JOE SEIDEL

STAT 245 HOMEWORK 0

1. *Approximate confidence intervals for Poisson Distribution*

Let $X_1, ..., X_n$ be independent random variables distributed according to a Poison(λ) distribution. Then the MLE of λ is $\hat{\lambda} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the two r.v.

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}} \text{ and } \frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/n}}$$

both have approximately $N(0, 1)$ for large n. Using the "pivotal method" derive two approximate confidence intervals for λ. What are the interval midpoints? Are the intervals guaranteed to comprise only nonegative numbers? Explain.

For $\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda/n}}$ see Prof. Gao's handout using Wilson's approach.

For $\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/n}}$ observe

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/n}} \rightsquigarrow N(0, 1)$$

is assymptotic pivotal and the CLT implies

$$\Pr\left(z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\hat{\lambda}}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha. \tag{1}$$

The inequality in equation (1) can be manipulated

$$\begin{aligned} z_{\frac{\alpha}{2}} &\leq \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\hat{\lambda}}} \leq z_{1-\frac{\alpha}{2}} \\ z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{n}} &\leq (\hat{\lambda} - \lambda) \leq z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{n}} \\ \hat{\lambda} - z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{n}} &\leq \lambda \leq \hat{\lambda} + z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{n}} \end{aligned}$$

.

Hence

$$\Pr(\hat{\lambda} - z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{n}} \leq \lambda \leq \hat{\lambda} + z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{n}}) \approx 1 - \alpha$$

and the confidence interval is

$$[\hat{\lambda} \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{n}}]$$

with midpoint $\hat{\lambda} = \overline{X}$. Furthermore, the interval does not guarantee comprising non-negative values. Consider $\hat{\lambda} = 1$ and small n.

Running 100000, with $n = 30$, $\alpha = .05$ and $\lambda = 1$, Wilson's confidence iterval does slightly better, .9751 vs .9291.

R code available q1.R

2. *Sample size determination*

Let X follow a Binomial(n, p) distribution and let $\hat{p} = \frac{\overline{X}}{n}$ be the maximum likelihood estimator of the success probability, p. Recall that the "Wald" $(1-\alpha)100\%$ confidence interval for p is of the form

$$[\hat{L}, \hat{U}] = \left[\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

For $\alpha = .05$ find the smalled integer $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ the confidence interval has length $\hat{U} - \hat{L} \leq 0.06$ regardless of the value $p \in [0, 1]$.

Through algebra observe

$$\begin{aligned}
\hat{U} - \hat{L} &= (\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) - (\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \\
&= 3.92\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\end{aligned}$$

then

$$\begin{aligned}
3.92\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &\leq 0.06 \\
\frac{\hat{p}(1-\hat{p})}{n} &\leq .01515^2 \\
n &\geq \frac{\hat{p}(1-\hat{p})}{0.01515^2}
\end{aligned}$$

Since $\hat{p}(1-\hat{p})$ is largest when $\hat{p} = .5$ we should use that value in the above inequality and conlude $n \geq 1098$.

3. *Approximate confidence intervals for Binomial distribution*

Let X have Binomial(n, p) distribution, and let $\hat{p} = \frac{X}{n}$ be the maximum likelihood estimator of the success probability p.

For the "Wald method", "Wilson method", and the arcsin transformation simulate in R. What proportion of the confidence intervals would we expect to contain $p = .1$ if the approximations are good. From simulations, which proportion of confidence intervals actually contain $p = .1$.

Given $\alpha = 0.05$ expect that $\frac{95}{100}$ of the intervals contain $p = 0.1$ if approximations are good.

Running $n = 100$ simulations, calculate the intervals and the proportions that contain $p = .1$.

1. The Wald interval

$$\left[\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Approximately 83% of the confidence intervals contained $p = .1$.

2. The Wilson interval

$$\left[\frac{\hat{p} + \frac{z^2}{2n} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}z^2 + \frac{z^4}{4n^2}}}{1 + \frac{z^2}{n}}\right] \text{ with } z = z_{1-\frac{\alpha}{2}}$$

Approximately 97% of confidence intervals simulated contained $p = .1$.

3. The arcsin transformation

$$\left[\sin^2(\arcsin(\sqrt{p} \pm \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}))\right]$$

Approximately 93% of confidene intervals contain $p = .1$.

By repeting with $n = 150$ the propertions get closer to 95%.

4. Distribution of a ratio

Show that if X_1 and X_2 are independent exponential random variables with parameter $\lambda = 1$, then $\frac{X_1}{X_2}$ follows an F-distribution. Also identify the degress of freedom.

Observe

$$f_X = f_{X_1} = e^{-x}$$
$$f_Y = f_{X_2} = e^{-y}$$

Let $U = X$ and $V = \frac{X}{Y}$ then $X = U$ and $Y = \frac{U}{V} = g(x)$.

$$f_{U,V} = f_{X,Y}(u, g(x))|g'(x)| = e^{-u(1+\frac{1}{v})}(\frac{u}{v^2}).$$

Then, find the marginal distribution of f_V by integrating out U.

$$\begin{aligned} f_V(v) &= \int_0^\infty e^{-u(1+\frac{1}{v})}(\frac{u}{v^2})du \\ &= \frac{1}{v^2}\int_0^\infty e^{-u(1+\frac{1}{v})}u\frac{1+\frac{1}{v})^2}{\Gamma(2)}du\frac{\Gamma(2)}{(1+\frac{1}{v})^2} \\ &= (\frac{1}{v^2})(1+\frac{1}{v})^{-2} \\ &= (1+v)^{-2} \sim F_{2,2} \end{aligned}$$

5. Do questions 16, 17, and 18 on p.241 in Rice

1. True or False?

   The center of a 95% confidence interval for the population mean is a random variable. TRUE

   A 95% confidence interval for μ contains the sample mean with probability 95%. FALSE: the interval is build around the sample mean so it it contains with probability 1.

   A 95% confidence interval contains 95% of the population. FALSE: A CI means that some percentange of samples constructed using indentical methods will contain the true parameter.

   Out of one hundred 95% confidence intervals for μ, 95 will contain μ. FALSE: It is actually a $\text{Binom}(100, .95)$ random variable.

2. A 90$ confidence interval for the average number of children per house based on a simple random sample is fund to be $(.7, 2.1)$. Can we conlcude that 90% of households have between .7 and 2.1 children?

   No. The correct interpretation of the interval would be: were the sample procedure repeated on numerous samples the fraction of calculated intervals that contain the true mean would tend toward 95%.

3. From independent surveys of two populations, 90% confidence intervals for the population means are constructed. What is the probability that neither interval contains the respective population mean? That both do?

   For both $\binom{2}{2}.9^2$. For niether $\binom{2}{0}.9^0(1-.9)^{.2}$.

6. Pivotal quantities and Normal distribution

Let $X_1, ..., X_n$ be iid as $N(\mu, \mu^2)$,where $\mu \in \mathbb{R}$ is an unknown parameter.

(a) Find pivotal(s) for μ.

Since the MLE, $\hat{\mu} = \overline{X}$ then

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\mu} = \sqrt{n}\Big(\frac{\overline{X}}{\mu} - 1\Big) \sim N(0, 1)$$

is a pivotal for μ.

Also, since the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

and a random variable we have the following result

$$\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\mu^2} = \frac{(n-1)s^2}{\mu^2} \sim \chi^2_{n-1}.$$

Another pivotal quantity for μ. There may be more, I do not know at this point.

(b) Let $\hat{\mu}$ be the MLE for μ. Find a function g such that

$$\sqrt{n}|g(\hat{\mu} - g(\mu)| \Rightarrow N(0,1).$$

First consider the likelihood function

$$f(\mu \mid X_1, ..., X_n) = \left(\frac{1}{\mu\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n \left(\frac{(X_i-\mu)^2}{2\mu^2}\right)}$$

and

$$l(\mu \mid X_1, ..., X_n) = -n\log(\mu) - \frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\mu^2}.$$

Where

$$\frac{\partial l}{\partial \mu} = -\frac{n}{\mu} + \sum_{i=1}^n \frac{x_i^2}{\mu^3} - \sum_{i=1}^n \frac{x_i}{\mu^2}$$

which when set to 0 gives

$$n\mu^2 + \mu\sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2 = 0$$

whose positive root will be the MLE of μ.

To find g, use Talyor expansion in conjunction with asymptotic normaliy. If there exists g such that

$$\sqrt{n}[g(\hat{\mu}) - g(\mu)] \to N(0,1)$$

then by Taylor expansion

$$\sqrt{n}g'(\mu)(\hat{\mu} - \mu) \approx \sqrt{n}(g(\hat{\mu}) - g(\mu)) \to N(0,1). \tag{2}$$

By asymptotic normalily we have

$$\sqrt{nI(\mu)}(\hat{\mu} - \mu) \to N(0,1) \tag{3}$$

where

$$\begin{aligned} I(\mu) &= -E\frac{\partial^2 l}{\partial \mu^2} \\ &= -\frac{1}{\mu^2} + \frac{3}{\mu^4}E(X^2) - \frac{2}{\mu^3}E(X) \\ &= \frac{3}{\mu^2} \end{aligned}$$

Fisher information.

Compairing left hands sides of (2) and (3)

$$g'(\mu) = \frac{\sqrt{3}}{\mu}$$

which implies $g(u) = \sqrt{3}\log(\mu)$.

(c) Comment on the confidence intervals for μ^2 constructed based on (a) and (b). Which one has smaller length.

The interval from (b) is smaller.

I got lazy and didn't derive this. See the solutions from thr TA