به نام خداوند جان و خرد

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

استنباط آماری

تمرین شماره ۱

نام و نام خانوادگی: علی خرم فر

شماره دانشجویی: ۸۱۰۱۰۲۱۲۹

آبان‌ماه ۱۴۰۲

فهرست مطالب

فهرست اشکال

۱_ پاسخ مسئله شماره ۱

۱-۱_ احتمال هشدار اشتباه

در این قسمت از سوال به دنبال احتمال فعال‌شدن هشدار در زمانی هستیم که اتفاق خطرناک رخ نداده و در شرایط نرمال هستیم. به این منظور از احتمال شرطی استفاده می‌کنیم. احتمال هشدار اشتباه در شرایطی خواهد بود که شرایط نرمال باشد. پس میزان احتمال وقوع هشدار در شرایط نرمال را در احتمال اینکه دستگاه در شرایط نرمال باشد ضرب می‌شود:

P(false alarm) = P(normal) × P(alarm | normal)

طبق فرضیات سوال احتمال وقوع هشدار در شرایط نرمال برابر است با:

P(alarm | normal) = 0.005

برای بدست آوردن احتمال اینکه دستگاه در شرایط نرمال قرار داشته‌باشد نیز، تفاضل مقدار احتمال شرایط خطرناک را از ۱ محاسبه می‌شود:

P(normal) = 1-P(dangerous) = 1 - 0.005 = 0.995

حال مقادیر جایگذاری می‌شود:

P(false alarm) = 0.995 × 0.005 = 0.004975

۱-۲_ احتمال وضعیت بحرانی

باتوجه به توصیف سوال این وضعیت در حالتی رخ می‌دهد که اولاً دستگاه در شرایط خطرناک باشد و ثانیاً هشدار دستگاه عمل نکرده و فعال نشود. در این قسمت از سوال نیز از احتمال شرطی کمک گرفته می‌شود:

P(UCC[۱]) = P(dangerous) × P(~alarm | dangerous)

طبق فرضیات سوال احتمال شرایط خطرناک برابر است با:

P(dangerous) = 0.005

برای بدست آوردن احتمال اینکه دستگاه در شرایط خطرناک بوده و هشدار فعال نشود نیز، تفاضل مقدار احتمال اینکه در شرایط خطرناک بوده و هشدار فعال شود را از ۱ بدست می‌آوریم. احتمال وقوع هشدار در شرایط خطرناک طبق فرضیات مسئله برابر ۰.۹۵ است:

P(~alarm | dangerous) = 1- P(alarm | dangerous) = 1 - 0.95 = 0.05

---

[۱] Unidentified critical condition

حال مقادیر جایگذاری می‌شود:

$$P(UCC) = 0.005 \times 0.05 = 0.00025$$

۳-۱_ دفعات وقوع حوادث در بازه ۱۰ ساله

طبق فرضیات و زیرمسئله‌های حل‌شده در قسمت قبل، احتمال وقوع شرایط بحرانی و همچنین هشدار اشتباه را در بازه روزانه محاسبه شده‌است. حال برای محاسبه تعداد دفعات مورد انتظار برای وقوع شرایط بحرانی و هشدار اشتباه در بازه ۱۰ سال مقادیر مذکور در تعداد روزهای این بازه ضرب می‌شود:

$$\text{False alarms expected in 10 years} = 0.004975 \times 10 \times 365 = 18.15875$$

$$\text{UCCs expected in 10 years} = 0.00025 \times 10 \times 365 = 0.9125$$

اثربخشی سیستم

باتوجه به محاسبات انجام شده و داده‌های قبلی مسئله، این سیستم از نظر تشخیص شرایط خطرناک دقت خوبی دارد. احتمال وقوع هشدار اشتباه مقدار کمی دارد. ولی باید توجه داشت که در سیستم‌های حیاتی که نیاز به دقت بالا دارند، این میزان دقت ممکن است باعث وقوع حوادث جبران ناپذیری شود. وقوع هشدارهای اشتباه ممکن است باعث شود که حساسیت کاربر دستگاه نسبت به وقوع شرایط خطرناک واقعی کاهش یابد. ولی احتمال عدم هشدار در شرایط بحرانی نیز در این سیستم قابل قبول است.

۲_پاسخ مسئله شماره ۲

۲-۱_ میانگین و واریانس جمع دو متغیر تصادفی

در این قسمت از مسئله قصد داریم میانگین و واریانس متغیر تصادفی Y_1 را در شرایطی محاسبه کنیم:

$$Y_1 = X_1 + X_2$$

محاسبه میانگین

میانگین یا انتظار یک متغیر تصادفی، امید ریاضی است. برای محاسبه میانگین از خواص امید ریاضی استفاده می‌شود. امید ریاضی خطی است.پس امیدهای داخلی از پرانتز خارج می‌شود. سپس میانگین X_1 و X_2 طبق فرضیات سوال جایگذاری می‌شوند.

$$E(Y_1) = E(X_1 + X_2) = E(X_1) + E(X_2) = m + m = 2m$$

محاسبه واریانس

باتوجه به نکات مطرح‌شده در محتوای درس، مجموع واریانس یک سری متغیر تصادفی با **شرط استقلال** برابر است با:

$$\mathrm{Var}(\sum_i a_{i\,x_i}) = \sum_i a^2 Var(X_i)$$

پس باتوجه به این نکته و فرضیات این سوال که هیچکدام از متغیرهای تصادفی ضریب ندارند و مستقل‌اند، واریانس مجموع آن‌ها برابر است با:

$\mathrm{Var}(Y_1) = \mathrm{Var}(X_1 + X_2) = \mathrm{Var}\,(X_1) + \mathrm{Var}\,(X_2) = \sigma^2 + \sigma^2 = 2\sigma^2$

۲-۲_ میانگین و واریانس دوبرابر متغیر تصادفی

در این قسمت از مسئله قصد دارد میانگین و واریانس متغیر تصادفی Y_2 را در شرایطی محاسبه شود که:

$Y_2 = 2X_1$

محاسبه میانگین

باتوجه به توضیحات مذکور در قسمت ۱ مسئله داریم:

$E(Y_1) = E(2X_1) = 2E(X_1) = 2m$

محاسبه واریانس

باتوجه به نکات مطرح‌شده در محتوای درس، در صورت برقراری شرایط زیر می‌شود ضریب را از واریانس خارج کرد و به توان ۲ رساند، همچنین مقدار جمع شده با متغیر تصادفی تاثیری در واریانس نخواهدداشت:

$$\mathrm{Var}(aX + b) = a^2 Var(X)$$

$\mathrm{Var}(Y_2) = \mathrm{Var}(2X_1) = 4\,\mathrm{Var}\,(X_1) = 4\,\sigma^2$

۳-۲_ تساوی واریانس Y_1 و Y_2

خیر، باتوجه به محاسبات انجام شده و خواص واریانس مذکور، واریانس Y_1 و Y_2 برابر نیست. برای محاسبه واریانس Y_1 ما شرط بر این گذشته‌شده که دو متغیر تصادفی که با هم جمع می‌شوند، مستقل هستند، هرچند که مقدار واریانس برابری دارند. در حالی که در Y_2 واریانس برای دوبرابر یک متغیر تصادفی

محاسبه شده‌است. دلیل این تفاوت در خواصی که در قسمت ۱ سوال مطرح شده قابل اثبات است و هرگاه واریانس یک متغیرتصادفی که یک عدد در آن ضرب شده محاسبه شود، مقدار ضریب به توان ۲ رسیده و از واریانس خارج می‌شود. در حالی که در قسمت ۱ سوال دو متغیر جمع شده و ضریب ندارند.

۴-۲_ محاسبه کوواریانس Y_1 و Y_2

باتوجه به فرضیات سوال داریم:

$Y_1 = X_1 + X_2$

$Y_2 = 2X_1$

$Cov(Y_1 , Y_2) = Cov(X_1 + X_2 , 2X_1)$

از طرفی باتوجه به خواص کوواریانس داریم:

$$\mathbf{Cov}(aX + bY, cZ) = ac\,\mathbf{Cov}(X , Z) + bc\,Cov(Y , Z)$$

$Cov(Y_1 , Y_2) = Cov(X_1 + X_2 , 2X_1) = 2Cov(X_1 , X_1) + 2Cov(X_2 , X_1)$

باتوجه به فرضیات مسئله متغیرهای تصادفی X_1 و X_2 مستقل هستند پس مقدار کوواریانس آن‌ها صفر خواهدبود. پس پاسخ نهایی برابر است با ۲ برابر کوواریانس متغیر تصادفی X_1 و X. از طرفی مقدار کوواریانس بین دو متغیر تصادفی یکسان برابر واریانس آن متغیر تصادفی خواهدبود.

$Cov(Y_1 , Y_2) = 2Cov(X_1 , X_1) = 2Var(X_1) = 2 \times \sigma^2 = 2\,\sigma^2$

۳_ پاسخ مسئله شماره ۳

۳-۱_ تابع احتمال روز تولد

برای تعریف تابع احتمال برای این قسمت از سوال با توجه به اینکه روز تولد هر نفر با نفر دیگر مستقل است، و احتمال تولد هر نفر 1/365 می‌باشد، برای n نفر تابع احتمال برابر است با:

$P(\Omega) = \left(\frac{1}{365}\right)^n$

A: At least one person in the group shares your birthday

برای محاسبه P(A) این مورد در نظر گرفته‌شود که ۳۶۵ حالت برای انتخاب روز تولد من وجود دارد. پس برای محاسبه فرمول برای P(A) که بیانگر این است که حداقل یک نفر در این گروه روز تولد یکسانی با من دارد، می‌توان با کمک مکمل‌گیری، احتمال اینکه n فرد تولد مشترک با من ندارند را از ۱ تفاضل گرفت. یعنی همه روزها به جز روز تولد من محاسبه شوند.

$P(A) = 1 - \left(\frac{364}{365}\right)^n$

برای محاسبه حداقل مقدار n برای شرایطی که $P(A) > 0.5$ شود، باید نامساوی زیر برقرار شود:

$$1 - \left(\frac{364}{365}\right)^n > 0.5 \rightarrow 0.5 > \left(\frac{364}{365}\right)^n \rightarrow Log(0.5) > nlog\left(\frac{364}{365}\right) \rightarrow n > \frac{Log(0.5)}{log\left(\frac{364}{365}\right)}$$

این نامساوی برای nهای بزرگترمساوی ۲۵۳ برقرار است.

بدون نیاز به محاسبات مشخص است که برای اینکه احتمال ۵۰ درصدی وجود تولد مشترک بین من و n نفر دیگر وجود داشته‌باشد، نیاز است حداقل به اندازه نیمی از روزهای سال فرد در گروه وجود داشته باشد.

B: At least two people in the group share a birthday

با کمک پایتون و اجرای کد مربوطه میزان n برابر ۴۱ تخمین زده‌شد. فایل برنامه در پوشه تمرین پیوست شد.

با ۳۰ آزمایش مقدار n برابر ۴۱ بود که تغییری ایجاد نشد.

برای تعیین فرمول برای P(B) که حداقل ۲ نفر در گروه تاریخ تولد مشترکی دارند،همانند فرمول P(A) از مکمل استفاده کرده و احتمال اینکه هیچ دو نفری تولد یکسان نداشته باشند، از ۱ تفاضل گرفته می‌شود:

$P(B) = 1 - \frac{365 \times 364 \times 363 \times \ldots \times 366-n}{365^n} = 1 - \frac{365!}{(365-n)!365^n}$

C: ”At least three people in the group share a birthday

با کمک پایتون و اجرای کد مربوطه میزان n برابر ۸۸ تخمین زده‌شد.

باتوجه به اینکه در دنیای واقعی، توزیع روزهای تولد مانند آزمایشی که انجام شد، یکنواخت نیست و ممکن است برخی روزهای سال باتوجه به دلایل مختلف(همچون ۳۱ شهریور و ۱ فروردین در ایران) و یا دلایل مذهبی و فرهنگی روز تولد افراد بیشتری یکسان باشد. پس نسبت به آزمایش انجام شده n کمتر خواهدبود.

۴_ پاسخ مسئله شماره ۴

۱-۴_ احتمال دختربودن هردو فرزند

به طور کلی برای احتمال جنسیت پسر یا دختر بودن ۲ فرزند ۴ حالت وجود دارد. در قسمت ۱ این مسئله جنسیت فرزند بزرگتر مشخص شده که دختر است پس ۲ حالت باقی می‌ماند. در فضای نمونه‌ای

جدید داریم اینکه فرزندان دختر/پسر باشند یا دختر/دختر. مطلوب سوال محاسبه احتمال برای حالت دختر/دختر است، پس داریم:

$P = \frac{1}{2}$

۲-۴_ احتمال پسربودن هردو فرزند

می‌دانیم که حداقل یکی از فرزندان پسر است. پس حالات این مسئله پسر/دختر و پسر/پسر خواهدبود. احتمال مورد نظر حالت پسر/پسر از میان ۲ حالت مذکور است. پس داریم:

$P = \frac{1}{2}$

البته این مسئله باتوجه به مبهم‌بودن نوع سوال از منظرهای دیگری نیز قابل بررسی بوده و نوع تفسیر سوال ممکن است درک ناصحیحی برای محاسبه احتمال به ما بدهد. در(Bar-Hillel & Falk, 1982) بررسی شده که می‌توان شرایطی را در نظر گرفت که حالات این مسئله پسر/دختر، دختر،پسر و پسر/پسر خواهدبود. مطلوب احتمال این مسئله حالتی است که حداقل یک فرزند پسر است. پس احتمال اینکه هر دو پسر باشند در حالتی که حداقل یکی از آن‌ها پسر باشد برابر است با:

$P= \frac{1}{3}$

۵_ پاسخ مسئله شماره ۵

۱-۵_ استدلال منطقی

در این سوال تصادفی رخ داده که توسط شاهدی عینی تاکسی به رنگ آبی را مشاهده کرده است. هرچند که این گمان به نسبت قوی است ولی قصد داریم با توجه به احتمالات تشخیص اشتباهی که در شرایط مسئله فرض شده است به نفع موکل مورد نظر در هیئت منصفه تردید به وجود آورده‌شود. به این منظور باید در متن اعلامیه که به هیئت منصفه اعلام می‌شود تا حد ممکن از فرموله‌بندی مسئله پرهیز کرده و از استدلال منطقی و جدول استفاده‌شود. باتوجه به داده‌های اعلام‌شده از وضعیت تشخیص شاهد این قضیه، او در تشخیص تاکسی‌های آبی دقت ۹۹درصدی دارد. هرچند که به احتمال ۲درصد نیز تاکسی‌های سبز را به عنوان آبی شناسایی می‌کند. باتوجه به اینکه در این شهر ۹۹ تاکسی سبز و فقط ۱ تاکسی آبی وجود دارد، احتمال اینکه او تاکسی سبز را، آبی تشخیص دهد بیشتر از آن است که تاکسی آبی را به درستی شناسایی کند. باید توجه شود که باید از تکنیک‌های توصیف آمار به نفع موکل بهره گرفت.

۵-۲_ متن پیشنهادی برای هیئت منصفه

هیئت محترم منصفه توجه داشته باشید که در شهر ما فقط ۱ تاکسی آبی در میان ۱۰۰ تاکسی وجود دارد. پس این موضوع را در نظر بگیرید که وهله اول فقط ۱ درصد احتمال این وجود دارد که یک تاکسی انتخابی در شهر آبی باشد. حالا به سراغ تنها شاهد ماجرا برویم.

توجه شما را به داده‌هایی که از نتیجه آزمایش بر روی شاهد بدست آورده‌ایم جلب می‌کنم. ما شرایط صحنه‌ی تصادف را برای شاهد بازسازی کردیم و متوجه شدیم که ایشان در ۲ درصد مواقع در تشخیص درست تاکسی سبز عاجز است. علاوه بر این او نه تنها رنگ تاکسی سبز را به درستی شناسایی نمی‌کند، آن را به رنگ آبی می‌بینند که دست بر قضا تاکسی متهم این پرونده به رنگ آبی است. حال این شرایط را در حالی در نظر بگیرید که فقط ۱ درصد تاکسی‌های این شهر آبی هستند و در ۹۹درصد موارد تاکسی‌ها سبز هستند. درست است که در بیشتر مواقع تاکسی‌های آبی را به درستی تشخیص می‌دهد. ولی احتمال اینکه تاکسی سبز به آبی تشخیص داده‌شود خیلی بیشتر است.

رنگ تاکسی	تعداد	احتمال حضور در حادثه	دقت شاهد
سبز	۹۹	۹۹ درصد	۲درصد تشخیص اشتباه به آبی
آبی	۱	۱ درصد	۱ درصد احتمال تشخیص اشتباه

شکل ۱ جدول ارائه شده به هیئت منصفه

۶_ پاسخ مسئله شماره ۶

۶-۱_ تابع توزیع احتمال (PMF[1]) برای S

$P_{X(x_0)} = P\{X = x_0\}$

۴ تاس وجود دارد. نوع هر تاس بر اساس تعداد اضلاع آن مشخص شده‌است. احتمال اینکه هر یک از این تاس‌ها انتخاب شود 1/4 است. از طرفی ۲ تاس ۸ضلعی وجود دارد پس احتمال انتخاب آن‌ها به صورت مجموع 1/2 است. متغیر تصادفی S بیانگر تعداد وجه‌های آن است.

$P(4) = P\{S = 4\} = \frac{1}{4}$

$P(6) = P\{S = 6\} = \frac{1}{4}$

$P(8) = P\{S = 8\} = \frac{1}{2}$

[1] Probability mass function

۲-۶_ احتمال R با کمک بیز

مقدار هربار تاس انداختن را با R مشخص شده‌است. حال با کمک تئوری بیز احتمال وقوع هریک از تاس‌ها وقتی که R=3 باشد محاسبه می‌شود.

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(S|R{=}3) = \frac{P(R = 3|S)P(S)}{P(R{=}3)}$$

برای محاسبه احتمال وقوع هر یک از تاس ها ابتدا $P(R = 3|S)$ را برای هریک از آن‌ها محاسبه کرده و در فرمول بیز جایگذاری می‌شود. قبل از محاسبه برای هر تاس مقدار $P(R = 3)$ را برای جایگذاری در آن‌ها محاسبه می‌شود. بر اساس قانون احتمال کل داریم:

$$P(R = 3) =$$

$$P(R = 3|\, S = 4)P(S = 4) + P(R = 3|S = 6)P(S = 6) + P(R = 3|S = 8)P(S = 8)$$

$$\frac{1}{4} \times \frac{1}{4} + \frac{1}{4} \times \frac{1}{6} + \frac{1}{2} \times \frac{1}{8} = \mathbf{0.16}$$

محاسبه وقوع تاس S=4

احتمال اینکه در تاس ۴وجهی عدد ۳ ظاهر شود:

$$P(R = 3|S = 4) = \frac{1}{4}$$

$$P(S = 4) = \frac{1}{4}$$

$$P(R = 3) = 0.16$$

$$P(S{=}4|R{=}3) = \frac{P(R = 3|S = 4)P(S{=}4)}{P(R{=}3)} = \frac{\frac{1}{4}\times\frac{1}{4}}{0.16} = 0.390625$$

محاسبه وقوع تاس S=6

احتمال اینکه در تاس ۶وجهی عدد ۳ ظاهر شود:

$$P(R = 3|S = 6) = \frac{1}{6}$$

$$P(S = 6) = \frac{1}{4}$$

$$P(R = 3) = 0.16$$

$$P(S{=}6|R{=}3) = \frac{P(R = 3|S = 6)P(S{=}6)}{P(R{=}3)} = \frac{\frac{1}{6}\times\frac{1}{4}}{0.16} = 0.26041$$

محاسبه وقوع تاس S=8

احتمال اینکه در تاس ۸وجهی عدد ۳ ظاهر شود:

$P(R = 3|S = 8) = \frac{1}{8}$

$P(S = 8) = \frac{1}{2}$

$P(R = 3) = 0.16$

P(S=8|R=3) $= \frac{P(R = 3|S = 8)P(S=8)}{P(R=3)} = \frac{\frac{1}{8}\times\frac{1}{2}}{0.16} = 0.390625$

هنگامی که R=6 باشد احتمال اینکه تاس ۴ وجهی انتخاب شده باشد صفر است. پس باید مقایسه‌ای بین ۸وجهی و ۶ وجهی انجام شود.

$P(R = 6) =$

$P(R = 6|\ S = 4)P(S = 4) + P(R = 6|S = 6)P(S = 6) + P(R = 6|S = 8)P(S = 8)$

$0 \times \frac{1}{4} + \frac{1}{4} \times \frac{1}{6} + \frac{1}{2} \times \frac{1}{8} = \mathbf{0.1041}$

P(S=6|R=6) $= \frac{P(R = 3|S = 6)P(S=6)}{P(R=6)} = \frac{\frac{1}{6}\times\frac{1}{4}}{0.1041} = 0.6003$

P(S=8|R=6) $= \frac{P(R = 3|S = 8)P(S=8)}{P(R=6)} = \frac{\frac{1}{8}\times\frac{1}{2}}{0.1041} = 0.40025$

احتمال اینکه تاس ۸ انتخاب شده‌باشد بیشتر است.

هنگامی که R=7 باشد احتمال اینکه تاس ۴ وجهی و یا ۶وجهی انتخاب شده باشد صفر است. پس بدون نیاز به محاسبات می‌توان گفت که تاس ۸ وجهی انتخاب شده است.

۷_پاسخ مسئله شماره ۷

۷-۱_ محاسبه انحراف معیار

برای محاسبه انحراف معیار یک متغیر تصادفی داریم:

$$\sigma_x = \sqrt{Var} = \sqrt{E(X - \mu)^2} = \sqrt{\sum_i (xi - \mu)^2\ Px(x_i)}$$

برای محاسبه انحراف معیار X که نشان دهنده نتیجه تاس ۴وجهی است، تابع جرم احتمال برای همه مقادیر برابر 1/4 است و همچنین میانگین حساب می‌شود. واریانس را حساب کرده و جذر آن محاسبه می‌شود. پس برای محاسبه انحراف معیار داریم:

$$\mu = \frac{1+2+3+4}{4} = 2.5$$

$$\sigma_x = \sqrt{\sum_i (xi - \ 2.5)^2 \frac{1}{4}} = \sqrt{\frac{1}{4}(\,(1-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (4-2.5)^2)}$$

$$= \ \sqrt{1.25} \ = 1.118033$$

برای محاســبه انحراف معیار Y که نشــان دهنده نتیجه تاس ۶وجهی اســت، تابع جرم احتمال برای همه مقادیر برابر 1/6 اســت و همچنین میانگین حســاب می‌شــود. واریانس را حســاب کرده و ریشــه دوم آن حساب می‌شود. پس برای محاسبه انحراف معیار داریم:

$$\mu = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\sigma_y = \sqrt{\sum_i (xi - \ 3.5)^2 \frac{1}{6}} =$$

$$\sqrt{\frac{1}{6}(\,(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2))}$$

$$= \ \sqrt{2.9166666} \ = 1.70782510814$$

برا محاسبه انحراف معیار Z داریم:

$$Z = \frac{X+Y}{2}$$

$$\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2} \ = \sqrt{1.25 + 2.91} \ = \ 2.039$$

۲-۷_ توابع PMF و CDF برای Z

برای مقدار PMF تابع Z ، احتمال را برای تمام مقادیر آن محاسبه می‌شود. PMF متغیر تصادفی X از آن‌جا که توزیع یکنواخت بوده برای تمامی مقادیر 1/4 و برای متغیر تصــادفی Y برای تمامی مقادیر 1/6 برمی‌گرداند.

باتوجه به اینکه مطلوب سوال یک PMF و CDF جامع است با کمک Excel مقادیر مختلف این دو تابع را محاسبه می‌شود و برای میانگین‌های مختلف یک احتمال را حساب می‌کنیم.اگر که دو مقدار میانگین یکسانی داشته باشند احتمال آن‌ها برای PMF جمع می‌شود. فایل مورد نظر در پوشه تمرین پیوست شد.

۳-۷_ احتمال سود یا ضرر در بازی

باتوجه به قوانین بازی اگر که X>Y باشد ۲برابر X سود حاصل می‌شود و اگر X<Y باشد ۱ دلار ضرر محاسبه می‌شود. برای محاسبه احتمال سود یا ضرر در ۶۰ دور بازی ابتدا میزان مورد انتظار متوسط از سود یا ضرر در یک دور از بازی حساب می‌شود.

به طور کلی باتوجه به مقادیر مختلف X و Y در این بازی ۲۴ حالت اولیه وجود دارد. از این ۲۴ حالت تعداد دفعات اینکه شرط اول برقرار باشد یعنی X>Y برابر است با ۶ حالت:

شکل ۲ جدول حالات برد

X	۲	۳	۳	۴	۴	۴
Y	۱	۱	۲	۱	۲	۳

پس احتمال برد در یک دور از بازی 6/24 یا 1/4 خواهدبود. از طرفی باتوجه به اینکه اگر مقدار Y مساوی یا بزرگتر از X باشد تفاضل ۶ حالت قبلی با ۲۴ تعداد حالات شرط دوم هستند. به همین ترتیب ۱۸ حالت هم برای باخت وجود دارد. پس احتمال باخت 18/24 یا 3/4 خواهدبود. در هر دور تاس ریخته می‌شود. میزان سودی که هر دور برنده‌شدن به طور متوسط بدست می‌آید برابر است با:

$$\frac{1}{6} \times 4 + \frac{2}{6} \times 6 + \frac{3}{6} \times 8 = 6.66$$

میزان ضـــرر در هر دور باخت نیز برابر ۱ خواهدبود. میزان مورد انتظار ســـود از یک دور بازی برابر است با:

$$E(W) = \frac{1}{4} \times 6.66 - \frac{3}{4} \times 1 = 0.915$$

برای ۶۰ دور بازی مستقل میزان سود احتمالی برابر ۵۴.۹ خواهدبود.

۸_پاسخ مسئله شماره ۸

۱-۸_ محاسبه تعداد کشمش‌ها

باتوجه به تابع ارائه‌شـده در مسـئله در رابطه با چگالی وجود کشـمش در هر قسـمت جعبه، برای محاسـبه تعداد کل کشـمش‌ها در یک جعبه بر روی تابع ارائه شـده روی بازه ۰ تا ۳۰ که ارتفاع جعبه ورودی است، انتگرال گرفته می‌شود:

$$\text{Number of raisins} = \int_0^{30} f(h)dh = \int_0^{30}(40-h)dh = \int_0^{30}(40-h)dh = [40h - \frac{h^2}{2}]_0^{30} = (40 \times 30 - \frac{30^2}{2}) - 0 = 1200 - 450 = 750$$

پس تعداد کشمش‌ها برابر ۷۵۰ خواهدبود.

۸-۲_ تابع چگالی احتمال PDF

متغیر تصادفی جدید H میزان ارتفاع یک کشمش انتخابی را می‌دهد برای محاسبه احتمال وجود کشمش در ارتفاع h از فرمول ارائه‌شده در مسئله استفاده می‌شود. f(h) چگالی کشمش را نسبت به ارتفاع خروجی می‌دهد.. برای اینکه از این تابع چگالی برای ساخت یک PDF استفاده شود باید نرمال سازی شود تا احتمال تمامی ۷۵۰ کشمش برابر ۱ شود. پس این تابع بر ۷۵۰ تقسیم می‌شود.

g(h) = $\frac{40-h}{750}$

۸-۳_ تابع توزیع تجمعی CDF

مطابق تعاریف درسی برای تبدیل PDF به CDF از آن انتگرال گرفته می‌شود.

G(h) = $\int_0^h \frac{40-h}{750} dh = [\frac{40h-\frac{h^2}{2}}{750}]_0^h = [-\frac{(h-80)h}{1500}]_0^h = -\frac{h^2-80h}{1500}$

۸-۴_ احتمال در ارتفاع دلخواه

باتوجه به مفروضات مسئله، ارتفاع جعبه ۳۰ سانتی‌متر است. پس یک‌سوم پایینی آن برابر با ارتفاع ۰ تا ۱۰ سانتی‌متری خواهدبود. پس احتمال وجود یک کشمش تصادفی برای h بین ۰ و ۱۰ محاسبه می‌شود. می‌توان از CDF برای محاسبه احتمال استفاده کرد:

P(h <= 10) =$-\frac{h^2-80h}{1500} = -\frac{10^2-800}{1500} = 0.466$

۹_پاسخ مسئله شماره ۹

۹-۱_ محاسبه توزیع توام

P(X=1) = P(X=-1) = $\frac{1}{2}$, P(Y=-1) = P(Y=-1) = $\frac{1}{2}$

طبق صورت مسئله P(X=1,Y=1)= c می‌باشد. برای محاسبه دیگر مقادیر توزیع توام طبق توزیع حاشیه‌ای داریم:

	Y=1	Y=-1
X=1	c	1-c
X=-1	1-c	c

شکل ۳توزیع توام دو متغیر تصادفی

Cov(X,Y) = E(XY) – E(X)E(Y)

E(X) = $1 \times \frac{1}{2} + -1 \times \frac{1}{2} = 0$

Var(X) = $E((X - E(X)^2)) = E(X^2) = 1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 1 , \sigma_X = 1$

$\mathrm{E(Y)} = 1 \times \frac{1}{2} + -1 \times \frac{1}{2} = 0$

$\mathrm{Var(Y)} = E((Y - E(Y)^2)) = E(Y^2) = 1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 1 \,, \sigma_Y = 1$

Cov(X,Y) = E(XY) = 1×1×c + 1×-1×(1-c) + -1×1×(1-c) + -1×-1×c = c+-1+c+-1+c+c = 4c -2

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{4c-2}{1} = 4c - 2$$

۲-۹_ تعیین استقلال دو متغیر تصادفی

باتوجه به مطالب درس، در صورت استقلال دو متغیر تصادفی رابطه زیر برقرار است:

P(X,Y) = P(X)P(Y)

$P(X{=}1,X{=}1) = c \rightarrow P(X{=}1)P(Y{=}1) = \frac{1}{2} \times \frac{1}{2} = c \rightarrow c = \frac{1}{4}$

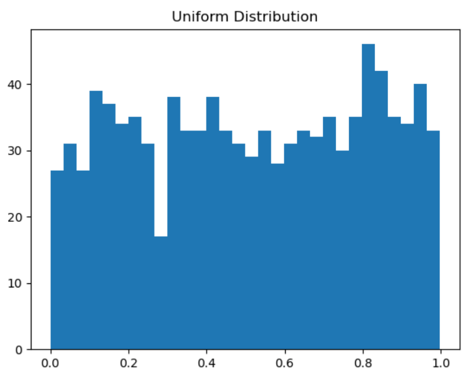
همچنین شرط اینکه همبستگی ۱۰۰درصد باشد، باید ضریب همبستگی برابر ۱ باشد.

$$4c - 2 = 1 \rightarrow 4c = 3 \rightarrow c = \frac{3}{4}$$

۱۰_ پاسخ مسئله شماره ۱۰

ابتدا اعداد تصادفی برای هرکدام از توزیع‌های قسمت اول با کمک کتابخانه numpy تولید شد. سپس برای هر کدام از توزیع‌ها نمودارهای مختلفی رسم شد که در شکل مربوطه به هر توزیع در ادامه مشاهده می‌شود.

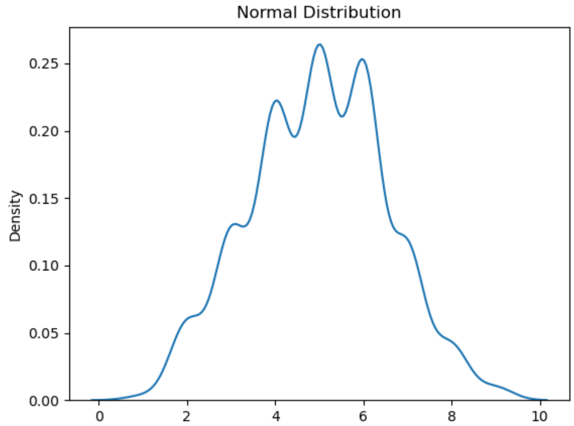
۱-۱۰_ توزیع یکنواخت



شکل ۴ نمودار توزیع یکنواخت

برای رسم این نمودار از کتابخانه Matplotlib استفاده شده و نوع نمودار hist می‌باشد. مشخص است که برای اعداد مختلف مقادیر تقریبا یکسانی تولید شده‌است.

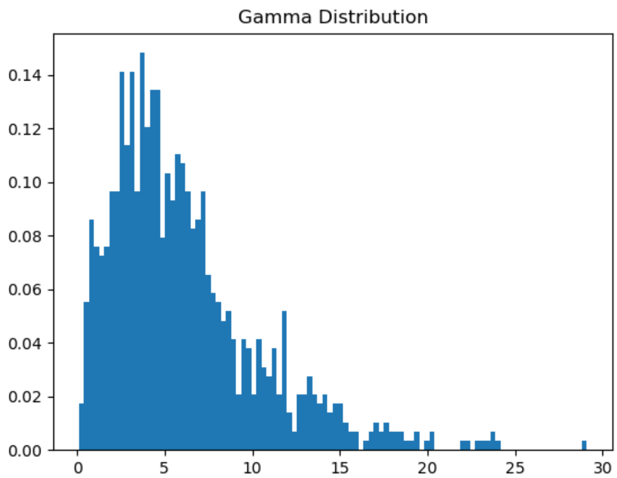
۲-۱۰_ توزیع نرمال



شکل ۵ نمودار توزیع نرمال

برای رسم این نمودار از کتابخانه Seaborn استفاده شده و نوع نمودار KDE می‌باشد. مشخص است که برای اعداد نزدیک به میانگین مقادیر بیشتری تولید شده‌است و شکل کلی نمودار به صورت زنگوله‌ای است.

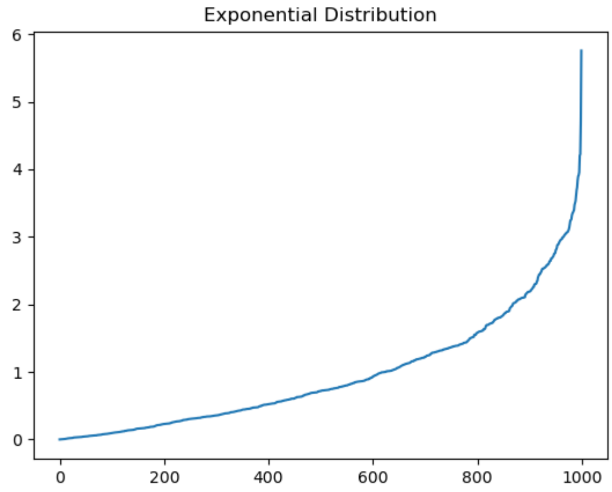
۳-۱۰_ توزیع گاما



شکل ۶نمودار توزیع گاما

برای رسم این نمودار از کتابخانه Matplotlib استفاده شده و نوع نمودار hist می‌باشد.

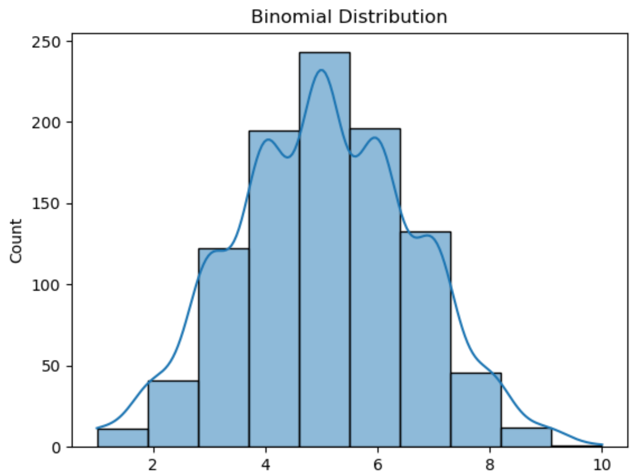
۴-۱۰_ توزیع نمایی



شکل ۷ نمودار توزیع نمایی

برای رسم این نمودار از کتابخانه Matplotlib استفاده شده و نوع نمودار Plot ساده می‌باشد.

۵-۱۰_ توزیع دوجمله‌ای

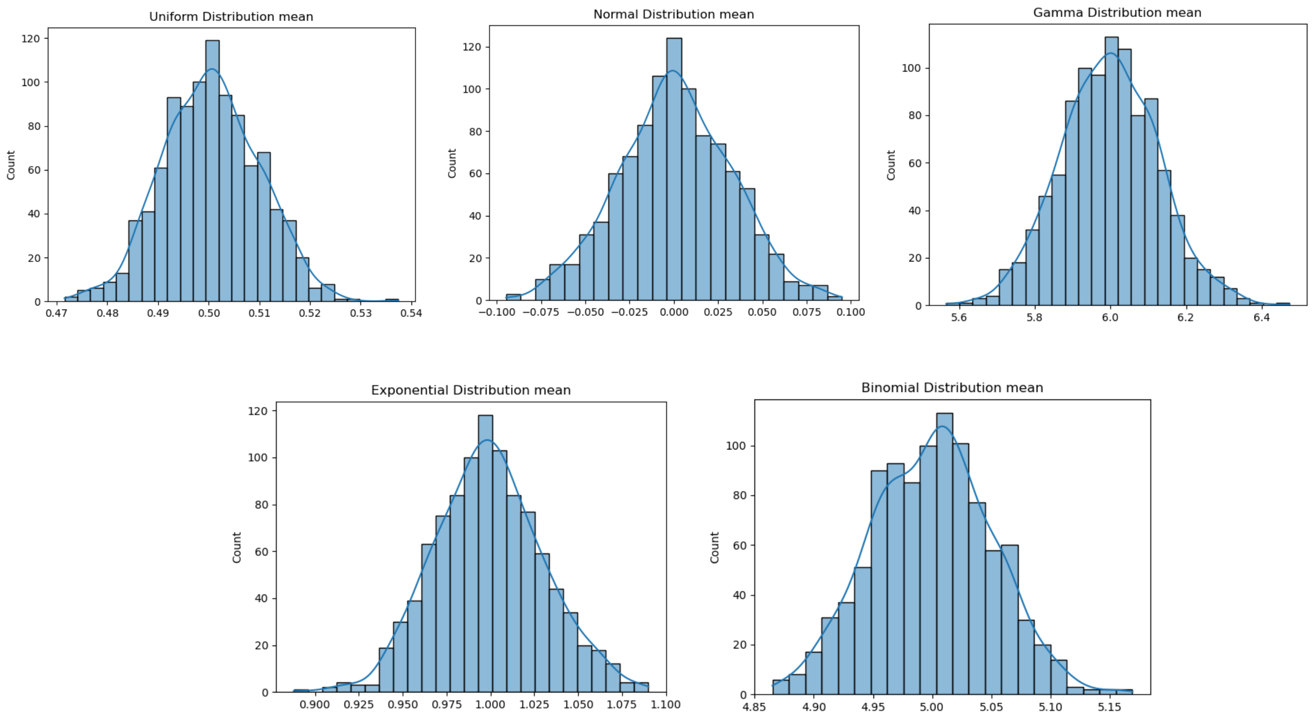


شکل ۸ نمودار توزیع دوجمله‌ای

برای رسم این نمودار از کتابخانه Seaborn استفاده شده و نوع نمودار histplot می‌باشد. Kde نیز برای این نمودار فعال شده‌است.

۱۰-۶_ میانگین توزیع‌های مختلف

در این قســـمت از مســـئله، برای هر کدام از توزیع‌های بخش قبلی با کمک یک حلقه ۱۰۰۰ بار از توزیع مورد نظر میانگین گرفته‌شد و پس از اضافه‌کردن به یک لیست، توزیع آن به کمک نمودار histplot به نمایش درآمد. تصاویر نمودارهای مختلف در ادامه قابل مشاهده هستند که با بررسی آن‌ها متوجه می‌شویم با افزایش تکرار و میانگین‌گیری از هر توزیع شکل توزیع میانگین آن‌ها نرمال است.



شکل ۹ نمودار میانگین توزیع‌های مختلف

۱۰-۷_ بررسی Dataset مسئله

مرحله Data Cleaning

روش‌های مختلفی برای این رویکرد وجود دارد. ابتدا با کمک کتابخانه pandas فایل دیتای مسئله را وارد می‌شوند. سپس از رویکردهای زیر برای Data Cleaning استفاده می‌شود.

- Remove duplicates: ممکن است در دیتاست ورودی برخی مقادیر تکراری باشند. به این منظور داده‌های تکراری را حذف کرده تا در نتیجه‌گیری‌های آماری دقت بالاتری وجود داشته‌باشد.
- Missing values: برخی از مقادیر دیتاست با علامت سوال مشخص شده‌اند. این داده‌ها که نامعلوم هستند باید به فرمت درست جایگذاری شوند. به این منظور از تابع replace استفاده

شـــده و مقدار np.nan که یک اســـتاندارد برای داده نامعلوم در پایتون اســـت جایگذاری می‌شود.
برای محاســـبات مربوط به شـــاخص‌های پراکندگی در ادامه ســـوال، تمامی ردیف‌هایی که قیمت ندارند نیز حذف می‌شوند. همچنین برای اینکه بتوان در برخی سوالات از ستون‌هایی که مقدار nan دارند اســـتفاده کرد، میتوان مقدار nan را با میانگین کل داده‌های آن ســـتون جایگزین کرد.

- Normalization: برای داده‌های مختلف مثل طول، عرض و ارتفاع خودرو در این دیتاســـت، داده‌ها نرمال‌سازی می‌شوند تا در بازه ۰ تا ۱ قرار بگیرند. همچنین فاکتور ریسک بیمه که از -۳ تا ۳ بوده به بازه ۰ تا ۱ نرمال می‌شوند.
- Variable Type: برخی ســـتون‌ها با متغیر مناسب توصـیف نشـده‌اند. برای مثال price باید float یا int باشد در صورتی که object است. پس این ستون به float تبدیل می‌شود.

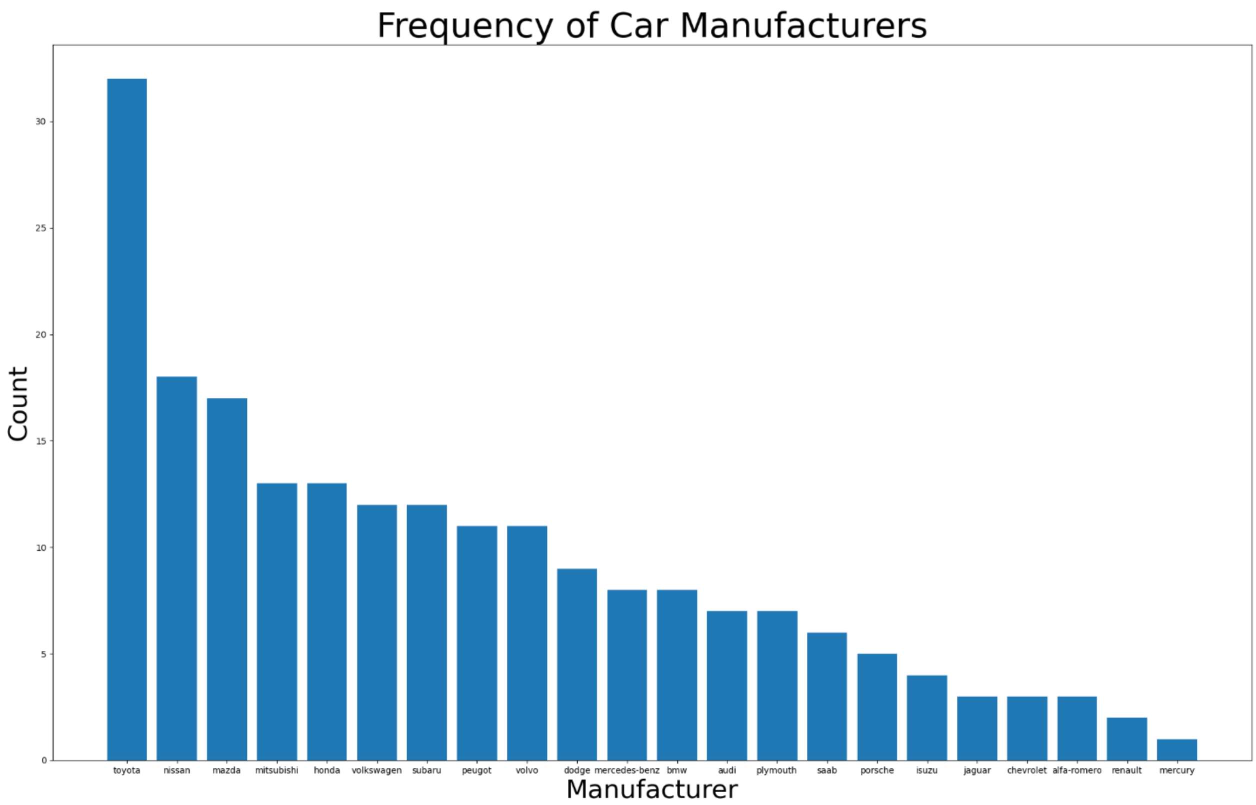
در نهایت از دیتاست پاکسازی شده خروجی گرفته تا در مراحل بعد استفاده شود.

توصیف ستون‌ها

- Symboling : فاکتور ریسک بیمه که با نرمال‌سازی انجام شده به بازه ۰ تا ۱ تبدیل شد.
- normalized-losses: : تفاوت بین فروش تک‌واحدی و عمده - متغیر numerical.
  The car's manufacturer: شرکت تولید کننده خودرو - متغیر categorical.
- fuel-type : نوع سوخت خودرو - متغیر categorical.
- aspiration : توربوشارژ یا سوپرشارژ بودن - متغیر categorical.
- num-of-doors: تعداد درب‌های خودرو - متغیر categorical.
- body-style : استایل بدنه خودرو - متغیر categorical.
- drive-wheels: نوع چرخ های خودرو - متغیر categorical.
- engine-location : محل قرارگیری موتور - متغیر categorical.
- wheel-base: فاصله بین محور جلو و عقب خودو - متغیر numerical.
- length: طول خودرو - متغیر numerical.
- width: عرض خودرو - متغیر numerical.
- height: ارتفاع خودرو - متغیر numerical.
- curb-weight: وزن خودرو بدون احتساب مسافران - متغیر numerical.
- engine-type: نوع موتور خودرو - متغیر categorical.
- num-of-cylinders: تعداد سیلندرهای خودرو - متغیر categorical.
- engine-size: سایز موتور خودرو - متغیر numerical.

- fuel-system: نوع سیستم سوخت‌رسانی خودرو - متغیر categorical.
- bore: اندازه سیلندرها به اینچ - متغیر numerical.
- stroke: اندازه فاصله جابجایی پیستون‌ها- متغیر numerical.
- compression-ratio: نسبت حجم پایین سیلندر باپایستون نسبت به بالامتغیر numerical.
- horsepower stroke: قدرت خودرو به اسب بخار - متغیر numerical.
- peak-rpm stroke: حداکثر دورموتور خودرو - متغیر numerical.
- city-mpg stroke: میزان مصرف سوخت در شهر برحسب هر ۱۰۰مایل- متغیر numerical.
- highway-mpg: میزان مصرف سوخت در جاده برحسب هر ۱۰۰مایل- متغیر numerical..
- price: قیمت خودرو به دلار- متغیر numerical.

داده‌های مختلفی در این دیتاســـت وجود دارند که فاکتورهای مختلفی از یک خودرو را نمایش می‌دهند. اینکه یک داده کاربرد داشــته باشــد یا خیر بســتگی به هدف اســتخراج داده از دیتاســت دارد.



شکل ۱۰ نمودار میزان تولید خودرو شرکت‌های مختلف

همانطور که در نمودار شکل ۱۰ مشخص است، شرکت Toyota بیشترین تولید خودرو را داشته است. برای رسم این نمودار ابتدا تمامی مقادیر شرکت‌های تولید کننده از دیتاست جمع‌آوری شد و سپس با کمک کتابخانه matplotlib و نمودار bar مقادیر استخراج‌شده از دیتاست به نمودار تبدیل شد.

محاسبه پراکندگی، چولگی و کشیدگی

در این قسمت از سوال مشخص نشده که برای کدام یک از متغیرهای دیتاست این مراحل طی شود. برای نمونه این شاخص‌ها برای ستون قیمت بررسی می‌شود. ابتدا ردیف‌هایی که قیمت ندارند حذف می‌شوند. این مرحله در قسمت پاکسازی انجام شد.

شاخص‌های پراکندگی با کمک متدهای کتابخانه pandas به آسانی قابل استخراج هستند. برای نمونه مقادیر خروجی برای ستون price به شرح زیر است.

Range: 40282.0

Variance: 63155863.443184026

Standard Deviation: 7947.066341939271

IQR: 8725.0

Skewness: 1.8096753390980749

Kurtosis: 3.2315368868295193

در این مثال دامنه تفاوت بزرگترین و کوچکترین مقادیر قیمت را خروجی می‌دهد. که به معنی این است که ارزان‌ترین خودرو با گران‌ترین آن ها حدود ۴۰ هزار دلار اختلاف قیمت دارد.

واریانس و انحراف معیار نیز پراکندگی نسبت به میانگین را نشان می‌دهند. انحراف معیار از جنس داده است و در این خروجی خدود ۷ هزار دلار است.

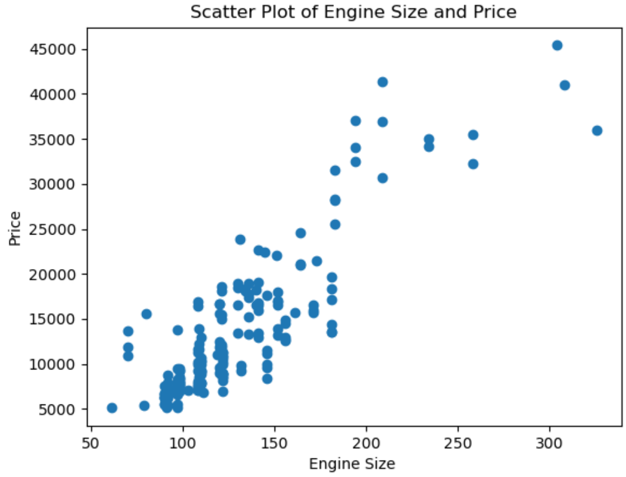
IQR تفاوت بین چارک ۷۵ و ۲۵ است. IQR لندازه‌گیری میزان پراکندگی داده‌ها در اطراف میلنه است.

چولگی نیز نشان می‌دهد داده‌ها به چه سمتی کج شده‌اند که در نمودار توصیف بهتری می‌توان نشان داد. کشیدگی نیز برای اندازه‌گیری پهن بودن یا تخت بودن توزیع است. که مقدار بالاتر یا کمتر از ۳ نسبت این ویژگی را به توزیع نرمال نشان می‌دهد.

۸-۱۰_ همبستگی دو متغیر

برای اینکه همبستگی مقادیر دو ستون از دیتاست بررسی شود از نمودار scatter استفاده می‌شود. خروجی برنامه پایتون به صورت زیر است. به این منظور ابتدا مقادیر ستون قیمت و ستون اندازه موتور

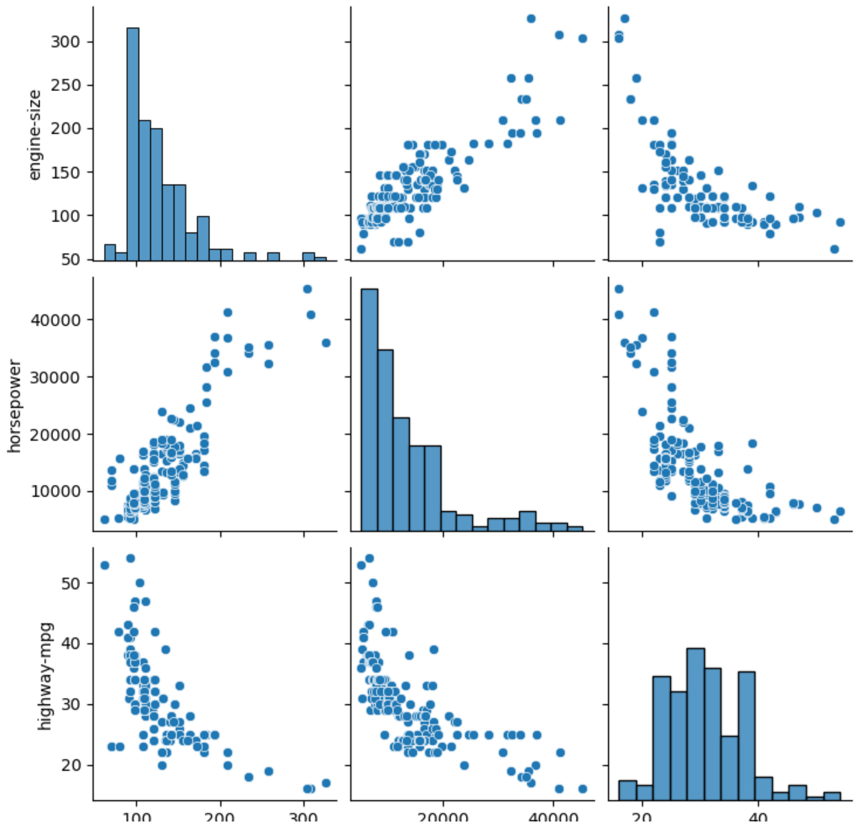
استخراج شده و سپس با کمک متد scatter از کتابخانه matplotlib نمودار تشکیل شده و با کمک متد show خروجی تشکیل می‌شود.



شکل ۱۱نمودار مقایسه همبستگی قیمت خودرو و اندازه موتور

از شکل ۱۱ می‌توان نتیجه گرفت که هرچه اندازه موتور خودرو بالاتر باشد، به احتمال زیاد قیمت آن هم بالاتر خواهدبود.

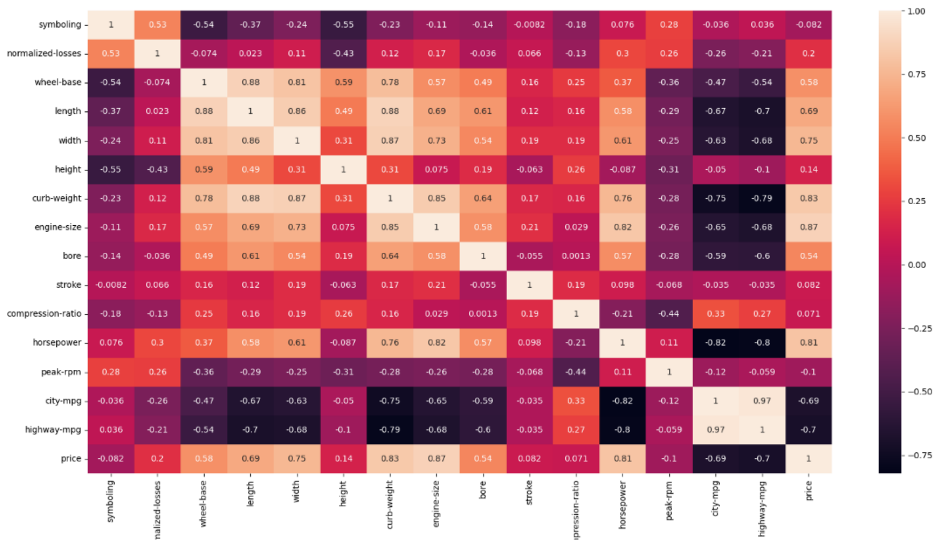
در ادامه pairplot برای ۳ متغیر اندازه موتور، قدرت موتور و مصرف سوخت آن در جاده رسم می‌شود.



شکل ۱۲ نمودار جفتی برای سه فاکتور اندازه موتور، قدرت و مصرف سوخت آن

به این منظور تمام داده‌های مربوط به ۳ متغیر مورد نظر را از دیتاســـت اســـتخراج کرده و با کمک کتبخانه seaborn و متد pairplot نمودار مربوطه تشکیل می‌شود.

برای رســم نمودار heatmap برای تمام داده‌های عددی ابتدا با کمک متد select_dtypes باید تمام ســتون‌هایی که داده عددی دارند انتخاب شــوند. به این منظور از متد np.number اســتفاده می‌شــود. ســپس ســتون‌های این مقادیر به عنوان خروجی برگردانده شده و برای این ســتون ها correlation محاسبه می‌شود. سپس نمودار heatmap از کتابخانه seaborn برای تشکیل نمودار بر روی correlation فراخوانی می‌شود.



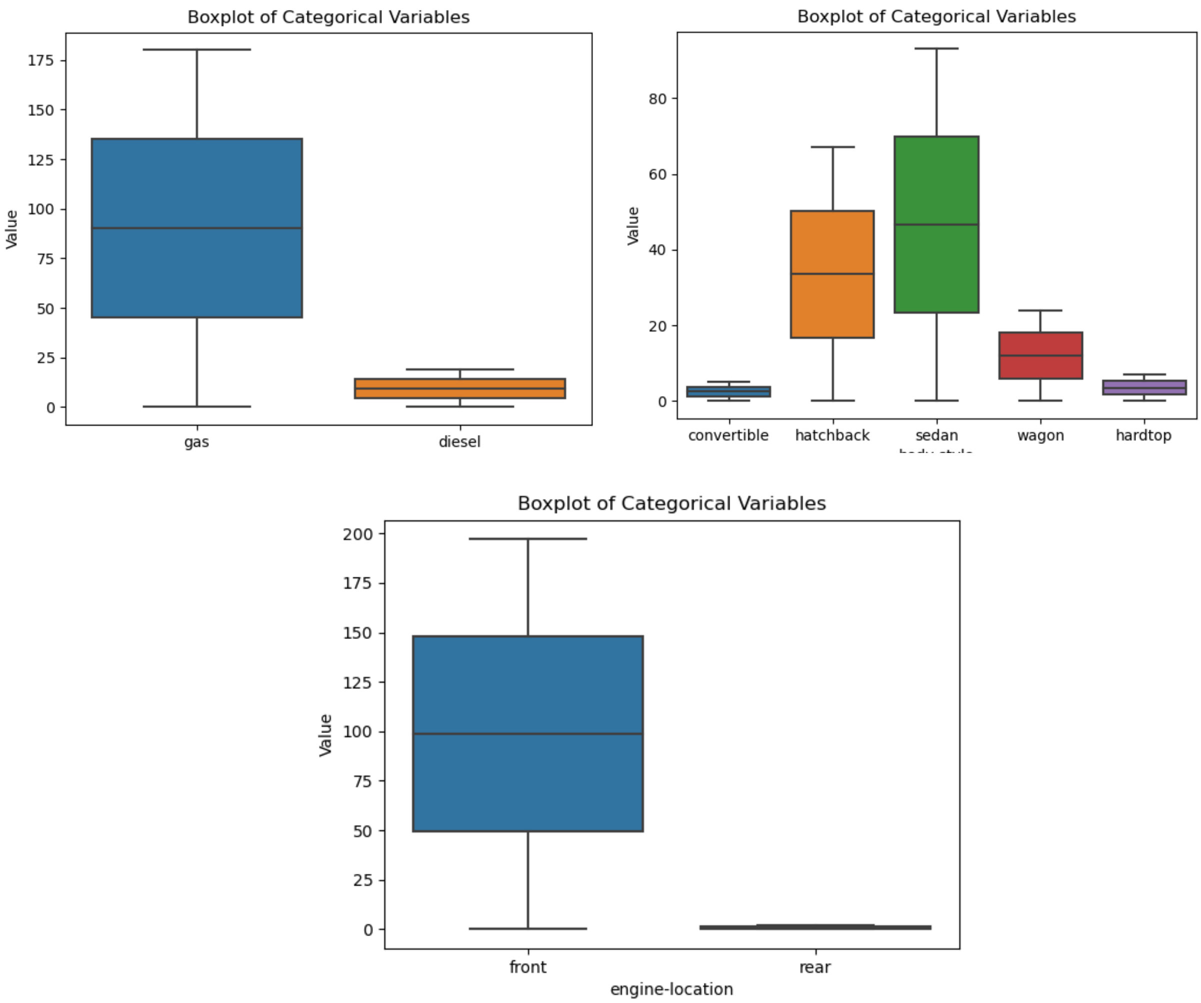
شکل ۱۳ نقشه گرمایی یا heatmap

برای محاســـبه شـــاخص‌های مختلف برای داده‌های categorical ســـه متغیر fuel-type، engine-location و body-style انتخاب می‌شــوند. این ۳ متغیر در یک لیســت قرار داده شــده و ابتدا برای هرکدام مقادیر q1 و q2 محاسبه می‌شود. محاسبه این مقادیر با نحوه محاسبه برای مقایر numerical در مرحله قبل متفاوت است. سپس IQR از تفاضل آن‌ها بدست می‌آید. محاسبه ویسکرها نیز به صورت زیر است:

lower_whisker = q1 - 1.5 * IQR

upper_whisker = q3 + 1.5 * IQR

ســپس مقادیر هرکدام از نوع‌هایی که در هر متغیر وجود دارد شــمرده شــده و با کمک متد boxplot از کتابخانه seaborn نمودار آن تشکیل می‌شود.



شکل ۱۴ نمودار جعبه‌ای برای ۳ متغیر مختلف

۱۱_ پاسخ مسئله شماره ۱۱

باتوجه به تعاریف مطلب درس، متغیر تصـــادفی مانند یک تابع عمل می‌کند که هرکدام از نتایج وقایع یک آزمایش تصادفی را به یک مقدار از اعداد حقیقی نگاشت می‌کند.

- **Population mean** : میانگین کل جامعه یک عدد ثابت اســـت. پس متغیر تصـــادفی محســـوب نمی‌شود.
- **Population size** : اندازه کل جامعه یک مقدار ثابت است. پس متغیر تصادفی محسوب نمی‌شود.
- **Sample size** : اندازه نمونه یک مقدار ثابت است پس متغیر تصادفی محسوب نمی‌شود.

- **Sample mean** : میانگین نمونه برابر است با محاسبه میانگین یک سری از نمونه‌های تصادفی که مقدار آن محاسبه می‌شود. بر اساس اینکه چه نمونه‌هایی از جامعه انتخاب شوند میانگین آن‌ها متغیر خواهدبود پس این مورد یک متغیر تصادفی است.
- **Variance of the sample mean** : واریانس میانگین نمونه بیانگر میزان پراکندگی میانگین‌های نمونه‌های مختلف یک جامعه بوده مقدار ثابتی است. پس متغیر تصادفی محسوب نمی‌شود.
- **The largest value in the sample** : بزرگترین مقدار در نمونه بستگی به مقادیر موجود در نمونه انتخاب شده‌دارد و متغیر است. پس متغیر تصادفی محسوب می‌شود.
- **Population variance** : واریانس جامعه یک مقدار ثابتی است و متغیر تصادفی محسوب نمی‌شود.

۱۲_ پاسخ مسئله شماره ۱۲

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:

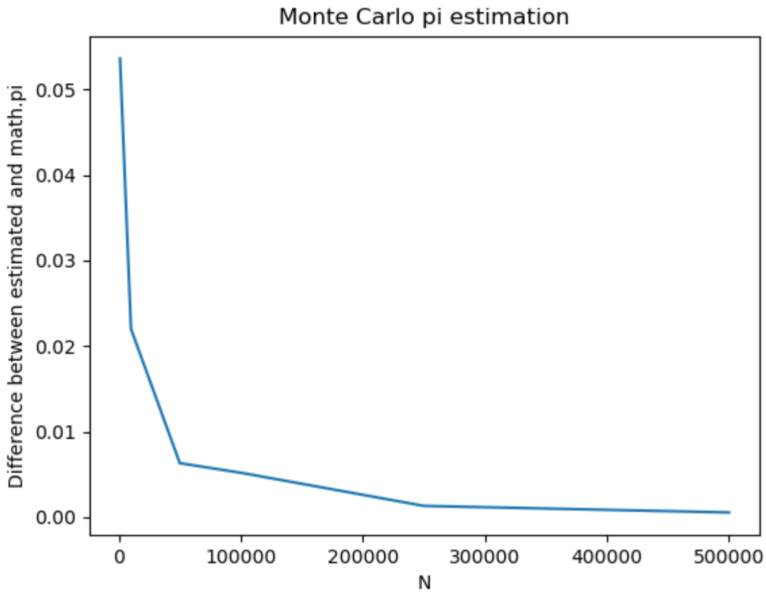
$$\frac{Sc}{Ss} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} \rightarrow \pi = \frac{4Sc}{Ss}$$

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

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

مراحل بالا با کمک پایتون پیاده‌سازی شد. فرض شد که r=1 پس مقدار مقادیر مربع x و y در صورت کمتر از ۱ بودن داخل دایره هستند.

به این منظور یک متد برای تخمین از روش بالا تعریف شد. در برنامه اصلی برای N های مختلف آزمایش تکرار شد.پس از شبیه‌سازی با مقادیر مختلف N مقدار عدد π تخمین زده‌شد. نتایج اختلاف به شرح زیر است.



شکل ۱۵ نتایج تخمین عدد پی به روش مونت کارلو

۱۳_ پاسخ مسئله شماره ۱۳

مسئله The Gambler's Ruin نشان می‌دهد که در یک شبیه‌سازی شرطبندی برای احتمال بدست‌آوردن یا از دست دادن سرمایه اولیه در اینجا تعریف شده‌است. برای شبیه سازی این مسئله یک متد با کمک برنامه‌نویسی پایتون تعریف شده که با دریافت ورودی‌های مربوط به تعداد دور شرطبندی، مقدار سرمایه اولیه، مقدار شرطبندی، مقدار هدف و همچنین احتمال برد، بررسی می‌کند که هربار با یک احتمال بر اساس ورودی چه میزان سود و یا ضرر روی سرمایه اولیه اعمال می‌شود.

در صورت اتمام سرمایه و یا رسیدن به هدف این تابع مقدار لیست مربوط به رهگیری تمام مقادیر سرمایه در طول شرطبندی را برمی‌گرداند. برهمین اساس آخرین خانه این لیست، آخرین سرمایه شرطبندی بوده و یا بیشترمساوی هدف و یا کمترمساوی ۰ خواهدبود. حالت سوم هم زمانی است که تعداد دورهای بازی تمام شده‌است و سرمایه در آخرین دور در این خانه قابل دسترسی است.

با شبیه‌سازی بازی بر روی مقادیر زیر و احتمال ۰.۵ پس از ۱۰۰۰ بار شبیه‌سازی مقادیر زیر از خروجی دریافت شد.

```
initial_stake = 1500
bet_amount = 200
win_probability = 0.5
target_amount = 3000
N_rounds = 100
N_simulations = 1000
```

خروجی آخرین دور:

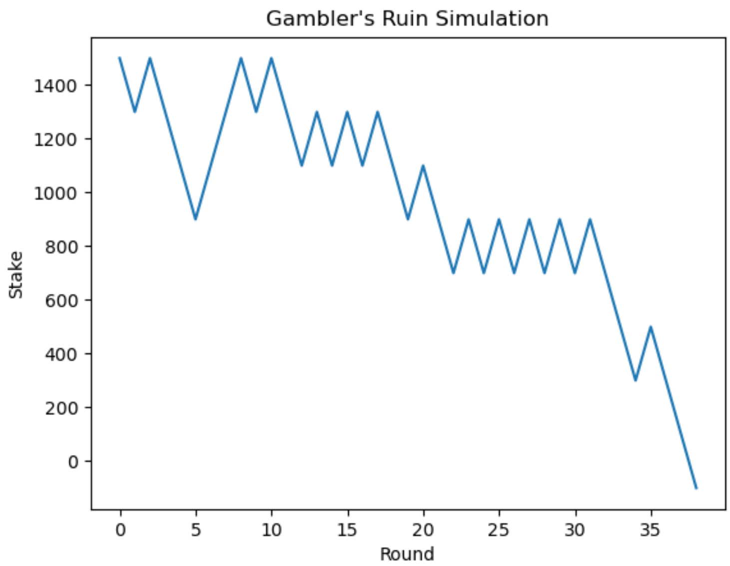
Tracking = [1500, 1300, 1500, 1300, 1100, 900, 1100, 1300, 1500, 1300, 1500, 1300, 1100, 1300, 1100, 1300, 1100, 1300, 1100, 900, 1100, 900, 700, 900, 700, 900, 700, 900, 700, 900, 700, 900, 700, 500, 300, 500, 300, 100, -100]

Final Stake= -100

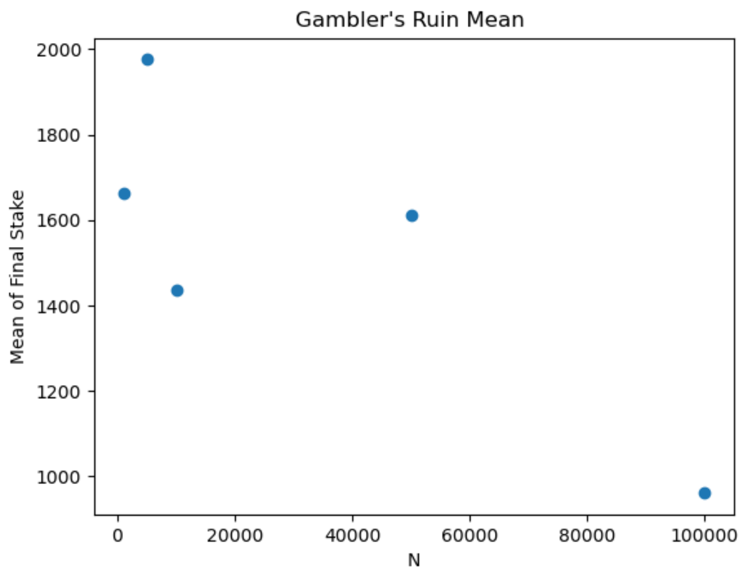
خروجی نتیجه‌گیری از تمام دورها:

Probability of Win= 0.425

همچنین برای رهگیری انجام شده نمودار زیر رسم شده‌است.



شکل ۱۷ رهگیری سرمایه در طول شرط‌بندی



شکل ۱۶ مقادیر میانگین سرمایه در شبیه‌سازی‌های مختلف

این آزمایش برای N های مختلف اجرا شـــد که نتیجه بررســـی میانگین‌های مختلف درشـــکل ۱۶ مقادیر میانگین سرمایه در شبیه‌سازی‌های مختلف قابل مشاهده است.

۱۴_ منابع

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition*, *11*(2), 109-122.