پویا شیخ الاسلامی ٨١٠١٠٠٣٤٩

تکلیف سری دوم — استنباط آماری

1) a)

$X, Y \sim N(0,1)$

طبق راهنمایی‌های کلاس داریم:

$\rho(X,Y) = corr(X,Y) = \dfrac{cov(X,Y)}{\sigma_X \sigma_Y} = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$

$$\begin{cases} \sigma_X = E(X - \mu_X) \\ \sigma_Y = E(Y - \mu_Y) \\ \sigma_{XY} = E(XY) - \underbrace{E(X)}_{\mu_X}\underbrace{E(Y)}_{\mu_Y} \end{cases}$$

$m = max(X,Y)$

$l = min(X,Y)$

$\rho(X,Y) = corr(m,l) = \dfrac{\cancel{E(m.l)} - E(m)\cdot E(l)}{\sqrt{\left[E\{m^2\} - E^2\{m\}\right]\cdot\left[E\{l^2\} - E^2\{l\}\right]}} \Longrightarrow$

می‌دانیم

داریم

$$\begin{cases} E\{m.l\} = E\{X.Y\} = \mu_x \mu_y = \boxed{0} \\ E\{m+l\} = E\{X+Y\} = E\{X\} + E\{Y\} = 0 \quad \text{(I)} \\ E\{m-l\} = E\{|X-Y|\} = \sqrt{\dfrac{1}{\pi}} \end{cases}$$

داریم

$$\begin{cases} \text{(I)} \quad E(m) + E(l) = 0 \\ \text{(II)} \quad E(m) - E(l) = \sqrt{\dfrac{1}{\pi}} \end{cases}$$

$A = X - Y \sim N(0,2)$

$E\{|X-Y|\} = E\{|A|\} = \int_{-\infty}^{+\infty} |a| f_A(a)\, da =$

$\dfrac{-1}{\sqrt{4\pi}} \int_{-\infty}^{0} a\, exp(0.5a^2)\, da + \dfrac{1}{\sqrt{4\pi}} \int_{0}^{\infty} a\, exp(-0.5a^2)\, da$

$= 2\sqrt{\left(\dfrac{1}{4\pi}\right)} = \boxed{\sqrt{\dfrac{1}{\pi}}}$

$$\begin{cases} E(m) = \sqrt{\dfrac{1}{\pi}} \\ E(l) = -\sqrt{\dfrac{1}{\pi}} \end{cases}$$

همچنین

$$\begin{cases} E(m^2) = E(m.m) = \underbrace{E(X^2)}_{\mu_x} \text{ or } \underbrace{E(Y^2)}_{\mu_y} = \boxed{0} \\ \text{به طور مشابه } E(l^2) = \boxed{0} \end{cases}$$

$$\rho = \frac{\frac{1}{n}}{\sqrt{E\{m^2\}E\{e^2\} - \frac{1}{n}E\{m^2\} - \frac{1}{n}E\{e^2\} + \frac{1}{n^2}}} = \frac{\frac{1}{n}}{\sqrt{\frac{1}{n^2}}} = \boxed{1}$$

ب)

$$\text{if} \begin{cases} P(A\cap B) = P(A).P(B) \\ \text{or} \\ P(A|B) = P(A) \\ P(B|A) = P(B) \end{cases} \Longrightarrow$$

نشان دهنده استقلال دو متغیر A,B نسبت به هم است

$$\Longrightarrow P(\underbrace{X\cap X}_{X}) = P(X).P(X) \Longrightarrow P(X) = P(X)^2 \Longrightarrow$$

$$P(X).(1-P(X)) = 0 \Longrightarrow P(X) = 0$$

$$P(X) = 1$$

پس اگر متغیر همیشگی باشد ($P=1$) و یا غیرممکن باشد ($P=0$) دارای چنین خاصیتی است.

ج) (مثال نقض)

فرض کنید در دو آزمایش متوالی از هم ابتدا (W) مراحل: دو مقدار صفر و 4950 را با احتمال 0.99 و 0.01 به ترتیب تولید می‌کند داریم:

$$E(W) = 0.99\times 0 + 0.01\times 4950 = 49.5$$

و سپس در فرایند Z دو مقدار صفر و یک را با احتمال 0.01 و 0.99 تولید می‌کند پس

$$E(Z) = 0.99\times 1 + 0.01\times 0 = 0.99$$

در اینجا چنین اتفاقی متضاد است $\Longleftarrow$ و ترتیب نیز برقرار است $\begin{cases} E(W) > E(Z) \\ P(Z>W) = 0.98 \end{cases}$

>0.98 (با احتمال 0.98)

د) فضای سمتی نداریم ... هر قاضی جداگانه رأی، محکومیت مجرم را داده است convicted (محکوم شده)

$K<n$

a) G: گناهکار بودن — being guilty

$$P(G|X=k) = \frac{P.c^k.(1-c)^{n-k}}{P.c^k(1-c)^{n-k} + (1-P).v^k.(1-v)^{n-k}}$$

judges — براساس توزیع

b) unanimous vote $\rightarrow u$

systematic error $\rightarrow e$

$X=n$

طبق قانون احتمال کل

$$P(G|u) = \frac{P(u|G)P(G)}{P(u)} = \frac{P.P(u|G)}{(1-P)P(u|G_c) + P P(u|G)} =$$

guilty — محکوم نشود — محکوم شود

طبق قانون احتمال کل بازنویسی کنیم:

$$P(u|G) = P(u|G,e).P(e|G) + P(u|G,e_c).P(e_c|G) = \boxed{s+(1-s)c^n}$$

$$P(u|G_c) = P(u|G_c,e)P(e|G_c) + P(u|G_c,e_c).P(e_c|G_c) = \boxed{s+(1-s)v^n}$$

$$P(G|u) = \frac{P(s+(1-s)c^n)}{P(s+(1-s)c^n) + (1-P)(s+(1-s)v^n)}$$

Scanned with CamScanner

c) ن. صعودی است چون $\frac{c^n}{c^n+r^n}$ ~ ، $r < c$ هرچند که نباید رابطه دیگری میان برخی با راقیرها موجود باشد که در این جا به آن پرداخته نشده است

قابل ملاحظه باشد n و r ؛ به طور کلی این تصمیم گیری از روی عالمانه بودن منطقی است چون سبب می شود که اگر در تصمیم گیری جهت گیری واهی بوده یعنی قضات

نسبت به حکم کردن یا نکردن متهم جهت گیری فکری داشته باشند اگر از حد (k) عبور کنند تصمیم آن قضات تاثیرش بر محکومیت فرد در نظر

دادگاه کم شود.

3) چرا توزیع پواسون برای اینجا مناسب است؟ فرآیند پواسون یک فرآیند نقطه‌ای شمارشی است که بیانگر وقوع رخدادهای تصادفی در یک فاصله مکانی یا

زمانی تعریف می‌شود و برای مدل سازی واپاشی رادیواکتیو، تماس تلفنی و اتصال داده از آن استفاده می‌شود و برای نشان دادن متغیرهای گسسته مانند تعداد داغ

و ... مناسب است اما خود توزیع پیوسته در زمان است. فرآیند پواسون بی‌حافظه است که یعنی وقوع یک پیامد به پیامدهای قبلی وابسته نیست (از هم مستقل اند)

ویژگی دیگر پواسون این است که متوسط تعداد پیامدها ثابت است و هم چنین یک پیامد دارای دو حالت رخداد و عدم رخداد است که در هر زمان

ب.ظ 2015/06/07 04:59 فقط یکی از آن‌ها رخ می‌دهد با توجه به آنچه گفته شد و وجود ویژگی‌ها نظیر استقلال پیامدها (بی‌حافظگی)، گسسته بودن متغیر

تعداد ایمیل و وجود نرخ ثابت (میانگین ثابت) از مدل پوآسون استفاده می‌کنیم.
در هر بازه

بخش دوم سوال: λ ، p ، $1-p$

$\sigma_w^2 \quad \sigma_p^2$

w , p فرآیندهای نظیر تعداد ایمیل‌های کاری و شخصی (working and personal)

$$\Longrightarrow \quad W \sim pois(\lambda p t)$$
$$P \sim pois(\lambda(1-p)t)$$

- مدت زمان لازم برای پاسخگویی به تمام ایمیل‌های کاری $T_w \triangleq$

قانون میانگین کل $\longrightarrow$ $E(T_w) = E(E(T_w|W)) = E(W\mu_w) = \mu_w \lambda p t$

مدت زمان لازم برای پاسخگویی به تمام ایمیل‌های شخصی $T_p \triangleq$

$\longrightarrow E(T_p) = E(E(T_p|P)) = E(P\mu_p) = \mu_p \lambda(1-p)t$

برای تمام ایمیل‌ها $\Longrightarrow E(T) = E(T_w + T_p) = E(T_w) + E(T_p) = \lambda t(\mu_w p + \mu_p(1-p))$

استقلال

برای واریانس داریم

$\longrightarrow Var(T_w) = Var(E(T_w|W)) + E(Var(T_w|W)) = Var(\mu_w W) + E(\sigma_w^2 W) = \mu_w^2 Var(W) + \sigma_w^2 E(W) = \boxed{\lambda p t(\mu_w^2 + \sigma_w^2)}$

بطور مشابه برای T_p $\Longrightarrow Var(T_p) = \boxed{\lambda(1-p)t(\mu_p^2 + \sigma_p^2)}$

$Var(T) = Var(T_w) + Var(T_p) = \lambda t(p(\mu_w^2 + \sigma_w^2) + (1-p)(\mu_p^2 + \sigma_p^2))$

درنهایت به علت استقلال T_p و T_w داریم:

④ $\leftarrow M_j$

یک شاخص برای اعداد است $\longleftarrow$

که اگر minimum محلی باشد، برابر یک شود: $A[j]$

داریم: $E[M_1 + M_2 + \cdots + M_n] = E[M_1] + \cdots + E[M_n]$

به خاطر تقارن $E[M_1] = E[M_n] =$ border (مرزها)، $E[M_2] = E[M_3] = \cdots = E[M_{n-1}] =$ middle (وسط)

n^3 حالت می‌تواند برای اعداد میانی (middle) رخ دهد از بین تمام این حالات، حالاتی که شرایط زیر را دارند مجازند:

$$\begin{cases} left = a, middle = b, right = c, b < a, b < c, a \neq c \\ left = a, middle = b, right = a \quad b < a \end{cases}$$

تعداد حالت‌های
$\binom{n}{3} = \frac{n!}{(n-3)!} = n(n-1)(n-2)$
= گزینش سه عدد از بین n عدد

$\binom{n}{2} = n(n-1)$ = گزینش دو عدد از n عدد

از بین این تعداد $\frac{1}{3}$ شرط این که b (middle) مینیمم محلی باشد را دارا است.

از بین این تعداد فقط در $\frac{1}{2}$ حالت‌ها (middle) مینیمم محلی است.

داریم $E[M_j] = \text{Middle} = \dfrac{(n-1)(n-2)/3 + (n-1)/2}{n^2}$

با توجه به اعدادی که در حد border قرار دارند n^2 جفت وجود دارد:
که $n(n-1)$ تا از آن‌ها متمایزند و در نصف حالت‌ها عدد اول از دومی بزرگ‌تر است (طبق آنچه گفته شد):

$$E[M_1] = E[M_n] = \text{border} = \frac{(n-1)/2}{n}$$

$$\left[\frac{n-1}{n} + (n-2)\,\frac{(n-1)(n-2)/3 + n-1/2}{n^2}\right] = \boxed{\frac{n+1}{3}}$$

$n \to \infty$ پاسخ نهایی

---

(5) 15 people
پاسخ این مسئله 0.5 است. به طور کوتاه پاسخ این شرح است: صندلی که آخرین نفر وارد می‌شود تنها حالتی که برای صندلی خالی وجود دارد همانی است که شماره صندلی درست باشد یا همان صندلی شخص اول. دلیل این است که اگر مثلاً صندلی شخص اُم دهم خالی باشد وقتی که نفر آخر وارد می‌شود پس همانی که نفرات دهم هم وارد می‌شد نشسته خالی بوده است؛ پس باید رفت و سر جایش می‌نشسته است به طور قطع. برای سایر افراد (پس انتخاب بین دو حالت داریم $p = 1/2$)

اثبات:

$\boxed{p(f_i) = 1/n}$

مشابه به هر مرحله $n = i$

P_l ← نفر آخر (last) روی صندلی درست (خودش) بنشیند.

P_{f_i} ← نفر اول (first) روی صندلی i اُم بنشیند.

قانون احتمال کل: $p(l) = \sum_{1}^{n} p(l \mid f_i)\,p(f_i)$

$p(l \mid f_1) = 1$ → همه از جمله نفر آخر روی صندلی خودشون می‌نشینند.

$p(l \mid f_2) = P_{n-1}$

$p(l \mid f_3) = P_{n-2}$

$\vdots$

$p(l \mid f_{n-2}) = P_3$

$\vdots$

$p(l \mid f_n) = P_1 = 0$ (P_2)

داریم $p(l) = \frac{1}{n}(1 + P_{n-1} + P_{n-2} + \cdots + P_2)$ $n \geq 2$

$n = 2 \quad p(l) = 1/2 = P_2$

$n = 3 \quad P_3 = 1/3(1 + P_2) = 1/2$

$n = 4 \quad P_4 = 1/4(1 + P_2 + P_3) = 1/2$

بطور مشابه:

$$P_{150} = \frac{1}{150}(1 + P_2 + \cdots P_{149}) = \boxed{\frac{1}{2}}$$

$n = 150$

---

6) شمارش روزهای هفته از دوشنبه

a)

فرزند دوم دوشنبه

فرزند اول دوشنبه

B_1B_2

B_2B_1 $B_2B_2 \ldots B_2B_7\ B_2G_1 \ldots B_2G_7$

B_3B_2

$\vdots$

B_7B_2

G_1B_2

$\vdots$

G_7B_2

فضای نمونه‌ای مورد نظر ← 26 حالت است.

$13 = 14 - 1$ ← تعداد حالات مطلوب که پسر باشد (فرزند دیگر)

$$P = \boxed{\frac{13}{26}}$$

---

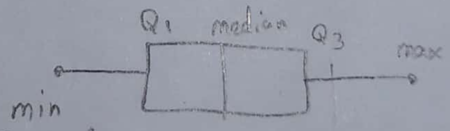
b)

مانند حالت a) است با این تفاوت که B_2B_2 حالت نامطلوب است و حذف می‌شود

$$p = \frac{13-1}{26} = \boxed{\frac{12}{26}}$$

---

7) a، b، c ← Code



d) شکل کلی یک نمودار جعبه‌ای به صورت زیر است که در آن

Q_1 چارک پایینی و Q_3 چارک بالایی باشد ($Q_3 - Q_1 = IQR$) یعنی IQR در واقع اختلاف چارک‌های بالا و پایین است. outlier ها داده‌هایی هستند که نسبت به توزیع داده‌های اصلی پرت محسوب می‌شوند یعنی نسبت به داده‌های دیگر فاصله‌ی بیشتری از آنها دارند. که این فاصله به صورت زیر تعریف می‌شود: (اگر از این فاصله داده بیشتر شود، داده outlier است.)

lower outer fence $= Q_1 - 3\,IQR$ → l.o.f

(l) ← در کد

upper outer fence $= Q_3 + 3\,IQR$ → u.o.f

(u) ← در کد

اگر داده در محدوده‌ی بین این دو نقطه نباشد outlier است.

$$\forall x \notin [\text{l.o.f}, \text{u.o.f}] \Rightarrow x \in \text{outlier}$$

که در کدها الگوریتم

---

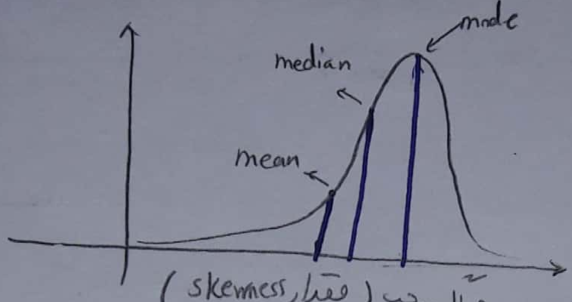
e) skewness (چولگی) نشان دهنده‌ی میزان تقارن یک متغیر تصادفی است (در نمودار توزیع PDF) که البته بیشتر برای متغیرهای تصادفی گسسته کاربرد دارد.

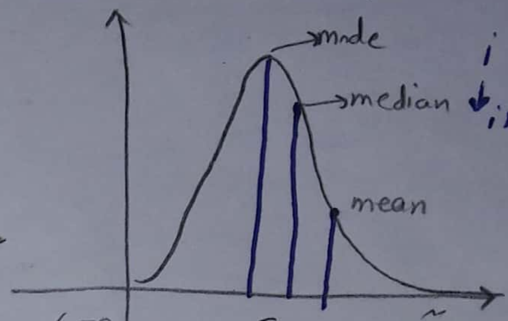
← فرمول رایج: $\text{skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$ ؛ اگر skewness مثبت باشد چولگی به راست داریم و اگر skewness منفی باشد چولگی به چپ داریم.

(8) a تا b ← کد

c) نحوه‌ی محاسبه‌ی outlier در سؤال هفت بررسی شد در این جا مقادیر در کد 7,66 که به دست آمدند.

e) i)





i) شکل کلی (در کد) ← چولگی به چپ دارد skewness منفی است. → -1.044596

ii) اگر skewness منفی باشد و چولگی به چپ داشته باشیم رابطه‌ی زیر برقرار است. (mean < median < mode) که پیشتر در شکل بالا به آن اشاره شد (با توجه به فرمول skewness) است. که در کد نیز مقادیر به دست آمده چنین است. mean < median

iii) میانه (median) و میانگین (mean) هر دو معیارهای خوب برای مرکز داده است که با توجه به نوع داده و ویژگی‌های آن می‌توان از یکی از آنها استفاده کرد. اگر داده توزیع متقارن باشد و چولگی نداشته باشیم از mean استفاده می‌کنیم ولی اگر skewness (و outlier) مقدار داشته باشد mean (میانگین) عامل مناسب‌تری است.

d) رابطه‌ی خطی با توجه به نمودار بین دو متغیر وجود ندارد. با توجه به نمودار نقطه‌ای

e) ارزش‌ها با توجه به نمودار جعبه‌ای (به طور تقریبی) و به طور قطعی در کد

$Q_1 = 12.25$
$Q_3 = 16.5$

هم‌چنین outlier نیز مقدار 5 دارد. (در کد)

f) با توجه به نمودار و مقدار skewness ≃ -0.74 (منفی) چولگی به چپ داریم؛ تفاوت نمودار هیستوگرام و چگالی (density) در این است که نمودار هیستوگرام چون بیشتر برای داده‌های گسسته به کار می‌رود نمی‌تواند پراکندگی داده‌ها را نشان دهد یعنی نمی‌توان با آن outlier داده‌ها را تشخیص داد حال آن که نمودار چگالی (density) چون برای نمایش داده‌های پیوسته کاربرد دارد ابزار چندان مناسبی برای تشخیص outlier‌ها نیست!

(9