CSE 544, Spring 2017, Probability and Statistics for Data Science

Assignment 3: Statistical Inference

(8 questions, 70 points total)

Due: 3/27, in class

I/We understand and agree to the following:

(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.

(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

(write down the name of all collaborating students on the line below)

---

1. MME for Gamma distribution **(Total 5 points)**

The Gamma(x, y) distribution has mean $x * y$ and variance $x * y^2$. Find the MME for x and y.

Gamma(x,y) → mean $= \boxed{xy = E[X]}$

Variance $= \boxed{xy^2 = Var(X)}$

Equating E[X] & Var(X) to corresponding sample mo

exp. var

we get, $E[X] = \frac{\sum X_i}{n} = xy$

$Var[X] = \frac{\sum (X_i - \bar{X})^2}{n} = xy^2$

$\boxed{x = \frac{\bar{X}}{y}}$

$xy^2 = \frac{\sum (X_i - \bar{X})^2}{n}$

$\overline{xy} = \bar{X}$

MME of y $\boxed{\hat{y} = \frac{\sum (X_i - \bar{X})^2}{n\bar{X}}}$

MME of x $\boxed{\hat{x} = \frac{\bar{X}}{\hat{y}} = \frac{n\bar{X}^2}{\sum (X_i - \bar{X})^2}}$

2. Properties of estimators

(Total 5 points)

Recall the three properties of estimators introduced in class: (i) bias($\hat{\theta}$) = E[$\hat{\theta}$] - θ, (ii) se($\hat{\theta}$) = $\sqrt{Var[\hat{\theta}]}$, and (iii) MSE($\hat{\theta}$) = E[($\hat{\theta}$ - θ)^2]. Find these quantities for $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim$ Poisson(θ).

$\boxed{\hat{\theta} = \frac{1}{n}\sum X_i}$ where $X_i \sim$ Poisson

(i) $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

$= E\left(\frac{1}{n}\sum X_i\right) - \theta$

$= \frac{1}{n}\sum_{i=1}^{n} E[X_i] - \theta = \frac{\sum E[X_i]}{n} - \lambda$

$= \frac{n\lambda}{n} - \lambda = \lambda - \lambda = 0$

($\because E[X_i]$ of Poisson $= \lambda$)

$\boxed{\text{In our case } \theta = \lambda}$

$= \cancel{\theta - \theta}$

$= \cancel{0}$

(ii) $se(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$

$= \sqrt{Var\left(\frac{1}{n}\sum X_i\right)} = \sqrt{\frac{1}{n^2}\sum_i^n Var(X_i)}$

$= \sqrt{\frac{\theta}{n}}$ ($\because$ Var of Poisson $= \lambda$)

$*$

(iii) $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

$= se^2 + bias^2$

$= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + E[\theta^2]$

$= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2$

$= Var(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2$

$= \left(\sqrt{\frac{\theta}{n}}\right)^2 + (bia(\hat{\theta}))^2$?

$= \frac{\theta}{n}$

$= \frac{\theta}{n} + 0$

$= \frac{\theta}{n}$

$* E(\hat{\theta} - \theta)^2$

3. Plug-in estimates **(Total 10 points)**

(a) The kurtosis for a random variable X is defined as $Kurt[X] = E[(X-\mu)^4] / \sigma^4$. Derive the plug-in estimate of the kurtosis. (3 points)

(b) Show that the plug-in estimator of the variance of X is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, where $\bar{X}_n$ is the sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. (3 points)

(c) Find the bias of $\hat{\sigma}^2$ in terms of the true variance, σ^2. (4 points)

(a) $$Kurt[X] = \frac{E[(X-\mu)^4]}{\sigma^4}$$

$$E[X] = x f(x)$$

$$= \frac{\int (x-\mu)^4 \, dF(x)}{\left(\int (x-\mu)^2 \, dF(x)\right)^2}$$

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \hat{\mu})^2$$

$$= \frac{\sum (X_i - \hat{\mu})^4}{n \, \hat{\sigma}^4}$$

$$= \frac{n \sum (X_i - \bar{X})^4}{\left(\sum (X_i - \bar{X})^2\right)^2}$$

Remove $\int dF(x) \to \frac{\sum}{n}$

$$\therefore \frac{\sum (X_i - \hat{\mu})^4}{n} = \int (x-\mu)^4 \, dF(x)$$

$\sigma^2 = \sum$

(b) $$\hat{\sigma}^2 = Var(X) = \int x^2 \, d\hat{F}_n(x) - \left(\int x \, d\hat{F}_n(x)\right)^2$$

$$E[X^2] - E[X]^2$$

$$= \frac{1}{n}\sum X_i^2 - \left(\frac{\sum X_i}{n}\right)^2$$

$$= \frac{1}{n}\sum X_i^2 - \bar{X}^2$$

$$= \frac{1}{n}\sum X_i^2 - 2\bar{X}^2 + \bar{X}^2$$

$$= \frac{1}{n}\sum X_i^2 - \frac{2\sum X_i \bar{X}}{n} + \frac{\sum \bar{X}^2}{n}$$

$$= \frac{1}{n}\sum (X_i - \bar{X})^2$$

(c) $E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]$

$$= E\left[\frac{1}{n}\sum\left((X_i-\mu)-(\bar{X}-\mu)\right)^2\right]$$

$$= E\left[\frac{1}{n}\sum(X_i-\mu)^2 - \frac{2}{n^2}\sum(\bar{X}-\mu)\sum(X_i-\mu) + \frac{1}{n}\sum(\bar{X}-\mu)^2\right]$$

$$= E\left[\frac{1}{n}\sum(X_i-\mu)^2 - \frac{2}{\cancel{n^2}}\cancel{n}(\bar{X}-\mu)(\cancel{n}\bar{X}-\cancel{n}\mu) + \frac{1}{\cancel{n}}\times\cancel{n}(\bar{X}-\mu)^2\right]$$

$$= E\left[\frac{1}{n}\sum(X_i-\mu)^2 - 2(\bar{X}-\mu)^2 + (\bar{X}-\mu)^2\right]$$

$$= E\left[\frac{1}{n}\sum(X_i-\mu)^2 - (\bar{X}-\mu)^2\right]$$

$$= E\left[\frac{\sum(X_i-\mu)^2}{n}\right] - E\left[(\bar{X}-\mu)^2\right]$$

$$= \sigma^2 - Var(\bar{X})$$

$$= \sigma^2 - Var\left(\frac{\sum X_i}{n}\right)$$

$$= \sigma^2 - \frac{1}{n^2}\sum Var\, X_i$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

$$E[\hat{\sigma}^2] - \sigma^2 = \frac{-\sigma^2}{n}$$

Scanned by CamScanner

4. Functionals of the empirical distribution function **(Total 10 points)**

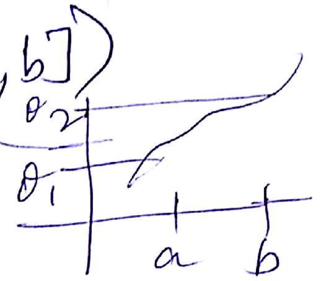
Let $X_1, X_2, \dots, X_n$ be i.i.d. RVs with true CDF F. Let $\hat{F}$ be their empirical distribution function. Let a and b be some numbers with a < b. Define $\theta = F(b) - F(a)$. Let $\hat{\theta} = \hat{F}_n(b) - \hat{F}_n(a)$.

(a) Find $\widehat{se}(\hat{\theta})$ (5 points)

(b) Find an approximate (1-α) CI for $\hat{\theta}$. (5 points)

$$\theta = F(b) - F(a)$$

$$\hat{\theta} = \hat{F}_n(b) - \hat{F}_n(a) = \frac{1}{n}\sum 1(X_i \in (a,b])$$



(a) $$Var(\hat{\theta}) = Var\left(\frac{1}{n}\sum 1(X_i \in (a,b])\right)$$

$$= \frac{1}{n^2} Var\left(\sum 1(X_i \in (a,b])\right)$$

$$= \frac{1}{n} Var(1(X_i \in (a,b]))$$

$$= \frac{1}{n}(F(b) - F(a))(1 - F(b) + F(a))$$

$$= \frac{\theta(1-\theta)}{n}$$

$$Var(\hat{F}_n) = \frac{F(x)(1-F(x))}{n}$$

$$\widehat{se} = \sqrt{\frac{\theta(1-\theta)}{n}}$$

(b) $$P\left(-z_\alpha \le \frac{\hat{\theta} - E[\hat{\theta}]}{se(\hat{\theta})} \le z_\alpha\right)$$

$$= Pr\left(-z_\alpha \le \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le z_\alpha\right)$$

$$= Pr\left(-z_\alpha\sqrt{\frac{\theta(1-\theta)}{n}} + \theta \le \hat{\theta} \le z_\alpha\sqrt{\frac{\theta(1-\theta)}{n}} + \theta\right)$$

CI for $\hat{\theta}$ is

$$L(x) = \theta - z_\alpha \sqrt{\frac{\theta(1-\theta)}{n}}$$

$$R(x) = \theta + z_\alpha \sqrt{\frac{\theta(1-\theta)}{n}}.$$

5. Histogram, meet the empirical distribution function **(Total 5 points)**

Let $X_1, X_2, \ldots, X_n$ be i.i.d. RVs with true CDF F with sample space [0, 1]. Let $\hat{F}_n$ be the associated empirical distribution function. Let $\hat{f}_n$ be the empirical pdf based on a histogram on the range [0, 1] with some bin size h, as in class. For some x in the range (0, 1), show that $\hat{f}_n(x) \approx d\hat{F}_n(x)$. Use the fact that the derivative of a function, g(), at x, is $\lim_{\Delta x \to 0} \frac{g(x+\Delta x)-g(x)}{\Delta x}$.

Histogram is

$$\hat{f}_n(x) = \frac{\sum 1(X_i \in (r_k, r_{k+1}])}{n(r_{k+1} - r_k)}$$

for $x \in (r_{k+1}, r_k]$

$$\hat{f}_n(x) = \frac{1}{nh} \sum 1(X_i \in (r_k, r_{k+1}])$$

$$= \frac{1}{nh} \sum \left(1(X_i \leq r_{k+1}) - 1(X_i \leq r_k)\right)$$

$$= \frac{1}{nh} \sum \left(1(X_i \leq x+h) - 1(X_i \leq x)\right)$$

$$= \frac{\frac{1}{n} \sum 1(X_i \leq x+h) - \frac{1}{n} \sum 1(X_i \leq x)}{h}$$

$$= \frac{\hat{F}_n(x+h) - \hat{F}_n(x)}{h}$$

$$\approx d\hat{F}_n(x)$$

Scanned by CamScanner

6. Normal-based CI for $\hat{F}_n$ **(Total 10 points)**

$X_1, X_2, \dots, X_n$ be i.i.d. RVs with true CDF F. Let $\hat{F}_n$ be their empirical distribution function.

(a) Derive Var[$\hat{F}_n$]. (4 points)

(b) Derive the Normal-based (1-α) CI for $\hat{F}_n$. (6 points)

(a) n indept Bernoulli RV, sum $= \sum I(X_k \le x)$
& success prob $= F(x)$

$\Rightarrow n\hat{F}_n(x) =$ ~~Bernoulli RV~~
Binomial RV

$\Rightarrow$ Variance of Binomial Distribution $= np(1-p)$

$\Rightarrow$ ~~Var~~

$n \operatorname{Var}(\hat{F}_n) = \frac{nF(x)(1-F(x))}{n}$ Var($\hat{F}_n$)

$=$

$\Rightarrow$

(b) (1-α) CI for $\hat{F}_n$

is $F(x) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{F(x)(1-F(x))}{n}}$