دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمارین نوبت دوم درس یادگیری ماشین آماری

دانشجو:

سیّد احمد نقوی نوزاد

شماره دانشجوئی:

۹۴۱۳۱۰۶۰

استاد:

دکتر نیک‌آبادی

پاییز ۹۴

سوال اول:

الف) برای تخمین پارامتر توزیع نمائی و محاسبه‌ی بایاس آن مطابق زیر از روش موسوم به درستنمائی بیشینه استفاده می‌نمائیم:

$$f(X_i;\beta)=\frac{1}{\beta}e^{-X_i/\beta},\quad n>0,\quad \beta>0$$

$$L_n(\beta)=\prod_{i=1}^{n}f(X_i;\beta)=\beta^{-n}e^{-S/\beta}\quad \text{where}\quad S=\sum_{i=1}^{n}X_i\quad\Rightarrow$$

$$\ell_n(\beta)=-n\log\beta-S/\beta\ \xrightarrow{\textit{derivative}\ \text{equal to zero}}\ \frac{-n}{\beta}+\frac{S}{\beta^2}=0\quad\Rightarrow$$

$$\frac{-\beta n+S}{\beta^2}=0\quad\Rightarrow\quad-\beta n+S=0\quad\Rightarrow\quad\beta=\frac{S}{n}=n^{-1}\sum_{i=1}^{n}X_i$$

$$\Rightarrow\quad\hat{\beta}=\overline{X}_n$$

but we do know that: $E_\theta(\overline{X}_n)=E_\theta(X)=\beta$

so we can conclude that the parameter is unbiased

ب) برای محاسبه بازه اطمینان ۹۳٪ ابتدا باید standard error تخمین پارامتر را مطابق زیر محاسبه نموده و در فرمول مربوطه قرار دهیم:

$$f(x;\beta)=\frac{1}{\beta}e^{-x/\beta},\quad n>0,\quad \beta>0\quad\rightarrow$$

$$\log f(x;\beta)=-\log\beta-x/\beta\quad\rightarrow\quad\frac{\partial^2\log f(x;\beta)}{\partial^2\beta^2}=\frac{1}{\beta^2}-\frac{2x}{\beta^3}$$

$$\rightarrow\quad \mathrm{I}_1(\beta)=-\mathrm{E}_\beta\left(\frac{\partial^2\log f(x;\beta)}{\partial^2\beta^2}\right)=-\mathrm{E}_\beta\left(\frac{1}{\beta^2}-\frac{2x}{\beta^3}\right)=\frac{1}{\beta^2}\quad\Rightarrow$$

$$\mathrm{I}_n(\beta)=n\mathrm{I}_1(\beta)=\frac{n}{\beta^2}\quad\Rightarrow\quad\hat{se}=\sqrt{1/\mathrm{I}_n(\hat{\beta})}=\hat{\beta}/\sqrt{n}=\overline{X}_n/\sqrt{n}$$

در نهایت بازه اطمینان مربوطه مطابق زیر خواهد بود:

$$\hat{\beta}\pm 1.812\times\hat{se}$$

برای پارامتر β=۰,۵ و ۱۰۰۰ نمونه بازه‌ی اطمینان به صورت زیر خواهد بود:

$$0.5019 \pm 1.812 \times 0.0159 \;=\; (0.4732\,,\, 0.5307)$$

ج) برای تعداد دفعات ۱۰۰۰۰ بار انجام همین آزمایش مشاهده می‌شود که پارامتر β به تعداد ۹۳۱۰ بار در بازه‌های اطمینان تخمینی تولیدشده قرار می‌گیرد که این مطلب نشانگر همان ۹۳٪ بودن میزان پوشش بازه‌ی اطمینان می‌باشد.

Beta parameter	0.4930
Confidence Interval	**(0.4648 , 0.5213)**
Occurrence No.	**9310**

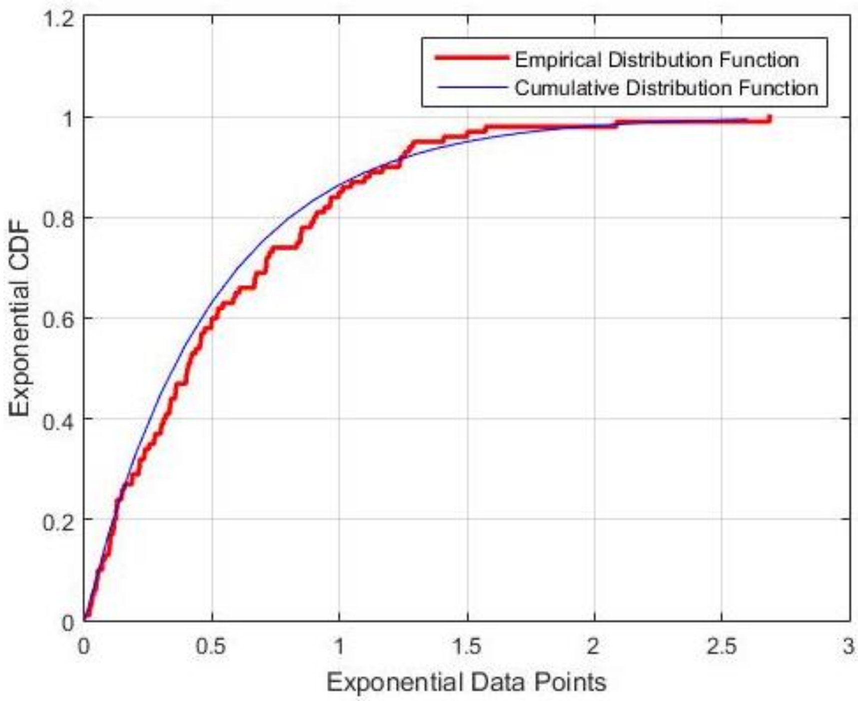
سوال دوم:

الف) در ابتدا ۱۰۰ نمونه داده تصادفی با توزیع نمائی تولید می‌کنیم و سپس با استفاده از رابطه زیر برگرفته از صفحه ۱۱۷ کتاب درسی تابع توزیع تجربی را محاسبه کرده و رسم می‌نمائیم و سپس نمودار CDF اصلی تابع نمائی را نیز بر روی نمودار قبلی رسم می‌نمائیم؛ نتایج در شکل زیر قابل مشاهده است:

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \le x)}{n} = \frac{\text{number of observations less than or equal to x}}{n}$$

$$where$$

$$I(X_i \le x) = \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{if } X_i > x. \end{cases}$$



ب) روابط مربوط به plug-in estimator های میانگین، واریانس و چولگی مطابق زیر می‌باشد:

Parameter	Plug-in Estimator		Symbol
Mean	$\frac{1}{n}\sum_{i=1}^{n} X_i$		$\bar{X}_n$
Variance	$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$	$\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$	$\hat{\sigma}^2$
Skewness	$\dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^3}{\hat{\sigma}^3}$		$\hat{\kappa}$

در انتها مقادیر واقعی و تخمین‌زده‌شده پارامترهای بالا را در جدولی مطابق زیر قرار می‌دهیم:

Parameter	True Value	Estimated Value	
Mean	0.5	0.4564	
Variance	0.25	0.1604	0.1588
Skewness	2	1.4143	

سوال سوم:

الف) در ابتدا تعداد ۱۰۰ نمونه داده تصادفی با توزیع نرمال با مقادیر پارامترهای $\mu = 5$ و $\sigma^2 = 1$ ایجاد کرده و مقدار MLE را برای میانگین داده‌ها و نیز تابع $\theta = e^{\mu}$ محاسبه می‌نمائیم:

$$\hat{\mu} = \bar{X}_n \quad , \quad \hat{\theta} = e^{\bar{X}_n}$$

سپس با استفاده از روش Bootstrap، به تعداد ۱۰۰۰۰۰ بار از بین همین ۱۰۰ عدد داده‌ی تصادفی تولیدشده با توزیع نرمال، تعداد ۱۰۰ عدد تصادفی را با جایگذاری انتخاب کرده و بر روی آن‌ها مقادیر قیدشده در بالا محاسبه می‌گردد. در نتیجه ما تعداد ۱۰۰۰۰۰ میانگین تخمینی جدید خواهیم داشت که می‌توانیم طبق رابطه‌ی زیر برای آن‌ها v_{boot} را محاسبه نمائیم:

$$v_{boot} = \frac{1}{B}\sum_{i=1}^{B}\left(T_{n,b}^{*} - \frac{1}{B}\sum_{r=1}^{B} T_{r,b}^{*}\right)^2$$

که در آن مقادیر $T_{n,i}^{*}$ همان مقادیر تخمین‌زده‌شده برای $\hat{\theta}$ می‌باشند. سپس با توجه به روابط زیر برای بازه‌های اطمینان خواهیم داشت:

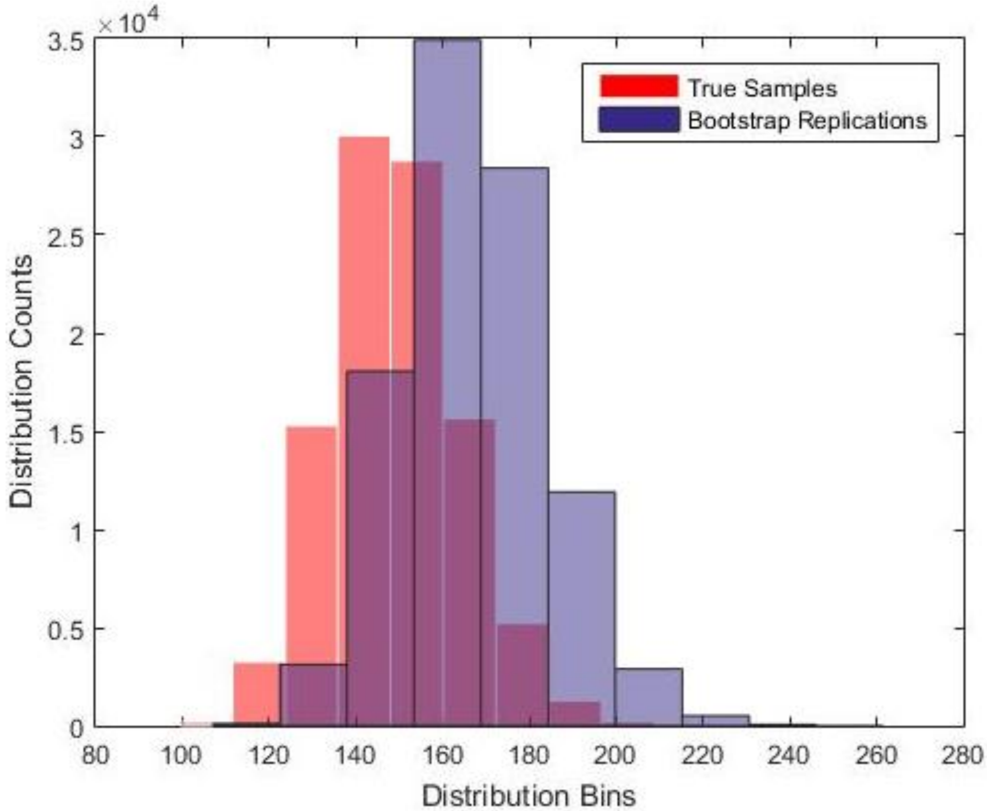
$$\text{Normal Interval} = \mathrm{T}_n \pm z_{\alpha/2}\,\hat{se}$$

$$\text{Pivotal Interval} = \left(2\hat{\theta}_n - \hat{\theta}^{*}_{1-\alpha/2} \;\; , \;\; 2\hat{\theta}_n - \hat{\theta}^{*}_{\alpha/2}\right)$$

$$\text{Percentile Interval} = \left(\theta^{*}_{\alpha/2} \; , \; \theta^{*}_{1-\alpha/2}\right)$$

Estimated Values	$\hat{\mu}$	5.1130
	$\hat{\theta}$	166.1733
Bootstrap Results	$\hat{se}$	16.9350
	Normal Interval	(132.9806 , 199.3660)
	Pivotal Interval	(129.5666 , 196.0596)
	Percentile Interval	(136.2869 , 202.7800)

ب) در این قسمت خواسته شده تا توزیع میانگین‌های داده‌های نرمال تولیدشده به صورت تصادفی به همان تعداد تکرار عمل Bootstrap (۱۰۰۰۰۰ بار، هربار ۱۰۰ داده تصادفی) با توزیع میانگین‌های حاصله از خود عمل Bootstrap (مطابق قسمت الف) بر روی نمودار مقایسه گردد که نتیج حاصله مطابق شکل زیر است:



همانطور که مشاهده می‌شود دو توزیع مشروح در بالا با یکدیگر تفاوت زیادی نداشته و مقادیر میانگین و واریانس دو توزیع در جدول زیر با یکدیگر مقایسه گشته‌اند:

Parameter	Bootstrap Replications	True Samples
Mean	149.1648	223.1146
Variance	167.0947	286.7987

سوال چهارم:

در اینجا برای اجرای آزمون موسوم به Wald برای بررسی برابری دو میانگین باید مقدار W را مطابق زیر محاسبه نمود که در آن از plug-in estimator های مربوط به میانگین‌های واقعی استفاده شده است:

$$W = \frac{\hat{\delta} - \theta_0}{\hat{se}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

$$\textit{where } s_1^2 \text{ and } s_2^2 \text{ are the sample variances.}$$

مقادیر حاصله برای W و p-value و confidence interval مطابق جدول زیر می‌باشد:

W	3.9446
p-value	0.0001
97% Confidence Interval	(0.0100 , 0.0344)

همانطور که از جدول بالا قابل مشاهده است مقدار p-value بسیار نزدیک به صفر است و بنا به توضیحات موجود در صفحه‌ی ۱۸۷ کتاب درسی، هرچه مقدار p-value کمتر باشد مدرکی محکم‌تر علیه فرضیه باطل (Null Hypothesis، در اینجا یکسان‌بودن دو نویسنده یعنی همان Mark Twain) خواهد بود. اما مطابق توضیحات مندرج در صفحه ۱۸۹ کتاب این نتیجه شاید از لحاظ آماری ارزشمند باشد اما اندازه‌ی اثر آن به دلیل کم‌بودن داده‌های موجود ناچیز بوده و به دنبال آن ما یک نتیجه‌ی آماری ارزشمند اما از لحاظ عملی بی‌ارزش و غیر قابل اتکاء خواهیم داشت، و در نتیجه عاقلانه خواهد بود که یک بازه‌ی اطمینان مثلاً ۹۷٪ نیز برای تفاضل میانگین‌ها مطابق جدول بالا اعلام نمائیم.

سوال پنجم:

مطابق مطلوب سؤال تعداد ۲۰ عدد داده‌ی تصادفی با توزیع Poisson و مقدار پارامتر $\lambda_0 = 1$ تولید نموده و سپس با استفاده از روابط مربوط به MLE مقادیر تخمینی $\hat{\lambda}$ و $\hat{se}$ را مطابق زیر محاسبه می‌نمائیم:

$$\textit{For} \text{ calculating } \hat{\lambda} \text{ we have:}$$

$$f(X_i;\lambda) = e^{-\lambda}\frac{\lambda^{X_i}}{X_i!}, \quad X_i \geq 0$$

$$L_n(\lambda) = \prod_{i=1}^{n} f(X_i;\lambda) = e^{-n\lambda}\frac{\lambda^S}{\prod_{i=1}^{n} X_i!} \quad \text{where} \quad S=\sum_{i=1}^{n} X_i$$

$$\rightarrow \ \ell_n(\lambda) = -n\lambda + S\log\lambda - \sum_{i=1}^{n}\log X_i! \xrightarrow[\textit{and} \text{ make it equal to } 0]{\textit{take} \text{ a derivative on } \lambda}$$

$$-n + \frac{S}{\lambda} = 0 \ \Rightarrow \ \hat{\lambda} = \frac{S}{n} = \bar{X}_n$$

$$\textit{For} \text{ calculating } \hat{se} \text{ we have:}$$

$$\log f(x;\lambda) = -\lambda + x\log\lambda - \log x! \ \rightarrow \ \text{s}(x;\lambda) = \frac{\partial \log f(x;\lambda)}{\partial\lambda} =$$

$$-1 + \frac{x}{\lambda} \ \Rightarrow \ \text{-s}'(x;\lambda) = x/\lambda^2 \ \rightarrow \ \text{I}_1(\lambda) = E_\lambda(\text{-s}'(x;\lambda)) = \lambda/\lambda^2 = 1/\lambda$$

$$\rightarrow \ \text{I}_n(\lambda) = n\text{I}_1(\lambda) = n/\lambda \ \Rightarrow \ \hat{se} = \sqrt{1/\text{I}_n(\lambda)} = \sqrt{\hat{\lambda}/n}$$

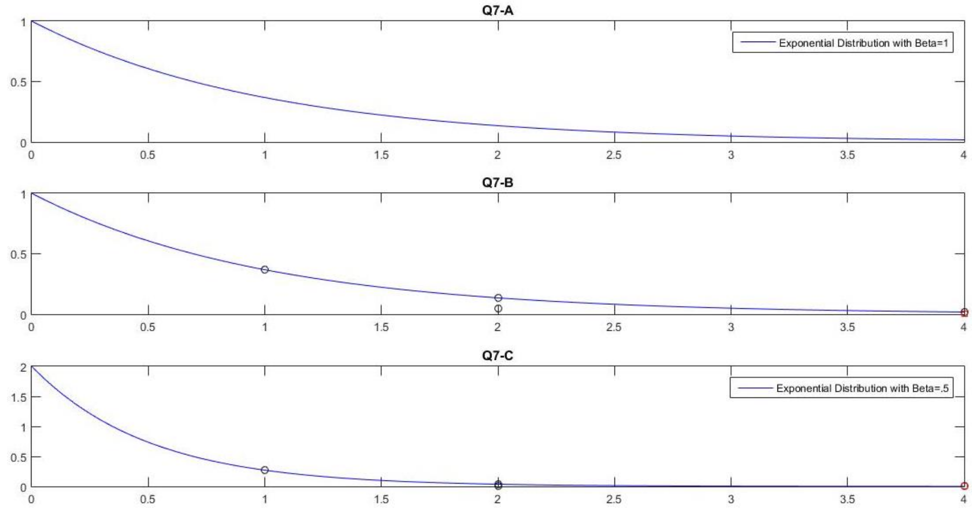
حال برای اجرای آزمون موسوم به Wald مقدار W را مطابق فرمول زیر محاسبه می‌نمائیم:

$$W = \frac{\hat{\lambda} - \lambda_0}{\sqrt{\hat{\lambda}/n}}$$

حال در اینجا عمل تولید داده‌های تصادفی را به تعداد ۱۰۰۰۰۰ بار تکرار کرده و هر بار مقادیر تخمینی را مطابق بالا محاسبه نموده و آزمون Wald را با فرض صحیح‌بودن $H_0: \ \lambda=\lambda_0$ روی آن‌ها انجام می‌دهیم و به عبارتی هر بار که $|W| > z_{\alpha/2}$ شد (در اینجا $z_{\alpha/2} = 1.96$) خطای نوع یک (type I error) رخ داده است و در واقع هدف از این آزمون بررسی این موضوع است که آیا مقدار فرض اولیه (Null Hypothesis) در تکرارهای متوالی در بازه‌ی اطمینان ۹۵٪ قرار می‌گیرد یا نه؛ که از آنجا که فرض اولیه صحیح در نظر گرفته شده انتظار آن است که این مسئله در ۹۵٪ مواقع برقرار بوده و در ۵٪ مواقع رد گردد که البته نتیجه‌ی حاصله از آزمایش نیز نرخ خطا را ۰،۰۵۳۰ یا همان ۵،۳٪ نشان می‌دهد که خود گواهی بر این موضوع می‌باشد.

سوال هفتم:

در هر سه مورد سؤال مطلوب مسئله رسم PDF توزیع نمائی در بازه‌ی [0,4] به ازای دو پارامتر متفاوت $\theta=1 \Rightarrow \beta=1$ و $\theta=2 \Rightarrow \beta=.5$ می‌باشد و نیز خواسته شده که مقادیر درستنمائی برای داده‌های مجموعه‌ی $X=\{1,2,4\}$ بر روی نمودار قبلی به ازای دو مقدار متفاوت پارامترهای قیدشده رسم گردد که نتایج در شکل زیر قابل مشاهده است:



همانطور که از شکل پیداست افزایش θ یا همان کاهش β سبب افزایش درستنمائی برای مجموعه داده‌ی آزمایشی قیدشده گردیده است چرا که مقادیر درستنمائی به مقادیر چگالی احتمال برای هر کدام از نقاط مجموعه‌ی فوق نزدیک‌تر گشته است.