STAT 513 fa 2019 Final Exam

Karl B. Gregory

Do not open this test until told to do so; no calculators allowed; no notes allowed; no books allowed; show your work so that partial credit may be given.

The table below gives some values of the function $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$:

z	0.841	1.282	1.645	1.96	2.326	2.576
$\Phi(z)$	0.80	0.90	0.95	0.975	.990	0.995

Some upper quantiles of some t-distributions:

ξ	0.10	0.05	0.025	0.01	0.005
$t_{15,\xi}$	1.3406	1.7531	2.1314	2.6025	2.9467
$t_{16,\xi}$	1.3368	1.7459	2.1199	2.5835	2.9208
$t_{17,\xi}$	1.3334	1.7396	2.1098	2.5669	2.8982
$t_{18,\xi}$	1.3304	1.7341	2.1009	2.5524	2.8784
$t_{19,\xi}$	1.3277	1.7291	2.0930	2.5395	2.8609
$t_{20,\xi}$	1.3253	1.7247	2.0860	2.5280	2.8453

Some upper quantiles of some chi-squared distributions:

ξ	0.10	0.05	0.025	0.01	0.005
$\chi^2_{1,\xi}$	2.71	3.84	5.02	6.63	7.88
$\chi^2_{2,\xi}$	4.61	5.99	7.38	9.21	10.60
$\chi^2_{3,\xi}$	6.25	7.81	9.35	11.34	12.84

A test statistic for something: $-2\sum_{i=1}^{K}\sum_{j=1}^{M} O_{ij}\log(O_{ij}/E_{ij})$.

For any $\mathbf{a} \in \mathbb{R}^{p+1}$, a $(1-\alpha)100\%$ confidence interval for $\mathbf{a}^T\hat{\boldsymbol{\beta}}$ is given by

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}.$$

Let $X_1, \dots, X_n$ be a random sample with likelihood function $L(\theta; X_1, \dots, X_n)$. Then for hypotheses of the form H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$ the likelihood ratio is

$$\text{LR}(X_1, \dots, X_n) = \frac{L(\theta_0; X_1, \dots, X_n)}{L(\hat{\theta}; X_1, \dots, X_n)},$$

where $\hat{\theta}$ is the MLE of θ.

1. Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Poisson}(\lambda)$, and recall that the pmf of the Poisson(λ)-distribution is given by

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

Suppose you wish to test H_0: $\lambda \leq 10$ versus H_1: $\lambda > 10$, and you have decided to reject H_0 if and only if $\sum_{i=1}^n X_i > C$ for some $C > 0$.

(a) Give an expression for the power function $\gamma(\lambda)$ using the fact that $\sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$.

Solution:

$$\begin{aligned}\gamma(\lambda) &= P_\lambda\left(\sum_{i=1}^n X_i > C\right)\\ &= P(Y > C), \quad Y \sim \text{Poisson}(n\lambda)\\ &= \sum_{y=\lceil C\rceil}^{\infty} \frac{e^{n\lambda}(n\lambda)^y}{y!}\end{aligned}$$

(b) Give an expression for the size of the test.

Solution: The size of the test will be given by the value of the power function $\gamma(\lambda)$ at $\lambda = 10$, so the size is given by

$$\sum_{y=\lceil C\rceil}^{\infty} \frac{e^{n10}(n10)^y}{y!}.$$

(c) Suppose that for some $\alpha \in (0, 1)$, C is chosen as the largest whole number such that the size is less than or equal to α. Make a drawing to show the shape of the power curve and add to the drawing a horizontal line at the height α and a vertical line at $\lambda = 10$.

Solution: The power curve should be monotonically increasing, and the horizontal and vertical lines should cross the power curve in approximately the same place.

(d) A <u>**larger**</u> (larger/smaller) value of C will give the test <u>**smaller**</u> (larger/smaller) size.

Solution:

(e) A <u>**larger**</u> (larger/smaller) value of C will give the test <u>**smaller**</u> (larger/smaller) power when $\lambda > 10$.

Solution:

(f) Recall that the mean and variance of the Poisson(λ)-distribution are both equal to λ and consider the test which rejects H_0 if and only if

$$\frac{\sqrt{n}(n^{-1}\sum_{i=1}^{n} X_i - 10)}{\sqrt{10}} > K$$

for some $K > 0$. Give the value of K such that the size of the test converges to $\alpha = 0.05$ as $n \to \infty$.

Solution: By the central limit theorem, when $\lambda = 10$,

$$\frac{\sqrt{n}(n^{-1}\sum_{i=1}^{n} X_i - 10)}{\sqrt{10}} \to \text{Normal}(0, 1)$$

in distribution as $n \to \infty$, so setting $K = 1.645$ will give the test size approaching $\alpha = 0.05$ as $n \to \infty$.

(g) Write down the likelihood function for λ based on the data $X_1, \ldots, X_n$.

Solution: The likelihood is given by

$$L(\lambda; X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} = \frac{e^{-n\lambda}\lambda^{n\bar{X}_n}}{\prod_{i=1}^{n} X_i!}.$$

(h) Write down the log-likelihood function.

Solution: The log-likelihood is given by

$$\ell(\lambda; X_1, \ldots, X_n) = -n\lambda + n\bar{X}_n \log\lambda - \sum_{i=1}^{n} \log X_i!$$

(i) Give the likelihood ratio corresponding to the hypotheses H_0: $\lambda = 10$ versus H_1: $\lambda \neq 10$.

Solution: The mle of λ is $\hat{\lambda} = \bar{X}_n$, so the likelihood ratio is given by

$$\text{LR}(X_1, \ldots, X_n) = \frac{\dfrac{e^{-n10}10^{n\bar{X}_n}}{\prod_{i=1}^{n} X_i!}}{\dfrac{e^{-n\bar{X}_n}\bar{X}_n^{n\bar{X}_n}}{\prod_{i=1}^{n} X_i!}} = e^{n\bar{X}_n - n10}(10/\bar{X}_n)^{n\bar{X}_n}.$$

(j) Find $-2\log\text{LR}$, where LR is the likelihood ratio.

Solution: We have

$$\begin{aligned}-2\log \mathrm{LR}(X_1,\ldots,X_n) &= -2\left[n\bar{X}_n - n10 + n\bar{X}_n\log 10 - n\bar{X}_n\log\bar{X}_n\right]\\ &= -2n\left[(\bar{X}_n - 10) + \bar{X}_n(\log 10 - \log\bar{X}_n)\right]\end{aligned}$$

(k) Give the rejection rule for the asymptotic likelihood ratio test of size 0.05.

Solution: According to the asymptotic likelihood ratio test we should reject H_0 if and only if $-2\log \mathrm{LR}(X_1,\ldots,X_n)$ exceeds the upper 0.05-quantile of the χ_1^2-distribution, which is 3.84.

2. Suppose you observe the data $(Y_1, x_{11}, x_{12}, x_{13}), \ldots, (Y_n, x_{n1}, x_{n2}, x_{n3})$, for which you assume the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1,\ldots,n,$$

where $\varepsilon_1,\ldots,\varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0,\sigma^2)$. Suppose $n = 20$ and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.053 & -0.008 & 0.000 & -0.010 \\ -0.008 & 0.084 & -0.021 & -0.037 \\ 0.000 & -0.021 & 0.107 & -0.036 \\ -0.010 & -0.037 & -0.036 & 0.124 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} 1.086 \\ -0.396 \\ -0.936 \\ 3.516 \end{bmatrix}, \text{ and } \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 = 17.87.$$

(a) Give the value of an unbiased estimator of σ^2.

Solution: Use $\hat{\sigma}^2 = 17.87/(20-3-1)$

(b) Give a 99% confidence interval for β_1.

Solution: Using $\mathbf{a} = (0,1,0,0)^T$ and the formula on the front of the exam, we have

$$\hat{\beta}_1 \pm t_{20-3-1,0.005}\hat{\sigma}\sqrt{0.084} = -0.396 \pm 2.9208\sqrt{17.87/(20-3-1)}\sqrt{0.084}.$$

(c) The height of the regression function at $x_{\text{new},1} = 0$, $x_{\text{new},2} = 1$, and $x_{\text{new},3} = 0$ is given by $\beta_0 + \beta_2$. Give a 95% confidence interval for $\beta_0 + \beta_2$. *Hint: Begin by finding* $\mathbf{a}$ *such that* $\mathbf{a}^T\boldsymbol{\beta} = \beta_0 + \beta_2$.

Solution: Use $\mathbf{a} = (1,0,1,0)^T$. This gives

$$\begin{aligned}&\hat{\beta}_0 + \hat{\beta}_2 \pm t_{20-3-1,0.025}\hat{\sigma}/\sqrt{20}\sqrt{0.053 + 0.107 + 2(0.000)}\\ &\quad = 1.086 - 0.936 \pm 2.1199\sqrt{17.87/(20-3-1)}/\sqrt{20}\sqrt{0.053 + 0.107 + 2(0.000)}\end{aligned}$$

3. Suppose tags are attached to 50 fish drawn from a lake. In one week, n fish will be drawn from the lake (one at a time, with replacement) and the number Y of these fish that have tags will be recorded. It is of interest to estimate N, the total number of fish in the lake.

(a) Treating N as a fixed, unknown parameter, assume $Y \sim \text{Binomial}(n, 50/N)$, since 50 out of the N fish have tags and the fish will be sampled with replacement.

i. Write down the likelihood function for N based on the data Y.

Solution:

$$L(N;Y) = \binom{n}{Y}\left[\frac{50}{N}\right]^{Y}\left[1-\frac{50}{N}\right]^{n-Y}.$$

ii. Give the log-likelihood function.

Solution:

$$\ell(N;Y) = \log\binom{n}{Y} + Y\log(50/N) + (n-Y)\log(1-50/N).$$

iii. Show that the maximum likelihood estimator of N is given by $\hat{N} = 50/(Y/n)$.

Solution: Taking the first derivative of the log-likelihood with respect to N and setting it equal to zero results in

$$\begin{aligned}
\frac{\partial}{\partial N}\ell(N;Y) &= \frac{Y}{50/N}\left(-\frac{50}{N^2}\right) + \frac{n-Y}{1-50/N}\left(\frac{50}{N^2}\right) \overset{\text{set}}{=} 0 \\
\iff \frac{Y}{50/N} &= \frac{n-Y}{1-50/N} \\
\iff \frac{Y}{n-Y} &= \frac{50/N}{1-50/N} \\
\iff \frac{Y/n}{1-Y/n} &= \frac{50/N}{1-50/N} \\
\iff Y/n &= 50/N \\
\iff N &= 50/(Y/n).
\end{aligned}$$

iv. Suppose you draw a sample of 10 fish from the lake and none of them have tags. What is your estimate of N according to the maximum likelihood estimator?

Solution: The maximum likelihood estimator would in this case take the value of ∞. Since we know *a priori* that this is impossible, maybe we should turn to Bayesian methods?

(b) Now regard N as a random variable; if we have tagged 50 fish in the lake, we know that $N \geq 50$, which is a bit of prior information. Based on this, let us assume the following hierarchical model:

$$Y|N \sim p(Y|N) = \binom{n}{Y}(50/N)^Y(1-50/N)^{n-Y} \quad \text{for } Y = 0, 1, \ldots, n$$
$$N \sim p(N) = \frac{e^{-\lambda}\lambda^N}{N!}\left[\sum_{i=50}^{\infty}\frac{e^{-\lambda}\lambda^i}{i!}\right]^{-1} \quad \text{for } N = 50, 51, 52, \ldots$$

So the prior distribution of N is the Poisson with mean λ, but conditional on $N \geq 50$.

i. Show that the conditional pmf of $N|Y$ is given by

$$p(N|Y) = \left[\frac{50}{N}\right]^Y\left[1-\frac{50}{N}\right]^{n-Y}\frac{\lambda^N}{N!}\left(\sum_{i=50}^{\infty}\left[\frac{50}{i}\right]^Y\left[1-\frac{50}{i}\right]^{n-Y}\frac{\lambda^i}{i!}\right)^{-1} \quad \text{for } N = 50, 51, 52, \ldots$$

Hint: Begin with $p(N|Y) \propto p(Y|N)p(N)$; discard constants; then note that it must sum to 1.

Solution: Write

$$p(N|Y) \propto \binom{n}{Y}(50/N)^Y(1-50/N)^{n-Y}\frac{e^{-\lambda}\lambda^N}{N!}\left[\sum_{i=50}^{\infty}\frac{e^{-\lambda}\lambda^i}{i!}\right]^{-1}$$
$$\propto (50/N)^Y(1-50/N)^{n-Y}\frac{e^{-\lambda}\lambda^N}{N!},$$

after discarding all constants which do not have to do with N. Now, since we need this to sum to 1 over the support of N, which is $N = 50, 51, 52, \ldots$, we can just divide the above expression by its sum taken over these values of N.

ii. Is the prior distribution of N a conjugate prior? Explain why or why not.

Solution: The prior is *not* a conjugate prior, because the posterior distribution is not in the same family (Poisson conditioned on being greater than some value) as the prior distribution.

iii. Propose a Bayesian estimator of N and write down an expression for how to compute it.

Solution: We might use the posterior mean of $N|Y$, which we would compute as

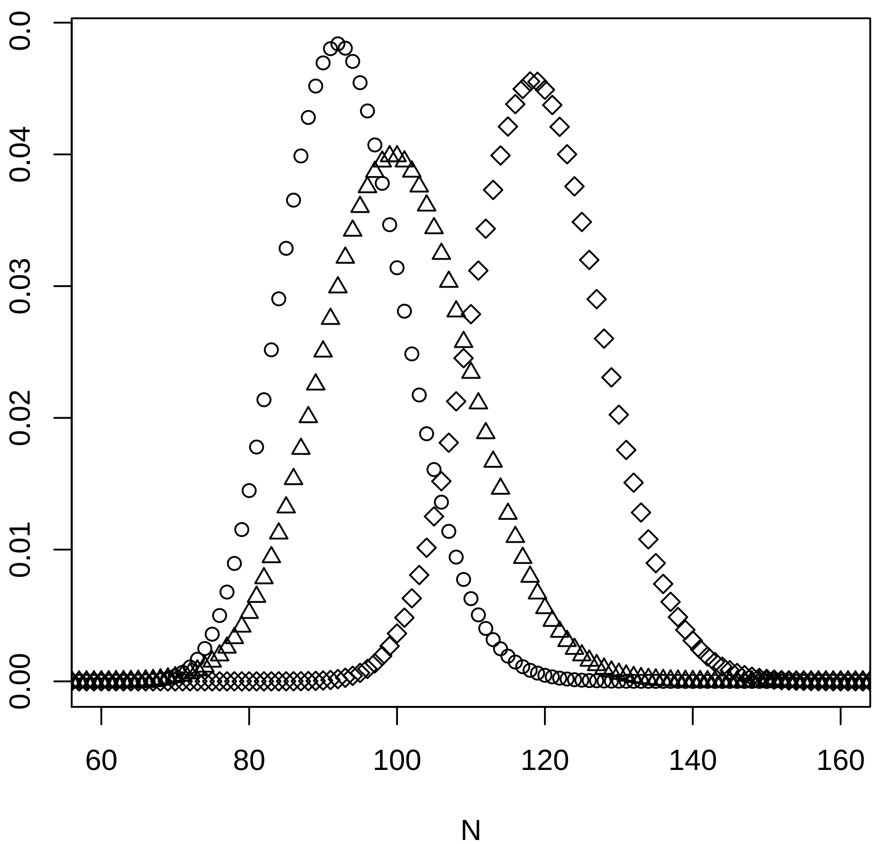
$$\mathbb{E}N|Y = \sum_{N=50}^{\infty} N \cdot \left[\frac{50}{N}\right]^Y\left[1-\frac{50}{N}\right]^{n-Y}\frac{\lambda^N}{N!}\left(\sum_{i=50}^{\infty}\left[\frac{50}{i}\right]^Y\left[1-\frac{50}{i}\right]^{n-Y}\frac{\lambda^i}{i!}\right)^{-1}$$

iv. Two researchers who believe in this hierarchical model with $\lambda = 100$ have independently drawn samples of $n = 40$ fish from the lake. Researcher 1 observed $Y = 5$ and researcher 2 observed $Y = 25$. Give the symbols (diamonds, circles, or triangles) with which each of the following is plotted in the figure below:

The pmf of the prior distribution.

The pmf of the posterior distribution of researcher 1.
The pmf of the posterior distribution of researcher 2.

Solution: Researcher 1, if using the mle, would estimate the total number of fish at $50/(5/40) = 400$, while researcher 2, if using the mle, would estimate the total number of fish at $50/(25/40) = 80$. The posterior of each researcher will represent a mixture of prior beliefs and evidence from the data, so that the prior will be pulled in the direction of the respective researchers' data findings. The prior mean is very close to $\lambda = 100$ (though it is not exactly 100 due to the condition that $N \geq 50$), so *the prior pmf is the one plotted with triangles; that of researcher 1 is plotted with diamonds; and that of researcher 2 is plotted with circles.*



4. Researchers in [1] studied the frequency of cross-racial friendships between individuals in the United States. Out of a sample of 270 people of ages 18-35 years, 19 reported having 0 cross-racial friendships; out of 453 people of ages 36-55 years, 29 reported having 0 cross-racial friendships; out of 280 people of ages 56 or more years, 22 reported having 0 cross-racial friendships. Regard the data as having been drawn as

$$\begin{aligned} Y_1 &\sim \text{Binom}(270, p_1), &\quad \text{with } Y_1 = 19 \text{ observed} \\ Y_2 &\sim \text{Binom}(453, p_2), &\quad \text{with } Y_2 = 29 \text{ observed} \\ Y_3 &\sim \text{Binom}(280, p_3), &\quad \text{with } Y_3 = 22 \text{ observed.} \end{aligned}$$

(a) Summarize the data in a contingency table, filling in the values for all the cells. Include row and column totals.

Solution:

	None	Not none	Total
18-35	19	251	270
36-55	29	424	453
≥ 56	22	258	280
Total	70	933	1003

(b) Give the table of expected values under the null hypothesis H_0: $p_1 = p_2 = p_3$. You do not have to evaluate the expected values, but give the expression for computing each one.

Solution:

	None	Not none	Total
18-35	70(270)/1003	933(270)/1003	270
36-55	70(453)/1003	933(453)/1003	453
≥ 56	70(280)/1003	933(280)/1003	280
Total	70	933	1003

(c) Give an expression for the test statistic of the asymptotic likelihood ratio test of H_0: $p_1 = p_2 = p_3$ versus H_1: p_1, p_2, *and* p_3 *are not all equal.*

Solution: The test statistic is given by

$$-2\left[19\log\left(\frac{19}{70(270)/1003}\right) + 251\log\left(\frac{251}{933(270)/1003}\right) + \cdots + 258\log\left(\frac{258}{933(280)/1003}\right)\right]$$

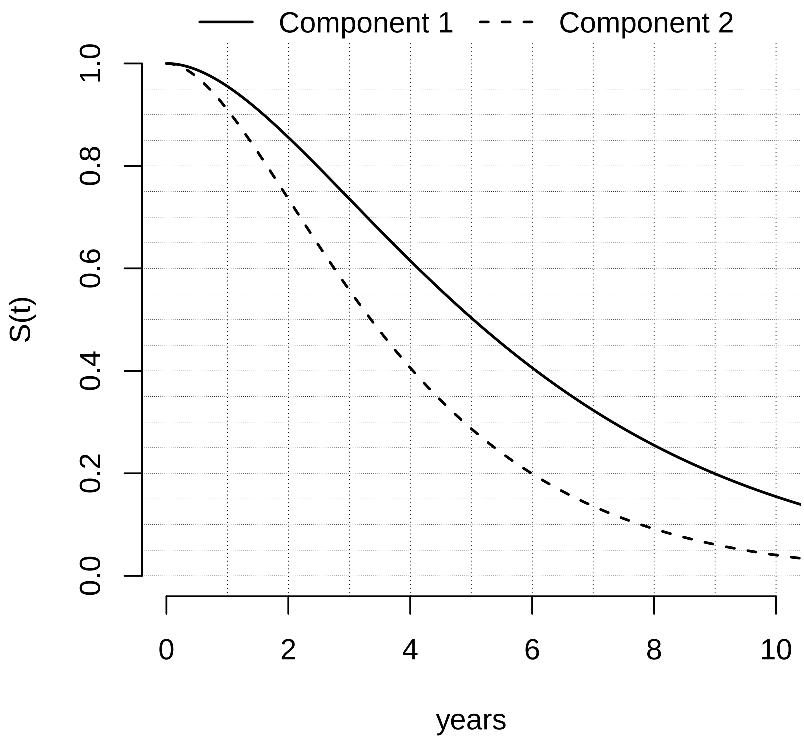
(d) The value of the test statistic from part (c) is 0.561. What is the critical value to which this should be compared in order to test the hypotheses in part (c) at significance level 0.01?

Solution: We should compare this value to the upper 0.01 quantile of the chi-squared distribution with degrees of freedom 2, since the table is a 3×2 table. This value is 9.21.

(e) Is there evidence to conclude, at the $\alpha = 0.01$ significance level, that younger people have more cross-racial relationships than older people? Explain your answer.

Solution: We do not reject the null hypothesis that $p_1 = p_2 = p_3$, so there is no evidence to claim that there is a difference between the age groups.

5. Let T_1 and T_2 be independent random variables representing the failure times of two components (component 1 and component 2, respectively) of a dishwasher. In order for the dishwasher to operate, both components must be functioning. The survival functions of T_1 and T_2 are plotted in the figure below.



(a) What is the probability that component 1 fails in the first 2 years?

> **Solution:** This is $P(T_1 \leq 2) = 1 - P(T_1 > 2) = 1 - 0.85 = 0.15$.

(b) What is the probability that the dishwasher will still function after 6 years?

> **Solution:** For the dishwasher to function after 6 years, the events $T_1 > 6$ and $T_2 > 6$ must both occur. We have
>
> $$P(T_1 > 6 \cap T_2 > 6) = P(T_1 > 6)P(T_2 > 6) = 0.4(0.2) = 0.08,$$
>
> where we have used the fact that T_1 and T_2 are independent.

(c) A 1-year warranty is offered with the dishwasher, under which the dishwasher will be replaced if *either* component fails during the 1-year period following the purchase. What is the probability that a customer may claim a replacement under the warranty?

Solution: We must find the probability of the event $P(T_1 < 1 \cup T_2 < 1)$. This is given by

$$\begin{aligned} P(T_1 < 1 \cup T_2 < 1) &= (P(T_1 < 1) + P(T_2 < 1) - P(T_1 < 1 \cap T_2 < 1) \\ &= (1 - 0.95) + (1 - 0.90) - (1 - 0.95)(1 - 0.90) \\ &= 0.05 + 0.10 - 0.05(0.10) \\ &= 0.145. \end{aligned}$$

References

[1] Deborah L Plummer, Rosalie Torres Stone, Lauren Powell, and Jeroan Allison. Patterns of adult cross-racial friendships: A context for understanding contemporary race relations. *Cultural Diversity and Ethnic Minority Psychology*, 22(4):479, 2016.