STA 215: Midterm Exam

Time: 1 hour 10 minutes

Name: **Solution Keys**

Qn	1	2	3	4	5	6	7	Total
Points								
Max	5	5	5	5	5	5	3	33

Traffic accident counts $X_1, \cdots, X_n$ of $n = 1000$ drivers from a county are modeled by the following *zero-inflated Poisson* distribution: $X_i \overset{\text{IID}}{\sim} g(x_i|\mu, \pi)$, $\mu > 0$, $\pi \in [0, 1]$ where

$$g(x_i|\mu,\pi) = \begin{cases} (1-\pi) + \pi e^{-\mu} & x_i = 0 \\ \pi e^{-\mu}\frac{\mu^{x_i}}{x_i!} & x_i = 1, 2, \cdots, \end{cases}$$

which is same as saying X_i's are IID and each X_i is zero with probability $1 - \pi$ and is drawn from $Poi(\mu)$ with probability π. For this discussion we focus on testing $H_0 : \pi = 1$, i..e, there is no zero-inflation.

1. Give an expression for the log-likelihood $\ell_x(\mu, \pi)$ which makes it obvious that $n_0(x)$ = number of x_i equaling zero and $\bar{x}$ **form a pair of sufficient statistics** for (μ, π). That is, your expression for $\ell_x(\mu, \pi)$, up to an additive constant, should include only $n, n_0(x)$ and $\bar{x}$ as summaries of data $x = (x_1, \cdots, x_n)$. [5 *points*]

$$\begin{aligned}\ell_x(\mu,\pi) &= \sum_{i=1}^{n} \log g(x_i|\mu,\pi) \\ &= \sum_{i=1}^{n}\{I(x_i = 0)\log(1-\pi+\pi e^{-\mu}) + I(x_i > 0)(\log\pi - \mu + x_i\log\mu - \log x_i!)\} \\ &= \text{const} + n_0(x)\log(1-\pi+\pi e^{-\mu}) + (n - n_0(x))(\log\pi - \mu) + n\bar{x}\log\mu\end{aligned}$$

2. Some algebra shows that a unique solution $(\hat{\mu}, \hat{\pi})$ exists to the first-order equations

$$\frac{\partial}{\partial\mu}\ell_x(\mu,\pi) = 0, \frac{\partial}{\partial\pi}\ell_x(\mu,\pi) = 0$$

whenever $\bar{x} > 0$ (i.e., not al x_i are zero) and that these $\hat{\mu}$, $\hat{\pi}$ also satisfy

$$\hat{\pi} = \frac{\bar{x}}{\hat{\mu}}, \quad \hat{\mu} = h_x(\hat{\mu})$$

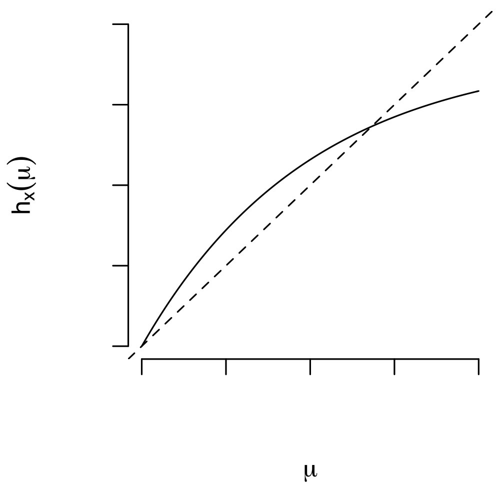

Figure 1: Plot of $h_x(\mu)$ for an x with $\bar{x} > 0$. The dashed line is the 45 degree line

where

$$h_x(\mu) = \frac{\bar{x}(1 - e^{-\mu})}{1 - \frac{n_0(x)}{n}}.$$

It is simple to check that whenever $\bar{x} > 0$, the function $h_x(\mu)$ is concave in μ with $h_x(0) = 0$, $\dot{h}_x(0) > 1$ and consequently a graph of $h_x(\mu)$ looks like the curve in Figure 1 (it cuts the 45 degree line precisely at two points, one being 0 and the other a positive number, and stays above the line only in between these two points)

Argue why the solution $(\hat{\mu}, \hat{\pi})$ **can not be the MLE** whenever $\frac{n_0(x)}{n} < e^{-\bar{x}}$ [however, the MLE does exist in this case]. [5 *points*]

When $n_0(x)/n < e^{-\bar{x}}$,

$$h_x(\bar{x}) = \frac{\bar{x}(1 - e^{-\bar{x}})}{1 - n_0(x)/n} < \bar{x}$$

so $\bar{x}$ must lie to the right of the non-zero solution $\hat{\mu}$ of $h(\mu) = \mu$. So $\hat{\mu} < \bar{x}$ and consequently, $\hat{\pi} = \bar{x}/\hat{\mu} > 1$ which is not possible because $\pi \in [0, 1]$.

3. When $X_i \overset{\text{IID}}{\sim} Poi(\mu)$, it follows from multivariate CLT that

$$\sqrt{n} \begin{pmatrix} \frac{n_0(X)}{n} - e^{-\mu} \\ \bar{X} - \mu \end{pmatrix} \xrightarrow{d} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} e^{-\mu}(1 - e^{-\mu}) & -\mu e^{-\mu} \\ -\mu e^{-\mu} & \mu \end{pmatrix} \right).$$

Argue that when $X_i \overset{\text{IID}}{\sim} Poi(\mu)$ we must have

$$\sqrt{n} \left(\frac{n_0(X)}{n} - e^{-\bar{X}} \right) \xrightarrow{d} N(0, \sigma(\mu)^2)$$

for some $\sigma(\mu) > 0$. [**I do not need a technical proof**. Just give an outline of how one would proceed to prove something like this. Bonus points for identifying the expression of $\sigma(\mu)^2$.] [5 *points*]

Apply Delta theorem with $g(u,v) = u - e^{-v}$ which gives,

$$g(\frac{n_0(X)}{n}, \bar{X}) = \frac{n_0(X)}{n} - e^{-\bar{X}}, \quad \text{and} \quad g(e^{-\mu}, \mu) = 0$$

which gives the desired result with $\sigma(\mu)^2 = \{\dot{g}(e^{-\mu}, \mu)\}^T \Sigma(\mu) \dot{g}(e^{-\mu}, \mu)$ where $\Sigma(\mu)$ is the covariance matrix in the statement of the theorem. The derivative equals

$$\dot{g}(u,v) = \begin{pmatrix} 1 \\ e^{-v} \end{pmatrix}$$

and hence

$$\begin{aligned} \sigma(\mu)^2 &= \begin{pmatrix} 1 & e^{-\mu} \end{pmatrix} \begin{pmatrix} e^{-\mu}(1-e^{-\mu}) & -\mu e^{-\mu} \\ -\mu e^{-\mu} & \mu \end{pmatrix} \begin{pmatrix} 1 \\ e^{-\mu} \end{pmatrix} = \begin{pmatrix} 1 & e^{-\mu} \end{pmatrix} \begin{pmatrix} e^{-\mu}(1-e^{-\mu}) - \mu e^{-2\mu} \\ 0 \end{pmatrix} \\ &= e^{-\mu}(1-e^{-\mu}) - \mu e^{-2\mu} \end{aligned}$$

4. Any ML test for $H_0 : \pi = 1$ is given by "reject H_0 if $2\log\Lambda(x) > c$" for some choice of the threshold $c \geq 0$, where

$$2\log\Lambda(x) = 2\left[\max_{\mu>0, \pi\in[0,1]} \ell_x(\mu,\pi) - \max_{\mu>0} \ell_x(\mu, 1)\right] = 2\left[\ell_x(\hat{\mu}_{\text{MLE}}, \hat{\pi}_{\text{MLE}}) - \ell_x(\bar{x}, 1)\right]$$

because, under H_0 (i.e. $\pi = 1$) the log-likelihood in μ is maximized at $\bar{x}$. However, the exact distribution of $2\log\Lambda(X)$ under H_0 is unknown and the usual chi-square approximation **does not work**. This is demonstrated in Figure 2 where $2\log\Lambda(x)$ is calculated for 10,000 samples of $x = (x_1, \cdots, x_n)$, each with $n = 1000$, simulated from a zero-inflated Poisson distribution with $\pi = 1$ and μ set as one of $1/3, 1$ or 3. The histograms of these simulated values do not match the pdf of $\chi^2(1)$.

Discuss what **causes the usual chi-square approximation to break down**. [Again, no technical proof is needed. Try to argue logically by making connections with parts (2) and (3).] [5 *points*]

The usual chi-square approximation needs that the MLE is given by the solution of the first order condition (and that the log-likelihood is nearly quadratic near this solution). But part (2) says that this is not the case whenever $n_0(x)/n < e^{-\bar{x}}$ and part (3) says that for large n, this happens with nearly 50% probability.

5. In part (3), the quantity $\sigma(\mu)$ is continuous in μ and so whenever $X_i \overset{\text{IID}}{\sim} Poi(\mu)$,

$$Z(X) = \frac{\sqrt{n}(\frac{n_0(X)}{n} - e^{-\bar{X}})}{\sigma(\bar{X})} \xrightarrow{d} N(0,1)$$

by the fact that $\bar{X} \xrightarrow{p} \mu$ (coupled with Slutsky's theorem). Figure 3 confirms this through a simulation study similar to what we did with $2\log\Lambda(x)$ above.

We could think of two (approximately) size-α tests for $H_0 : \pi = 1$:

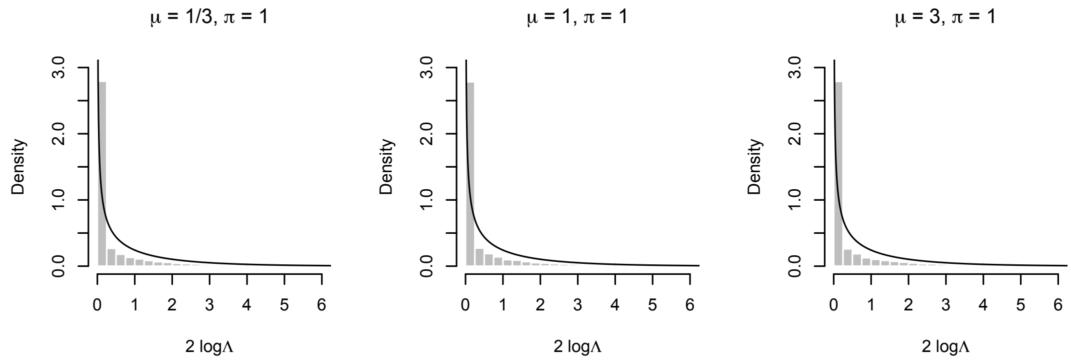

Figure 2: The distribution of $2\log\Lambda(X)$ under 3 parameters (μ, π) each satisfying $H_0 : \pi = 1$. The solid line is the pdf of $\chi^2(1)$.

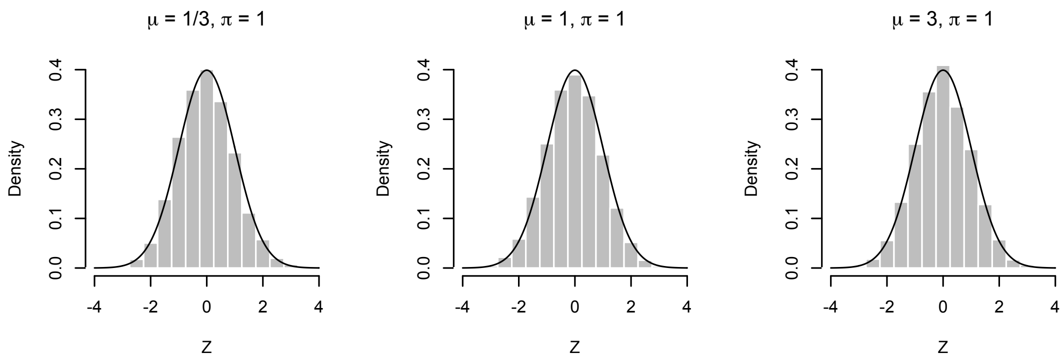

Figure 3: Distribution of Z under 3 choices of (μ, π) each satisfying $H_0 : \pi = 1$. The solid line is the pdf of $N(0, 1)$.

(a) Test 1: reject H_0 if $|Z(x)| > z(\alpha)$ or

(b) Test 2: reject H_0 if $Z(x) > z(2\alpha)$

Justify why Test 2 is more appropriate. Write your answer with clear logic, but no technical proof is required. Here $z(\alpha) = \Phi^{-1}(1 - \alpha/2)$ where Φ is the standard normal CDF. [Hint: what happens to $Z(x)$ when H_0 is not true? You may find this inequality useful: $1 - \pi + \pi e^{-\mu} > e^{-\pi\mu}$ whenever $0 < \pi < 1$ and $\mu > 0$.] [5 *points*]

Use the one-sided test (Test 2). For $\pi < 1$ and large n by WLLN, $n_0(X)/n \approx 1-\pi+\pi e^{-\mu}$ and $\bar{X} \approx (1-\pi)\cdot 0+\pi\mu = \pi\mu$ and hence $e^{-\bar{X}} \approx e^{-\pi\mu}$. So $\frac{n_0(X)}{n} - e^{-\bar{X}} \approx 1-\pi+\pi e^{-\mu} - e^{-\mu}$ which is a positive number. Hence $Z(X)$ is more likely to take positive values in this case. So only large positive values of $Z(X)$ indicate strong evidence against H_0 and hence it makes more sense to use the one-sided test than the two sided one [in other words, more power].

6. Another approximately size-α test for $H_0 : \pi = 1$ is the so called *over-dispersion test*

given by:

$$\text{reject } H_0 \text{ if } O(x) = \sqrt{\frac{n-1}{2}}\left(s_x^2/\bar{x} - 1\right) > z(2\alpha)$$

which again relies on the result that when $X_i \overset{\text{IID}}{\sim} Poi(\mu)$, $O(X) \overset{d}{\to} N(0,1)$. Simulations of $O(x)$ under the null give very similar pictures as in the case of $Z(x)$ in part (5).

However simulating $Z(x)$ and $O(x)$ under (μ, π) taken from outside the null show some differences. Figure 4 reports histograms of $Z(x)$ and $O(x)$ simulated under a zero-inflated Poisson distribution with $\pi = 0.95$ and $\mu \in \{1/3, 1, 3\}$.

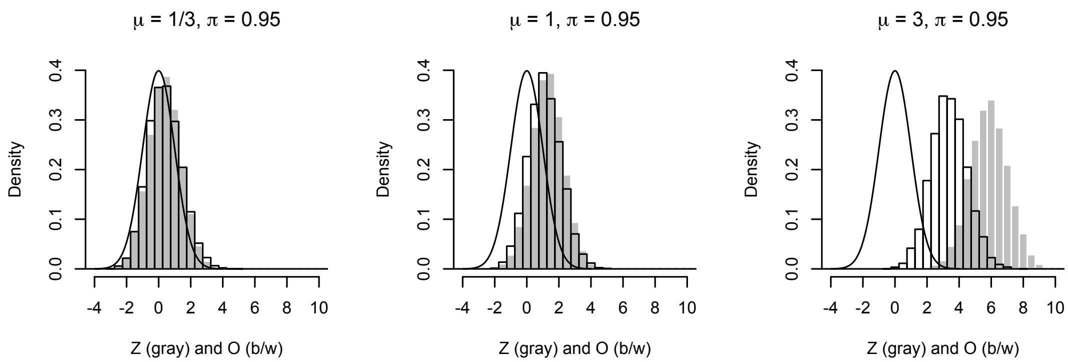

Figure 4: Distribution of Z under 3 choices of (μ, π) each with $\pi = 0.95$. Gray solid histogram is for $Z(x)$ and the histogram with black outline and white interior is for $O(x)$. The solid line is the pdf of $N(0,1)$.

Which test would you prefer using – the test based on $Z(x)$ [Test 2 from part (5)] or the test based on $O(x)$? Explain your choice. [No proof needed, give a clear logical argument.] [5 *points*]

The test based on $Z(x)$ because the above histograms reveal that it has more power than the test based on $O(x)$ with the same size.

7. Could you point out any reason for the difference we see in part (6)? Does one statistic **make better use of data** than the other? Justify your answer. [3 *points*]

   $Z(x)$ is derived from the sufficient statistics $(n_0(x), \bar{x})$ while $O(x)$ is not (it does not use $n_0(x)$). So $Z(x)$ makes better use of data.