تمرین سری چهارم آمار و احتمال

آزمون فرض و نیکویی برازش

پاسخ سؤال سوم

این تابع بر اساس داده های مشاهده شده از یک نمونه ساخته می شود. نشان می دهد که چند درصد از داده های نمونه کمتر یا مساوی یک مقدار مشخص مثل (x) هستند.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_i (X_i \leq x)$$

n: تعداد کل مشاهدات در نمونه

X_i: مقادیر داده های نمونه

$\mathbb{I}(X_i \leq x)$: درصورتی که X_i $x \geqslant$ باشد برابر ۱ و درغیر این صورت برابر صفر خواهد بود.

در واقع برای محاسبه ی $F_n(x)$ در یک نقطه ی خاص:

تعداد مشاهداتی که کوچکتر مساوی با x هستند رو می شماریم

این تعداد رو به تعداد کل مشاهدات تقسیم می کنیم.

این ماجرا گسسته و پیوسته هم نداره و به تعداد نمونه حالتش داره از اون جایی که نمونه گیری ما قطعاً گسسته است پس همواره با این $\sum$ خسته کار جمع میشه :)

پاسخ سؤال پنجم

الف) فرضیه ها:

فرض صفر: بیانگر ادعای محافظه‌کارانه یا وضعیت موجود هست.

$H_0: P \geq \alpha$

در اینجا نسبت دانش آموزانی است که دچار آلرژی هستند

$H_0:$ حداقل ۱۰٪ از دانش آموزان از آلرژی رنج می‌برند. $H_0: P \geq 0.1$

فرض یک: فرض مقابل و رد کننده‌ی فرض صفر است. $H_1: P < \alpha$

$H_1:$ کمتر از ۱۰٪ از دانش آموزان از آلرژی رنج می‌برند. $H_1: P < 0.1$

این یک آزمون فرضیه‌ی یک طرفه است زیرا $P < 0.1$ بیان شده (left_tailed test)

ب)

آماره‌ی آزمون مناسب برای این مسئله

Z_test است به دلایل زیر:

1. نوع داده: باینری است. برای داده‌های باینری که توزیع دو جمله ای استفاده می‌شود اما به دلیل تعداد نمونه‌های بالا به توزیع نرمال میل خواهد کرد که در این وضعیت بهتر است که تست های t و یا Z استفاده کرد.

2. انحراف معیار جامعه معلوم است. به علت اینکه P_0 را داریم می‌توانیم انحراف معیار رو بدست بیاریم $\sqrt{P_0(1-P_0)}$

3. حجم نمونه: چون حجم نمونه بالاست شرایط استفاده از تقریب نرمال رو مهیا میکنه.

4. آزمون تک نسبتی: به علت اینکه نسبت دانش آموزان با ویژگی آلرژی در حال بررسی است و برای آزمون های نسبت تست Z مناسب است.

5. برقراری شروط:

$$5 \leq P_0 \cdot n \Rightarrow 5 \leq 22.5 \checkmark$$

$$5 \leq n \cdot (P_0 - 1) \Rightarrow 5 \leq 202.5 \checkmark$$

به همین دلایل استفاده از آزمون Z برای این مسئله کاملاً مناسب است.

یافتن مقدار P_Value مناسب

مراحل زیر را طی می‌کنیم:

1. محاسبه‌ی آماره‌ی آزمون Z
2. محاسبه‌ی P_Value
3. تفسیر P_Value

① $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$ s.t. $\begin{cases} \hat{p} = \dfrac{x}{n} = \dfrac{21}{225} = 0.093 \\ p_0 = 0.1 \\ n = 225 \end{cases}$ $\Rightarrow Z = \dfrac{0.1 - 0.093}{0.02} = -0.33$

② طبق قضیه‌ی حد مرکزی درصورتی که n بزرگ باشد، توزیع نسبت یا $\hat{p}$ به سمت نرمال میل می‌کند به همین دلیل آماره‌ی آزمون Z از توزیع نرمال $N(0,1)$ پیروی خواهد کرد.

$$P\_value = \mathbb{P}(Z \leq z) \rightarrow P\_value = \mathbb{P}(Z \leq -0.33)$$

$$P\_Value = CDF(z) = CDF(-0.33) = 0.3094$$

③ مقدار P_Value بیانگر این است که اگر فرض صفر درست باشد، احتمال اینکه یک آماره‌ی آزمون Z با این مقدار و یا نادرتر مشاهده شود برابر با 30.94٪ است.

$$0.05 = \alpha < 0.3094 = P\_value$$

1C

از آن جایی که P_Value بزرگتر از α است ما نمی‌توانیم فرض صفر را رد کنیم.

فرض صفر: نسبت دانش‌آموزانی که آلرژی دارند، حداقل ۱۰ درصد است.

بنابراین با 95 درصد اطمینان، نمی‌توانیم ادعا کنیم که کمتر از ۱۰ درصد دانش‌آموزان از آلرژی رنج می‌برند. یا به عبارت دیگر شواهد کافی برای حمایت از این ادعا که نسبت آلرژی کمتر از ۱۰ درصد است وجود ندارد.

پاسخ سوال ششم

۹)

فرض صفر H_0 : میانگین مقاومت پارچه μ حداقل 18 است. $\quad H_0: \mu \geq 18$

فرض یک H_1 : میانگین " " μ کمتر از 18 است. $\quad H_1: \mu < 18$

از آن جایی که فرض ما یک طرفه است پس آزمون یک طرفه می زنیم (left_tailed test)

میانگین μ_0 : میانگین جامعه تحت فرض صفر $\quad \mu_0 = 18$

انحراف معیار : در صورت سوال داده شده $\quad \sigma = 0.5$

سطح معنا داری : $\quad \alpha = 0.05$

شرایط محاسبه برای Z تست

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} , \quad \begin{cases} \bar{x} = \frac{1}{10}(17.2 + \dots + 17.3) = 17.25 \\ \mu_0 = 18 \\ \sigma = 0.5 \\ n = 10 \end{cases}$$

$$\Rightarrow Z = \frac{17.25 - 18}{\frac{0.5}{\sqrt{10}}} \cong -4.11$$

محاسبه P_Value

$$P\_Value = CDF(Z) = CDF(-4.11) = 0.00197$$

در صورتی که $P\_Value < \alpha$ فرض صفر رد می شود ✓ ← فرض صفر رد می شود

در صورتی که $P\_Value > \alpha$ فرض صفر را رد نمی کنیم.

پس میانگین مقاومت بار پارچه حداقل 18 نیست ← بار را قبول نمی کنیم

علت استفاده از آزمون Z : 1 داشتن واریانس ← از z تست استفاده می کنیم

2 کوچک بودن جامعه اما در صورت سوال ذکر شده که از توزیع نرمال پیروی می کند

(ط

توی این بخش یه جورایی داریم آزمون فرضی که توی بخش قبلی زدیم رو با یه آزمون فرض دیگه صحت سنجی می‌کنیم.

H_0 : احتمال پذیرش $\mu \geq 18$

H_1 : احتمال عدم پذیرش $\mu < 18$

حالا اگر میانگین واقعی 17.5 باشد، ممکنه که آزمون فرض این میانگین را به مقدار 18 تشخیص بده و بارش پذیرفته شود
همون مقادیری که احتمال پذیرش رو حساب کنیم، یعنی احتمال اینکه فرض H_0 رد نشود.

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{-0.5}{\frac{0.5}{\sqrt{10}}} \simeq -3.162$$

درسته یک لحظه با توجه به اینکه سطح اطمینان $\alpha = 0.05$ هست و توزیع نرمال در این داستان

$$Z_\alpha = -1.645$$

ما H_0 رو رد می‌کنیم اگه $Z < -1.645$

و پذیرش H_0 در حالتی که $Z > -1.645$

$$P(\text{reject } H_0) = P(Z < -3.162) = 0.000782$$

$$P(\text{accept } H_0) = 1 - 0.000782 = 0.9992$$

در نتیجه با احتمال 99.92 درصد باری که میانگین واقعیش 17.5 باشه از آزمون ما با موفقیت عبور می‌کنه و قبول میشه!!!

این مدل خطا به خطای β معروفه یا خطای نوع دوم.

درسته که میانگین واقعی 17.5 است اما به علت اینکه از توزیع نرمال با واریانس 0.5 از این داده‌ها و حجم نمونه‌مان که در 10 تاست

دچار این خطا می‌شیم.

(c

هدف در این بخش سؤال این است که اندازه‌ی نمونه (n) لازم برای این آزمون رو پیدا کنیم.

فرض می‌کنیم که میانگین واقعی برابر 17.5 است در حالی که فرض صفر میانگین بیش از 18 می‌خواد
برای اینکه با احتمال 99 درصد (توان آزمون $1-\beta = 0.99$) بتونیم این میانگین 17.5 را از 18 تشخیص بدیم به چند تا نمونه نیاز داریم؟

$\mu_0 = 18$ $\qquad \sigma = 0.5$ $\qquad 1-\beta = 0.99$

$\mu_{actual} = 17.5$ $\qquad \alpha = 0.05$ $\qquad H_0: \mu \geq 18 \ , \ H_1: \mu < 18$

z_α : از جدول نرمال استاندارد $= 1.645$

z_β : $\beta = 0.01 \Rightarrow$ از همین جدول $= 2.33$

$$\Rightarrow n = \left(\frac{(z_\alpha + z_\beta)\cdot\sigma}{\mu_0 - \mu_{actual}}\right)^2 \Rightarrow n = \left(\frac{(3.975)\times 0.5}{18-17.5}\right)^2 = (3.975)^2 = 15.77$$

گرد می‌کنیم به سمت بالا و به عدد 16 می‌رسیم برای تعداد نمونه :)

علت استفاده از رابطه برای n، آزمون فرضیه برای تفاوت میانگین‌ها است که به ما می‌گه چند تا نمونه لازم داریم تا با توان مشخص 99 درصد را
تفاوت میانگین‌ها رو از هم به‌درستی تشخیص بدیم.

عدد 16 هم معقول به نظر می‌رسه چون که با 16 تا نمونه در انحراف معیار 0.5 تأیید پوشش داده می‌شود، مطمئناً باید یک نمونه بزرگتر
می‌داشتیم که $\sqrt{n}$ در دست رابطه‌ی Z ظاهر می‌شه، عدد بزرگتری به ما بده و توانایی ما رو بهتر کرده کمک کنه :)

پاسخ سؤال هشت

a)

باید از توزیع χ^2 استفاده کرد زیرا رابطه‌ی مستقیمی بین واریانس نمونه s^2 و واریانس جامعه σ^2 را به ما می‌دهد. از طرفی برای داده‌هایی که از توزیع نرمال می‌آیند که طول‌های ماهی‌ها همچنین خصوصیات χ^2 از جمله واریانس که چولگی به سمت راست دارد با آن تطابق دارد

برای تخمین بازه‌ی اطمینان $(1-\alpha)$ برای جامعه با واریانس σ^2 از توزیع χ^2 استفاده می‌کنیم و رابطه‌ی آن به شکل زیر است:

$$\left(\frac{s^2(n-1)}{\chi^2_{\alpha/2}} , \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}} \right)$$

s.t.

$n = 25$

$s^2 = (4.25)^2$

$\chi^2_{\alpha/2}$ = مقدار برای سطح معناداری $\frac{\alpha}{2}$ $\longrightarrow$ $\chi^2_{\alpha/2} = \chi^2_{0.025,24} \simeq 39.364$

$\chi^2_{1-\alpha/2}$ = مقدار برای سطح معناداری $1-\frac{\alpha}{2}$ $\longrightarrow$ $\chi^2_{1-\alpha/2} = \chi^2_{0.975,24} \simeq 12.401$

$$\frac{s^2(n-1)}{\chi^2_{\alpha/2}} = \frac{18.0625 \times 24}{39.364} \simeq 11.028 \quad , \quad \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}} = \frac{18.0625 \times 24}{12.401} \simeq 34.979$$

$$\Rightarrow (11.028, 34.979)$$

علت متقارن نبودن: توزیع خی 2 برای درجه آزادی کم چولگی به راست دارد

این عدم تقارن نشان دهنده‌ی نا اطمینانی بیشتر به مقادیر بزرگتر واریانس است.

b)

$$s = \sqrt{\sigma^2} \Rightarrow (11.028, 34.979) \rightsquigarrow (3.32, 5.91)$$

انحراف معیار طول واقعی ماهی‌ها در جامعه بین 3.32 و 5.91 است که یک بازه‌ی نامتقارن می‌باشد $\longleftarrow$ علت عدم تقارن این است که انحراف معیار ریشه‌ی دوم واریانس است و خود واریانس نامتقارن می‌باشد.

در واقع با اطمینان 95 درصد انحراف معیار واقعی ماهی‌ها در بازه‌ی (3.32, 5.91) قرار دارد. انحراف معیار نشان می‌دهد که طول ماهی‌ها چه مقدار از میانگین پراکنده است. و این بازه یک تخمین برای میزان پراکندگی طول ماهی‌ها در کل جامعه است.

c)

فرض صفر H_0: واریانس جامعه برابر با مقدار 16.27 است $\qquad H_0: \sigma^2 = 16.27$

فرض یک H_1: 〃 〃 〃 〃 نیست $\qquad H_1: \sigma^2 \neq 16.27$

این آزمون دو طرفه است

آماره‌ی آزمون ما $\chi^2_0 = \frac{s^2(n-1)}{\sigma_0^2}$ است به طوری که

$n = 25$

$s^2 = 18.0625$

$\sigma_0^2 = 16.27$

$$\Rightarrow \chi^2 = \frac{18.0625 \times 24}{16.27} \simeq 26.078$$

سطح معناداری $\alpha = 0.05 \Rightarrow$

$\chi^2_{\alpha/2} = \chi^2_{0.025,24} = 39.364$

$\chi^2_{1-\alpha/2} = \chi^2_{0.975,24} = 12.401$

$$\checkmark \quad \chi^2_{1-\alpha/2} \le \chi^2 < \chi^2_{\alpha/2}$$

$\Rightarrow$ نمی‌توانیم فرض صفر را رد کنیم چون در بازه هست :)

هدف بازه‌ی اطمینان: بازه‌ی اطمینان یک تخمین از محدوده‌ی مقادیر ممکن برای یک پارامتر جامعه (مثل میانگین یا انحراف معیار) ارائه می‌دهد.
هدف آزمون فرضیه: آزمون فرضیه بررسی می‌کند که یک مقدار برای یک پارامتر جامعه پذیرفته شده است یا نه

ارتباط این دو: بازه‌ی اطمینان محدوده‌ای را مشخص می‌کند (با یک سطح اطمینان) که قابل قبول است. آزمون فرضیه بررسی می‌کند که آیا مقدار یک پارامتر خاص در این بازه قرار می‌گیرد یا خیر. به نوعی مکمل همدیگه هستن

در مثال ما مقدار فرضی 16.27 اینچ در بازه‌ی اطمینان قرار داشت و آزمون فرضیه هم نشان داد که این مقدار قابل قبول است.