به نام خداوند جان و خرد

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

استنباط آماری

تمرین شماره ۲

نام و نام خانوادگی: علی خرم فر

شماره دانشجویی: ۸۱۰۱۰۲۱۲۹

آذرماه ۱۴۰۲

فهرست مطالب

۱_ پاسخ مسئله شماره ۰

۱-۱_ پاسخ قسمت a

در این قسمت از سوال باید احتمال عدد ۶ از ۱۵ تا ۲۰ بار در یک تاس با کمک قضیه حد مرکزی محاسبه شود. در این سوال ما از توزیع نرمال کمک گرفته می‌شود زیرا که در پرتاب ۱۰۰ بار سکه به توزیع نرمال نزدیک می‌شود.

متغیر تصادفی X به عنوان تعداد دفعاتی که عدد ۶ ظاهر می‌شود در نظر گرفته می‌شود. پس این سوال به دنبال مقدار $P(15 < X < 20)$ می‌باشد. اگر که در فرض سوال به استفاده از قضیه حد مرکزی اشاره نشده باشد این مقدار به کمک توزیع دوجمله‌ای قابل محاسبه خواهدبود که در آن احتمال آمدن ۶ و یا نیامدن آن مورد بررسی قرار می‌گیرد.

قضیه حد مرکزی

قضیه حد مرکزی [1] یکی از مهم‌ترین قضیه‌های نظریه آمار است که بیان می‌کند اگر از یک جمعیت آماری با میانگین μ و انحراف معیار σ به تعداد زیاد نمونه بگیریم، توزیع حاصل از میانگین نمونه ها تقریبی از یک توزیع نرمال خواهد بود. در واقع، میانگین توزیع شکل گرفته برابر با میانگین جامعه اصلی است که از آن نمونه برداری کردیم. انحراف معیار توزیع حاصل نیز برابر با نسبت انحراف معیار توزیع اصلی به جذر اندازه نمونه است.

پس در این سوال ابتدا باید مقدار میانیگن و واریانس محاسبه شده تا پس از تبدیل به توزیع نرمال استاندارد، مقدار مورد نظر محاسبه شود. باتوجه به توزیع پرتاب تاس از توزیع دوجمله‌ای پیروی می‌کند مقادیر مورد نظر باتوجه به این توزیع محاسبه می‌شوند.

در این آزمایش احتمال p که همان آمدن عدد ۶ است برابر $\frac{1}{6}$ و تعداد دفعات پرتاب تاس یا n برابر ۱۰۰ می‌باشد.

میانگین و واریانس در توزیع دوجمله‌ای:

$$\mu = E[X] = np = 100 \times \frac{1}{6} = 16.66$$

$$\sigma^2 = Var[X] = np(1-p) = 100 \times \frac{1}{6}\left(1 - \frac{1}{6}\right) = \frac{100}{6} \times \frac{5}{6} = 13.88$$

---

[1] Central Limit Theorem

توزیع نرمال استاندارد:

$$P(15 \le X \le 20) = P\left(\frac{15-\mu}{\sigma} \le Z \le \frac{20-\mu}{\sigma}\right) = P\left(\frac{15-16.66}{\sqrt{13.88}} \le Z \le \frac{20-16.66}{\sqrt{13.88}}\right)$$

$$P(-0.445 \le Z \le 0.896)$$

مقدار بالا با کد پایتون زیر و یا با کمک جدول مربوط به توزیع نرمال استاندارد قابل محاسبه است:

```
probability = norm.cdf(0.896) - norm.cdf(-0.445)
```

با اجرای کد بالا مقدار ۰.۴۸۶۷ در خروجی چاپ شد.

۲-۱_ پاسخ قسمت b

برای محاسبه احتمال اینکه جمع اعداد مشاهده‌شده در ۱۰۰ پرتاب کمتر از ۳۰۰ باشد، متغیر تصادفی Y به عنوان جمع اعداد در نظر گرفته می‌شود. پس در این قسمت از سوال باید با کمک قضیه حد مرکزی، مقدار $P(Y < 300)$ محاسبه شود. این ۱۰۰ پرتاب مستقل از هم بوده و کافی است که ابتدا مقدار میانگین و واریانس برای یک بار پرتاب محاسبه شده و مقدر را برای ۱۰۰ بار تکرار آن تبدیل کنیم.

اگر X برابر با مقدار ظاهرشده در یکبار پرتاب تاس باشد، آنگاه واریانس و میانگین آن به صورت زیر خواهدبود:

$$\mu = E[X] = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\sigma^2 = Var[X] = E[X^2] - E[X]^2 = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} - 3.5^2$$

$$= 15.166 - 12.25 = 2.916$$

حال پارامترهای لازم برای تبدیل به نرمال استاندارد در محاسبه ۱۰۰ پرتاب از آن‌جا که ۱۰۰ پرتاب با هم جمع می‌شوند:

$$\mu = E[Y] = 100 \times E[X] = 350$$

$$\sigma^2 = Var[Y] = 100 \times Var[X] = 291.6$$

توزیع نرمال استاندارد:

$$P(Y \le 300) = P\left(Z \le \frac{300-\mu}{\sigma}\right) = P\left(Z \le \frac{300-350}{\sqrt{291.6}}\right) = P(Z \le -2.928) = 0.001705$$

۲_پاسخ مسئله شماره ۱

در این مسئله مصاحبه‌کننده فقط جواب بله یا خیر را ثبت می‌کند و انتظار دارد با توجه به اینکه مصاحبه‌شونده امید دارد که مصاحبه‌کننده نمی‌داند کدام سوال پاسخ داده می‌شود،احتمال بیشتری وجود دارد تا پاسخ صادقانه‌ای به سوال بدهد. R به عنوان نسبتی از نمونه‌ای از جمعیت هستند که پاسخ بله می‌دهند. p احتمال پاسخگویی به سوال ۱ است. q نسبتی از جمعیت باشند که خصوصیت A را دارند و r احتمالی است که مصاحبه‌شونده به آن پاسخ مثبت دهد.

۱-۲_ پاسخ قسمت a

در این قسمت باید اثبات شود که:

$$r = (2p - 1)q + (1 - p)$$

باتوجه به احتمال شرطی زیر، $P(yes)$ احتمالی است که مصاحبه‌شونده پاسخ بله دهد. که این احتمال به احتمال انتخاب‌شدن سوال و همچنین احتمال پاسخ بله آن بستگی دارد:

$$P(yes) = P(yes|question1).P(question1) + P(yes|question2).P(question2)$$

در این سوال باتوجه به اینکه مصاحبه‌گر از روش *Randomize Response* استفاده کرده‌است، مقادیر احتمال پرسش سوال ۱ برابر p و احتمال پرسش سوال ۲ مکمل آن یا *1-p* خواهدبود:

$$P(question1) = p$$

$$P(question2) = 1 - p$$

و همچنین در صورت سوال مطرح شده که احتمال پاسخ مثبت $P(yes)$ برابر با r خواهدبود. پس کافی است فرضیات جدید را در فرض مطرح شده در سوال جایگزین کنیم:

$$P(yes) = r = P(yes|question1).p + P(yes|question2).(1 - p)$$

در صورت سوال مطرح شده که q برابر نسبتی از نمونه است که خصوصیت A را دارند. یعنی اگر که سوال ۱ از آن‌ها پرسیده شود امید داریم که پاسخ مثبت دهند.پس می‌توان $P(yes|question1)$ را با این

مقدار جایگزین کرد. همچنین مکمل آن یعنی افرادی که خصوصیت A را ندارند برابر 1-q خواهدبود که این مقدار نیز با $P(yes|question2)$ جایگزین می‌شود. پس عبارت نهایی برابر است با:

$$r = P(yes|question1).p + P(yes|question2).(1-p) = q.p + (1-q)(1-p)$$

$$q.p + (1-q)(1-p) \rightarrow q.p + (1-p) - q + q.p \rightarrow q(2p-1) + (1-p)$$

با عمل فاکتورگیری بالا، حکم مسئله اثبات شد:

$$r = (2p-1)q + (1-p)$$

۲-۲_ پاسخ قسمت b

اگر که مقدار r مشخص باشد پس در معادله بالا معلوم است و q مجهول خواهدبود. کافیست که q به یک طرف معادله منتقل شده و معادله جدید پاسخ این سوال خواهدبود:

$$r = (2p-1)q + (1-p) \rightarrow -(2p-1)q = (1-p) - r \rightarrow q = -\frac{(1-p)-r}{(2p-1)}$$

$$q = \frac{r-(1-p)}{(2p-1)}$$

۳-۲_ پاسخ قسمت c

در این قسمت از سوال باید نشان داده‌شود که $E(R) = r$ و باتوجه به مقدار تخمینی $\hat{Q}$ برای q باید اثبات شود این تخمین *unbiased* است.

عبارت $E(R)$ باتوجه به اینکه R برابر نسبتی از یک نمونه‌ی جمعیت است که پاسخشان بله است، برابر با میانگین پاسخ‌های بله در نمونه‌های مختلف است. باید اثبات شود که امیدریاضی آن برابر r است که احتمال این است که یک مصاحبه‌شونده پاسخ بله بدهد. بدیهی است که وقتی پرسش از افراد مختلف انجام می‌شود پاسخشان بله یا خیر است که انجام این آزمایش برای تمامی افراد یک نمونه از توزیع دوجمله‌ای پیروی می‌کند که پارامترهای آن برابر r احتمال p است که یعنی پاسخ‌دهنده جواب بله داده است.

X به عنوان تعداد پاسخ‌های بله در یک نمونه درنظر گرفته می‌شود که توزیع آن دوجمله‌ای بوده و در نتیجه امید ریاضی آن nr است. باتوجه به اندازه نمونه n متوسط نسبت پاسخ‌های مثبت نسبت به اندازه n که همان R است برابر است با:

$$E(R) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{nr}{n} = r$$

هرچند باتوجه به همگرایی در احتمال توزیع R حول یک عدد متمرکز خواهدشد و باتوجه به اینکه R نسبت پاسخ‌دهندگان است که پاسخشان بله است پس مقدار متوسط آن به r نزدیک می‌شود.

برای اینکه تخمین *unbiased* باشد از آن امید ریاضی گرفته‌شده و با مقدار واقعی آن مقایسه می‌شود. در این سوال $\hat{Q}$ برابر مقدار تخمینی q است. مقدار r نیز مشخص نیست و تخمینی از آن در نظر گرفته می‌شود که در صورت سوال به R اشاره شده‌است.

$$\hat{Q} = \frac{\hat{r} - (1-p)}{(2p-1)} = \frac{R - (1-p)}{(2p-1)}$$

پس از مقدار تخمینی امید ریاضی گرفته می‌شود:

$$E(\hat{Q}) = E\left(\frac{R-(1-p)}{(2p-1)}\right) = \frac{E(R) - E((1-p))}{E((2p-1))}$$

در قسمت قبل اثبات شد که مقدار متوسط R برابر r خواهدبود. و مقدار متوسط عدد نیست برابر خود عدد است.

$$E(\hat{Q}) = \frac{E(R) - E((1-p))}{E((2p-1))} = \frac{r-(1-p)}{(2p-1)}$$

باتوجه به رابطه $q = \frac{r-(1-p)}{(2p-1)}$ در قسمت b مقدار بالا برابر q است. پس باتوجه به اینکه $E(\hat{Q}) = q$ پس اثبات می‌شود این تخمین *unbiased* است.

۴-۲_ پاسخ قسمت d

در این قسمت از سوال باید عبارت زیر برای واریانس نسبت پاسخ‌های بله برای نمونه بااندازه n اثبات شود:

$$Var(R) = \frac{r(1-r)}{n}$$

در قسمت قبل بررسی شد که X به عنوان احتمال پاسخ بله در نظر گرفته‌شد و نتیجه گرفته‌شد که توزیع آن دوجمله‌است. در توزیع دوجمله‌ای واریانس برابر np(1-p) خواهد بود که در اینجا p برابر است با احتمال جواب بله که همان r است.

$$Var(X) = np(1-p) \to nr(1-r)$$

واریانس R برابر است با واریانس X به نسبت n :

$$Var(R) = Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} = \frac{nr(1-r)}{n^2} = \frac{r(1-r)}{n}$$

۲-۵_ پاسخ قسمت e

در این قسمت از مسئله باید فرمول واریانس برای مقدار تخمینی $\hat{Q}$ محاسبه شود. کافی است از عبارتی که برای $\hat{Q}$ در قسمت c بدست آمده واریانس گرفته شود:

$$\hat{Q} = \frac{R-(1-p)}{(2p-1)}$$

$$Var(\hat{Q}) = Var\left(\frac{R-(1-p)}{(2p-1)}\right) = Var\left(\frac{R}{(2p-1)} - \frac{(1-p)}{(2p-1)}\right) =$$

مقدار احتمال پرسش به عنوان عددی ثابت در نظر گرفته شده پس واریانس آن صفر است:

$$Var(\hat{Q}) = Var\left(\frac{R}{(2p-1)}\right) - Var\left(\frac{(1-p)}{(2p-1)}\right) \rightarrow Var(\hat{Q}) = Var\left(\frac{R}{(2p-1)}\right) - 0$$

$$= \frac{Var(R)}{(2p-1)^2}$$

در این مرحله کافی‌است واریانس R که در قسمت d محاسبه شده جایگذاری شود:

$$Var(\hat{Q}) = \frac{Var(R)}{(2p-1)^2} = \frac{1}{(2p-1)^2} \times \frac{r(1-r)}{n}$$

۳_ پاسخ مسئله شماره ۲

۳-۱_ پاسخ قسمت a

در این قسمت از سوال باید اندازه‌ای برای n که اندازه نمونه است، تعیین شود تا خطای استاندارد یا SE کمتر از ۰.۰۱ شود. در این سوال که از توزیع دوجمله‌ای پیروی می‌کند یا شخص بیمار است یا خیر. پس مقدار σ در آن برابر است با $\sqrt{p(1-p)}$:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

برای بیماری ۱ باتوجه به فرمول بالا، مقدار SE برابر است:

$$SE_1 = \sqrt{\frac{p_1(1-p_1)}{n}} = \sqrt{\frac{0.03(1-0.03)}{n}}$$

که مقدار آن باید کمتر از ۰.۰۱ شود. همین مورد برای بیماری ۲ نیز وجود دارد:

$$SE_2 = \sqrt{\frac{p_2(1-p_2)}{n}} = \sqrt{\frac{0.40(1-0.40)}{n}}$$

پس برای محاسبه n یک نامساوی تشکیل داده می‌شود و از بین n های بدست آمده از هر دو بیماری ناچار به انتخاب n بزرگتر برای خطای کمتر از ۰.۰۱ هستیم.

$$SE_1 = \sqrt{\frac{0.03(1-0.03)}{n}} < 0.01 \quad , SE_2 = \sqrt{\frac{0.40(1-0.40)}{n}} < 0.01$$

محاسبه حداقل n برای بیماری ۱:

$$\sqrt{\frac{0.03(1-0.03)}{n}} < 0.01 \rightarrow \frac{0.03(1-0.03)}{n} < 0.0001 \rightarrow \frac{0.03(1-0.03)}{0.0001} < n$$

$$\frac{0.0291}{0.0001} < n \rightarrow 291 < n$$

پس باید حداقل n با اندازه ۲۹۲ انتخاب شود.

محاسبه حداقل n برای بیماری ۲:

$$\sqrt{\frac{0.40(1-0.40)}{n}} < 0.01 \rightarrow \frac{0.40(1-0.40)}{n} < 0.0001 \rightarrow \frac{0.40(1-0.40)}{0.0001} < n$$

$$\frac{0.24}{0.0001} < n \rightarrow 2400 < n$$

پس باید حداقل n با اندازه ۲۴۰۱ انتخاب شود.

پس اگر مقدار SE کمتر از ۰.۰۱ برای هردو مطلوب باشد، باید مقدار n بزرگتر از ۲۴۰۰ باشد.

۲-۳_ پاسخ قسمت b

اگر که مطلوب SE کمتر از ۱۰٪ مقدار واقعی در هر بیمار باشد، از همان فرمول قسمت قبلی برای محاسبه حداقل n استفاده می‌شود. ابتدا محاسبه SE هر بیماری با شرایط جدید انجام می‌شود:

$SE_1 = 0.1 \times 0.03 = 0.003$

$SE_2 = 0.1 \times 0.40 = 0.04$

محاسبه حداقل n برای بیماری ۱:

$$\sqrt{\frac{0.03(1-0.03)}{n}} < 0.003 \rightarrow \frac{0.03(1-0.03)}{n} < 0.000009 \rightarrow \frac{0.03(1-0.03)}{0.000009} < n$$

$$\frac{0.0291}{0.000009} < n \rightarrow 3233.33 < n$$

پس باید حداقل n با اندازه ۳۲۳۴ انتخاب شود.

محاسبه حداقل n برای بیماری ۲:

$$\sqrt{\frac{0.40(1-0.40)}{n}} < 0.04 \rightarrow \frac{0.40(1-0.40)}{n} < 0.0016 \rightarrow \frac{0.40(1-0.40)}{0.0016} < n$$

$$\frac{0.24}{0.0016} < n \rightarrow 150 < n$$

پس باید حداقل n با اندازه ۱۵۱ انتخاب شود.

پس اگر مقدار SE کمتر از ۱۰٪ مقدار واقعی در هر بیمار مطلوب باشد، باید مقدار n بزرگتر از ۳۲۳۴ باشد.

۴_پاسخ مسئله شماره ۳

a. غلط

قضیه حد مرکزی بیان می‌کند که وقتی اندازه نمونه به اندازه کافی بزرگ شود، توزیع میانگین نمونه‌ای یک توزیع نرمال با میانگین μ و واریانس $\frac{\sigma^2}{n}$ خواهدبود. بازه‌ی اطمینان با کمک توزیع نرمال بدست آمده و با z *scoring* به نرمال استاندارد برای احتمال خطای تخمین μ در میانگین نمونه‌ای به کار می‌رود. پس شاید برعکس این عبارت صحیح تر به نظر برسید.

b. صحیح

در آزمون فرض دوطرفه ، هر دو سمت توزیع در نظر گرفته می‌شود، پس احتمال این مورد وجود دارد که فرض صفر در آزمون دو طرفه رد شود ولی در یک طرفه رد نشود.

c. غلط

فرض قضیه حد مرکزی تعداد کافی نمونه با اندازه‌های بزرگ است. اگر اندازه نمونه‌ها کوچک باشد ممکن است که توزیع نرمال نشود و این تضمین وجود نخواهدداشت. اما با افزایش اندازه نمونه، توزیع به نرمال نزدیک می‌شود.

d. صحیح

در یک توزیع با شیب مثبت، دم در راست قرار دارد و مقدار میانگین از مد و میانه بیشتر خواهدبود زیرا که مقادیر بزرگتر تاثیر بیشتری در میانگین دارند ولی این مورد برای مد و میانه برقرار نخواهدبود.

e. غلط

بازه اطمینان یک بازه برای پوشش پارامتری که تخمین زده می‌شود یک احتمال ارائه می‌دهد. برای یک *SE* هرچه که مقدار *Confidence Interval* بیشتر شود، بدیهی است که بازه‌ی پوشش داده‌شده بیشتر خواهدبود.

f. غلط

موضوعی که از بازه اطمینان اشتباه برداشت می‌شود این است که احتمال حضور آن پارامتر در بازه مشخص‌شده را بیان می‌کند. در صورتی که بازه اطمینان یک بازه برای پوشش پارامتری که تخمین زده می‌شود یک احتمال ارائه می‌دهد، یعنی اگر بازه‌هایی انتخاب شوند، احتمال اینکه آن پارامتر را پوشش دهند ۹۵ درصد است.

g. غلط

تمامی بازه‌های اطمینان ۹۵٪ به معنی این هستند که احتمال ۹۵ درصدی وجود دارد که پارامتر مورد نظر را پوشش دهند و اندازه نمونه تاثیری بر این احتمال وجود ندارد بلکه فقط بازه را کوچکتر می‌کند.

h. غلط

در بازه اطمینان، پارامتر میانگین نمونه مورد بحث نبوده و باید گفته شود که به احتمال ۹۵ درصد میانگین جامعه در بازه تعیین شده پوشش داده می‌شود.

i. صحیح

این تعریف را در پاسخ‌های قبلی به وضوح بیان شده و تعریف صحیحی است.

۵_پاسخ مسئله شماره ۴

برای بدست‌آوردن بازه اطمینان از رابطه زیر استفاده می‌شود:

$$CI = \bar{X} \pm Z\frac{S}{\sqrt{n}}$$

در این سوال باتوجه به بازه اطمینان ۹۵٪ مقدار Z برابر ۱.۹۶ ، مقدار انحراف معیار S برابر ۱۰ و همچنین تفاوت بازه اطمینان بین دوگروه کوچکتر و یا برابر با ۲ فرض شده و مطلوب است که مقدار n برای این حالت محاسبه شود. برای محاسبه بازه اطمینان اخلاف میانگین دو گروه از رابطه زیر استفاده می‌شود:

$$CI\ for\ Difference\ in\ Means = \bar{X}_1 - \bar{X}_2 \pm Z\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

که $\bar{X}_1$ و $\bar{X}_2$ میانگین نمونه‌ای برای دو گروه مستقل است. در سوال بالا مقدار n برای هردو گروه برابر است. پس مقادیر مذکور در رابطه بالا جایگذاری می‌شوند:

$$CI = \bar{X}_1 - \bar{X}_2 \pm Z\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = (\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{10^2}{n} + \frac{10^2}{n}}$$

در مسئله بیان شده که این اختلاف باید حداکثر ۲ واحد باشد

$$2 \geq 1.96\sqrt{\frac{10^2}{n}+\frac{10^2}{n}} \rightarrow 4 \geq 3.8416\frac{200}{n} \rightarrow 4 \geq \frac{768.32}{n} \rightarrow n \geq \frac{768.32}{4}$$

$$n \geq 192.08$$

پس مقدار n باید حداقل ۱۹۳ باشد تا بازه اطمینان اختلاف میانگین دو گروه حداکثر ۲ باشد.

در این سوال ابهامی وجود دارد. اگر منظور یعنی طول بازه باید حداکثر ۲ باشد که طول بازه برابر است با ۲ برابر مقداری که از PE کم و زیاد می‌شود. پس:

$$2 \geq 1.96\sqrt{\frac{10^2}{n}+\frac{10^2}{n}}+1.96\sqrt{\frac{10^2}{n}+\frac{10^2}{n}} \rightarrow 2 \geq 3.92\sqrt{\frac{200}{n}} \rightarrow n \geq 768$$

پس اگر حالت دوم صحیح باشد مقدار n باید حداقل ۷۶۸ باشد تا بازه اطمینان اختلاف میانگین دو گروه حداکثر ۲ باشد.

۶_پاسخ مسئله شماره ۵

در این سوال باید اثبات شود که اگر که بازه اطمینان اختلاف میانگین‌ها شامل صفر نباشد، فرض صفر رد می‌شود.

در این مسئله از آزمون t با دو نمونه مستقل برای تعیین اینکه اختلاف معناداری بین میانگین دو جامعه وجود دارد یا خیر استفاده می‌شود که در آن بازه اطمینان برای برای تفاوت در میانگین نقشی اساسی دارد. باتوجه به فرض صفر H0 در مسئله، $\mu_1 = \mu_2$ پس اختلاف میانگین واقعی جمعیت‌ها برابر صفر است.

یک آزمون t دونمونه‌ای طبق شرایط سوال برای دو نمونه مستقل با میانگین $\bar{X}_2$ و $\bar{X}_1$، خطای استاندارد S_1 و S_2 و اندازه نمونه n_1 و n_2:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

یک بازه اطمینان که صفر را شامل نشود، بیانگر این است که اختلاف قابل توجهی میان میانگین‌های جامعه‌ها وجود دارد. $t_{\frac{\alpha}{2}}$ بیانگر مقدار ویژه برای توزیع t است.

مقدار $Margin\ Of\ Error(ME)$ به مقدار $Point\ Estimation$ اضافه و کم می‌شود.

$$\bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

اگر که کران بالا منفی و یا کران پایین مثبت باشد، بازه‌ی اطمینان شامل صفر نخواهدشد.

بازه اطمینان کاملا بالاتر از صفر باشد، پس کران پایین بازه اطمینان بزرگتر از صفر است:

$$\bar{X}_1 - \bar{X}_2 > \ t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

بازه اطمینان کاملا پایین‌تر از صفر باشد، پس کران پایین بازه اطمینان کمتر از صفر است:

$$\bar{X}_1 - \bar{X}_2 < \ -t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

اگر فرض صفر رد شود، اما بازه اطمینان شامل صفر باشد، تناقض وجود خواهدداشت.ME و t از یک رابطه هستند. اگر توجهی به علامت ME نداشته باشیم و قدرمطلق آن‌ها را مقایسه کنیم در صورتی که مقدار t از ME بزرگتر باشد، مقدار PE باید از ME بزرگتر باشد که نشان می‌دهد بازه اطمینان شامل صفر نخواهدبود و مطلوب مسئله اثبات می‌شود.

همچنین

$$P((\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \leq \ \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}})$$

باتوجه به $\mu_1 > \mu_2$ و $\mu_1 < \mu_2$ نتیجه می‌شود که صفر شامل این بازه نیست و فرض صفر رد می‌شود.

۷_پاسخ مسئله شماره ۶

در این سوال تخمینی برای میانگین جامعه μ ارائه شده که به صورت زیر است:

$$\bar{X}_C = \sum_{i=1}^{n} c_i X_i$$

برای اینکه نشان‌داده شود که تخمین $unbiased$ است باید امید ریاضی تخمین محاسبه شده و برابر با میانگین جامعه μ باشد.

$$E(\bar{X}_C) = E\left(\sum_{i=1}^{n} c_i X_i\right) = \mu \rightarrow \sum_{i=1}^{n} c_i E(X_i) \rightarrow \sum_{i=1}^{n} c_i \mu = \mu \rightarrow \sum_{i=1}^{n} c_i \mu = \mu \rightarrow \mu \sum_{i=1}^{n} c_i = \mu$$

$$\sum_{i=1}^{n} c_i = 1$$

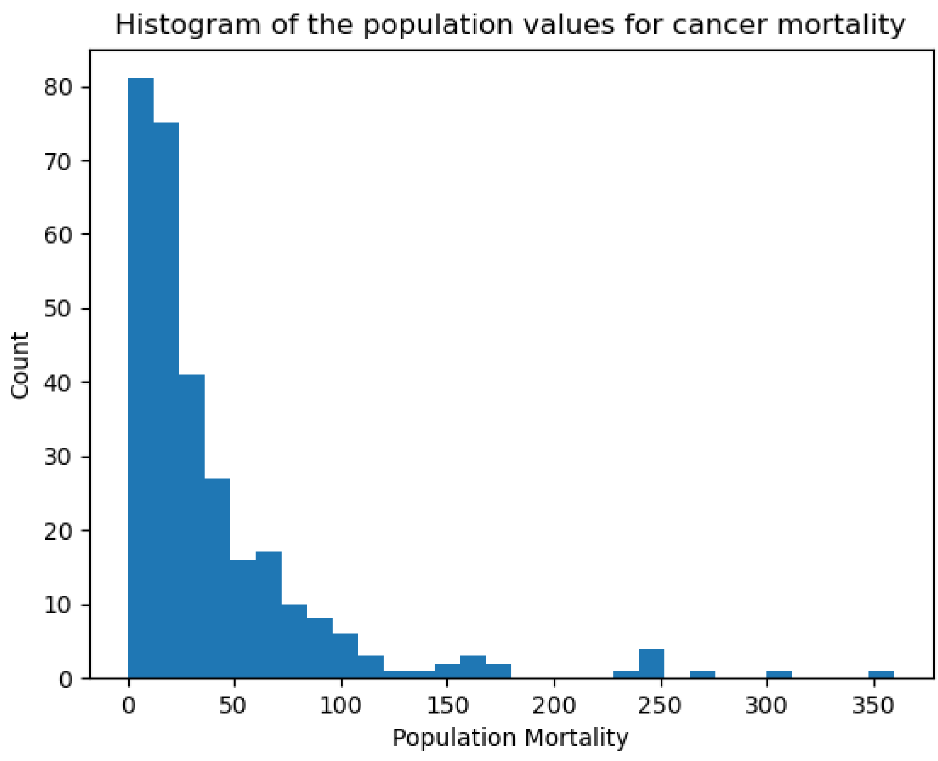
در این صورت مقدار تخمین برابر با μ خواهدبود در نتیجه تخمین *unbiased* است.

۸_پاسخ مسئله شماره ۷

تمامی کدهای این مسئله به همراه خروجی هر قسمت در فایل تمرین پیوست شد.

۸-۱_ پاسخ قسمت a

برای نرخ مرگ بر اثر سرطان نمودار هیستوگرام به شکل زیر رسم شد.



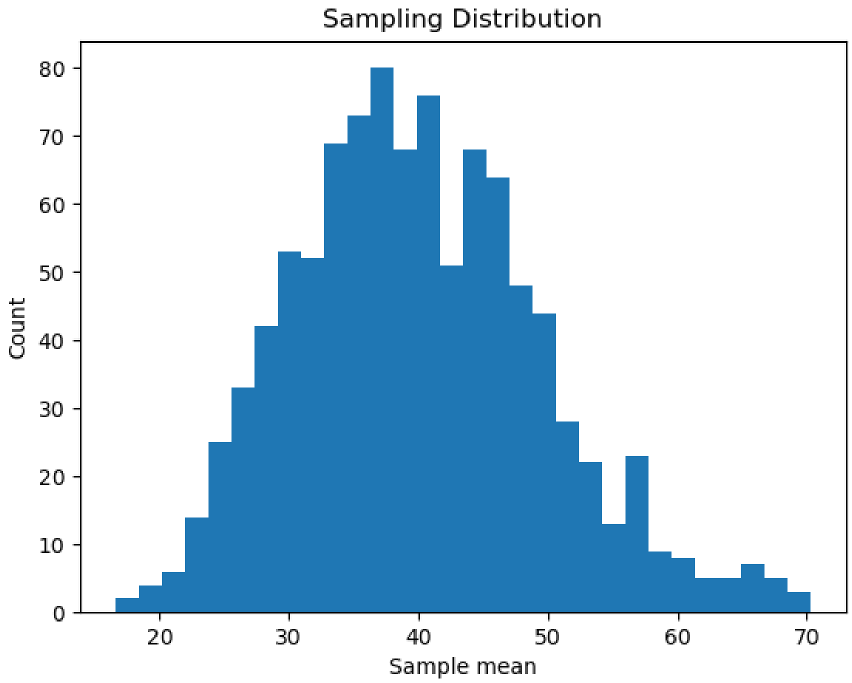
شکل ۱ هیستوگرام تعداد مرگ بر اثر سرطان

۸-۲_ پاسخ قسمت b

برای این قسمت از مسئله با کمک تابع کتابخانه pandas مقادیر مورد نظر از داده استخراج شد و خروجی در فایل تمرین پیوست شد.

۸-۳_ پاسخ قسمت c

برای حل این مسئله ۱۰۰۰ نمونه با اندازه ۲۵ برداشته شد و برای همه آن‌ها میانگین نمونه‌ای در یک لیست ذخیره شد. سپس هیستوگرام مورد نظر به شکل زیر دریافت شد.



شکل ۲ توزیع نمونه‌ای برای ۱۰۰۰ نمونه با اندازه ۲۵

۸-۴_ پاسخ قسمت d

در کتابخانه pandas به صورت پیشفرض نمونه‌برداری با جایگذاری است پس نیاز به کد خاصی برای این مورد وجود ندارد. پس به این منظور نمونه تصادفی با اندازه ۲۵ برداشته شد و پس از محاسبه میانگین باتوجه به فرمول موجود در کتاب مرجع درس، در تعداد ضرب شد تا مقدار کل محاسبه شود. خروجی و کد مربوطه نیز پیوست شد.

۸-۵_ پاسخ قسمت e

در این قسمت از سوال واریانس و انحراف معیار به صورت *unbiased* با کمک نمونه‌برداری تخمین زده شد که خروجی و کد مربوطه پیوست شد.

۸-۶_ پاسخ قسمت f

در این قسمت از سوال، ابتدا لیست دوتایی برای بازه اطمینان در نظر گرفته شد و پس از آن با توجه به فرمول بازه اطمینان مقدار ابتدا و انتهای بازه محاسبه شد. سپس با میانگین واقعی بررسی شد که پارامتر مورد نظر پوشش داده شده است. همین عمل برای تعداد کل نیز انجام شد. کد و خروجی مربوطه پیوست شد.

۸-۷_ پاسخ قسمت g

در این قسمت تمامی مراحل قبل مجددا برای نمونه برداری با اندازه ۱۰۰ انجام شد. مشاهده شد که میانگین مرگ و تعداد کل نسبت به نمونه ۲۵تایی تخمین کمتری دارد و تفاوتی قابل توجه بین آن‌ها وجود دارد. همچنین طول بازه اطمینان کمتر شد.

۸-۸_ پاسخ قسمت h

در این قسمت از سوال درخواست شده که از ratio estimator برای تخمین میزان جمعیت ناشی از مرگ سرطان استفاده شود.

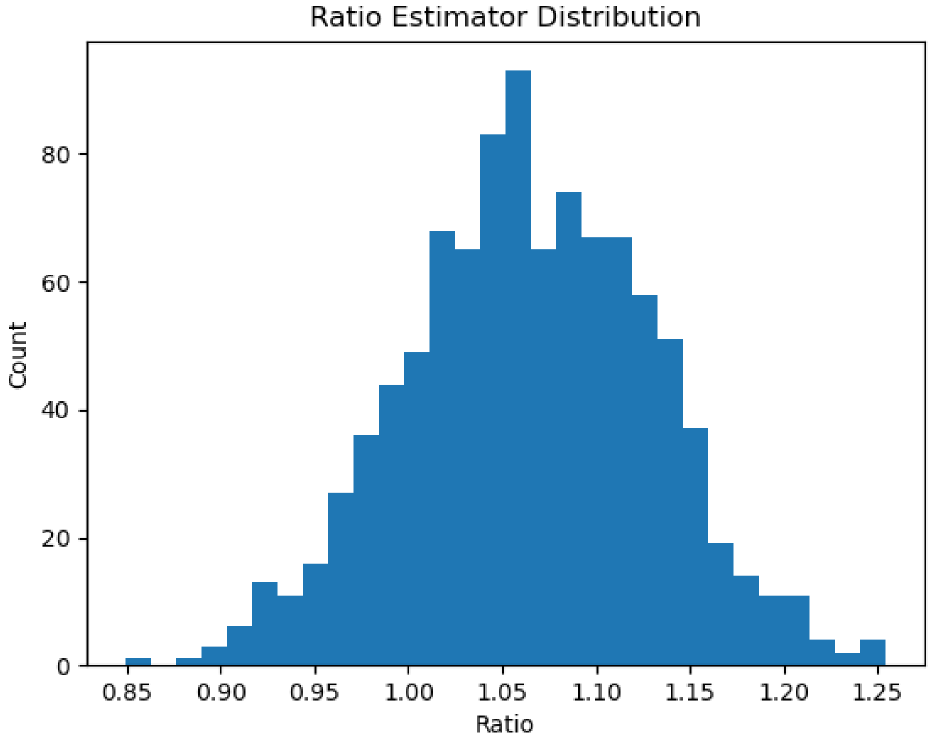
Ratio Estimator یک روش در آمار است که برای تخمین یک ویژگی یا متغیر از جمعیت به کمک نسبت یک ویژگی دیگر استفاده می‌کند. این روش معمولاً در مواردی مورد استفاده قرار می‌گیرد که دو ویژگی یا متغیر به یکدیگر مرتبط هستند یا به عبارت دیگر، وقتی که نسبت دو ویژگی در جمعیت ثابت است. اگر در این سوال نسبت مرگ به جمعیت در نمونه را در تعداد جمعیت نمونه ضرب کنیم، تخمینی از کل مرگ‌های ناشی از سرطان بدست می‌آید.

اگر که جمعیت هر شهر داده شود ولی نسبت سرطان به جمعیت مشخص و ثابت نباشد، احتمالا این تخمین دقیق نخواهدبود. چرا آمار نشان داده که شهرستان‌های بزرگ مثل تهران نسبت به شهرستان‌های کوچک نسبت سرطان به جمعیت بیشتر است. پس باید این نسبت ثابت باشد تا تخمین دقیقی ارائه شود.

۸-۹_ پاسخ قسمت i

در این قسمت از سوال یک نمونه‌برداری از داده‌ها انجام شد و سپس با کمک میانگین مرگ و جمعیت نسبت مرگ به جمعیت برای Ratio Estimator محاسبه شد. سپس با ضرب تعداد در این Ratio و ذخیره

حاصل در لیست، توزیع مربوطه به صورت زیر رسم شد. مشاهده می‌شود که احتمال خطا در بعضی مقادیر مثلا بالای ۸۰ نسبت به روش قبلی بررسی شده وجود دارد.



شکل ۳توزیع Ratio Estimator

۸-۱۰_ پاسخ قسمت j

در این قسمت نیز مانند قسمت‌های قبلی، با کمک Ratio Estimator مقادیر mean و totoal تخمین زده شد که باتوجه به خروجی با مقادیر روش قبلی تفاوت محسوسی دارند. خروجی و کد مربوطه پیوست شد.

۸-۱۱_ پاسخ قسمت k

در این قسمت از سوال، ابتدا لیست دوتایی برای بازه اطمینان در نظر گرفته شد و پس از آن با توجه به فرمول بازه اطمینان مقدار ابتدا و انتهای بازه محاسبه شد. سپس با میانگین واقعی بررسی شد که پارامتر مورد نظر پوشش داده شده است. همین عمل برای تعداد کل نیز انجام شد. کد و خروجی مربوطه پیوست شد.

۸-۱۲_ Stratified Sampling

نمونه‌برداری تراکم‌بندی‌شده یا Stratified Sampling یک روش نمونه‌برداری است که در آن جامعه مورد مطالعه به بخش‌های کوچکتر، یا همان دسته‌های استراتا، تقسیم می‌شود و سپس از هر دسته به طور

جداگانه نمونه‌برداری می‌شود. هدف اصلی این روش، اطمینان حاصل کردن از نمایندگی صحیح دسته‌های مختلف جامعه در نمونه‌ای که برای تحلیل استفاده می‌شود است.

پس ابتدا شهرها را بر اساس اندازه جمعیت به چهار دسته تقسیم می‌کنیم. سپس از هر دسته به طور تصادفی شش مشاهده را نمونه‌برداری می‌کنیم و برآوردهای میانگین جمعیت و مجموع مرگ را تشکیل می‌دهیم.

۹_پاسخ مسئله شماره ۸

۹-۱_ توضیح کد ارائه شده

مطلوب این مسئله تقریب عدد π به روش شبیه‌سازی مونت کارلو می‌باشد. باتوجه به اینکه مساحت دایره برابر با πr^2 بوده و همچنین مساحت مربعی که ضلع آن دوبرابر r باشد برابر با $2r \times 2r = 4r^2$ می‌باشد، در این حالت دایره‌ای درون مربع در نظرگرفته می‌شود، اگر مساحت دایره بر مساحت مربع تقسیم شود رابطه زیر برقرار است:

$$\frac{Sc}{Ss} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} \rightarrow \pi = \frac{4Sc}{Ss}$$

در این شبیه‌سازی بجای مساحت مربع و دایره نقاط درون آن‌ها جایگزین می‌شود. یک نقطه در صورتی درون دایره است که:

$$x^2 + y^2 = r^2$$

برای شبیه‌سازی بالا یک کد در صورت مسئله ارائه شده‌است که آن را برای n های ۱۰، ۱۰۰۰۰ و ۱۰۰۰۰۰۰۰ اجرا می‌کنیم. نتایج خروجی به صورت زیر است:

برای مقدار n برابر ۱۰ خروجی برابر با ۳.۲ ، n برابر با ۱۰۰۰۰ خروجی برابر ۳.۱۲۵۶ و برای n برابر با ۱۰۰۰۰۰۰۰ مقدار خروجی برابر با ۳.۱۴۱۶۵۵ چاپ شد. که نشان دهنده این است که با افزایش n تخمین دقیق تری انجام شده است.

۹-۲_ محاسبه مساحت بیضی

برای محاسبه مساحب بیضی، یک متد برای تخمین در R نوشت شد که کد مربوطه در فایل تمرین پیوست شد.

یک نمونه خروجی آن برای بیضی با قطرهای ۱۷ و ۲۳ برابر ۱۲۲۶.۳۱۷ برای n برابر ۲۰۰۰۰۰ تخمین زده شد که نزدیک به مقدار واقعی مساحت ۱۲۲۶.۳۶ می‌باشد.

۱۰_ پاسخ مسئله اضافه

۱-۱۰_ شرح مسئله

متغیر تصادفی X از توزیعی نمایی پیروی می‌کند و تابع چگالی احتمال آن برای $x \geq 0$ و پارامتر $\alpha > 0$ به صورت زیر است:

$$f(x) = \frac{1}{\alpha} e^{-\frac{x}{\alpha}}$$

یک متغیر تصادفی Z با مقدار ثابت T به صورت زیر تعریف می‌شود:

$$\mathrm{Z} = \begin{cases} \mathrm{X}\,, & \mathrm{if\,X} \leq \mathrm{T} \\ \mathrm{T}\,, & \mathrm{o.w} \end{cases}$$

مطوب سوال محاسبه میانگین و واریانس این متغیر تصادفی است.

۲-۱۰_ محاسبه میانگین

$$E(Z) = \int_{-\infty}^{+\infty} z.f(z)dz$$

پس ابتدا تابع چگالی احتمال برای Z محاسبه می‌شود.

باتوجه به اینکه تابع چگالی احتمال PDF برابر با مشتق CDF است، پس :

$$f(z) = \frac{d}{dz} F(z)$$

برای z های کوچکتر از T

$$F(z) = P(Z \leq z) = P(X \leq z) = \int_0^Z \frac{1}{\alpha} e^{-\frac{x}{\alpha}} dx$$

برای z های بزرگتر از T

$$F(z) = P(Z \leq z) = 1$$

تابع چگالی احتمال برای z های بزرگتر از T برابر صفر و برای z های کوچکتر از T با مشتق CDF:

$$f(z) = \frac{d}{dz} \int_0^Z \frac{1}{\alpha} e^{-\frac{x}{\alpha}} dx = \frac{1}{\alpha} e^{-\frac{Z}{\alpha}}$$

حال کافیست مقادیر بالا در رابطه میانگین جایگذاری شوند

برای محاسبه انتگرال از وب‌سایت *Symbolab* استفاده شده‌است:

$$E(Z) = \int_{-\infty}^{+\infty} z.f(z)dz = \int_0^T z.\frac{1}{\alpha}e^{-\frac{Z}{\alpha}}\,dz + \int_T^{+\infty} T.\frac{1}{\alpha}e^{-\frac{Z}{\alpha}}\,dz =$$

$$-Te^{-\frac{T}{\alpha}} - \alpha e^{-\frac{T}{\alpha}} + \alpha + Te^{-\frac{T}{\alpha}} = \alpha - \alpha e^{-\frac{T}{\alpha}} = \alpha(1 - e^{-\frac{T}{\alpha}})$$

۳-۱۰_ محاسبه واریانس

برای محاسبه واریانس از رابطه زیر استفاده می‌کنیم:

$$Var(Z) = E(Z^2) - E(Z)^2$$

پس ابتدا $E(Z^2)$ محاسبه می‌شود.

برای محاسبه انتگرال از وب‌سایت *Symbolab* استفاده شده‌است:

$$E(Z^2) = \int_{-\infty}^{+\infty} z^2.f(z)dz = \int_0^T z^2.\frac{1}{\alpha}e^{-\frac{Z}{\alpha}}\,dz + \int_T^{+\infty} T^2.\frac{1}{\alpha}e^{-\frac{Z}{\alpha}}\,dz$$

$$= -T^2e^{-\frac{T}{\alpha}} - 2\alpha e^{-\frac{T}{\alpha}}(\alpha + T) + 2\alpha^2 - T^2e^{-\frac{T}{\alpha}} - T^2$$

حال $E(Z)^2$ محاسبه می‌شود:

$$E(Z)^2 = \left(\alpha\left(1 - e^{-\frac{T}{\alpha}}\right)\right)^2$$

$$Var(Z) = E(Z^2) - E(Z)^2 = -2T^2e^{-\frac{T}{\alpha}} - 2\alpha e^{-\frac{T}{\alpha}}(\alpha + T) + 2\alpha^2 - T^2 - \left(\alpha\left(1 - e^{-\frac{T}{\alpha}}\right)\right)^2$$