استنباط آماری

پاییز ۱۴۰۴

تمرین دوم

مرتضی ملکی‌نژاد شوشتری - ۸۱۰۱۰۴۲۵۶

سوال ۱

A: Pipeline Pressure Monitoring System

a

داریم:

$$p(\bar{x_n} - z_{\frac{\alpha}{2}}\sigma_{\bar{x_n}} < \mu < \bar{x_n} + z_{\frac{\alpha}{2}}\sigma_{\bar{x_n}}) \approx 1 - \alpha$$

$\sigma_{\bar{x_n}} = \dfrac{s(n)}{\sqrt{n}} = 2.5 \Rightarrow p(75358 - 2.5z_{2.5} < \mu < 75358 + 2.5z_{2.5}) \xrightarrow{z_{2.5}=1.96} p(75353.1 < \mu < 75362.9) \Rightarrow$
95% confidence interval: $[75353.1, 75362.9]$

کار بهتر این است که بجای توزیع Z از توزیع T استفاده شود،ولی باتوجه به اینکه n از ۳۰ بزرگ‌تر است و توزیع T تفاوت کمی با Z دارد و در درس نیز از توزیع Z استفاده شده، از توزیع Z استفاده شد.

b

گزینه ۴. چیزی که بازه اطمینان می‌گوید این است که احتمال ای‌نکه این بازه شامل میانگین باشد چقدر است. به بیان ملموس تر، در توزی‌عهایی که می‌توانند همچنین نمونه برداری ای داشته باشند، میانگین ۹۵ درصد آن‌ها در این بازه است.

c

عملا باید حاصل $\dfrac{z_{2.5} * s(n)}{\sqrt{n}}$ از یک کوچکتر شود. پس داریم:

$$\sqrt{n} \geq 25 * 1.96 \Rightarrow \sqrt{n} \geq 49.0 \Rightarrow n \geq 2401$$

پس برای اینکه حدود بازه اطمینان از میانگین ۱ واحد اختلاف داشته باشند با فرض این که در تلاش‌های مجدد نیز واریانس نمونه ۲۵ باشد، باید سایز نمونه حداقل ۲۴۰۱ باشد.

B: Smart Vending Machine Calorie Monitoring

a

طبق قضیه CLT و باتوجه به اینکه n از ۳۰ بزرگتر است، توزیع Sample mean یک توزیع نرمال می‌باشد. برای میانگین توزیع، دو گزینه در اختیار است؛ 100 و 102.25 مقدار 102.25 از یک نمونه‌برداری بدست آمده و نمی‌تواند مقدار میانگین را بدرستی تخمین بزند. دراینجا با فرض درست بودن ادعا توزیع بدست می‌آید و سپس در قسمت بعد با بازه اطمینان میزان درستی این فرض چک می‌شود. برای واریانس ولی چون واریانس جامعه موجود است، می‌توان انحراف معیار Sample mean را براحتی بدست آورد:

$$\sigma = 5 \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{6}$$

پس درنهایت می‌توان گفت توزیع Sample mean یک توزیع از نوع $Normal(100, \frac{5}{6})$ می‌باشد.

b

باید مقدار $\sigma_{\bar{x}} z$ را حساب کرد. مقدار انحراف معیار درقسمت قبل محاسبه شد؛ مقدار Z نیز طبق جدول برابر 1.96 می‌باشد. پس داریم:

$$\sigma_{\bar{x}} z = \frac{5}{6} * 1.96 = 1.6\bar{3} \Rightarrow P(98.3\bar{3} < \mu < 101.6\bar{3}) \approx 0.95$$

همچنان احتمال دارد که این نمونه از پنج درصد مواردی باشد که در بازه نمی‌گنجد، ولی حداقل می‌توان به آن مشکوک شد. برای نظر دادن با احتمال بالاتر نیاز به نمونه‌های بیشتری می‌باشد.

c

صرفا ضریب z بیشتر می‌شود و بازه بزرگ‌تر می‌شود.

$$\frac{5}{6} * 2.33 \approx 1.94 \Rightarrow P(-1.94 < \mu - 100 < 1.94) \approx 98 \Rightarrow P(98.6 < \mu < 101.94) \approx 98$$

d

چون هدف یافتن بازه میانگین جامعه از روی میانگین estimator می‌باشد،برای این که احتمال قرارگیری میانگین در بازه بیشتر باشد، بازه باید بخش بیشتری از مقادیر ممکن را پوشش دهد.

C: Smart Health Kiosk Temperature Screening

a

بازه گفته شده یک بازه متقارن حول میانگین است. فاصله دوسر بازه از میانگین برابر $98.93 - 98.2 = 0.73$ می‌باشد که برابر با انحراف معیار می‌باشد. طبق قاعده $68 - 95 - 99.7$ احتمال وجود داده در این بازه حدود ۶۸ درصد می‌باشد.

b

داریم:

$$Z_Score = \frac{x-\mu}{\sigma} = \frac{96.8-98.2}{0.73} = \frac{1.4}{0.73} \approx 1.92$$

که نزدیک به 2σ می‌باشد. یعنی احتمال کمتر بودن دمای یک فرد از این مقدار حدود 2.5 درصد می‌باشد. بدون دانستن توزیع های شرطی افراد برای سالم بودن نمی‌توان به طور دقیق گفت که فرد سالم است یا خیر، ولی با توجه به احتمال کم این دما،بررسی بیشتر می‌تواند مفید باشد.

c

درواقع باید مقدار α را در نامساوی $P(t < \alpha) \leq 0.03$ چون توزیع نرمال است می‌توان آن را به توزیع نرمال استاندارد تبدیل کرد:

$$P(\frac{t-\mu}{\sigma} < \alpha) \leq 0.03 \Rightarrow P(\frac{t-98.2}{0.73} < -1.88) = 0.3 \Rightarrow \frac{t-98.2}{0.73} \leq -1.88 \Rightarrow 98.2 - t \geq 1.88 * 0.73 \Rightarrow$$
$$t \leq 98.2 - 1.88 * 0.73 \Rightarrow max(t) = 96.8276$$

d

a

خیر درست نیست. درواقع این می‌گوید به احتمال ۹۵ درصد، افراد بین ۵.۲ تا ۳.۸ درصد سود خواهند کرد؛ ولی چون این بازه درواقع امید سود را تخمین می‌زند، اعضای جامعه سرمایه‌گذاران نیستند، بلکه پروژه می‌باشد و فقط یک پروژه می‌باشد، یعنی افراد درصد سود یکسانی خواهند داشت (یا ضرر) و توصیف دقیقتر این بازه این می‌باشد که «با احتمال ۹۵ درصد سوددهی پروژه بین ۵.۲ تا ۳.۸ درصد خواهد بود.»

b

این توصیف از توصیف قبلی کمی دقیق‌تر می‌باشد ولی کلمه «احتمال» درست معنی نشده. احتمال بودن سود در آن بازه به این معنی نیست که به ازای هر ۱۰۰ تکرار «دقیقا» ۹۵ مورد آن در آن بازه باشد، ممکن است با داده هر ۱۰۰ سال نیز افراد همچنین سودی نکنند . مثل یک سکه که با این که احتمال رو شدن هر طرف برابر است، ممکن است ۱۰ بار پشت سر هم رو بیاید. همچنین ۹۵ درصد موارد این بازه را تولید نخواهند کرد، بلکه هرکدام ممکن است بازه مختلفی را تولید کنند.

c

احتمال این که سود بین ۵.۲ تا ۳.۸ درصد باشد ۹۵ درصد می‌باشد. یعنی از هر ۱۰۰ پروژه **به طور میانگین** سود ۹۵ پروژه در این بازه خواهد بود.

E: Smart Waste Management and Recycling Program

a

$$\hat{\sigma}^2 = p_{sample}(1 - p_{sample}) = 0.62 * 0.38 = 0.2356$$

b

$$SE(\hat{p}) = \sqrt{\frac{\hat{\sigma}}{n}} = \sqrt{\frac{0.2356}{500}} \approx 0.0217$$

c

طبق CLT می‌توان انتظار داشت که توزیع p در نمونه‌ها یک توزیع نرمال باشد. در صورت سوال میانگین توزیع (0.62) حساب شده و در قسمت b نیز انحراف معیار توزیع حساب شده. پس داریم:

$$P(\mu_n - \sigma_n z < \mu < \mu_n + \sigma_n z) \Rightarrow P(0.62 - 0.0217 * 1.96 < \mu < \mu + 0.0217 * 1.96)$$
$$\Rightarrow P(0.5775 < \mu < 0.6625) \approx 0.95$$

(چون انحراف معیار نمونه جایگذاری شده نیازی به تقسیم بر $\sqrt{n}$ نیست.)

باتوجه به بازه بدست آمده با احتمال خوبی هدف انجام شده (البته نمی‌توان به طور قطعی یا حتی با احتمال ۹۵ درصد این را گفت.)

سوال ۲

A: Theoretical Foundation

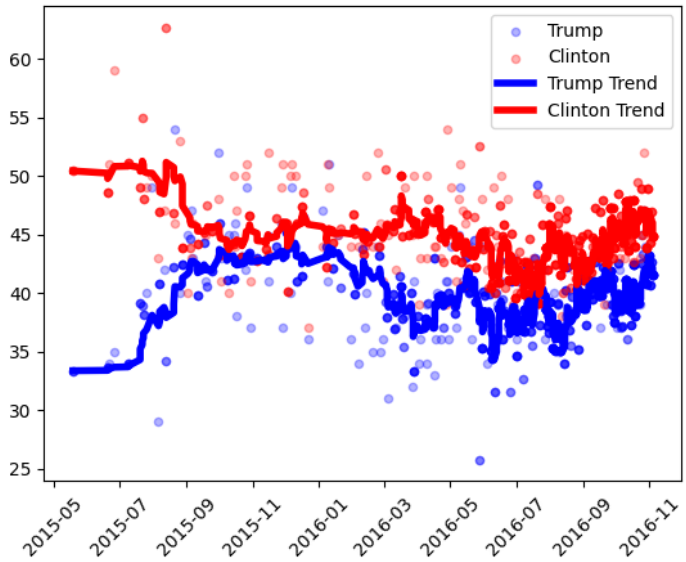
$$\sigma Z = 1.96 * \sqrt{\frac{p(1-p)}{N}} \Rightarrow CI = [p - 1.96\sqrt{\frac{p-p^2}{N}}, p + 1.96\sqrt{\frac{p-p^2}{N}}]$$

B: Monte Carlo Verification

با شرایط گفته شده، در بین آزمایش‌های انجام شده، 95.727 درصد موارد در بازه محاسبه شده بودند.

C: Time Series Visualization

برای محاسبه ترند می‌توان از میانگین نمایی استفاده کرد.



شکل ۱: نمودار پراکندگی و میل درصد طرفداران هر نماینده.

در شکل ۱ می‌توان دید که با گذر زمان نرخ تغییرات بیشتر است و علاوه بر این، تاثیر مکمل بودن درصد ها در نمودار مشخص است.

E: Total Sample Size

درمجموع اطلاعات 1940931 رای دهنده در مجموعه آمده است.

F: Aggregate Proportions

جدول ۱: تعداد و نسبت طرفداران هر کاندید

	Trump	Clinton	Total
Count	787,694.91	885,299.02	1,940,931
Proportion	40.58%	45.61%	100%

G: Confidence Intervals for Both Candidates

با استفاده از فرمولی که در بخش اول حساب شد می‌توان بازه اطمینان را حساب کرد.

Trump : $CI \approx [0.405, 0.407]$

Clinton : $CI \approx [0.455, 0.457]$

سوال ۳

در هرمورد باید بررسی شود با میل کردن n به سمت بینهایت، مقدار $\frac{1}{n}S_n$ به سمت $1-p$ میل می‌کند یا خیر. در حالت کلی می‌توان گفت:

$S_n = \sum x_i(1-p_i)$

که در آن p_i ها از توزیع برنولی با متغیر p هستند. در هرمورد باید چک کرد تعداد متغیرها به سمت بینهایت میل می‌کند یا خیر و همچنین واریانس S_n مقدار متناهی باشد.

۱. در اینجا عملا هدف میانگین گرفتن n متغیر تصادفی از توزیع برنولی است پس:

$$\frac{1}{n}E(S_n) = \frac{1}{n}\sum(1-p) = \frac{n(1-p)}{n} = 1-p$$

واریانس نیز یک کسر با مخرج $\sqrt{n}$ و صورت ثابت ($p(1-p)$) می‌باشد پس با زیاد شدن n به صفر میل می‌کند. پس قانون اعداد بزرگ در اینجا صدق می‌کند.

۲. در اینجا نیز با زیاد شدن n تعداد متغیرهای تصادفی زیاد می‌شوند. همچنین واریانس و امید مشابه قسمت قبل حساب می‌شوند با این تفاوت که تعداد متغیرهای تصادفی برابر $\frac{n}{2}$ است.

$$\frac{1}{n}E(S_n) = \frac{1}{n}\sum_{i=1}^{\frac{n}{2}} 2(1-p) = 1-p$$

واریانس اینجا نیز مشابه قسمت قبل است (می‌توان فرض کرد که این توزیع درواقع حاصل دوبار جمع کردن توزیع قبل است و چون n به سمت بینهایت می‌رود مسئله‌ای پیش نمی‌آید و دوبرابر صفر، صفر است.) و برابر صفر می‌شود. پس قانون اعداد بزرگ دراینجا صدق می‌کند.

۳. در اینجا با زیاد شدن n تعداد متغیرهای تصادفی تغییری نمی‌کند، پس قانون اعداد بزرگ در اینجا صدق نمی‌کند زیرا عملا تعداد متغیرهای تصادفی ثابت و برابر دو می‌باشد. حتی واریانس نیز به سمت صفر میل نمی‌کند.

$$\mu = E(x) = (1-p)^2 n + 2p(1-p)(\frac{n}{2}) + 0 * p^2 = n(1-p)(1-p+p) = n(1-p)$$

$$Var(S_n) = E((x-\mu)^2) = (1-p)^2(n-\mu)^2 + 2p(1-p)(\frac{n}{2}-\mu)^2 + p^2\mu^2 = \frac{n^2}{2}p(1-p)$$

$$\text{mean} : Var(\frac{S_n}{n}) = \frac{Var(S_n)}{n^2} = p(1-p)$$

که یعنی با زیاد شدن n واریانس میانگین به صفر میل نمی‌کند پس قانون اعداد بزرگ برقرار نیست. (به غیر از حالات خاص $p = 0$ و $p = 1$)

۴. اینجا نیز مشابه قسمت قبل چون تعداد متغیرهای تصادفی ثابت و برابر یک است قانون اعداد بزرگ صدق نمی‌کند. واریانس نیز برابر است با:

$$\mu = E(X) = \sum xf(x) = n(1-p)$$

$$Var(S_n) = E((X-\mu)^2) = (1-p)(n-n(1-p))^2 = (1-p)(pn)^2 = n^2 * p^2 * (1-p)$$

$$\text{mean} : Var(\frac{S_n}{n}) = \frac{Var(S_n)}{n^2} = p^2(1-p)$$

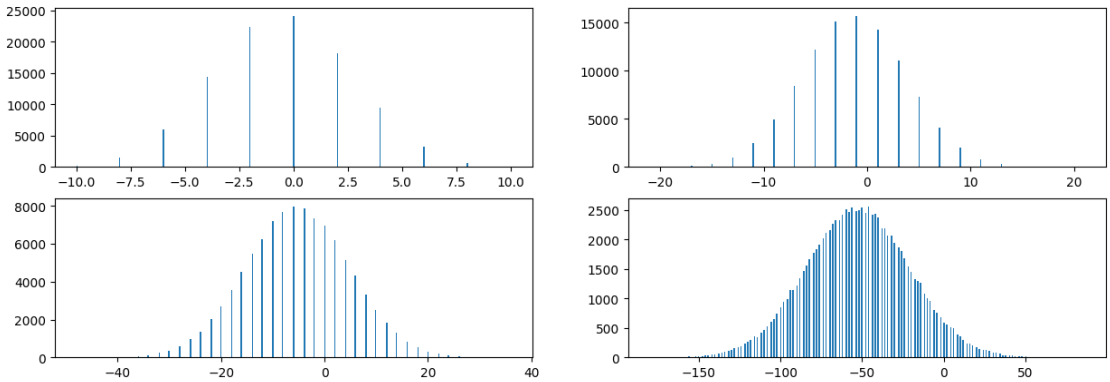
که در اینجا نیز واریانس به سمت صفر میل نمی‌کند و قانون صدق نمی‌کند (به غیر از حالات خاص $p = 0$ و $p = 1$)

□

سوال ۴

بخش ۲

توزیع سود شبیه یک توزیع نرمال به نظر می‌آید . مقدار میانگین با زیاد شدن N کاهش می‌یابد که مطابق انتظار است زیرا احتمال باخت از احتمال برد بیشتر است (به اندازه) $\frac{2}{38}$) و با دفعات بیشتر بازی این ضرر جمع می‌شود. مقدار انحراف معیار نیز افزایش می‌یابد.



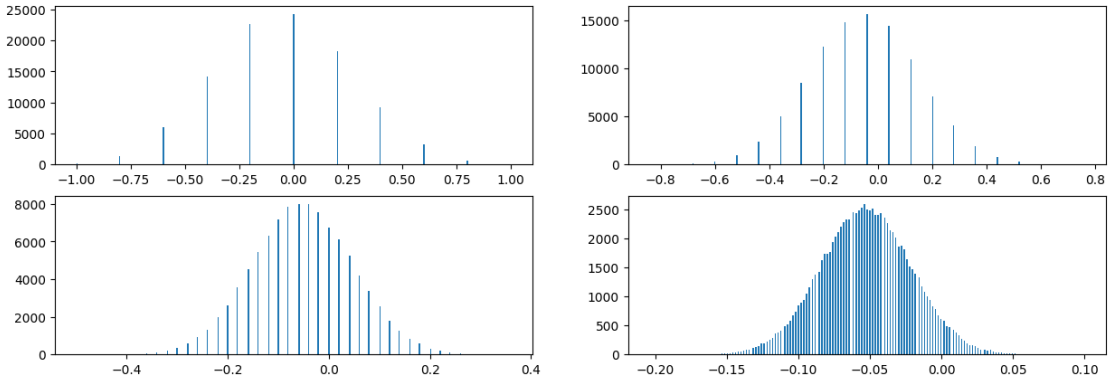
شکل ۲: توزیع میزان پول نهایی

جدول ۲: نتایج شبیه‌سازی برای میزان پول نهایی

N	Average	Standard Deviation
10	-0.52810	3.16515
25	-1.28590	4.98919
100	-5.23976	10.01691
1000	-52.76080	31.46681

بخش ۳

میانگین‌ها تقریبا برابر شد و انحراف معیار به سمت صفر میل می‌کند.



شکل ۳: توزیع میزان پول نهایی که با N نرمال شده است.

جدول ۳: نتایج شبیه‌سازی برای پول نهایی که با n نرمال شده است.

N	Average	Standard Deviation
10	-0.053252000	0.315180305
25	-0.052440800	0.199517434
100	-0.052237200	0.099651046
1000	-0.052664400	0.031606469

بخش ۴

در هر راند احتمال برد برابر $\frac{18}{38}$ و احتمال باخت برابر $\frac{20}{38}$ می‌باشد. چون راندها مستقل از هم هستند، می‌توان امید و واریانس یک راند را بدست آورد و از روی آن امید و واریانس را برای N راند حساب کرد.

$$E(X) = 1 * \frac{18}{38} + (-1) * \frac{20}{38} = \frac{-2}{38} = \frac{-1}{19}$$

$$E((X-\mu)^2) = (1-\mu)^2 * (\frac{18}{38}) + (-1-\mu)^2 * \frac{20}{38} = \frac{1}{19}(9 * (\frac{20}{19})^2 + (10 * (\frac{18}{19})^2) = \frac{1}{19^3}(3600 + 3240) = \frac{6840}{6859} \approx 0.9972$$

پس امید و واریانس برای یک متغیر تصادفی پیدا شد. حال باید امید و واریانس جمع N متغیر تصادفی پیدا شود که باتوجه به استقلال آن‌ها،صرفا کافیست جمع شوند.

$$\text{mean of n rounds} = n\mu = \frac{-n}{19}$$

$$\text{variance of n rounds} = n\sigma = \frac{6840n}{6859} \Rightarrow \text{std of n rounds} = \sqrt{\frac{6840n}{6859}}$$

جدول ۴: مقادیر محاسبه شده به طور نظری

N	Average	Standard Deviation
10	-0.5263	3.1579
25	-1.3158	4.9931
100	-5.2632	9.9861
1000	-52.6316	31.5789

با مقایسه جدول ۴ با ۲ می‌توان دید اندکی خطا ناشی از نمونه‌گیری توزیع وجود دارد ولی مقادیر به هم نزدیک هستند.

بخش ۵

البته معمولا برای n های بزرگ‌تر از ۳۰ از قضیه حد مرکزی استفاده می‌شود، با این حال اگر توزیع پول نهایی را طبق حد مرکزی نرمال بدانیم، در قسمت قبل واریانس و میانگین آن حساب شد. پس با استفاده از جدول ۴ داریم:

$$X \sim Normal(-1.3158, 4.9931) \Rightarrow \frac{X+1.3158}{4.9931} \sim Normal(0,1) \Rightarrow P(X>0) = P(Z > \frac{13158}{49931}) \approx 1 - Z(0.26) \xrightarrow{Z(0.26)=0.60257} P(X>0) \approx 0.3974$$

در شبیه‌سازی انجام شده، احتمال بدست آمده برابر 0.39414 شد که به مقدار محاسبه شده نزدیک است.

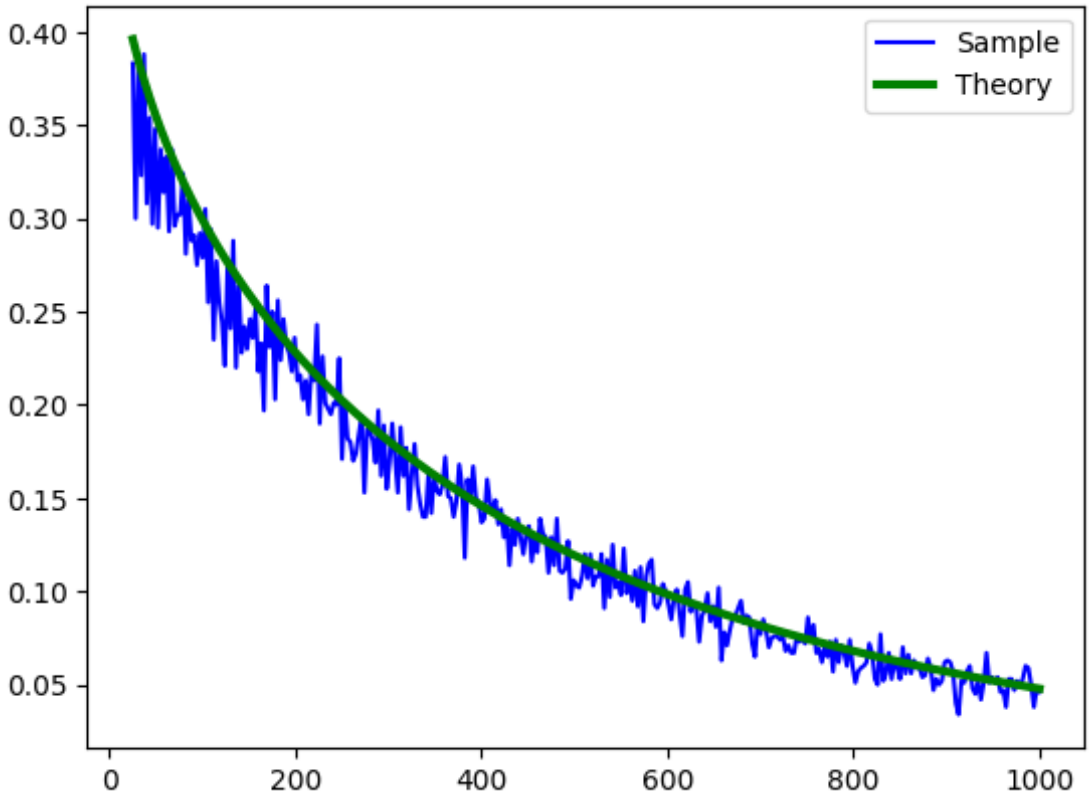
بخش ۶

قبل از شبیه‌سازی نیز می‌توان فرم نمودار را حدس زد. حالت کلی سوال قبل برابر است با:

$$X_n \sim Normal(\mu_n, \sigma_n) \Rightarrow P(X_n > 0) = P(Z > \frac{-\mu_n}{\sigma_n}) = 1 - Z(\frac{-\mu_n}{\sigma_n})$$

از طرفی در بخش ۴ مقادیر میانگین و واریانس به طور نظری بدست آمدند پس داریم:

$$P(X_n > 0) = 1 - Z(\frac{\frac{n}{19}}{\sqrt{\frac{6840n}{6859}}}) = 1 - Z(\frac{\sqrt{\frac{6859n}{6840}}}{19}) \approx 1 - Z(0.0527\sqrt{n})$$



شکل ۴: نمودار احتمال ضرر کازینو طبق شبیه‌سازی. خط سبز نمایانگر مقدار محاسبه شده توسط فرمول بدست‌آمده است.

سوال ۵

زده نشده.

سوال ۶

A: Multi-Stage Polling Design

۱

$$ME = 1.96 * \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{\sigma}{ME} \Rightarrow n = \frac{\sigma^2}{ME^2} \Rightarrow n = \begin{cases} \text{Urban} & 81 \\ \text{Suburban} & 100 \\ \text{Rural} & 77.4 \xrightarrow{n \in \mathbf{N}} 78 \end{cases}$$

۲

- در جدول ۵ تعداد نمونه هر ناحیه آمده. برای هر سطر درصد جمعیت آن در عدد 2000 ضرب شد.
- با استفاده از فرمول بازه خطا داریم:

جدول ۵: تعداد پاسخ‌دهنده‌ها به نسبت جمعیت

Region	Sample Size
Urban	900
Suburban	700
Rural	400

$$ME = \begin{cases} \text{Urban} & 1.96 * \frac{0.18}{30} = 0.0176 \\ \text{Suburban} & 1.96 * \frac{0.15}{\sqrt{700}} \approx 0.0111 \\ \text{Rural} & 1.96 * \frac{0.22}{20} \approx 0.0215 \end{cases}$$

- خیر،مقدار ME برای همه نواحی کمتر از مقدار تعیین شده است. (که مطابق انتظار است زیرا سایز نمونه‌ها از اعداد بیان شده در بخش اول بیشتر است.)

۳

$$ME = 1.96 * \frac{\sigma}{\sqrt{n}} = \frac{constant}{\sqrt{n}} \Rightarrow \frac{ME_1}{ME_2} = \frac{\sqrt{n_2}}{n_1} \xrightarrow{ME_2 = kME_1} k = \frac{\sqrt{n_2}}{\sqrt{n_1}} \Rightarrow n_2 = k^2 n_1$$

- طبق قضیه اثبات شده، برای نصف کردن مقدار ME باید تعداد نمونه‌ها را چهار برابر کرد.

- باتوجه به اینکه تنها مقادیر تغییر پذیر، درصد سایز نمونه برای هر بخش و مقدار ME است، باید بین سایز نمونه بخش‌های مختلف Trade off کرد یا به بیان دیگر روی ME بخش‌های مختلف Trade off کرد.

سوال ۷

A: Understanding Stratification

۱

- باتوجه به اینکه نمونه‌گیری از نوع SRS است، گروه‌ها هیچ مزیتی نسبت به هم ندارند، پس امید هر گروه، متناسب است با تعداد اعضای آن تقسیم بر تعداد همه خانوارها:

$$E(X_i) = \sum_{i=1}^{N_i} p_i = N_i p = N_i * \frac{1}{N} = \frac{N_i}{N}$$

برای پیدا کردن امید تعداد اعضای هر گروه باید امید هر متغیر را جمع کرد با فرض مستقل بودن نمونه‌برداری، در نمونه‌برداری با سایز n هر عضو نمونه یک متغییر تصادفی مستقل است و می‌توان امید آن ها را باهم جمع کرد یعنی عملا مقدار بدست آمده را در سایز نمونه ضرب کرد.

$$\begin{cases} N_1 = 5,000 \Rightarrow 10\% \Rightarrow 100 \\ N_2 = 25,000 \Rightarrow 50\% \Rightarrow 500 \\ N_3 = 20,000 \Rightarrow 40\% \Rightarrow 400 \end{cases}$$

- در SRS می‌توان گفت که امید میانگین نمونه یک تخمین Unbiased از میانگین جامعه می‌باشد؛ ولی در شرایطی که هدف پیدا کردن یک آماره کلی نیست،بلکه بررسی توزیع‌های هر بخش می‌باشد یا دسته‌بندی داده است،چون از گروهی مثل گروه اول داده کمی موجود است، نمی‌توان خیلی بررسی انجام داد یا حتی احتمال از دست دادن یک گروه بالا است.

  درواقع، در شرایطی که جامعه همگن نیست و تفاوت زیادی وجود دارد، واریانس نمونه‌برداری SRS بالا خواهد بود.

- بله ممکن است. با فرض جایگذاری (برای استقلال) داریم:

$$P(X_i \notin Y_1) = 1 - P(X_i \in Y_1) = 1 - 0.1 = 0.9$$

باتوجه به استقلال، می‌توان مقدار احتمال هر متغیر را ضرب کرد:

$$P(X_j \notin Y_1 \forall j) = \prod_{j=1}^{1000} 0.9 = 0.9^{1000}$$

۲

$$W_1 = \frac{1}{10}$$
$$W_2 = \frac{1}{2}$$
$$W_3 = \frac{2}{5}$$

۳

- داریم:

$$n_1 = 100$$
$$n_2 = 500$$
$$n_3 = 400$$

- در اینجا نیز همانند SRS امید میانگین نمونه با امید جامعه برابر است، ولی برخلاف SRS که اطلاعات گروه‌ها را به کلی نادیده می‌گیرد، در اینجا اطمینان حاصل می‌شود تا از هر گروه به نسبت اعضای آن در نمونه‌برداری باشد، به این ترتیب احتمال نادید گرفتن یک گروه نیست و واریانس نمونه‌برداری نیز کاهش می‌یابد.

- خیر. در حالتی که دو گروه با تعداد یکسان موجود است ولی واریانس یک گروه از گروه دیگر بیشتر است، اگر از هر گروه به تعداد مشابهی نمونه برداشته شود، واریانس Estimator به نسبت زیاد می‌شود؛ زیرا گروهی که واریانس بیشتری دارد، یعنی داده‌های آن پراکندگی بیشتری دارند و نبود تعداد کمی از اعضای آن می‌تواند تخمین را جابجا کند. در شرایطی که واریانس گروه‌ها برابر باشد، این نمونه‌گیری بهینه است ولی اگر واریانس نیز مثل میانگین بطور یکنواخت پخش نشده باشد،نمونه‌گیری واریانس بالایی خواهد داشت.

B: Optimal Allocation Theory

۱

- اگر یک گروه تعداد اعضای بالایی داشته باشد، نمونه‌گیری برای این‌که بتواند نماینده جامعه باشد، باید از گروه اعضای بیشتری داشته باشد وگرنه گروه با تعداد بیشتر اثر کمتری در نمونه دارد.
- گروهی که واریانس بیشتری دارد یعنی اعضا در آن باهم خیلی یکدست نیستند و عملا برای این‌که واریانس نمونه‌برداری کمتر باشد، باید تعداد بیشتری از گروه را انتخاب کرد. در طرف مقابل نیز، اگر گروهی واریانس کمی داشته باشد،نیازی نیست تا خیلی از آن نمونه گرفت، با نمونه‌های کمی می‌توان اطلاعات گروه را داشت. برای مثال یک گروه با واریانس صفر (همه اعضا یکسان باشند) اگر صرفا یک عضو از آن در نمونه‌برداری باشد، داده گروه بدست آمده است، با اعضای بیشتر صرفا اعضای نمونه‌برداری هدر می‌روند!

۲

$$n_h^{opt} = n*\frac{W_h\sigma_h}{\sum_k W_h\sigma_h}, \sum_k W_h\sigma_h = 5,000*80,000+25,000*30,000+20,000*15,000 = 1e6(400+750+300)$$

$$n_1^{opt} = 1000*\frac{40}{145}, n_2^{opt} = 1000*\frac{75}{145}, n_3^{opt} = 1000*\frac{30}{109}$$

۳

جدول ۶: تعداد اعضای هر گروه در نمونه‌برداری‌های مختلف

Allocation Method	n_a	n_b	n_c
Proportional	100	500	400
Equal	333	333	333
Optimal (Neyman)	276	517	207

برای محاسبه SE داریم:

$$SE = \sqrt{\sum_h W_h^2 \frac{\sigma_h^2}{n_h}}$$

که اگر مقادیر را در فرمول بگذاریم داریم:

جدول ۷: مقدار SE برای نمونه‌برداری‌های مختلف

Allocation Method	SE
Proportional	49,395,748
Equal	54,313,902
Optimal (Neyman)	45,853,031

طبق جدول ۷ می‌توان دید کمترین میزان SE برای روش Neyman می‌باشد که دور از انتظار نیز نیست، زیرا در روش Neyman مقدار n با مقدار $W*\sigma$ نسبت مستقیم دارد، ولی در روش Proportional مقدار n به σ ربطی ندارد و برای همین یک کلاس با واریانس بالا می‌تواند مقدار SE را بالا ببرد.

C: Practical Implications

۱

نمونه‌گیری بر اساس نسبت شاید «عادلانه» باشد، ولی باعث می‌شود تا نمونه های مختلف تفاوت زیادی باهم داشته باشند زیرا در یک گروه ممکن است تنوع بیشتری باشد. با نمونه‌گیری به اندازه مساوی از این گروه، نمی‌توان دید خوبی از جامعه داشت. در نمونه‌گیری با نسبت برابر نمی‌توان به نمونه به چشم نماینده جامعه نگاه کرد زیرا اگر نمونه جامعه را نمایندگی می‌کند، نباید خیلی تفاوت بکند.

کار منطقی‌تر این است که به شیوه‌ای نمونه‌گیری کرد که نمونه‌ها در آن با هم تفاوت زیادی نداشته باشند. برای این که نمونه‌گیری «عادلانه» باشد، باید امید آن برابر با میانگین جامعه باشد (نه لزوما سایز نمونه‌ها برابر باشد) برای این‌که دید خوبی از جامعه داشت، منطقی است که از گروهی که تنوع بیشتری دارد، نمونه‌های بیشتری گرفته شود.

۲

داریم:

$$SE = \sqrt{\sum W_h^2 \frac{\sigma_h^2}{n_h}} \approx 115,012,076$$

(با استفاده از کد ارائه شده در نوت بوک حساب شده است)

در محاسبه خطا، دخیل بودن W واضح است. اگر دو گروه پراکندگی یکسانی دارند، گروه با تعداد بیشتر باعث می‌شود نمونه‌گیری نادقیق تر باشد، به همین علت نمی‌توان فقط بر واریانس تمرکز کرد.

۳

طبق فرمول Neyman نسبت سایز نمونه‌ها ثابت است یعنی سایز نمونه هر گروه نصف می‌شود. با توجه به فرمول SE می‌توان دید که با نصف شدن همه n ها، مقدار SE $\sqrt{2}$ برابر می‌شود. یعنی $64,845,978$

- همانطور که درقسمت قبل محاسبه شد $\sqrt{2}$
- با نصف‌کردن نمونه‌گیری، مقدار ME در $\sqrt{2}$ ضرب می‌شود. به بیان دیگر، میزان خطا در $\sqrt{2}$ ضرب می‌شود.

سوال ۸

A: The Known Variance Case

۱

داریم:

$$CI : P(\bar{x} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$Z_{0.975}\frac{\sigma}{\sqrt{n}} = 1.96 * \frac{0.05}{8} = 0.01225 \Rightarrow P(12.06775 < \mu < 12.9225) = 0.95$$

۲

عملا مشابه بخش قبل باید عمل کرد،صرفا مقدار Z بجای 1.96 برابر است با 2.57 پس:

$$Z_{0.995}\frac{\sigma}{\sqrt{n}} = 2.57 * \frac{0.05}{8} = 0.0160625 \Rightarrow P(12.0639375 < \mu < 12.0960625) \approx 0.99$$

۳

* چون فقط مقدار Z تغییر می‌کند، نسبت عرض بازه‌ها برابر است با نسبت Z ها:

$$\frac{Width\ 99}{Width\ 95} = \frac{Z_{0.0995}}{Z_{0.975}} \approx \frac{2.57}{1.96} \approx 1.31$$

* چون هدف داشتن بازه با اطمینان بالاتر است، نیاز است که بازه مقادیر بیشتری را شامل شود. درحالت کلی می‌توان گفت اگر بازه a زیرمجموعه بازه b باشد، $P(x \in a) \leq p(x \in b)$ زیرا هر حالتی که x عضو a باشد، عضو b نیز هست. یعنی بازه‌های بزرگتر احتمال بیشتری دارند.

* مطابق عدد حساب شده، حدودا 1.3 است.

۴

شیوه درست استفاده از واریانس جامعه می‌باشد. در شرایطی که واریانس جامعه موجود نیست می‌توان از واریانس نمونه استفاده کرد و در بازه اطمینان بجای توزیع Z از توزیع T Student استفاده کرد که باعث می‌شود به دلیل ندانستن واریانس جامعه اطمینان کمتر شود و به همین علت بازه اطمینان کمی بزرگتر می‌شود.

جدول ۸: مقادیر بازه اطمینان برای سیگماهای مختلف

σ	α	CI Width	Left Bound	Right Bound
0.05	0.95	0.012 25	12.067 75	12.092 25
0.05	0.99	0.016 06	12.063 94	12.096 06
0.06	0.95	0.014 70	12.065 30	12.094 70
0.06	0.99	0.019 27	12.060 72	12.099 28
0.07	0.95	0.017 15	12.062 85	12.097 15
0.07	0.99	0.022 49	12.057 51	12.102 49

B: Unknown Variance - Comparing Approaches

۱

داریم:

$$Width = 1.96 * \frac{0.08}{5} = 1.96 * \frac{0.08}{5} = 0.03136 \Rightarrow CI : [11.92864, 11.99136]$$

۲

داریم:

$$Width = 2.64 * \frac{0.08}{5} = 0.04224 \Rightarrow CI : [11.91776, 12.00224]$$

۳

* مورد محاسبه شده با استفاده از توزیع T Student
* داریم:

$$\frac{Width_T}{Width_Z} = \frac{T_24, 0.025}{Z_0.025} = \frac{2.64}{1.96} \approx 1.35$$

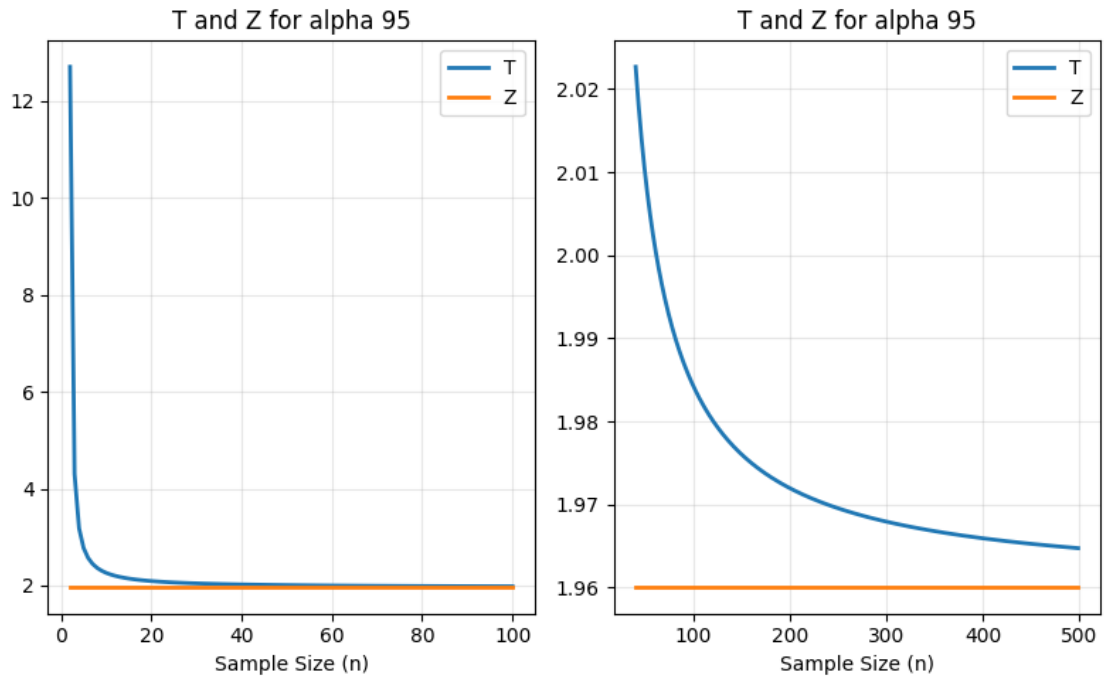
* چون مقدار واریانس به‌طور دقیق در اختیار نیست،ممکن است واریانس از واریانس توزیع بیشتر باشد و میانگین در بازه‌ای که با استفاده از توزیع Z بیان شد نباشد ،به بیان دیگر اطمینان کم می‌شود، برای این‌که همچنان با اطمینان 95 درصد بتوان بازه را گفت، باید بازه را بزرگتر گفت.

۴

* نسبت عرض بازه‌ها فقط به مقدار t و Z بستگی دارد پس داریم:

$$\frac{Width_T}{Width_Z} = \frac{T_99, 0.025}{Z_0.025} = \frac{1.984}{1.96} \approx 1.012$$

* با تعداد بالا، دقت واریانس نیز بیشتر می‌شود و واریانس نمونه‌برداری به واریانس توزیع نزدیک می‌شود پس دقت بیشتر می‌شود (ولی هیچگاه به Z نمی‌رسد. صرفا درحالت حدی برابر می‌شود.)
* با گذر از حدود 30 مقدار Z و T به هم به نسبت نزدیک است. با مقدار حدود 240 از 0.01 کمتر می‌شود.



شکل ۵: نمودار مقدار T به ازای زیاد شدن n

C: Interpretation and Misconceptions

۱

غلط است. میانگین مقداری ثابت است. شیوه درست این است: «این بازه به احتمال 95 درصد شامل میانگین است.»

۲

درست است. این تعریف کاملا مطابق تعریف ارائه شده در بخش قبل است. (میانگین مقداری ثابت دارد و این بازه است که رفتاری شبیه به متغیر دارد)

۳

اشتباه است. این موضوع کاملا به شکل توزیع و سایز نمونه بستگی دارد برای مثال، اگر n در نمونه‌گیری به سمت بینهایت برود، عرض بازه اطمینان به صفر میل می‌کند ولی قاعدتا 95 درصد داده‌ها دقیقا برابر میانگین نیستند.

۴

اشتباه است. بعد از نمونه‌گیری مقدار sample mean یک عدد است و کامل معلوم است. اگر منظور از sample mean درواقع امید آن باشد، چون امید آن برابر میانگین است، این گزاره شبیه مورد اول می‌شود که همچنان غلط است.

۵

بلی. با داده بیشتر با اطمینان بیشتری می‌توان صحبت کرد و باعث می‌شود بازه اطمینان کوچک تر شود.

D: Sample Size for Desired Precision

۱

$$ME = Z * \frac{\sigma}{\sqrt{n}} \Rightarrow n = (\frac{Z\sigma}{ME})^2$$

باتوجه به فرضیات مسئله داریم:

$$n = (\frac{1.96 * 0.08}{0.02})^2 \approx 61.4 \xrightarrow{n \in \mathbb{N}} n \geq 62$$

یعنی حداقل 62 بطری را باید نمونه بگیرند.

۲

* بازه قبلی درست است. بازه اطمینان برحسب واریانس نمونه‌گیری سنجیده می‌شود که خود بر حسب واریانس جامعه بدست می‌آید. واریانس نمونه ممکن است کمتر و یا بیشتر باشد ولی تاثیری در بازه اطمینان ندارد. در شرایطی که واریانس جامعه در دسترس نیست از واریانس نمونه و توزیع T Student استفاده می‌شود.
* همانطور که در بخش قبل بیان شد، نیازی به نمونه‌گیری اضافه نیست.

۳

می‌دانیم نسبت سایز نمونه با واریانس یک نسبت مربع است. یعنی اگر در اینجا واریانس یک و نیم برابر شده است،سایز نمونه باید $1.5^2 = 2.25$ برابر بشود. پس داریم:

$$n \approx 2.25 * 61.4 = 138.15 \xrightarrow{n \in \mathbb{N}} n \geq 139$$

یعنی حداقل 139 بطری باید نمونه گرفته شود.

سوال ۹

A: Proving the Central Limit Theorem via MGFs

۱

$$E(X)=0, Var(X)=1 \Rightarrow E(X^2)=E(X)^2+Var(X)^2=0+1=1$$

برای محاسبه MGF متغیر X چون می‌دانیم مشتق i ام تابع MGF برابر ممان i ام است داریم:

$$M_X(t) \approx M_X(0)+tM'_X(0)+t^2M''_X(0)=M_X(0)+tE(X)+\frac{t^2}{2}E(X^2)$$
$$\xrightarrow{E(X)=0,E(X^2)=1} M_X(t) \approx M_X(0)+0+\frac{t^2}{2} \xrightarrow{M_X(0)=E(e^0)=E(1)=1} M_X(t) \approx 1+\frac{t^2}{2}$$

۲

داریم:

$$M_{Z_n}(t)=M_{\sqrt{n}\bar{X}_n}=E(e^{t\sqrt{n}\bar{X}})=E(exp(t\sqrt{n}\bar{X}))=E(exp(t\sqrt{n}(\frac{\sum X_i}{n})))=E(\prod exp(t\sqrt{n}(\frac{X_i}{n})))=$$
$$E(\prod exp(t(\frac{X_i}{\sqrt{n}})))=E(\prod_{i=1}^{n} e^{\frac{tX_i}{\sqrt{n}}})$$

چون متغیرها مستقل هستند می‌توان امید ضرب‌ها را به ضرب امید تبدیل کرد و چون یکسان هستند، می‌توان فقط برای یک متغیر حساب کرد و سپس آن را بتوان n رساند:

$$M_{Z_n}(t)=E(exp(\frac{tx}{\sqrt{n}}))^n=M_X(\frac{t}{\sqrt{n}})^n$$

که این همان چیزی است که میخواستیم.

۳

با استفاده از دو رابطه قبلی داریم:

$$lim_{n\to\infty}M_{Z_n}(t)=lim_{n\to\infty}(1+\frac{(\frac{t}{\sqrt{n}})^2}{2})^n=(1+\frac{t^2}{2n})^n$$
$$\xrightarrow{m=\frac{t^2}{2}} lim_{n\to\infty}M_{Z_n}(t)=(1+\frac{m}{n})^n=e^m=e^{\frac{t^2}{2}}$$

۴

دیدیم که MGF میانگین یک سری متغیر iid با میانگین صفر و واریانس یک با MGF توزیع نرمال استاندارد برابر شد. پس می‌توان گفت که توزیع میانگین n متغیر iid با میانگین صفر و واریانس یک یک توزیع نرمال است.

در حالاتی که میانگین برابر صفر نیست یا واریانس برابر یک نیست، می‌توان Z Score کرد. به بیان دیگر:

$$Y_i=\frac{X_i-\mu}{\sigma} \Rightarrow E(Y)=0, Var(Y)=1 \Rightarrow \bar{Y_n} \sim Normal(0,1)$$
$$\Rightarrow \frac{\bar{X_n}-\mu}{\sigma} \sim Normal(0,1) \Rightarrow \bar{X_n} \sim Normal(\mu,\sigma)$$

که در این فرمول μ و σ میانگین و انحراف معیار Sample mean هستند که رابطه آن را با میانگین و انحراف معیار جمعیت داریم:

$$\mu_{\bar{X_n}}=\mu_{population}, \sigma_{\bar{X_n}}=\frac{\sigma_{population}}{\sqrt{n}}$$

پس به طور کلی برای هر توزیعی می‌توان گفت که توزیع میانگین نمونه‌ها برابر است با

$$\bar{X}_n \sim Normal(\mu, \frac{\sigma}{\sqrt{n}})$$

اگر n به سمت بینهایت برود. در عمل به ازای n های بزرگتر از 30 می‌توان با دقت خوبی از این رابطه استفاده کرد.

B: The Cauchy Anomaly (When CLT Breaks)

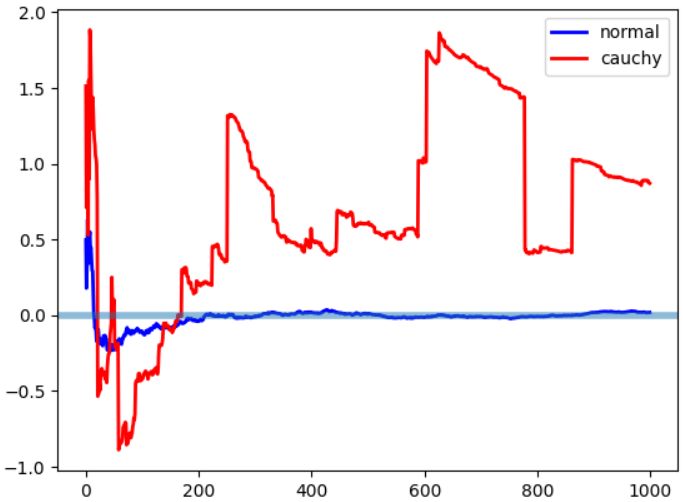
۱

خیر. زیرا براساس ویژگی بیان شده، میانگین n متغیر تصادفی با توزیع کوشی برابر با فقط یک متغیر است،پس عملا n تاثیری ندارد، از طرفی خود توزیع کوشی هم میانگین ندارد پس با میل کردن n به بینهایت، میانگین به یک عدد میل نمی‌کند.

۲

خیر کار نمی‌کند زیرا واریانس معین نیست و عملا مقدار ضرب را نمی‌توان بدست آورد. در درجه دوم حتی اگر واریانس معین بود، این بازه با استفاده از LLN بدست آمده که می‌گوید توزیع میانگین به شکل نرمال است ولی در توزیع کوشی چون واریانس معین نیست، از CLT پیروی نمی‌کند و نمی‌توان بازه اطمینان را به شکل توزیع نرمال استاندارد تبدیل کرد.

۳



شکل ۶: نمودار sample mean برای توزیع نرمال و کوشی

در شکل ۶ می‌توان دید که توزیع نرمال مطابق انتظار به مقدار میانگین جامعه (صفر) متمایل می‌شود ولی توزیع کوشی در مقادیر بالا نیز (مثلا ۸۰۰) نوسان دارد و به جایی متمایل نشده است.

۴

قضیه CLT برای هر توزیعی که واریانس محدود داشته باشد صدق می‌کند. اگر توزیع واریانس محدود و معین نداشته باشد نمی‌توان درباره CLT یا محاسبه بازه اطمینان با استفاده از توزیع نرمال استاندارد حرفی زد. علاوه بر این اگر توزیع واریانس بالایی داشته باشد، دیرتر Converge می‌کند و نیاز به n بالایی است (البته معمولا به ازای n برابر 30 توزیع را نرمال می‌گیرند.)

سوال ۱۰

A: Conceptual Understanding

۱

* توزیع نمونه توزیع اعدادی است که در یک نمونه بدست آمده‌اند. برای مثال اعدادی که در نمونه مشاهده می‌شوند. با نمونه‌گیری می‌توان میانگین و واریانس و دیگر آماره‌های توزیع نمونه را بدست آورد.
* توزیع نمونه‌گیری برابر توزیعی است که در نمونه‌گیری‌های مختلف بدست می‌آید. برای مثال وقتی در Confidence interval گفته می‌شود این بازه به احتمال ۹۵ درصد شامل میانگین جامعه می‌باشد، بازه‌ها از روی نمونه‌گیری‌های مختلف ساخته می‌شوند. با یک نمونه‌گیری نمی‌توان آماره‌های این توزیع مثل میانگین و ... را بدست آورد. (با آماره‌های جامعه می‌توان) بلکه می‌توان احتمالات مختلف آن را توضیح داد.

۲

* آزمایش اول توزیع نمونه را نشان می‌دهد. (برای یک نمونه خاص)
* آزمایش دوم توزیع نمونه‌گیری را نشان می‌دهد.
* آزمایش دوم. توزیع آزمایش اول به نوع توزیع بستگی دارد.

۳

* اشتباه. توزیع نمونه به توزیع جامعه بستگی دارد. CLT درباره توزیع نمونه‌گیری می‌باشد.
* درست است. طبق CLT این گزاره درست است.
* اشتباه. با زیاد شدن n واریانس نمونه به واریانس جامعه نزدیک و نزدیک‌تر می‌شود.
* درست است. (البته به شرط محدود بودن واریانس جامعه) با زیاد شدن n نمونه‌گیری‌های مختلف به‌هم نزدیک خواهند شد و واریانس کمتری خواهد داشت.

B: Detailed Example

۱

داریم:

$$\mu = E(X) = 0.3 + 0.4 + 1.5 = 2.2$$
$$\sigma = E((X-\mu)^2) = 0.3 * 1.2^2 + 0.2 * 0.2^2 + 0.5 * 0.8^2 = 0.3 * 1.44 + 0.2 * 0.04 + 0.5 * 0.64 = 0.76$$

۲

$$\begin{cases} A = (1,1) => P(A) = 0.09 \\ A = (1,2) => P(A) = 0.06 \\ A = (1,3) => P(A) = 0.15 \\ A = (2,1) => P(A) = 0.06 \\ A = (2,2) => P(A) = 0.04 \\ A = (2,3) => P(A) = 0.1 \\ A = (3,1) => P(A) = 0.15 \\ A = (3,2) => P(A) = 0.1 \\ A = (3,3) => P(A) = 0.25 \end{cases}$$

۳

$$\begin{cases} A = (1,1) => \mu_{sample} = 1 \\ A = (1,2) => \mu_{sample} = 1.5 \\ A = (1,3) => \mu_{sample} = 2 \\ A = (2,1) => \mu_{sample} = 1.5 \\ A = (2,2) => \mu_{sample} = 2 \\ A = (2,3) => \mu_{sample} = 2.5 \\ A = (3,1) => \mu_{sample} = 2 \\ A = (3,2) => \mu_{sample} = 2.5 \\ A = (3,3) => \mu_{sample} = 3 \end{cases}$$

۴

عملا باید معکوس بخش ۳ باتوجه به احتمالات حساب شده در بخش ۲ حساب شود:

$$P(\bar{x} = 1) = 0.09$$
$$P(\bar{x} = 1.5) = 0.12$$
$$P(\bar{x} = 2) = 0.34$$

$P(\bar{x} = 2.5) = 0.2$

$P(\bar{x} = 3) = 0.25$

۵

باتوجه به توزیع $\bar{X}$ داریم:

$E(\bar{X}_2) = 0.09 + 0.18 + 0.68 + 0.5 + 0.75 = 2.2 =$
$E(\bar{X}_2 - \mu) = 0.09 * 1.2^2 + 0.12 * 0.7^2 + 0.34 * 0.2^2 + 0.2 * 0.3^2 + 0.25 * 0.8^2 = 0.09 * 1.44 + 0.12 * 0.49 + 0.34 * 0.04 + 0.2 * 0.09 + 0.25 * 0.64 = 0.38$

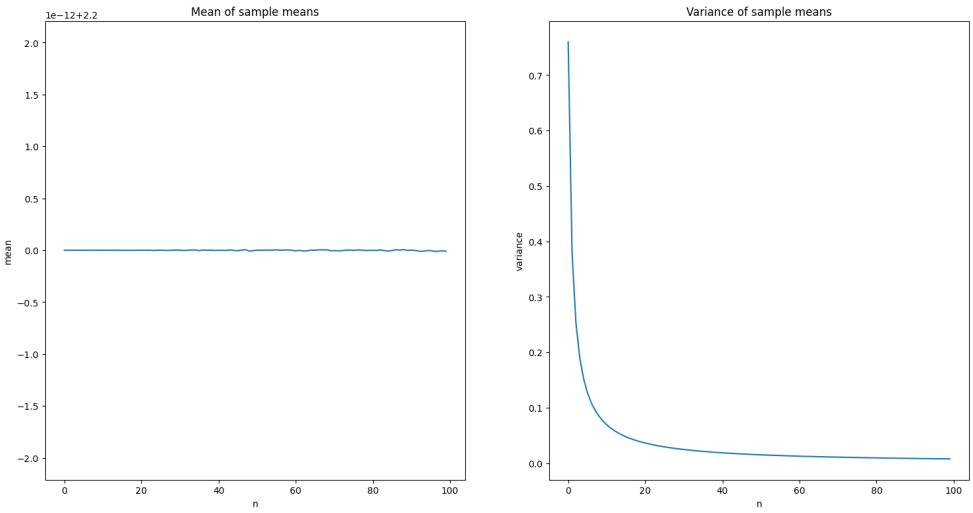
پس می‌توان دید که :

$$E(\bar{X}_2) = \mu_{population}, Var(\bar{X}_2) = \frac{\sigma_{population}}{2}$$

C: CLT Connection

۱

* توزیع مقادیر بیشتری را شامل می‌شود و طبق قضیه حد مرکزی به توزیع نرمال نزدیک می‌شود.
* تغییری نمی‌کند (به عنوان یک شهود، می‌توان دید که علیرغم این‌که توزیع مقادیر بیشتری را شامل می‌شود ولی حول میانگین بصورت متقارن تغییر می‌کند.) برای اثبات نیز طبق فرمول می‌توان دید که میانگین Sample mean ارتباطی با n ندارد و برابر میانگین جامعه می‌باشد..
* کم می‌شود (باز به عنوان شهود می‌توان این‌طور دید که با شامل شدن مقادیر بیشتر، فاصله بین نقاط تمرکز در توزیع اصلی پر می‌شود و واریانس کم می‌شود.) برای اثبات نیز طبق فرمول میانگین Sample mean می‌توان دید که واریانس این توزیع برابر است با: $\frac{\sigma_{population}}{\sqrt{n}}$



شکل ۷: نمودار واریانس و میانگین Sample Mean به ازای n

۲

از آن جا که عملگرهای خطی روی انتگرال (و بالطبع امید) تاثیر ندارند می‌توان عبارت را در مخرج ضرب کرد و با میانگین جمع کرد. پس باتوجه به فرمول‌های میانگین و واریانس توزیع داریم:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \Rightarrow \bar{X} \sim \frac{\sigma}{\sqrt{n}}N(0,1)+\mu \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

با توجه به قضیه حد مرکزی، توزیع $\bar{X}_n$ از نوع نرمال می‌باشد.

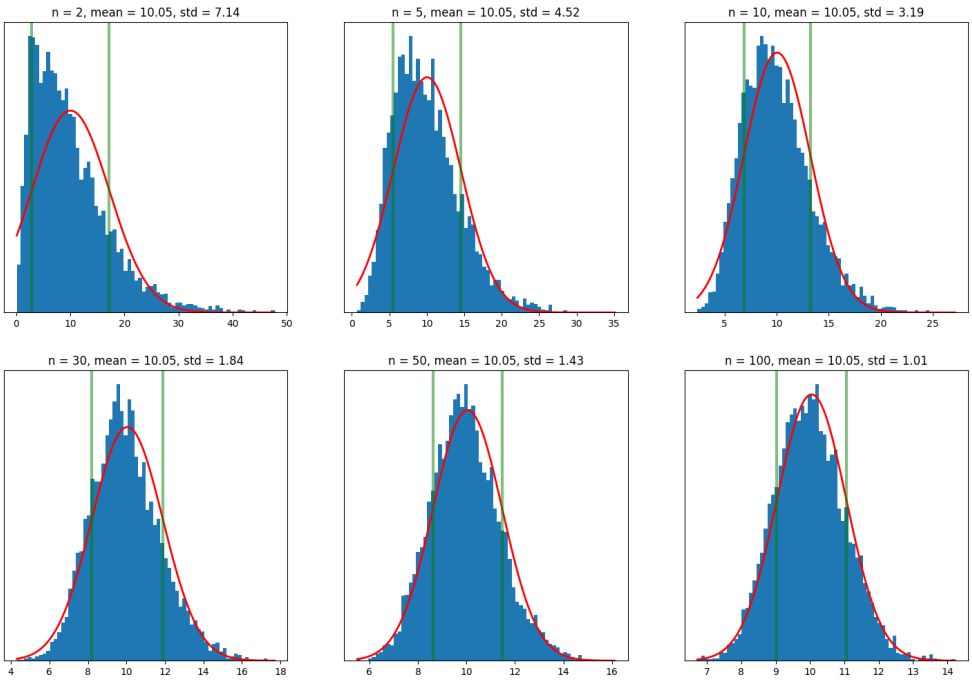
۳

با زیاد شدن سایز نمونه،همچنان در حال مشاهده یک نمونه‌برداری از یک توزیع خاص هستیم، پس با یک نمونه بزرگ، شکل توزیع نمونه شبیه به توزیع جامعه می‌شود و به سمت نرمال نمی‌رود و CLT صدق نمی‌کند.

CLT درباره توزیع نمونه‌برداری صحبت می‌کند. توزیع میانگین نمونه‌ها. با این‌که خود میانگین نمونه باتوجه به توزیع بدست می‌آید (برای مثال در بخش‌های قبل به ازای $n = 2$ حساب شد.) اما توزیع این میانگین‌ها به شکل نرمال خواهد بود.

سوال ۱۱

A: CLT Visual Demonstration



شکل ۸: نمودار هیستوگرام Sample mean

* به ازای $n = 30$ توزیع شبیه به نرمال به نظر می‌رسد. به‌ازای n های کمتر، توزیع شبیه به نرمال است ولی متقارن نیست ولی در $n = 30$ توزیع تقریبا متقارن است.

* به‌ازای n های بزرگتر، دامنه توزیع (مجموعه مقادیری که در نمونه مشاهده می‌شوند) بیشتر می‌شود ولی واریانس توزیع کم می‌شود (این را می‌توان از فاصله بین دوخط عمودی دید).

* خیر. با زیاد شدن n میانگین همچنان (مطابق انتظار) روی میانگین جامعه می‌ماند.

* می‌توان دید که توزیع X_n به سمت یک توزیع نمایی با میانگین جامعه و انحراف معیار $\frac{\sigma}{\sqrt{n}}$ می‌رود. که CLT هم همین را بیان می‌کند.

B: Confidence Interval Coverage

* ازآنجایی که CLT در بینهایت به صورت کامل صدق می‌کند،در $n = 30$ تقریبا می‌توان گفت که توزیع نرمال است ولی خطاهایی وجود دارد. همچنین خطای نمونه‌گیری نیز وجود دارد (با احتمال کم).

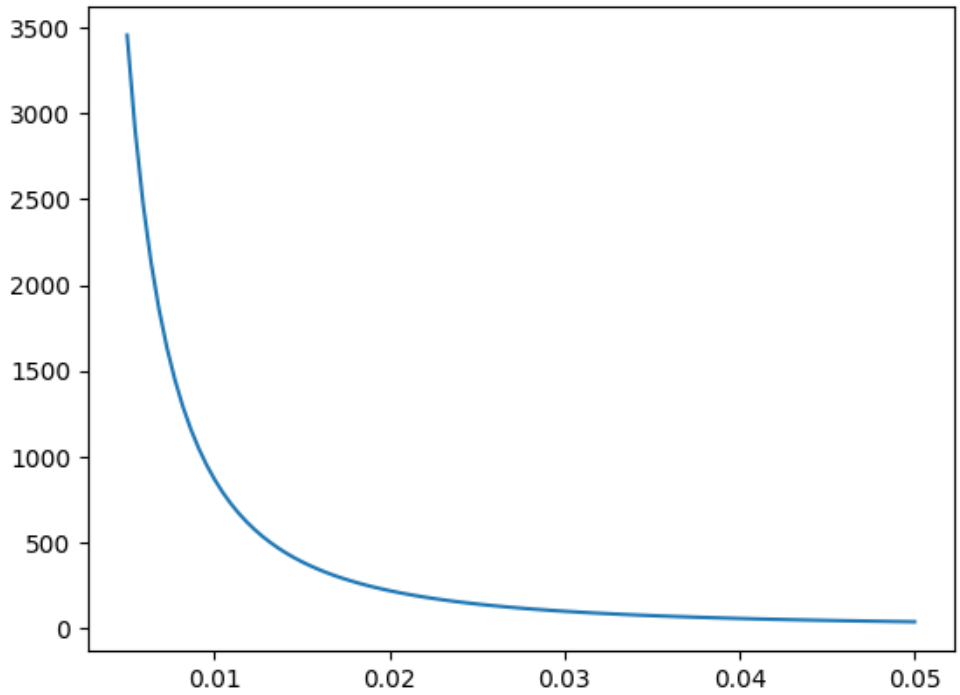
* چون CLT بیشتر به حد بینهایت خود نزدیک می‌شود انتظار می‌رود که Coverage اندکی بهتر شود. البته چون CLT در $n = 30$ نیز با دقت خوبی اعمال می‌شود، انتظار می‌رود تا روند بهبود بسیار کم باشد.

* زیرا برای اطمینان بیشتر، بازه بزرگتری درنظر گرفته می‌شود و میانگین جامعه، دفعات بیشتری در بازه می‌افتد.

جدول ۹: تعداد داده مورد نیاز به ازای پارامترهای مختلف بازه اطمینان

σ	Margin of Error	90% CI	95% CI	99% CI
0.1	0.01	271	385	664
0.1	0.02	68	97	166
0.2	0.01	1083	1537	2659
0.2	0.02	271	385	664

C: Sample Size Calculator



شکل ۹: میزان تعداد داده مورد نیاز برحسب حاشیه خطا

* طبق فرمول بدست امده در تابع، برای سایز داده برحسب حاشیه خطا داریم:

$$n = (\frac{Z_{0.5+\frac{\alpha}{2}} * \sigma}{ME})^2 = (\frac{constant}{ME})^2 = \frac{constant}{ME^2}$$

این رابطه ($\frac{1}{x^2}$) به وضوح خطی نیست و به همین علت نمودار نیز خطی نیست.

* با میل کردن ME به سمت صفر، n به سمت بینهایت می‌رود، برای همین برای حاشیه خطای کمتر نیاز به داده بیشتری داریم. شهود احتمالی نیز مطابق این است زیرا برای اینکه بتوان با احتمال یکسان بازه را کوچک‌تر کرد، باید داده بیشتری داشت که واریانس نمونه‌گیری کاهش یابد.

* برای نظرسنجی اول باید بزرگ‌ترین بازه قابل قبول را انتخاب کرد و سپس بر اساس آن اقدام به نظرسنجی کرد. اگر نیاز به حاشیه خطای در حد ۰۳.۰ داشت می‌توان با نمونه برداری با سایز تنها 96 یک بازه اطمینان با حاشیه خطای 0.03 ارائه داد. اما اگر با همین شرایط نیاز باشد حاشیه خطا 0.01 باشد، به نمونه‌ای با سایز 865 نیازمندیم. زیرا یک رشد معکوس با توان ۲ دارد. پس منطقی است که اول بزرگ‌ترین میزان قابل قبول حاشیه خطا را بیابیم و سپس به میزان مورد نیاز (یا برای اطمینان کمی بیشتر) نمونه برداری صورت بگیرد.

منابع خارج از درس