportion of true null hypothesis when p-value $\in (0.01, 0.05)$ increases.

4. Consider the Berkeley freshman admissions data on page 97 of the class notes.

   (a) Using the top 6 cells of the table, calculate a 2x3 table counting numbers of students according to the population (CA residents, non-residents, international) and admission status (yes, no).

   (b) Calculate the MLEs for the corresponding probabilities under the null hypothesis of independence of the two variables, and the MLEs with no restrictions. Report them in the same table format. (You may use the formulas given in class; you do not need to rederive the MLEs.)

   (c) Calculate the likelihood ratio test statistic and Pearson's χ^2 statistic. Also report the p-values for each, using the limiting distributions of the test statistics under the null hypothesis of independence.

   Solution:

   (a) The table is as follows.

	CA residents	Non-resdients	International	Total
Yes	11,252	1,110	666	13,028
No	26,830	5,199	3,593	35,622
Total	38,082	6,309	4,259	48,650

   (b) Define variables in the table in (a) as follows

	$Y=0$	$Y=1$	$Y=2$	
$Z=0$	X_{00}	X_{01}	X_{02}	$X_{0\cdot}$
$Z=1$	X_{10}	X_{11}	X_{12}	$X_{1\cdot}$
	$X_{\cdot 0}$	$X_{\cdot 1}$	$X_{\cdot 2}$	n

   and the corresponding probabilities as

	$Y=0$	$Y=1$	$Y=2$	
$Z=0$	p_{00}	p_{01}	p_{02}	$p_{0\cdot}$
$Z=1$	p_{10}	p_{11}	p_{12}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	$p_{\cdot 2}$	1

We treat $X = (X_{00}, X_{01}, X_{02}, X_{10}, X_{11}, X_{12})$ as a sample from a multinomial distribution. Under the null hypothesis that Y (population: CA resident/non-resident/international) and Z (admission status: yes/no) are independent, $p_{ij} = p_i p_j$.
The unconstrained MLEs are

$$\hat{p}_{ij} = X_{ij}/n, \quad i = 0, 1; j = 0, 1, 2,$$

which are summarized as follows.

	$Y=0$	$Y=1$	$Y=2$	
$Z=0$	$\hat{p}_{00} = 0.2312$	$\hat{p}_{01} = 0.0228$	$\hat{p}_{02} = 0.0137$	$\hat{p}_{0\cdot}$
$Z=1$	$\hat{p}_{10} = 0.5515$	$\hat{p}_{11} = 0.1069$	$\hat{p}_{12} = 0.0739$	$\hat{p}_{1\cdot}$
	$\hat{p}_{\cdot 0}$	$\hat{p}_{\cdot 1}$	$\hat{p}_{\cdot 2}$	1

Under H_0, the constrained MLEs are

$$\hat{p}_{0ij} = \hat{p}_{0i\cdot}\hat{p}_{0\cdot j} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n},$$

as in the following table.

	$Y=0$	$Y=1$	$Y=2$	
$Z=0$	$\hat{p}_{000} = 0.2096$	$\hat{p}_{001} = 0.0347$	$\hat{p}_{002} = 0.0234$	$\hat{p}_{00\cdot} = 0.2678$
$Z=1$	$\hat{p}_{010} = 0.5731$	$\hat{p}_{011} = 0.0950$	$\hat{p}_{012} = 0.0641$	$\hat{p}_{01\cdot} = 0.7322$
	$\hat{p}_{0\cdot 0} = 0.7828$	$\hat{p}_{\cdot 1} = 0.1297$	$\hat{p}_{\cdot 2} = 0.0875$	1

(c) From the formula on Page 100 of the lecture notes, the LRT statistic is

$$\lambda = 2\sum_{i=0}^{1}\sum_{j=0}^{2} X_{ij}\log\left(\frac{nX_{ij}}{X_{i\cdot}X_{\cdot j}}\right) = 743.2693,$$

and the Pearson's χ^2 statistic is

$$T = \sum_{i=0}^{1}\sum_{j=0}^{n}\frac{(X_{ij} - n\hat{p}_{0ij})^2}{n\hat{p}_{0ij}} = 689.8647.$$

Both test statistics have a limiting χ^2_2 distribution under the null hypothesis, where the degrees of freedom is $(2-1)\times(3-1) = 2$.
Their corresponding p-values are

$$\text{LRT: p-value} = P(\lambda > 743.2694) \approx 3.99 \times 10^{-162}.$$

$$\text{Pearson: p-value} = P(\lambda > 689.8647) \approx 1.58 \times 10^{-150}.$$

This shows that we have very strong evidence against the null hypothesis that the population and the admission status are independent.

5. (Multinomial tests of homogeneity) When Jane Austen died, her novel Sandition was incomplete. Someone else finished the novel and it was published. Morton (1978) examined word frequencies to see if the new author was distinguishable from Austen. The data are as follows:

Word	Sense and Sensibility	Emma	Sandition I (Austen)	Sandition II (New author)
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4
Totals	375	440	202	196

Treat each column as an independent Multinomial sample.

(a) Construct the likelihood ratio statistic and calculate the p-value for the null hypothesis that the first three columns (by Austen) have the same set of probabilities for each column.

(b) Now sum the first three columns to give a single column of counts for Austen and another for the new author. Construct the likelihood ratio statistic and calculate the p-value for the null hypothesis that probabilities are the same across authors.

Solution:

(a) Define the variables corresponding to the first three columns as follows:

Word	Sense and Sensibility	Emma	Sandition I (Austen)
a	X_{11}	X_{12}	X_{13}
an	X_{21}	X_{22}	X_{23}
this	X_{31}	X_{32}	X_{33}
that	X_{41}	X_{42}	X_{43}
with	X_{51}	X_{52}	X_{53}
without	X_{61}	X_{62}	X_{63}
Totals	n_1	n_2	n_3

The corresponding probabilities are in the following table.

Word	Sense and Sensibility	Emma	Sandition I (Austen)
a	p_{11}	p_{12}	p_{13}
an	p_{21}	p_{22}	p_{23}
this	p_{31}	p_{32}	p_{33}
that	p_{41}	p_{42}	p_{43}
with	p_{51}	p_{52}	p_{53}
without	p_{61}	p_{62}	p_{63}
Totals	1	1	1

The null hypothesis is

$$H_0 : p_{1i} = p_{2i} = p_{3i}, \quad i = 1, \ldots, 6.$$

$$H_1 : p_{1i}, p_{2i} \text{ and } p_{3i} \text{ are not all equal for some } i.$$

The unconstrained MLEs are

$$\hat{p}_{ij} = X_{ij}/n_i, \quad i = 1, \ldots, 6; j = 1, 2, 3.$$

Under H_0, the constrained MLEs are

$$\hat{p}_{0i1} = \hat{p}_{0i2} = \hat{p}_{0i3} = \frac{X_{i1} + X_{i2} + X_{i3}}{n_1 + n_2 + n_3}, \quad i = 1, \ldots, 6.$$

The likelihood ratio statistic is

$$\lambda = 2 \log \frac{L(\hat{p})}{L(\hat{p}_0)} = 2 \sum_{i=1}^{6} \sum_{j=1}^{3} X_{ij} \log \frac{\hat{p}_{ij}}{\hat{p}_{0ij}} = 12.587,$$

which has an asymptotic χ^2_{10} under H_0. The degrees of freedom is $(6-1)\times 3-(6-1)=10$.

The p-value is $P(\lambda > 12.587) \approx 0.248$. Therefore, we do not have a strong evidence to reject the null hypothesis that the first three columns have the same set of probabilities for each column.

(b) Now we want to compare the works by Austen to works of the new author. The variables we are looking at are

Word	Austen	New Author
a	$X_1 = X_{11} + X_{12} + X_{13}$	X_1^*
an	$X_2 = X_{21} + X_{22} + X_{23}$	X_2^*
this	$X_3 = X_{31} + X_{32} + X_{33}$	X_3^*
that	$X_4 = X_{41} + X_{42} + X_{43}$	X_4^*
with	$X_5 = X_{51} + X_{52} + X_{53}$	X_5^*
without	$X_6 = X_{61} + X_{62} + X_{63}$	X_6^*
Totals	$n = n_1 + n_2 + n_3$	n^*

The corresponding probabilities are

Word	Austen	New Author
a	p_1	p_1^*
an	p_2	p_2^*
this	p_3	p_3^*
that	p_4	p_4^*
with	p_5	p_5^*
without	p_6	p_6^*
Totals	1	1

The null hypothesis is

$$H_0 : p_i = p_i^*, \quad i = 1, \ldots, 6.$$

$$H_1 : p_i \neq p_i^* \quad \text{for some } i.$$

The unconstrained MLEs are

$$\hat{p}_i = X_i/n, \quad \hat{p}_i^* = X_i^*/n^* \quad i = 1, \ldots, 6.$$

Under H_0, the constrained MLEs are

$$\hat{p}_{0i} = \hat{p}_{0i}^* = \frac{X_i + X_i^*}{n + n^*}, \quad i = 1, \ldots, 6.$$

The likelihood ratio statistic is

$$\lambda = 2 \log \frac{L(\hat{p})}{L(\hat{p}_0)} = 2 \left(\sum_{i=1}^{6} X_i \log \frac{\hat{p}_i}{\hat{p}_{0i}} + \sum_{i=1}^{6} X_i^* \log \frac{\hat{p}_i^*}{\hat{p}_{0i}^*} \right) = 31.737,$$

which has an asymptotic χ_5^2 under H_0. The degrees of freedom is $(6 - 1) \times 2 - (6 - 1) = 5$.

The p-value is $P(\lambda > 31.737) \approx 6.699 \times 10^{-6}$. Therefore, we have a strong evidence to reject the null hypothesis that probabilities are the same across authors.