سؤال اول : H_0 : نتایج بهبودی بیماران به نوع بیمارستان وابسته نبوده ،

H_1 : نتایج بهبودی بیماران به نوع بیمارستان وابسته هست.

برای پی بردن اینکه ۲ متغیر اسمی استقلال دارند از آزمون chi square استفاده میکنیم

Expected values:

① بهبودی کامل / خصوصی $= \frac{55 \times 60}{110} = 30$

② بهبودی کامل / دولتی $= \frac{50 \times 55}{110} = 25$

③ بهبودی جزئی / خصوصی $= \frac{35 \times 60}{110} = 19.09$

④ بهبودی جزئی / دولتی $= \frac{50 \times 35}{110} = 15.91$

⑤ عدم بهبود / خصوصی $= \frac{20 \times 60}{110} = 10.91$

⑥ عدم بهبود / دولتی $= \frac{20 \times 50}{110} = 9.09$

$$\chi^2 = \sum \frac{(\text{مقدار مورد انتظار} - \text{مشاهدات})^2}{\text{مقدار مورد انتظار}}$$

① $= \frac{(30-30)^2}{30} = 0$

② $= \frac{(25-25)^2}{110} = 0$

③ $= \frac{(20-19.09)^2}{19.09} = \frac{(.91)^2}{19.09} \approx 0.043$

④ $\frac{(15-15.91)^2}{15.91} \approx \frac{(-.91)^2}{15.91} \approx 0.053$

⑤ $= \frac{(10-10.91)^2}{10.91} = \frac{(-0.91)^2}{10.91} = 0.078$

⑥ $\frac{(10-9.09)^2}{9.09} \approx \frac{(0.91)^2}{9.09} \approx 0.091$

$$\chi^2 = 0 + 0 + 0.043 + 0.053 + 0.078 + 0.091 = 0.266$$

$$df = (r-1)(c-1) = (3-1)(2-1) = 2$$

r تعداد ردیف ، c تعداد ستون

با توجه به اینکه مقدار $df = 2$ ، $\alpha = 0.05$ مقدار بحرانی طبق جدول chi square برابر با 5.991 و $\chi^2 = 0.266$

در نتیجه با توجه $\chi^2 < 5.991$ ⇐ فرض صفر رد نمی شود

نکته ای که در این آزمون بود تعداد نمونه ها کم بود و نتوانست رابطه ی آماری بین دو متغیر ، نوع بیمارستان را نشان دهد در این آزمون مقادیر expected value نباید کمتر از 5 باشند ، در هیچ کدام نبود و می توان گفت این آزمون معتبر است

سؤال دوم : H_0 : هیچ تفاوتی در عملکرد دانش آموزان در این ۲ فرمت وجود ندارد

H_1 : تفاوت معنا داری در این ۲ فرمت وجود دارد

از paird t-test برای این سؤال استفاده میکنیم زیرا این آزمون مناسب برای مقایسه ۲ گروه وابسته (۲ فرمت) می باشد

ابتدا اختلاف ۲ فرمت را برای هر دانش آموز بدست آورده و میانگین آن را حساب میکنیم

$$\bar{D} = \frac{5+3+3+3+3+2-1+2+3+3+4+2}{12} = \frac{32}{12} = 2.67 \quad , \quad SD = \sqrt{\frac{1}{n-1}\sum (D_i - \bar{D})^2}$$

$$SD_D = \sqrt{\frac{(5-2.67)^2 + 4(3-2.67)^2 + (2-2.67)^2 + (-1-2.67)^2 + (2-2.67)^2 + (3-2.67)^2 + (3-2.67)^2 + (4-2.67)^2 + (2-2.67)^2}{11}} = \sqrt{2.164} \approx 1.47$$

$$t = \frac{D}{SD/\sqrt{n}} = \frac{2.67 \times 3.464}{1.47} \approx 6.29$$

با توجه به مقادیر $df = 12-1 = 11$ و $\alpha = 0.05$ مقدار بحرانی برابر است با 2.201 ، چون مقدار بدست آمده از آزمون t بزرگتر از 2.201 شد در نتیجه فرض صفر رد می شود

سؤال نهم: الف. آلرم ۲۵ ا ۱ ا ۱۷ . . .

سؤال سوم:

A A A A A A A A A A — A

عدد: ۷۵، ۷۷، ۷۸، ۷۹، ۸۰، ۸۱، ۸۲، ۸۳، ۸۴، ۸۵، ۸۵، ۸۶، ۸۶ و ۸۷ و ۸۸ ، ۸۸ ، ۸۸ ، ۸۸

رتبه: ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸، ۹، ۱۰٫۵، ۱۰٫۵، ۱۲٫۵، ۱۲٫۵، ۱۴، ۱۶ ۱۶ ۱۶ ۱۶

A

عدد: ۸۹ ، ۸۹ ، ۹۰ ، ۹۰ ، ۹۰ ، ۹۱، ۹۲

رتبه: ۱۸٫۵ ۱۸٫۵ ۲۱ ۲۱ ۲۱ ۲۳ ۲۴

برای ۴ مورد ۸۸ رتبه به صورت: $16=\frac{15+16+17}{3}$

" ۳ مورد ۹۰ " " : $21=\frac{20+21+22}{3}$

استفاده از آزمون Mann Whitney برای ۲ نمونه مستقل:

مجموع رتبه‌های گروه A: $U_1 : 1+2+3+4+5+6+7+8+9+10.5+16+21=92.5$

مجموع رتبه‌های گروه B: $U_2 : 10.5+12.5+12.5+14+16+16+18.5+18.5+21+21+23+24=207.5$

محاسبه میانگین و انحراف معیار

بنا بر فرمول‌های صورت سؤال

$$\mu_W=\frac{12\times(12+12+1)}{2}=\frac{12\times25}{2}=150$$

$$\sigma_W=\sqrt{\frac{12\times12\times25}{12}}=\sqrt{300}\simeq17.32$$

و $U_1=207.5-92.5=115$

محاسبه‌ی Z-score:

$$Z=\frac{U_1-\mu_W}{\sigma_W}=\frac{115-150}{17.32}\simeq\boxed{-2.02}$$

با توجه به $\alpha=0.05$ مقدار ناحیه بحرانی در جدول Z برابر ∓1.96 و در اینجا $Z=-2.02$ شد یعنی کمتر از -1.96 پس فرض صفر رد می‌شود و نتیجه می‌شود تفاوتی بین سطح تدریس ۲ گروه A و B وجود دارد.

سؤال چهارم:

نفر	تحصیلات	درآمد	رتبه تحصیلات	رتبه درآمد	d = اختلاف رتبه تحصیلات - رتبه درآمد	d^2
۱	۱۰	۲٫۵	۵	۶	-۱	۱
۲	۱۲	۳	۷٫۵	۷٫۵	۰	۰
۳	۱۵	۴	۱۱	۱۰	۱	۱
۴	۸	۱٫۸	۱	۱	۰	۰
۵	۱۶	۴٫۵	۱۲	۱۲	۰	۰
۶	۱۱	۲٫۸	۶	۵	۱	۱
۷	۹	۲٫۱	۳	۳	۰	۰
۸	۱۴	۴٫۲	۱۰	۹	۱	۱
۹	۱۳	۳٫۹	۹	۸	۱	۱
۱۰	۱۲	۳	۷٫۵	۷٫۵	۰	۰
۱۱	۱۴	۴	۱۰	۱۰	۰	۰
۱۲	۱۰	۲٫۶	۵	۴	۱	۱

$\sum d_i^2=6$

$$r_S=1-\frac{6\times6}{12(12-1)^2}=1-\frac{36}{1716}\simeq1-0.021=0.979 \rightarrow$$

با توجه به اینکه این مقدار نزدیک به ۱ هست یعنی همبستگی قوی بین تحصیلات و درآمد وجود دارد و فرض صفر رد می‌شود

قطعاً عوامل خارجی در این مورد تأثیر دارد بسیاری از رشته‌های تحصیلی هستند که میزان تحصیل در آن‌ها اهمیت نداشته و بازار کاری متفاوت دارند، حتی خیلی از افراد توی شغلی که هستن ربطی به مدرک تحصیلشان ندارد ولی طبق این اطلاعات بدست آمده در سؤال نشان دهنده‌ی یک رابطه‌ی معنادار بین تحصیلات و درآمد هست

سؤال پنجم: H_0: میانگین IQ برابر با ۱۰۷ یعنی $\mu=۱۰۷$

H_1: میانگین IQ بیشتر است از ۱۰۷ یعنی $\mu>۱۰۷$

$$\bar{x}=\frac{۱۰۰+۹۰+۱۳۵+۱۰۸+۱۰۷+۱۱۹+۱۲۷+۱۰۹+۱۰۵+۱۱۲+۹۵+۱۲۳+۹۸+۱۱۰+۱۲۰}{۱۵}=\frac{۱۶۵۸}{۱۵}=۱۱۰{,}۵۳$$

$$S=\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$$

$$\Rightarrow S=\sqrt{\frac{1}{۱۴}\Big((۱۰۰-۱۱۰{,}۵۳)^۲+(۹۰-۱۱۰{,}۵۳)^۲+(۱۳۵-۱۱۰{,}۵۳)^۲+(۱۰۸-۱۱۰{,}۵۳)^۲+(۱۰۷-۱۱۰{,}۵۳)^۲+(۱۱۹-۱۱۰{,}۵۳)^۲+(۱۲۷-۱۱۰{,}۵۳)^۲}$$

ادامه

$$\Rightarrow \sqrt{(۱۰۹-۱۱۰{,}۵۳)^۲+(۱۰۵-۱۱۰{,}۵۳)^۲+(۱۱۲-۱۱۰{,}۵۳)^۲+(۹۵-۱۱۰{,}۵۳)^۲+(۱۲۳-۱۱۰{,}۵۳)^۲+(۹۸-۱۱۰{,}۵۳)^۲+(۱۱۰-۱۱۰{,}۵۳)^۲+(۱۲۰-۱۱۰{,}۵۳)^۲}$$

$$=\sqrt{\frac{(-۱۰{,}۵۳)^۲+(-۲۰{,}۵۳)^۲+(۲۴{,}۴۷)^۲+(-۲{,}۵۳)^۲+(-۳{,}۵۳)^۲+(۸{,}۴۷)^۲+(۱۶{,}۴۷)^۲+(-۱{,}۵۳)^۲+(-۵{,}۵۳)^۲+(۱{,}۴۷)^۲}{۱۴}}$$

ادامه

$$\Rightarrow \sqrt{\frac{(-۱۵{,}۵۳)^۲+(۱۲{,}۴۷)^۲+(-۱۲{,}۵۳)^۲+(-۰{,}۵۳)^۲+(۹{,}۴۷)^۲}{۱۴}}=\sqrt{\frac{۲۱۷۲{,}۴۵}{۱۴}}=\sqrt{۱۵۵{,}۱۸}\approx ۱۲{,}۴۶$$

استفاده از آزمون t:

$$t=\frac{\bar{X}-\mu}{\sqrt{n}/S}=\frac{۱۱۰{,}۵۳-۱۰۷}{S/\sqrt{n}}=\frac{۳{,}۵۳}{\frac{۱۲{,}۴۶}{\sqrt{۱۵}}}=\frac{۳{,}۵۳}{۳{,}۲۱}=۱{,}۰۹$$

مقدار بحرانی برای $\alpha=۰{,}۰۵$ و $df=۱۵-۱=۱۴$ برابر است با ۱٫۷۶۱ در نتیجه با توجه به اینکه مقدار بدست آمده از آزمون t برابر است با ۱٫۰۹ و از ۱٫۷۶۱ کمتر هست نمی‌توان فرض صفر را رد کرد

سؤال ششم: از نرمال بودن چیزی گفته نشده mean whitney U

$$\bar{d}=\frac{۵}{۱۰}=۰{,}۵$$

X	Y	Diff	abs(diff)	Rank
۱۲۰	۱۱۸	2	۲	۳
۱۳۴	۱۳۶	-۲	۲	۳
۱۲۸	۱۳۰	-۲	۲	۳
۱۴۰	۱۳۷	۳	۳	۸
۱۴۵	۱۴۳	۲	۲	۳
۱۳۲	۱۳۰	۲	۲	۳
۱۳۸	۱۴۱	-۳	۳	۸
۱۲۵	۱۲۲	۳	۳	۸
۱۳۰	۱۳۳	-۳	۳	۸
۱۳۵	۱۳۲	۳	۳	۸
		d=۵		

مرتب سازی: ۳ ۳ ۳ ۳ ۳ ۲ ۲ ۲ ۲ ۲

رتبه :

میانگین رتبه برای ۲: $\frac{۱+۲+۳+۴+۵}{۵}=۳$

میانگین رتبه برای ۳: $\frac{۱۰+۹+۸+۷+۶}{۵}=۸$

جمع رتبه‌های مثبت: $۳+۸+۳+۳+۸+۸=۳۳$

جمع رتبه‌های منفی: $-۳-۳-۸-۸=-۲۲$

$$\Rightarrow \min(|-۲۲|,۳۳)=۲۲$$

به ازای $n=۱۰$ و $\alpha=۰{,}۰۵$ مقدار بحرانی طبق جدول Wilcoxon، ۸ می‌شود و با توجه به اینکه مقدار بدست آمده ۲۲ هست، بیشتر از ۸ هست فرض صفر رد نمی‌شود یعنی تفاوت معناداری بین ۲ برند وجود ندارد

سوال هفتم: H_0: هیچ تفاوت معناداری بین میانگین تعداد تلاش‌ها در گروه‌ها وجود ندارد

H_1: تفاوت معناداری بین " " " " وجود دارد.

با استفاده از mean-whitney U چون از نرمال بودن در سوال چیزی گفته نشده.

$$\text{coached group}: \begin{cases} 1 \quad 2 \quad 3 \quad 4 \quad 6 & : \text{Rank} \\ 10 \quad 11 \quad 14 \quad 15 \quad 18 & : \text{Point} \end{cases} \Rightarrow 1+2+3+4+6=16$$

$$\text{independ group}: \begin{cases} 5 \quad 7 \quad 8 \quad 9 & : \text{Rank} \\ 17 \quad 19 \quad 20 \quad 22 & : \text{point} \end{cases} \Rightarrow 5+7+8+9=29$$

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1+1)}{2} - R_1 = 5\times 4 + \frac{5\times(5+1)}{2} - 16 = 20+15-16=19$$

$$U_2 = 5\times 4 + \frac{4\times(4+1)}{2} - 29 = 1$$

$$\Rightarrow U = \min(U_1, U_2) = \min(19, 1) = 1$$

با توجه به تعداد نمونه های ۵ و ۴ و $\alpha = 0.05$ مقدار بحرانی طبق جدول آزمون mean-whitney U برابر ۲ بوده. با توجه $1 < 2$ در نتیجه تفاوت معناداری بین میانگین تلاش ها بوده و فرض صفر رد می شود.

سوال هشتم: a)

$$F_n(x) = \frac{1}{n}\sum I(X_i \leq x) \qquad I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{else} \end{cases} \Rightarrow$$ توزیع برنولی و $p = F(x)$

$$\Rightarrow Var(F_n(x)) = Var\left(\frac{1}{n}\sum I(X_i \leq x)\right) = \frac{1}{n^2}\sum var(I(X_i \leq x))$$

$$= \frac{1}{n^2} \times n \times F_n(x)(1 - F_n(x)) = \frac{F_n(x)(1-F_n(x))}{n}$$

b) نامساوی DKW inequality:

$$P(\sup|F_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}, \quad \varepsilon = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}$$

بازه اطمینان: برای سطح اطمینان ۹۵ درصد داریم:

$$P\left(\sup|F_n(x) - F(x)| < \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}\right) \geq 1-\alpha$$

$$F_n(x) = \frac{\text{Point هر Rank}}{\text{جمع کل Point ها}}$$

c) برای همگرایی یکنواخت باید داشته باشیم $\sup|F_n(x) - F(x)| \xrightarrow{n\to\infty} 0$

به صورت احتمالی: $P(\sup_x|F_n(x) - F(x)| > \varepsilon) \xrightarrow{n\to\infty} 0$

از نامساوی DKW استفاده می کنیم

$$P(\sup_x|F_n(x) - F(x)| > \varepsilon) \leq 2e^{-n\varepsilon^2}$$

طبق سمت راست نامساوی اگر $n\to\infty$ میل کند عبارت $2e^{-n\varepsilon^2}$ به صفر میل می کند

$$\Rightarrow P(\sup_x|F_n(x) - F(x)| > \varepsilon) \xrightarrow{n\to\infty} 0 \Rightarrow P(\sup_x|F_n(x) - F(x)| \leq \varepsilon) \underset{n\to\infty}{=} 1-0 \rightarrow$$

این عبارت نشان می دهد که عبارت $\sup|F_n(x) - F(x)|$ با احتمال نزدیک به ۱ کمتر از ε هست. یعنی به ازای هر مقدار $\varepsilon > 0$ تابع $F_n(x)$ به صورت یکنواخت نزدیک به $F(x)$ می شود

d) با توجه به رابطه $\varepsilon \propto \frac{1}{\sqrt{n}}$ وقتی $n\to\infty$ باند اطمینان کوچکتر می شود در نتیجه eCDF به صورت دقیق‌تری $F(x)$ را تقریب می زند

که n کوچک باشد باند اطمینان بزرگتر شده و ممکن است eCDF تقریب خوبی از $F(x)$ نزدهد.

سوال نهم: توزیع گاما: $f(x,k,\theta)=\frac{1}{\Gamma(k)\theta^{k}}x^{k-1}e^{-x/\theta}$ طبق صورت سوال $Q(p)=F^{-1}(p)$

$$F(x,k,\theta)=\int_0^x \frac{t^{k-1}e^{-t/\theta}}{\theta^k\Gamma(k)}\,dt$$

با توجه به اینکه $F(Q(p),k,\theta)=p$ در نتیجه

$$p=\int_0^x \frac{t^{k-1}e^{-t/\theta}}{\theta^k\Gamma(k)}\,dt$$

c) اگر داده‌های تبدیل شده دقیقاً با کوانتایل های تئوری تطابق داشته باشند یعنی برای کوانتایل مشاهده شده (تجربی) داریم $q_i=T(x_i)$ در نتیجه نقاط در نمودار Q-Q قرار می‌گیرند یعنی در واقع نقاط روی یک خط راست با شیب 1 و عرض از مبدأ صفر یعنی همین نقاط در نمودار Q-Q به طور کامل هم‌راستا با $y=x$ خواهند بود طبق فرض $Q(p_i)=T(x_i) \Rightarrow x=\underbrace{Q(p_i)}_{\text{کوانتایل نظری}}=\underbrace{T(x_i)}_{\text{مقادیر تبدیل شده}}=y$

d) انتخاب نادرست فرضیات توزیعی می‌تواند باعث انحراف نقاط از خط مستقیم در نمودار Q-Q شود. اگر توزیع داده‌ها متفاوت از فرضیه باشد ممکن است در نمودار Q-Q چاله هایی در بعضی از قسمت های نمودار ایجاد شود. همچنین توزیع های خاص ممکن است حساسیت بیشتری به مقادیر پرت نشان دهند، باعث شوند که این نقاط از الگوی خطی نمودار Q-Q فاصله بگیرند.

سوال دهم: آزمون KS تست برای مقایسه‌ی ۲ توزیع تجربی به صورت زیر تعریف می‌شود

$D_n=\sup_x|F_A(x)-F_B(x)|$ ، $F_A(x)$ و $F_B(x)$ توابع توزیع تجمعی برای داده های شبیه سازی شده از ۲ شعبه A, B هستند

به صورت کلی $F(x)$ برای هر شعبه به صورت $F(x)=\frac{\text{تعداد مقادیر کوچکتر از } x}{\text{تعداد کل مقادیر}}$

$\sqrt{n}D_n \xrightarrow{n\to\infty} \sup_x|B(F(t))|$

H_0: هر دو توزیع یکسان هستند H_1: دو توزیع متفاوت هستند

$\sqrt{n}D_n$ به توزیع کولموگروف اسمیرنوف همگرا می‌شود یعنی:

* (1) $P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}\,D\le t\right)=H(t)$

فرض صفر رد می‌شود اگر $\sqrt{\frac{n_1n_2}{n_1+n_2}}\,D>H^{-1}(1-\alpha)$ باشد و مقدار بحرانی از توزیع کولموگروف $H^{-1}(1-\alpha)$ برای سطح معناداری $\alpha=0.05$ به صورت $H^{-1}(0.95)=1.358$ محاسبه می‌شود یعنی اگر $\sqrt{\frac{n_1n_2}{n_1+n_2}}\,D_n>1.358$ شود، فرض صفر رد می‌شود

$\Rightarrow D_{critical}=1.358\sqrt{\frac{n_1+n_2}{n_1n_2}}$

c) باید اثبات کنیم که توزیع آماره آزمون D_n به توزیع $F(x)$ وابسته نباشد (تحت فرض صفر) یعنی توزیع آماره فقط به تعداد نمونه وابسته است برای این کار توزیع آماره به توزیع یکنواخت تبدیل می‌کنیم به صورت: $U_i=F(x)$ یعنی

$$P(U_i\le u)=P(F(x)\le u)=P(X_i\le F^{-1}(u))=F(F^{-1}(u))=u$$

در نتیجه مقایسه $F_n(x)$ با $F(x)$ معادل مقایسه eCDF مربوط به U_i با تابع یکنواخت روی بازه $[0,1]$ است و یعنی آماره آزمون D_n به توزیع اصلی $F(x)$ وابسته نیست

همچنین طبق (1): $\lim P(\sqrt{n}D_n\le t)=\overbrace{1-2\sum(-1)^{j-1}e^{-2j^2t^2}}^{H(t)}$ طبق این عبارت هم می‌توان ثابت کرد که توزیع آماره به $F(x)$ وابسته نیست و به اندازه نمونه وابسته است

برای حالت ۲ نمونه‌ای هم با ۲ تبدیل $U_i=F(x_i)$ ، $V_i=F(y_i)$ می‌توان گفت U و V روی بازه $[0,1]$ خواهند بود، و $D_n=\sup|F_{n_1}(x)-F_{n_2}(x)|$ نیز فقط به توزیع یکنواخت وابسته است نه به توزیع اولیه $F(x)$

d) آزمون KS به علت استفاده بیش از حد از بیشترین اختلاف بین توزیع های ECDF به تفاوت در دم های توزیع حساس
در دم ها داده های کمتری وجود داشته که باعث می شود اختلاف نسبی بین ECDF ها بزرگ شده و D_n را تحت تاثیر قرار دهد
شاید این حساسیت در بعضی از موارد مانند ریسک کردن با مسائل مالی خوب باشد ولی ممکن است تفاوت های موجود در بخش
وسط توزیع را در نظر نگیرد.

سوال یازدهم:

$$KDE: \hat{\pi}_h(x) = \frac{1}{n}\sum \frac{1}{h} K\left(\frac{x-x_i}{h}\right), \quad MISE(h)=\int_{-\infty}^{\infty}(Bias^2+Var)\,dx$$

$K(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$ $\qquad$ $K(u)$: تابع کرنل که $\int K(u)du=1$

h: پهنای باند کرنل

$$MISE=E[ISE];\ ISE=\int_{-\infty}^{\infty}[\hat{\pi}(x)-\underbrace{\pi(x)}_{\text{چگالی واقعی}}]^2dx = E\left[\int_{-\infty}^{\infty}\hat{\pi}^2(x)dx-2\int_{-\infty}^{\infty}\hat{\pi}(x)\pi(x)dx+\int_{-\infty}^{\infty}\pi^2(x)dx\right]$$

$$\Rightarrow MISE=\frac{1}{nh}\int K(u)^2du+\frac{h^4}{4}\left(\int u^2K(u)du\right)\int \pi''(x)^2dx$$
بعدش باید فرمول بایاس، واریانس حساب شه که در نهایت

$$\Rightarrow MISE=\frac{1}{2\sqrt{\pi}\,nh}+\frac{h^4}{4}\int \pi''(x)^2dx \Leftarrow \int K(u)^2du=\frac{1}{2\sqrt{\pi}} \Leftarrow K(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$$
$$\int u^2K(u)du=1$$
کرنل گوسی

«شرح در اینترنت»

c) برای پیدا کردن h^* بهینه باید $\frac{d}{dh}MISE(h)=0$

$$\Rightarrow \frac{d}{dh}MISE(h)=\frac{d}{dh}\left(\frac{1}{2\sqrt{\pi}nh}\right)+\frac{d}{dh}\left(\frac{h^4}{4}\underbrace{\int \pi''(x)^2dx}_{C\to\text{ثابت نسبت به } h}\right)=0$$

$$\frac{d}{dh}\left(\frac{1}{2\sqrt{\pi}nh}\right)=-\frac{1}{2\sqrt{\pi}nh^2} \qquad \frac{d}{dh}\left(\frac{h^4}{4}C\right)=\frac{C}{4}\times 4h^3=Ch^3$$

$$\Rightarrow -\frac{1}{2\sqrt{\pi}nh^2}+Ch^3=0 \Rightarrow Ch^3=\frac{1}{2\sqrt{\pi}nh^2}\Rightarrow Ch^5=\frac{1}{2\sqrt{\pi}nC}\Rightarrow h=\sqrt[5]{\frac{1}{2\sqrt{\pi}nC}},\ C=\int \pi''(x)^2dx$$

با توجه به اینکه اگر h خیلی کوچک باشد واریانس زیاد، بایاس کم می شود و اگر h زیاد باشد واریانس کم، بایاس زیاد می شود با تقارن داده
h^* بهینه AMISE به حداقل می رسد

d) در KDE با داده های چندقله انتخاب پهنای باند (h) اهمیت دارد زیرا اگر پهنای باند کوچک باشد بایاس افزایش پیدا کرده، اگر پهنای باند خیلی
بزرگ باشد ممکن است کلاً قله ها را در نظر نگیرد یعنی قله ها را هم و صاف کند و نشود تشخیص داد به طور کلی انتخاب h^* بهینه بستگی به
تعادل میان بایاس و واریانس دارد

سوال دوازدهم: میانگین برش خورده به این معناست که درصدی از داده ها در دو طرف توزیع (بیشترین ها و کمترین ها) حذف شود
که در اینجا ۱۰ درصد از ۲۰ داده یعنی ۲ داده انتهای توزیع و ۲ داده ابتدای توزیع حذف شوند و میانگین نمونه های باقی مانده
حساب شود: اعداد حذف شده {۳۴۰، ۳۳۵، ۲۲۵، ۲۲۰} $\qquad$ قسمت a, b)

Trimmed mean: $\frac{۲۳۰+۲۴۰+۲۵۰+۲۶۰+۲۷۰+۲۷۵+۲۸۰+۲۸۵+۲۹۰+۲۹۵+۳۰۰+۳۱۰+۳۱۵+۳۲۰+۳۲۵+۳۳۰}{۱۶}=\frac{۴۵۷۵}{۱۶}=\boxed{۲۸۵,۹۳}$

مزیت استفاده از این روش این است که این روش حساسیت کمتری نسبت به مقادیر بیرونی یا همون outliers دارد یعنی وقتی اعداد خیلی بزرگ یا
خیلی کوچک داریم حذف می شود تا عدد بدست آمده نماینده بهتری برای داده ها باشد

میانگین حسابی $=\frac{۴۵۷۵+۲۲۰+۲۲۵+۳۳۵+۳۴۰}{۲۰}=\frac{۵۶۹۵}{۲۰}=\boxed{۲۸۴,۷۵}$

ادامه سوال دوازدهم :

c) با اضافه شدن عدد ۴۵۰ میانگین برش خورده !!

۲۲۰ + ۲۲۵

$$\text{میانگین حسابی} = \frac{۲۳۰ + ۲۴۰ + ۲۵۰ + ۲۶۰ + ۲۷۰ + ۲۷۵ + ۲۸۰ + ۲۸۵ + ۲۹۰ + ۲۹۵ + ۳۰۰ + ۳۱۰ + ۳۱۵ + ۳۲۰ + ۳۲۵ + ۳۳۰ + ۴۵۰}{۱۷} = \frac{۵۶۷۵}{۲۱} = ۲۷۰,۸$$

تقریباً می‌توان گفت میانگین حسابی (MPa) تحت تأثیر مقدار بیرونی ۴۵۰ قرار گرفته یعنی میانگین حسابی حساس به مقادیر بیرونی است

$$\text{Trimmed mean} = \frac{۲۳۰ + ۲۴۰ + ۲۵۰ + ۲۶۰ + ۲۷۰ + ۲۷۵ + ۲۸۰ + ۲۸۵ + ۲۹۰ + ۲۹۵ + ۳۰۰ + ۳۱۰ + ۳۱۵ + ۳۲۰ + ۳۲۵ + ۳۳۰ + ۳۳۵}{۱۷} = ۲۸۸,۸$$

اما می‌توان گفت میانگین برش خورده مقاومت بیشتری نسبت به مقادیر بیرونی دارد

d) با توجه به اینکه داده‌ها را مرتب کرده و α درصد از ابتدا و انتها حذف می‌کنیم وقتی α به ۵۰ درصد نزدیک می‌شود تقریباً همه‌ی داده‌ها حذف شده و فقط داده‌های وسط توزیع برای محاسبه‌ی میانگین استفاده می‌شود در نتیجه این داده‌های باقی مانده به میانه توزیع مربوط می‌شوند یعنی اگر تعداد n زوج باشد ۲ مقدار وسطی میانه بوده و اگر فرد باشد مقدار وسط میانه توزیع خواهد بود

سوال سیزدهم : a)

$$\text{m-estimate} : y^* = \arg\min \sum_{i=1}^{n} \psi(x_i, y)$$

تابع وزنی ψ به عنوان تابع هزینه تعریف می‌شود به ۲ صورت :

1. میانه: که در این صورت برابر است با قدر مطلق انحراف یعنی $|x_i - y|$
2. میانگین: که تابع ضرر برابر است با مربع انحراف یعنی $(x_i - y)^2$

اثبات :

$$\sum (x_i - y)^2 \Rightarrow \frac{\partial}{\partial y} \sum (x_i - y)^2 = 0$$
$$\Rightarrow 2\sum x_i - y = 0$$
$$\Rightarrow 2\sum x_i - 2ny = 0 \Rightarrow y = \frac{\sum x_i}{n} = \text{میانگین}$$

b) 1. میانه (حداقل کردن کردن قدر مطلق انحراف) : با توجه به فرد بودن تعداد نمونه‌ها عدد وسط یعنی ۱۱۰ میانه خواهد بود

2. میانگین (حداقل کردن مربع انحراف) :

$$\text{میانگین} = \frac{۹۵ + ۹۵ + ۱۰۰ + ۱۰۰ + ۱۰۵ + ۱۰۵ + ۱۱۰ + ۱۱۰ + ۱۱۵ + ۱۱۵ + ۳۰۰ + ۳۰۵ + ۳۱۰ + ۳۱۵ + ۳۲۰}{۱۵} = \frac{۲۶۰۰}{۱۵} = ۱۷۳,۶ \ \mu m$$

c) اگر تابع وزنی را به صورت $\psi(x_i, y) = |x_i - y|^p$ تعریف شود وقتی $p>1$ باشد مقادیر پرت یا فاصله‌ی تأثیر بیشتری در تابع ضرر می‌گذارند همانطور که در قسمت قبل دیدیم وقتی $p=2$ شد به میانگین حسابی رسیدیم و می‌دانیم در میانگین حسابی اعداد پرت تأثیر زیادی می‌گذارند یعنی اگر $p>1$ باشد تأثیر بیشتری بر تابع ضرر خواهند داشت

d) با تنظیم وزن‌ها می‌توان مقاومت تخمین را در داده‌های پرت بیشتر کرد در این روش تعامل پرت به دلیل انحراف معیار زیاد از معمول وزن کمتری دریافت می‌کنند و سهم آنها در محاسبات کاهش می‌یابد و برای انحراف معیارهای کم وزن 1 دریافت می‌کنند که باعث می‌شود مقادیر کم هم تأثیری در تخمین بگذارند به طور کلی میانگین به داده‌های پرت حساسیت کمتری پیدا می‌کند

سوال پانزدهم :

$$p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{\text{تعداد کل موفقیت‌ها}}{\text{تعداد کل شرکت کننده‌ها در دو کمپین}}$$

نرخ موفقیت کلی

$$p_1 = \frac{\text{تعداد موفقیت کمپین اول}}{\text{تعداد کل شرکت کنندگان کمپین اول}}$$

$$p_2 = \frac{\text{تعداد موفقیت کمپین دوم}}{\text{تعداد کل شرکت کنندگان کمپین دوم}}$$

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$CI = (p_1 - p_2) \pm Z_{\alpha/2} \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{P-value} = 2\varphi(-|z|)$$

a قسمت ۲ پاسخ داده شده است !!