به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

آمار و احتمال مهندسی

گزارش پروژه‌ی نهائی

نام و نام خانوادگی	سید علیرضا جاوید
شماره دانشجویی	810198375
تاریخ ارسال گزارش	1399/10/24

دی 1399

فهرست

سوال 1 – ضریب اویلر–ماسکرونی

(آ) می خواهیم عدد بزرگ n بر عدد تصادفی r تقسیم کنیم پس **باید متغیر تصادفی خود را r انتخاب کنیم** از آنجایی که می دانیم $r \leq n$ است و دارای توزیع یکنواخت است داریم :

$$P_R(r) = \frac{1}{n} \qquad (1) \qquad\qquad \epsilon_r = \left\lceil \frac{n}{r} \right\rceil - \frac{n}{r} \qquad (2)$$

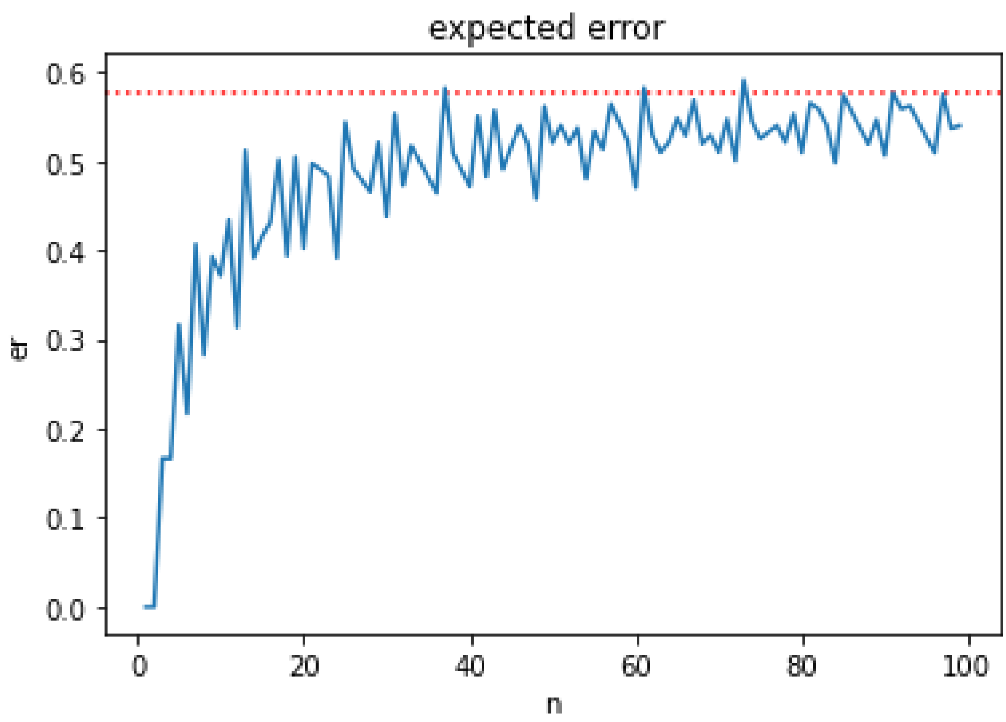
ϵ_r را نیز طبق فرض سوال نوشتیم.

(ب) حال می خواهیم امید ریاضی ϵ_r را برای n بزرگ حساب کنیم :

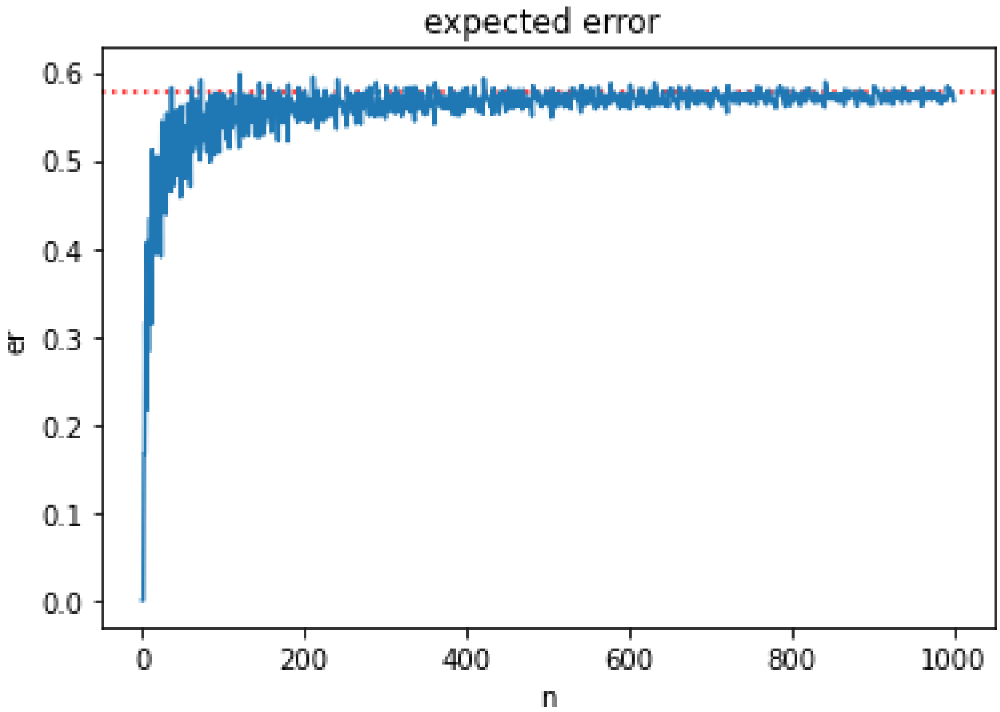
$$\lim_{n\to\infty} E(\varepsilon_r) = \lim_{n\to\infty} \sum_{r=1}^{n} \varepsilon_r \times P_R(r) = \lim_{n\to\infty} \sum_{r=1}^{n} \left(\left\lceil \frac{n}{r} \right\rceil - \frac{n}{r} \right) \times \frac{1}{n} \qquad (3)$$

سعی میکنیم به کمک کد پایتون همگرایی رابطه 3 به ثابت اویلر-ماسکرونی نشان دهیم.

به ازای 2 مقدار n=100 و n=1000 نمودار های زیر را بدست می آوریم با توجه به شکل های زیر نتیجه میشود که مقدار امید خطا به کندی به مقدار تقریبی 0.5772 یا همان ثابت اویلر-ماسکرونی همگرا می شود :



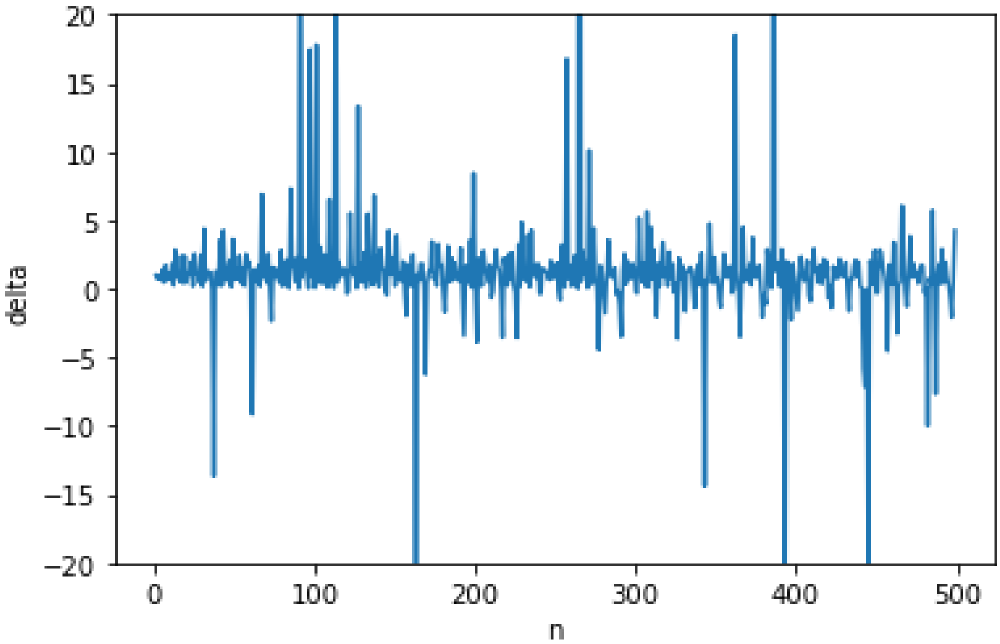
شکل **1-a :** به ازای **n=100** همچنین ثابت اویلر-ماسکرونی با خط چین نشان داده شده است



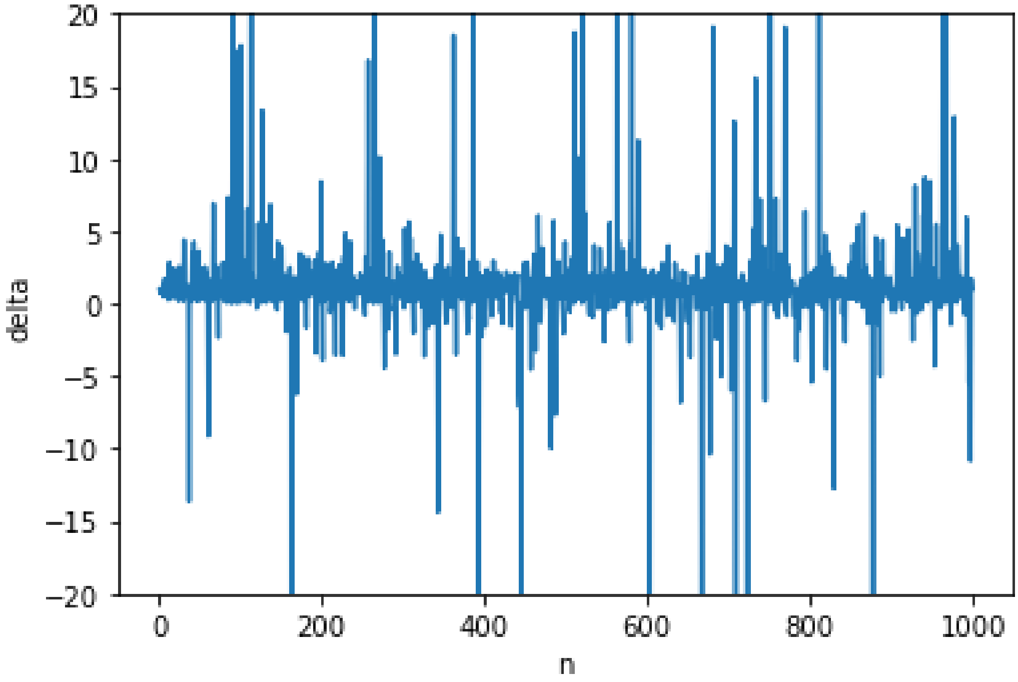
شکل **1-b :** به ازای **n=1000** همچنین ثابت اویلر-ماسکرونی با خط چین نشان داده شده است

حال شبیه سازی رابطه 4 را برای n=100 و n=1000 تکرار می کنیم :

$$\delta_n = \frac{\hat{\varepsilon}_r(n+1)-\gamma}{\hat{\varepsilon}_r(n)-\gamma} \qquad (4)$$



شکل **2-a :** نمودار $\boldsymbol{\delta_n}$ به ازای **n=100**



شکل **2-b :** نمودار δ_n به ازای **n=1000**

سوال 2 – مساله ی نیوتن–پیپس

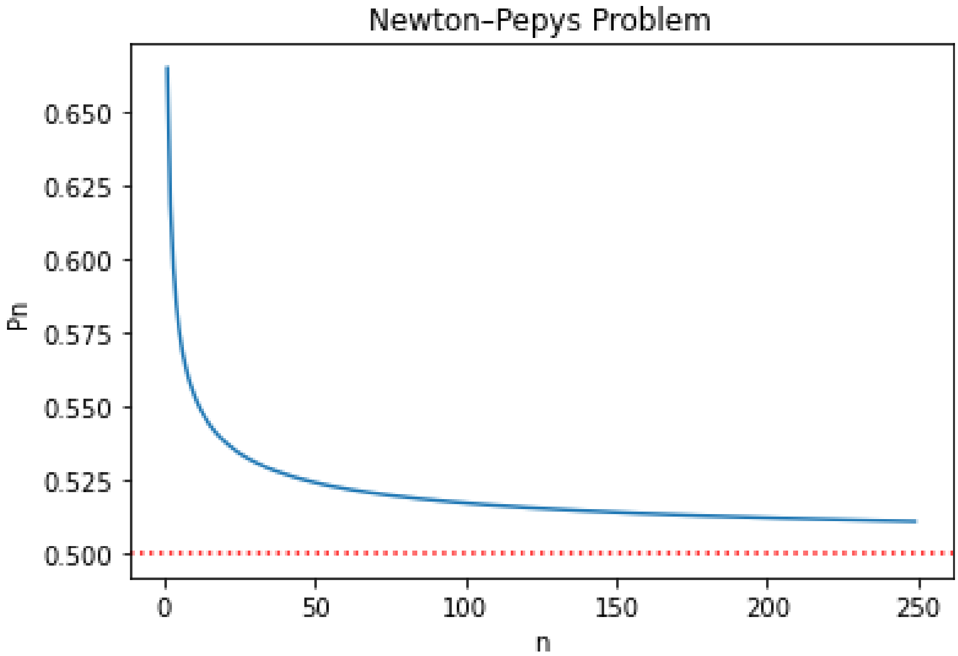
(آ) نیوتن در استدلال خود احتمال آنکه در یک گروه بیش از یک 6 بیاید را لحاظ نکرده است. برای تحلیل این مسئله می توان گفت در پرتاب 6 تاس وقتی که تنها یک تاس بین 1 تا 5 نیاید شرط مورد نظر براورده شده اما در پرتاب 12 تاس علاوه بر آنکه 12 تاس بین 1 تا 5 بیاید اگر یکی از 12 تاس نیز 6 بیاید باز شرط مورد نظر برآورد نشده و همین طور برای پرتاب 18 تاس که حالت 2 بار 6 نیز مطلوب نیست . پس بطور شهودی می توان گفت که با بیشتر شدن تاس ها تعداد حالات نامطلوب نیز بیشتر می شود و احتمالات حداقل 6 کمتر می شود که با نتایج بخش بعدی نیز سازگاری دارد.

(ب) پرتاب تاس به ازای متغیر تصادفی n دارای توزیع 2 جمله ای $Bin\left(6n\,,\ \frac{1}{6}\right)$ است حال بوسیله آن P_n را می یابیم :

$$P_{Bin} = \binom{6n}{i} \times (\tfrac{1}{6})^{i} \times (\tfrac{5}{6})^{n-i} \rightarrow$$

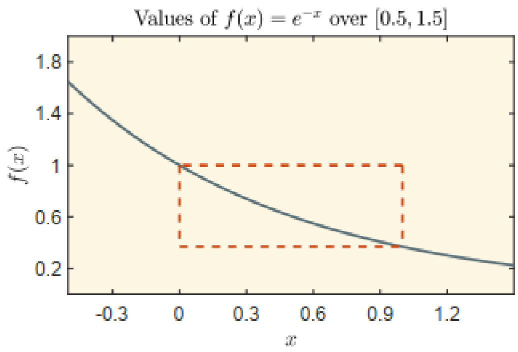
$$P_n\,(\text{n}) = 1 - \sum_{i=0}^{n-1} \binom{6n}{i} \times (\tfrac{1}{6})^{i} \times (\tfrac{5}{6})^{n-i} \qquad (5)$$

مطابق رابطه 5 شبیه سازی را انجام داده و با بزرگ کردن n همگرایی آن را به 0.5 در نمودار نشان می دهیم :



شکل 3 : نمودار همگرایی **Pn** به 0.5

سوال 3 – تخمین عدد نپرین با روش مونته کارلو



شکل 4 : منحنی تابع e^{-x}

آ)

$$\text{مساحت زیر منحنی در مستطیل} = \int_0^1 \int_{e^{-1}}^{e^{-x}} dy\, dx = 1 - 2e^{-1} \qquad (5)$$

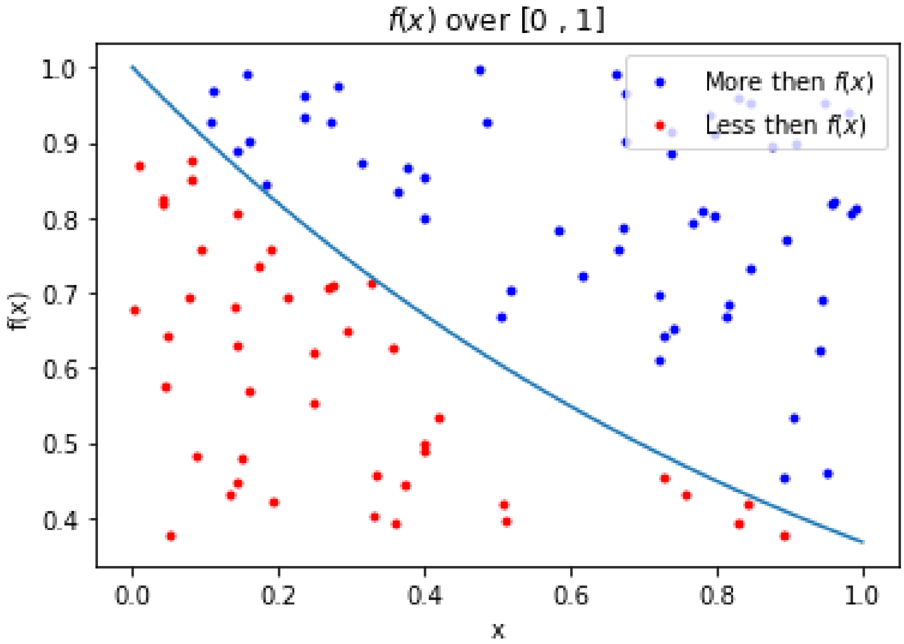
ب) از رابطه 5 استفاده می کنیم :

$$S = 1 - 2e^{-1} \rightarrow e = \frac{2}{1-S} \qquad (6)$$

ج) از رابطه 5 و نسبت داده شده استفاده میکنیم :

$$\lim_{n\to\infty} \frac{n_{red}}{n} = \frac{S}{S_{rect}} = \frac{1-2e^{-1}}{1-e^{-1}} = \alpha \quad \rightarrow \quad \hat{e} = \frac{2-\alpha}{1-\alpha} \qquad (7)$$

با استفاده از رابطه 7 در پایتون بدست می آوریم :



شکل 5 : پخش شدن نمونه ها روی فضا

$$\hat{e} = 2.785714$$

د) در تخمین بیشینه درست نمایی[1] برای توزیع $N(\mu,\sigma^2)$ داریم :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (7) \qquad\qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (8)$$

با فرض 40 نمونه تصادفی مشاهده شده به کمک روابط 7 و 8 و کد پایتون بدست می آوریم :

$$\hat{\mu} = \frac{1}{40}\sum_{i=1}^{40} x_i = 2.735041$$

$$\hat{\sigma}^2 = \frac{1}{40}\sum_{i=1}^{40}(x_i - 2.735041\,)^2 = 0.025887$$

پس برای n=100 توزیع تخمینی ما $N(2.735041,0.025887)$ است .

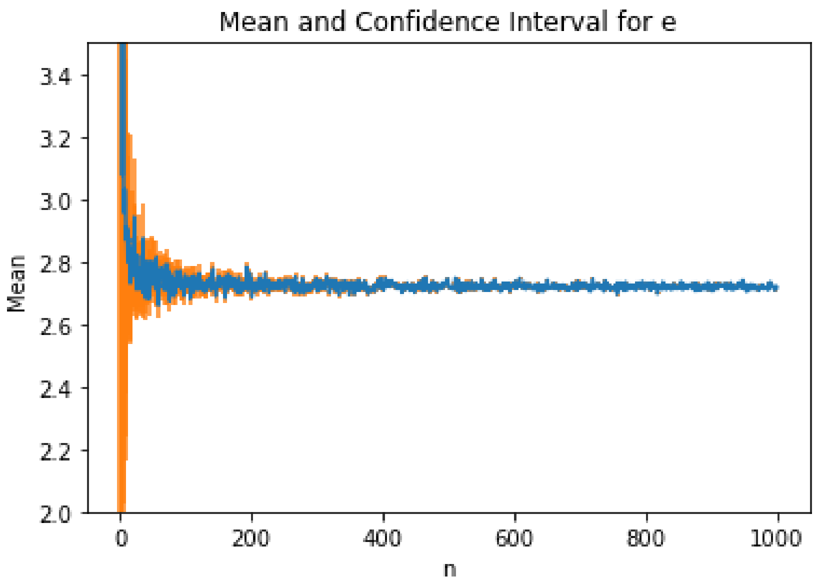
ه) از آنجایی که واریانس نمونه های ما مشخص نیست پس برای تعیین بازه اطمینان[2] برای میانگین داریم:

$$\left(\bar{X} - \frac{S}{\sqrt{n}}z_{1-\frac{\alpha}{2}}\,,\bar{X} + \frac{S}{\sqrt{n}}z_{1-\frac{\alpha}{2}}\,\right) \qquad (9)$$

که در آن S^2 واریانس نمونه است . همچنین به کمک جدول مقادیر تابع گوسی و با توجه با اینکه بازه اطمینان 95 درصد مطلوب است و $\alpha = 0.05$ است پس داریم :

$$z_{1-\frac{\alpha}{2}} = \ z_{0.975} = 1.96$$

مطابق رابطه 9 می توان دید به ازای n های بزرگ عدم قطعیت در بازه اطمینان به 0 میل می کند چون واریانس به تقریب مقدار ثابتی دارد, با رسم نمودار مقادیر امید ریاضی را برحسب تعداد نمونه ها و بازه اطمینان 95 درصد برای هر نمونه این فرضیه تایید می شود : (البته این نمودار برای $n > 30$ معتبر است)

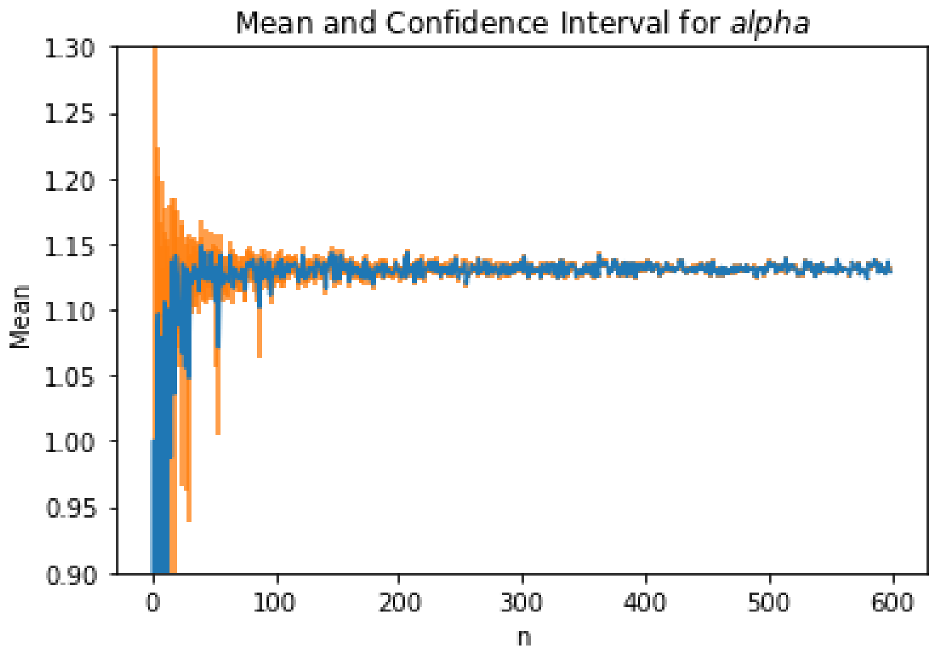


شکل 6 : نمودار مقادیر امید ریاضی را برحسب تعداد نمونه ها و بازه اطمینان 95 درصد برای هر نمونه

[1] Maximum Likelihood Estimation (MLE)
[2] Confidence Interval

سوال 4 – سری فیبوناچی تصادفی

آ) بازه ی اطمینان 95 درصدی را مطابق فرمول (9) که در بخش 3 ارائه شد برای α مشخص می کنیم :



شکل 7 : نمودار میانگین α را برحسب **n** و بازه اطمینان 95 درصد برای هر داده

$\hat{\alpha} = 1.128627$

که با مقدار تئوری 1.1319882 همخوانی مناسبی دارد .

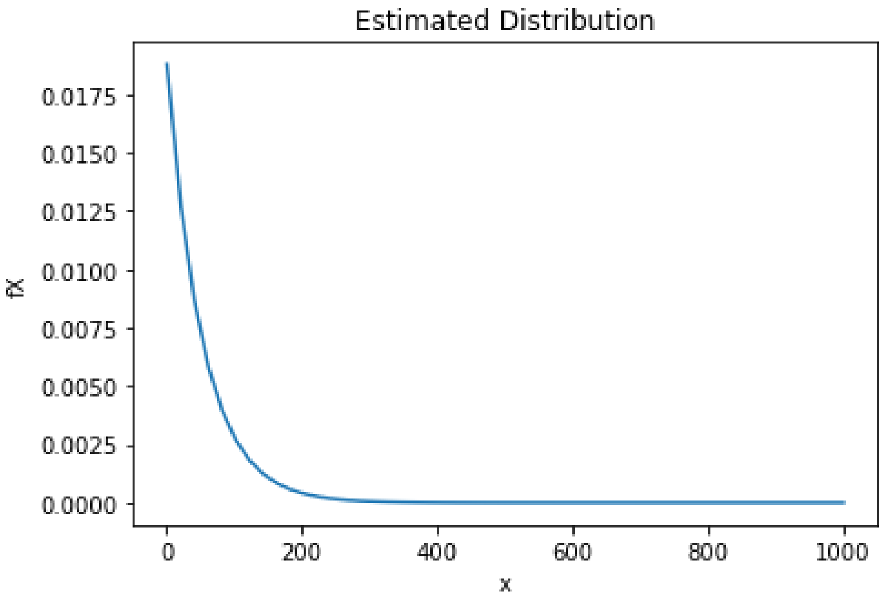
ب)

$$f(x_i|\lambda) = \lambda e^{-\lambda x_i} \qquad\qquad f\left(x_{1,2,\dots,n}\middle|\lambda\right) = \lambda^n \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

$$LL(\lambda) = \sum_{i=1}^{n} \ln\left(\lambda^n \times \lambda e^{-\lambda x_i}\right) = \mathrm{n} \times \ln(\lambda) - \lambda \times \sum_{i=1}^{n} x_i$$

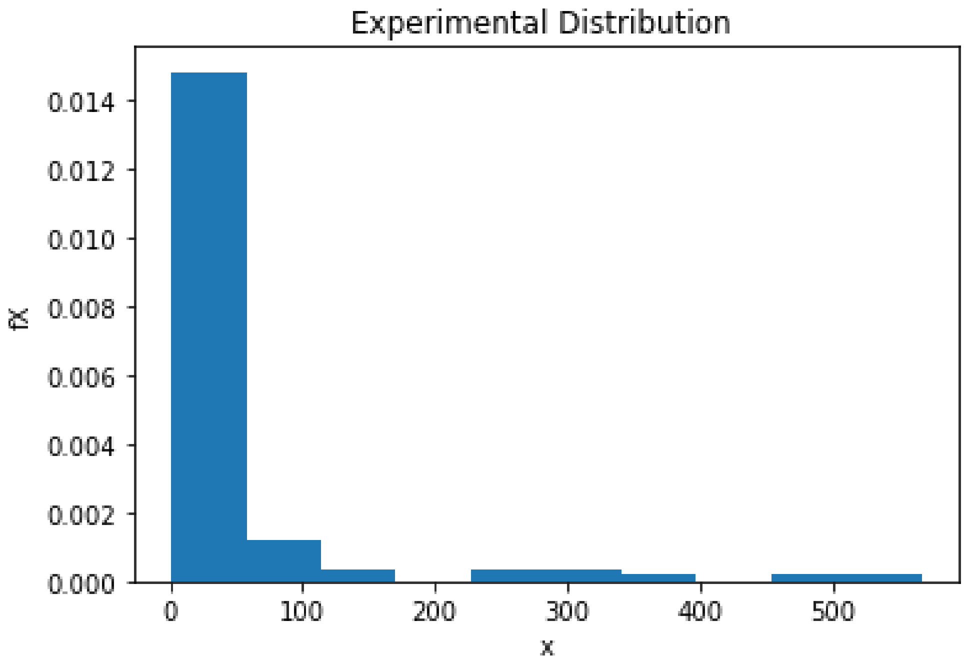
$$\frac{\partial LL}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \qquad \rightarrow \qquad \lambda = \frac{n}{\sum_{i=1}^{n} x_i} \qquad (10)$$

با قرار دادن رابطه (10) با 40 نمونه برای $f_{Max}(25)$ در پایتون و بدست آوردن λ , توزیع نمایی با پارامتر $\lambda = 0.021441$ بدست می آید :



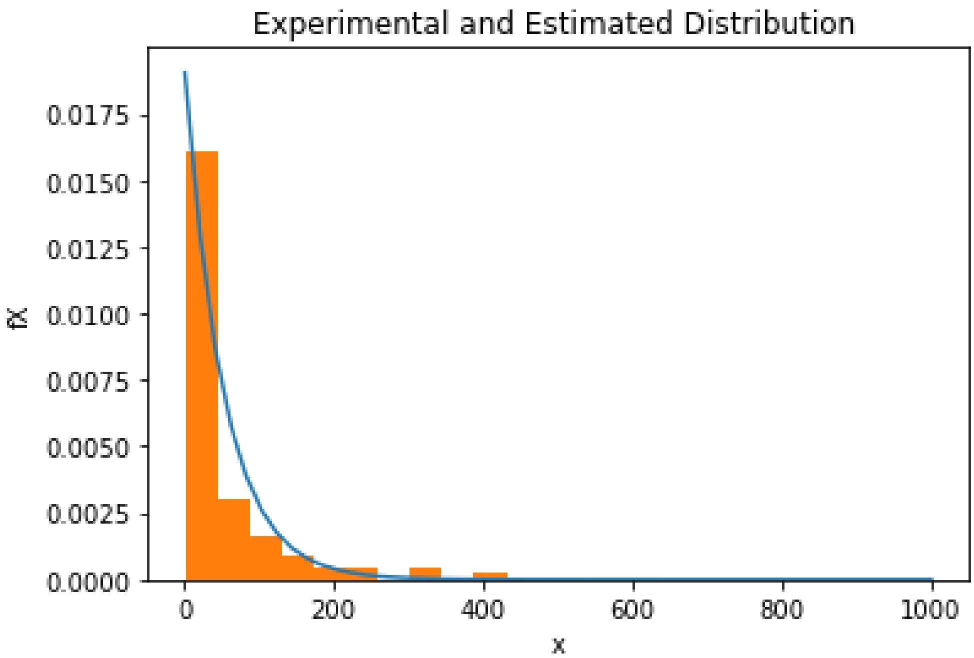
شکل 8 : نمودار توزیع تخمین زده شده برای $f_{Max}(25)$

ج) به کمک نمودار هیستوگرام می توان تخمین تابع چگالی براساس هیستوگرام توزیع بصورت شکل زیر نشان داد .



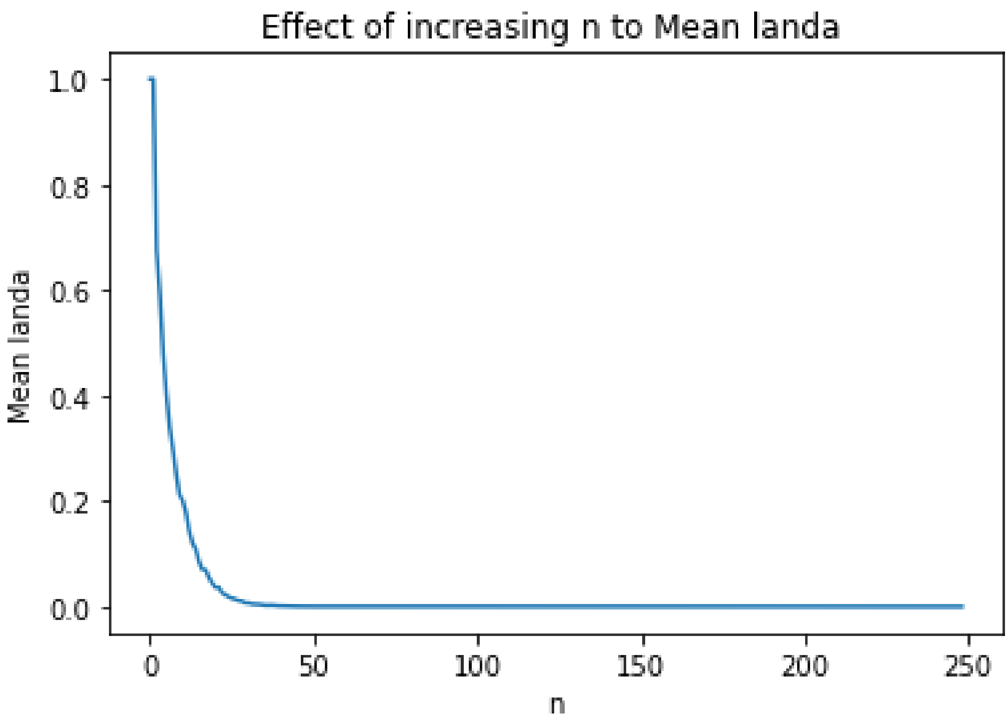
شکل 9 : نمودار برازش تابع چگالی تجربی داده ها برای $f_{Max}(25)$

همچنین برای نشان دادن این تابع چگالی تجربی به همراه تابع تخمین زده شده در قسمت قبل داریم :



شکل 10 : نمودار برازش تابع چگالی تجربی و تخمین زده شده داده ها برای $f_{Max}(25)$

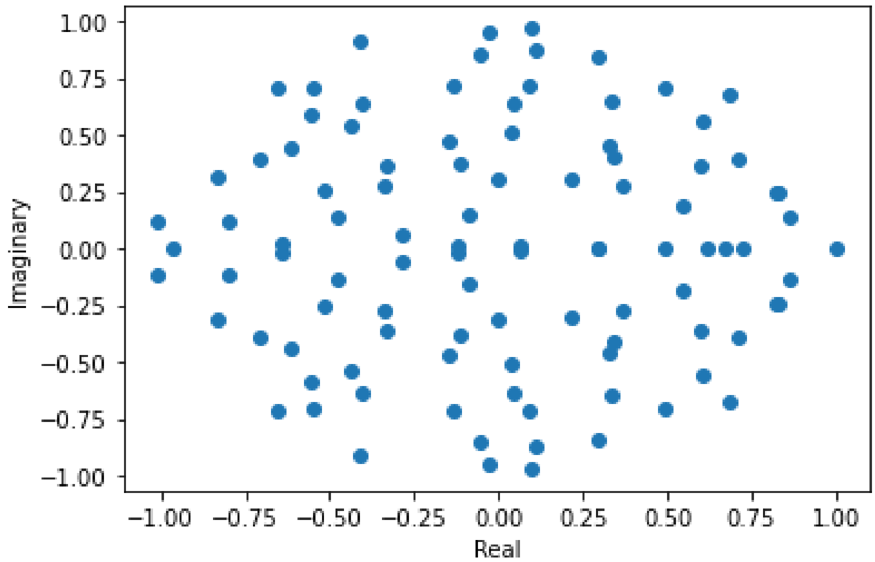
د) با توجه به شکل 11 که اثر افزایش n را بر امید ریاضی λ نشان میدهد می توان گفت که با افزایش n پارامتر تخمین یا همان لاندا به 0 میل می کند و از نظر شهودی با افزایش n مقدار max تابع با سرعت بیشتری افزایش یافته و همچنین با توجه به رابطه 10 در بخش ج نتیجه گرفته شده صحیح بنظر می رسد.



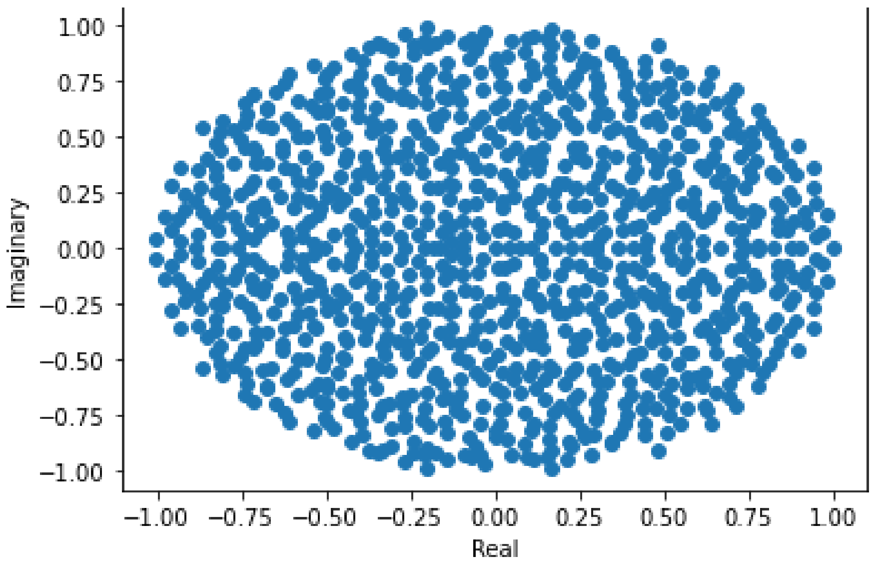
شکل 11 : نمودار اثر افزایش **n** بر امید ریاضی **λ**

سوال 5 – قاعده ی دایره ای برای مقادیر ویژه

آ) در مرتبه اول برای روی توزیع $N\left(0,\sqrt{\frac{1}{n}}\right)$ به ازای n = 100 و n = 1000 نمایش می دهیم :

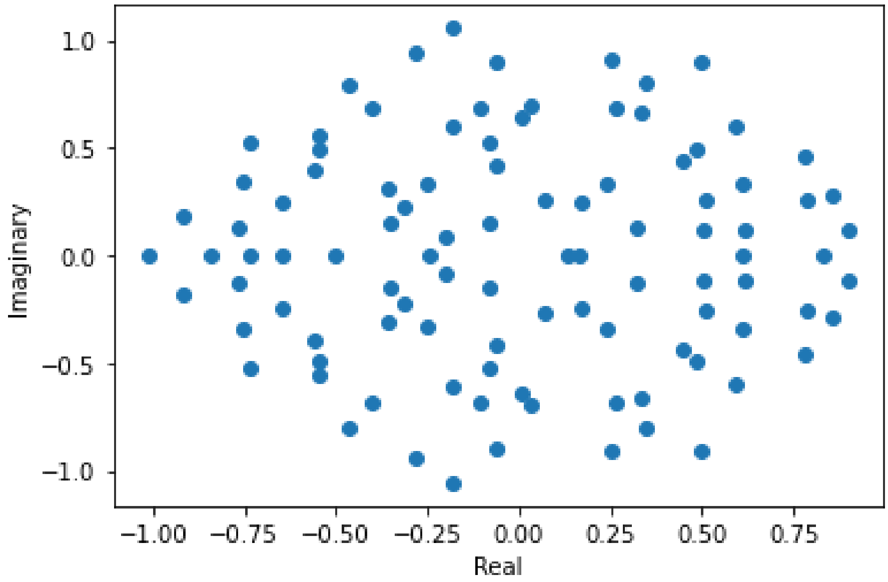


شکل **12-a** : توزیع مقادیر ویژه به ازای **n = 100** برای ماتریس نرمال

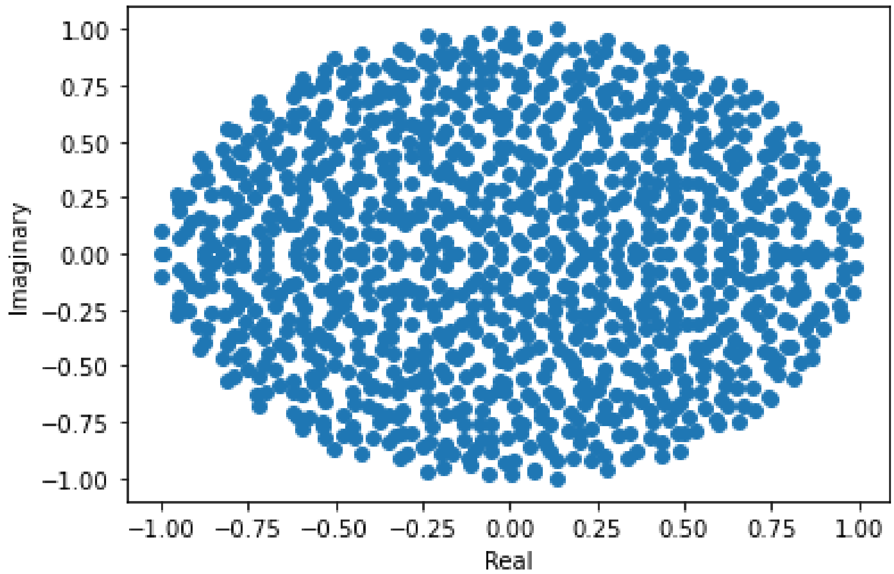


شکل **12-b** : توزیع مقادیر ویژه به ازای **n = 1000** برای ماتریس نرمال

بطور مشابه برای توزیع $U\left(-\sqrt{\frac{3}{n}},\sqrt{\frac{3}{n}}\right)$ با میانگین 0 و واریانس $\frac{1}{n}$ داریم :



شکل **13-a** : توزیع مقادیر ویژه به ازای **n = 100** برای ماتریس یونیفرم

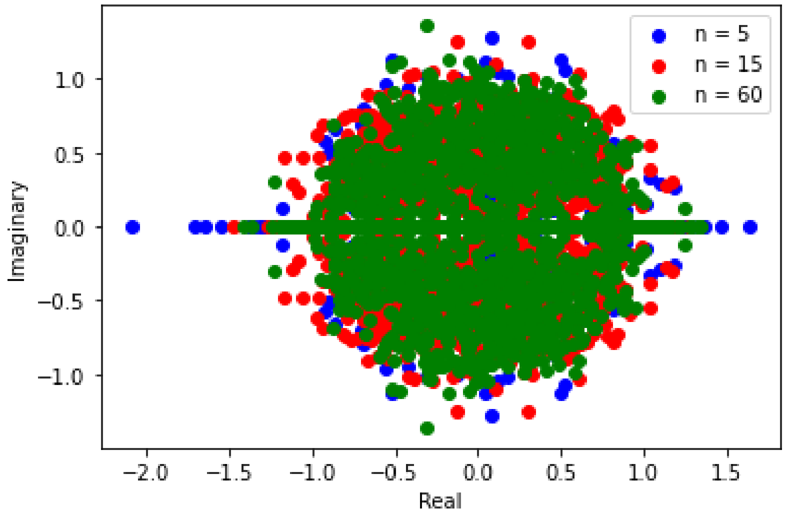


شکل **13-b** : توزیع مقادیر ویژه به ازای **n = 1000** برای ماتریس یونیفرم

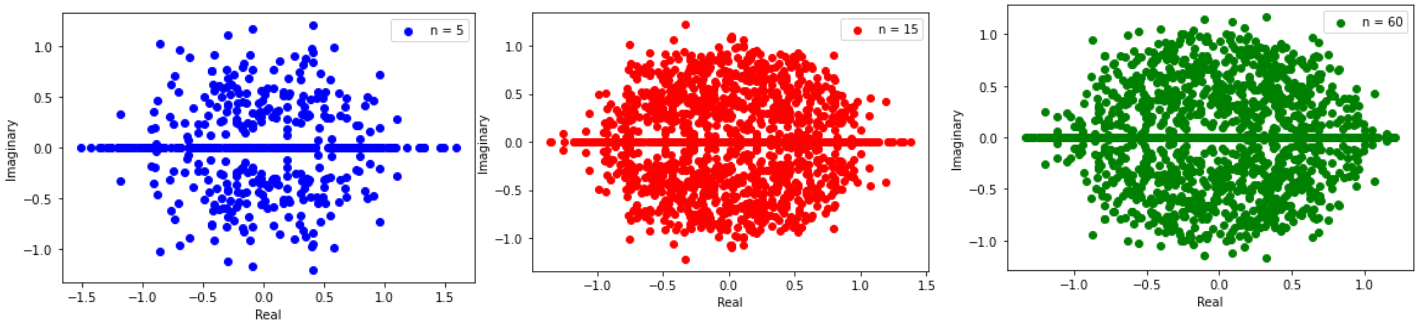
از شکل های بالا می توان درستی قاعده دایره برای مقادیر ویژه را نتیجه گیری کرد.

ب)

مطابق شکل به با نمونه گیری از 100 ماتریس با n=60 , n=15 , n=5 توزیع مقادیر ویژه بصورت یکنواخت حول محور y توزیع شده که اثبات آنرا نیز ارائه می کنیم :



شکل **14** : توزیع متقارن نمونه گیری شده 100 ماتریس نرمال با **n** های متفاوت



شکل **15** : 3 توزیع بصورت شکل های جدا برای شفافیت بیشتر

اثبات توزیع یکنواخت :

Suppose all Aij are real and λ = a + jb

→ AV = λV → $\overline{AV} = \overline{\lambda V}$ → $A\bar{V} = \bar{\lambda}\bar{V}$

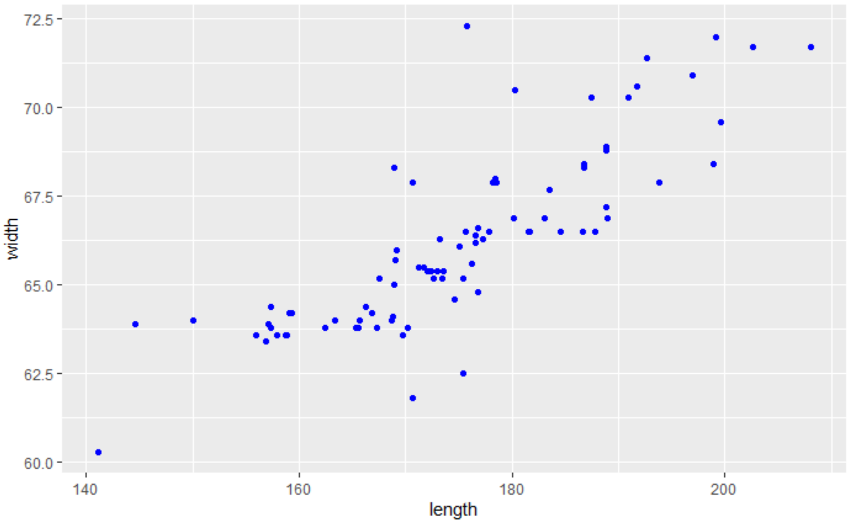
→$\bar{\lambda}$ ∈ Eigevalue of matrix A →

λ = a ± jb are both Eigevalues of matrix A

سوال 6 – داده بازی

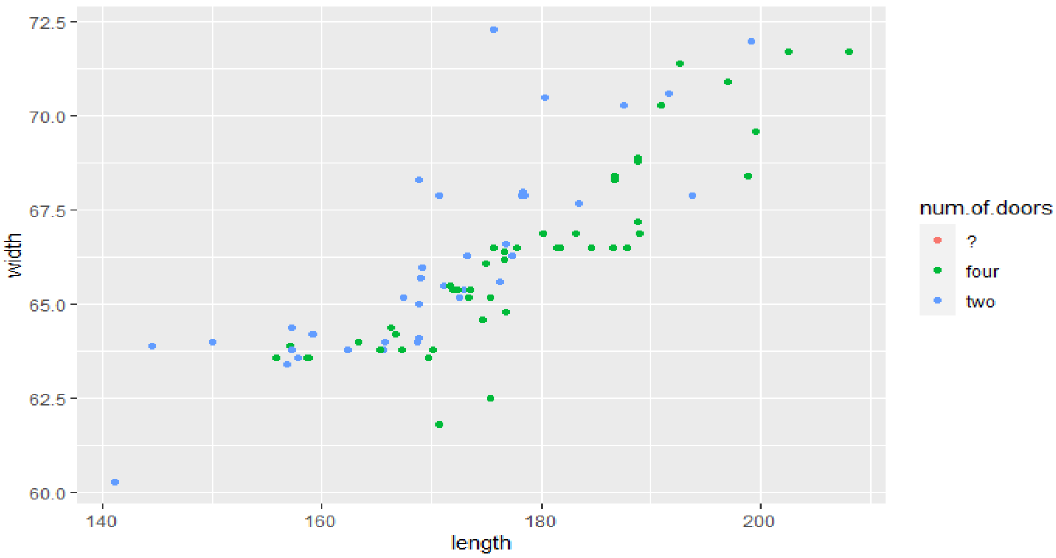
آ) با رسم نمودار هر بخش در مورد فرض های داده شده بحث می کنیم :

i) با توجه به شکل 16 می توان گفت کوواریانس[1] میان بلند بودن و عریض بودن ماشین مثبت است و تقریبا این گزاره صحیح است .



شکل **16** : نمودار پراکندگی طول و عرض ماشین

ii) پراکندگی خودروهای چهار در گوشه سمت راست بیشتر است اما مشخص است که تعداد قابل توجهی خودرو دو در نیز موجود است . با تقریب می توان گفت این گزاره صحیح است.

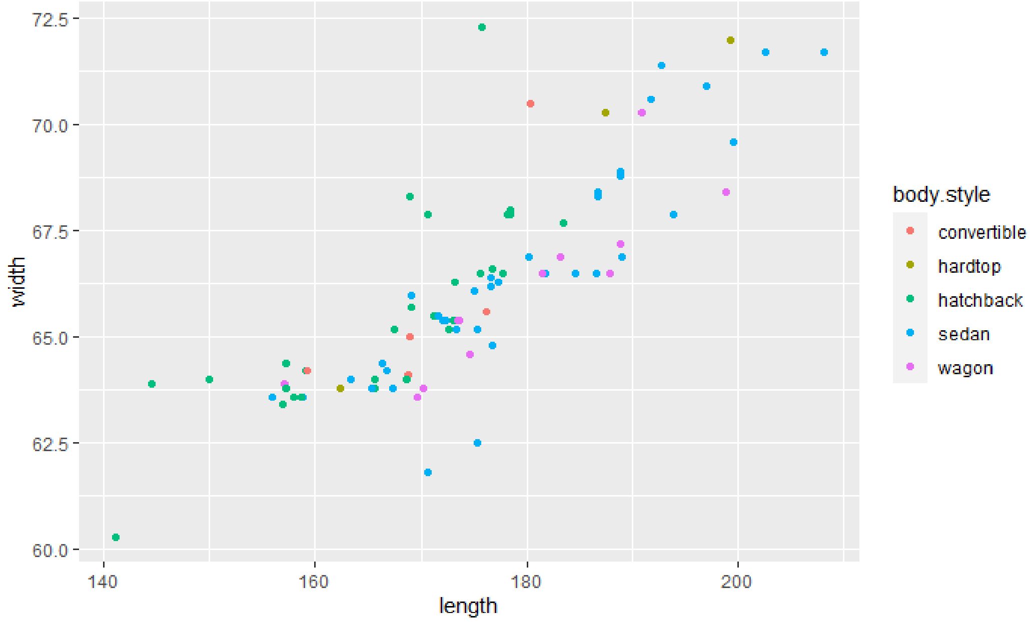


شکل **17** : نمودار پراکندگی طول و عرض ماشین همراه مشخص کردن تعداد در های نمونه

---

[1] Covariance

iii) با دقت در شکل 18 مشخص است که پراکندگی این داده ها در میانه های جدول بیشتر است و این عبارت نادرست است .



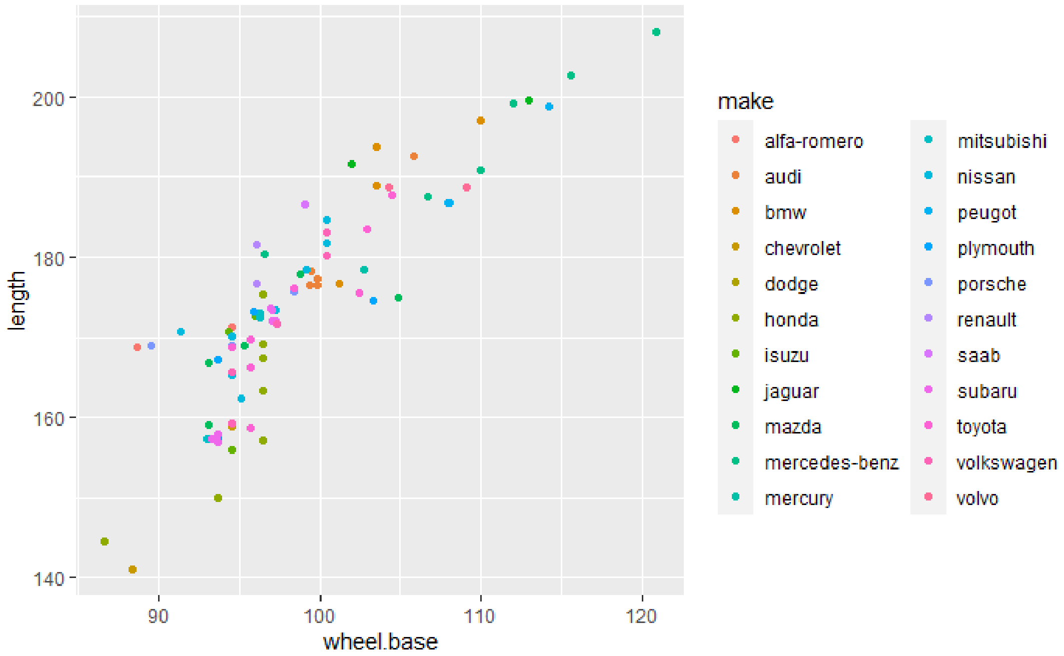
شکل **18** : نمودار پراکندگی طول و عرض ماشین همراه مشخص کردن نوع بدنه خودرو

iv)تقریبا برای تمام انواع موتور در نمودار پراکندگی با افزایش عرض خودرو ها سایز موتور آنان نیز افزایش میابد پس این گزاره درست است.

شکل **19** : نمودار پراکندگی ظرفیت موتور خودرو و عرض ماشین همراه مشخص کردن نوع موتور

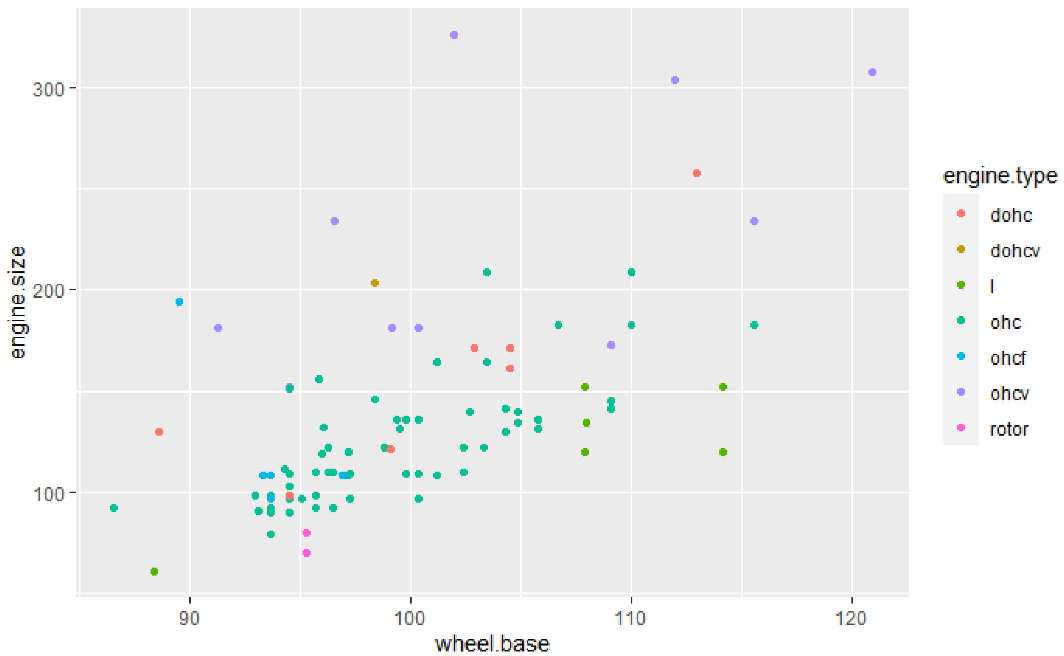
ب) 3 فرض را مطابق زیر مطرح کرده و با تحلیل داده ها صحت آنها را بررسی می کنیم :

i) طول هر خودرو با فاصله بین 2 چرخ رابطه مستقیم دارد : مطابق شکل 20 این گزاره صحیح است .



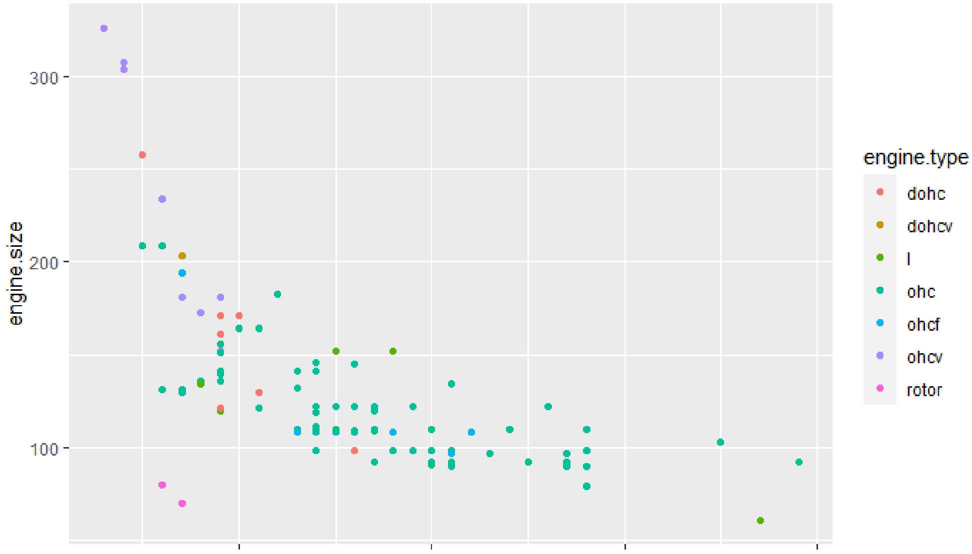
شکل **20** : نمودار پراکندگی طول خودرو و فاصله میان دو چرخ همراه مشخص کردن شرکت سازنده

ii) فاصله بین دو چرخ با اندازه موتور خودرو رابطه مستقیم دارد : مطابق شکل 21 برای موتور های نوع ohc رابطه نسبتا خطی وجود دارد اما برای دیگر موتور ها داده های کمی داریم و پراکندگی آنها نیز ما را به نتیجه درستی هدایت نمی کند .



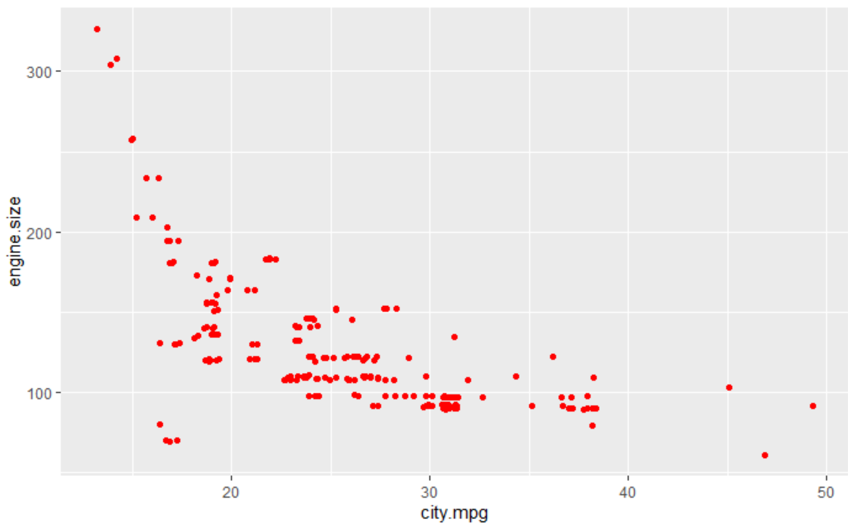
شکل **21** : نمودار پراکندگی اندازه موتور خودرو و فاصله میان دو چرخ همراه مشخص کردن نوع موتور

iii) میزان شاخص MPG[1] در شهر با اندازه موتور نسبت عکس دارد زیرا حدس میزنیم بازدهی موتور های کوچکتر بیشتر است : مطابق نمودار پراکندگی بطور کلی این عبارت صحیح است .



شکل **22** : نمودار پراکندگی اندازه موتور خودرو و شاخص **MPG** همراه مشخص کردن نوع موتور

ج) لغزش یا jitter در نمودار نویز هایی به نمودار ما اضافه می کند که موجب می شود تصمیم گیری و پیش بینی ما برای داده های جدید یا داده هایی که اکنون مقادیر آن را نداریم بهتر شود و مفاهیم دریافت شده از نمودار های پراکندگی را بهبود می دهد. برای شهود بیشتر آخرین نمودار (شاخص MPG شهر و اندازه موتور) را با این مفهوم بهبود می دهیم :

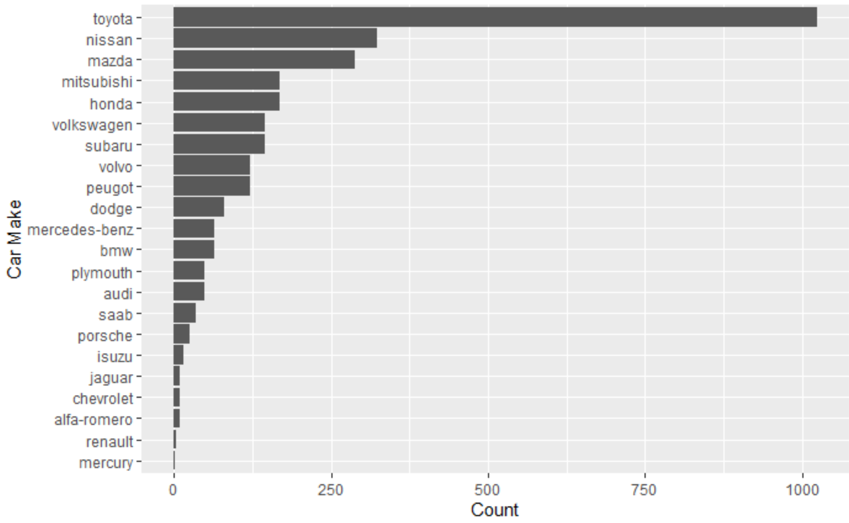


شکل **23** : نمودار پراکندگی **jitter** اندازه موتور خودرو و شاخص **MPG**

[1] Miles per Gallon

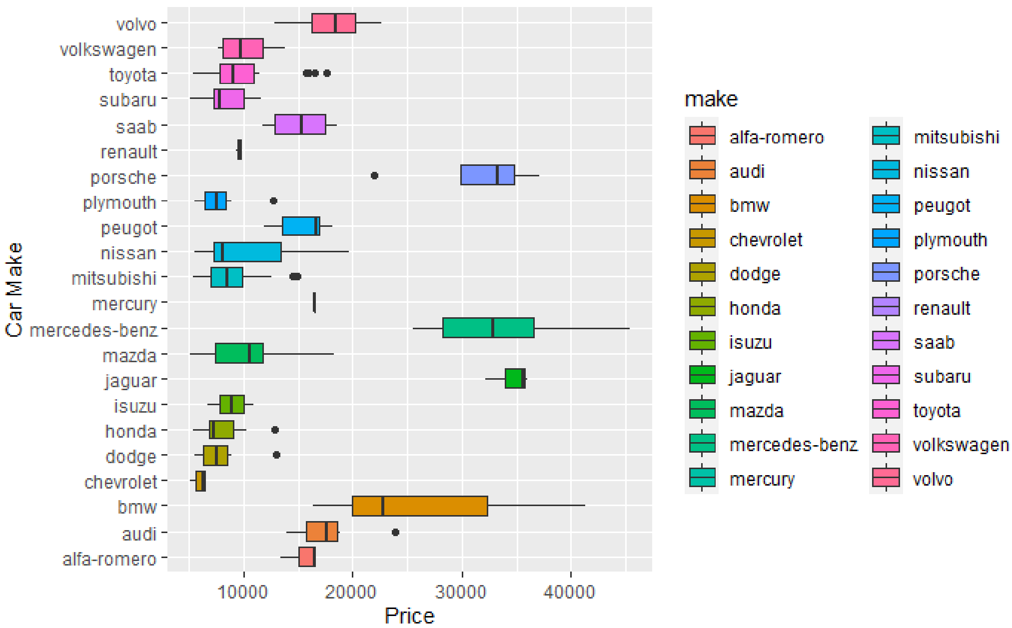
د) 6 نمودار مطابق زیر برای آنالیز داده ها انتخاب کرده ایم و سپس نتیجه گیری های خود را با توجه به آنها ارائه می دهیم .

i) از نمودار میله ای شکل 24 نتیجه می شود داده های ما برای ماشین های Toyota بیشتر است پس استنباط های آماری ما برای این نمونه از بقیه معتبر تر می باشد .



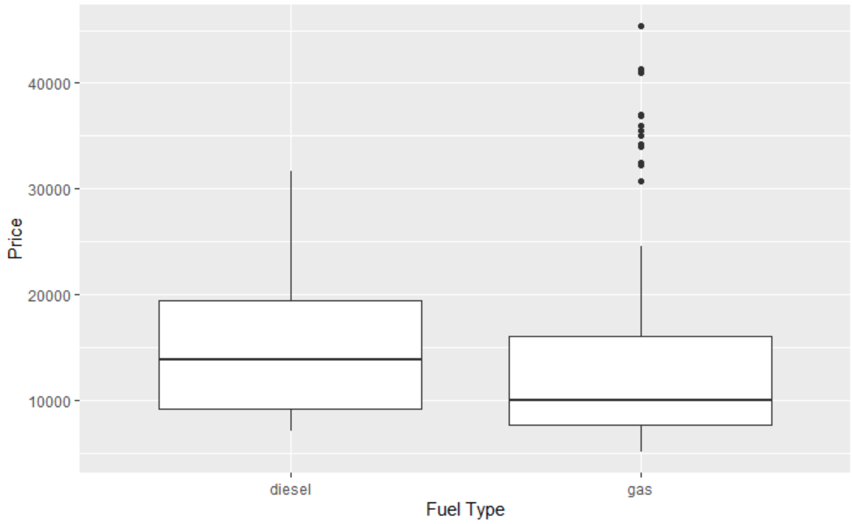
شکل **24** : نمودار میله ای فراوانی خودرو های کمپانی ها در نمونه

ii) به وسیله نمودار جعبه ای شکل 25 گستردگی و تمرکز قیمت خودرو های ساخته شده توسط شرکت های مختلف چگونه است که با دقت می توان دید که خودرو های کمپانی های jaguar و Mercedes Benz و Porsche گران قیمت تر هستند و همچنین خودرو های BMW رنج قیمت های بیشتری دارند.



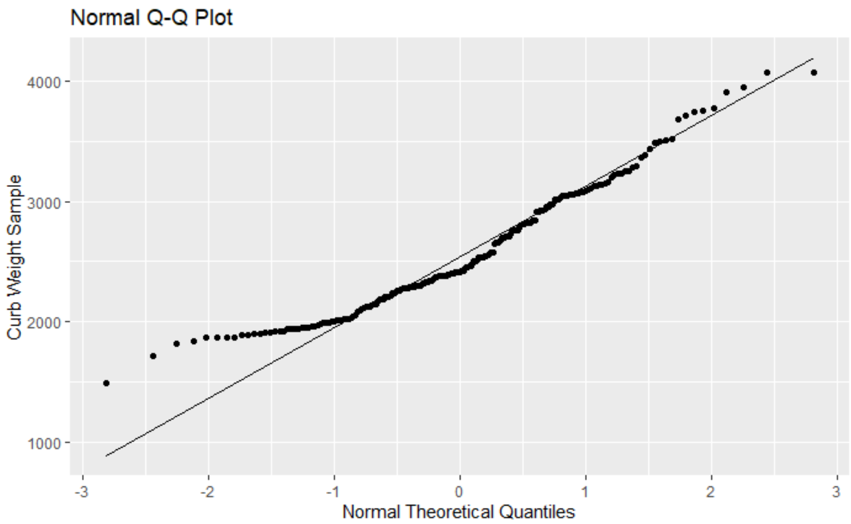
شکل **25** : نمودار جعبه ای قیمت و کمپانی سازنده خودرو

iii) به وسیله نمودار جعبه ای شکل 26 گستردگی و تمرکز قیمت خودروها بر اساس نوع سوخت آنها مشخص شده است . این نمودار به خوبی مشخص می کند که خودرو های دیزلی میانگین قیمت بالاتری دارند . این نشان می دهد که چرا خودرو های گازسوز فروش بیشتری دارند .



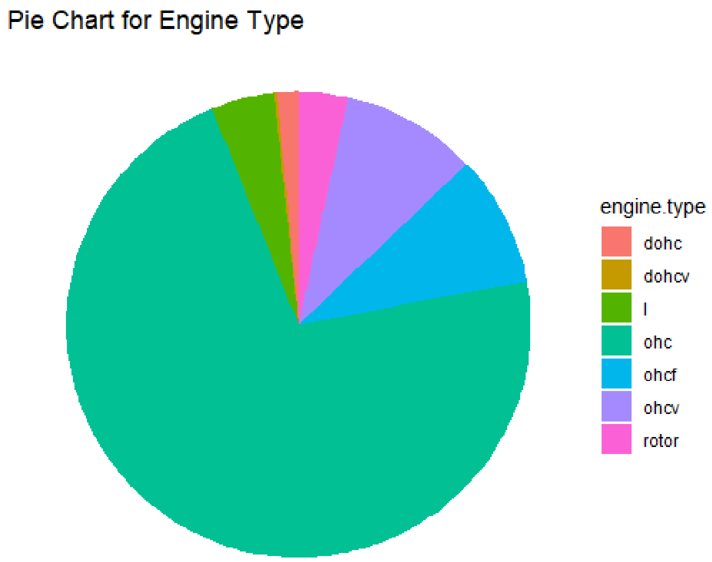
شکل **26** : نمودار جعبه ای قیمت و نوع سوخت

iv) Q-Q Plot را برای جرم کل یا Curb weight رسم میکنیم. مطابق شکل 27 با تقریب خوبی می توان گفت که توزیع Curb weight نرمال است که بسیار نکته مهم و جالبی است .



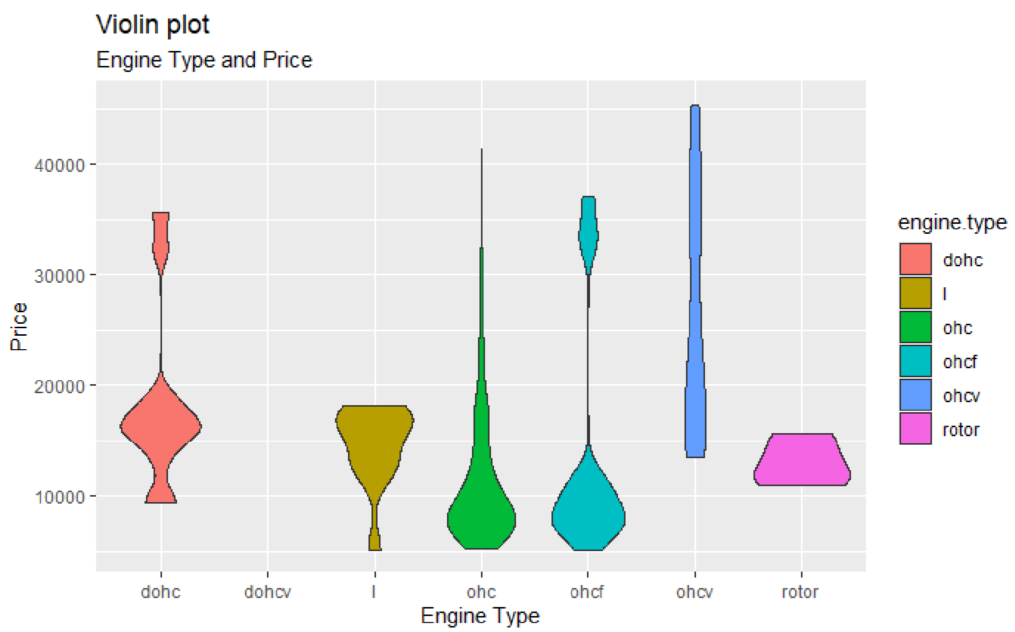
شکل **27** : **Q-Q Plot** برای جرم کل خودرو (**Curb weight**)

v) نمودار دایره ای ساده شکل 28 فراوانی انواع موتور های خودرو را نشان می دهد. بطور شهودی مشخص است داده های موتور ohc بسیار بیشتر است پس استنباط های آماری ما برای این نمونه از بقیه معتبر تر می باشد .



شکل 28 : نمودار دایره ای فرآوانی انواع موتور های خودرو

vi) با توجه به نمودار ویولنی[1] شکل 29 گستردگی و تمرکز قیمت خودروها بر اساس نوع موتور آنها مشخص است. میتوان نتیجه گرفت موتور نوع ohcf گستردگی و تنوع قیمت بیشتری دارد, موتور rotor گستردگی کمی دارد و بیشتر در خودرو های ارزان قیمت یافت می شود . موتور ohc هم که با توجه به قسمت قبل گفته شد داده های بیشتری از آن داریم تمرکز بیشتری روی خودرو های ارزان قیمت دارد.



شکل 29 : نمودار ویولنی انواع موتور های خودرو و قیمت آن

ه)

[1] Violin plot

i) بر طبق ادعای این سایت میانگین قیمت خودرو در سال 1985 (دیتاست ما برای این سال است) برابر $11,833 است حال با آزمون فرض آزمایش می کنیم :(n = 205)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	SD
5118	7775	10295	13207	16500	45400	4	7947.066

فرض صفر : متوسط قیمت خودرو $11,833 است. $H_0: \mu = 11833$

فرض متقابل : متوسط قیمت خودرو بیشتر یا کمتر از $11,833 است. $H_A: \mu \neq 11833$

$$\bar{X} \sim N(\mu = 11833\, , \frac{S}{\sqrt{n}} = \frac{7947}{\sqrt{201}} \approx 560.54)$$

→ test statistic: Z = $\frac{13207-11833}{560.54}$ =2.45

→ p-value = $P(Z > 2.45) + P(Z < -2.45) = 0.01428$

از آنجا که $\text{p-value} < 0.05$ شواهد قوی بر ضد فرض صفر وجود دارد پس آن را رد می کنیم و ادعای سایت غلط است .

ii) بر طبق ادعای این سایت شاخص MPG شهر برابر 25 است حال با آزمون فرض آزمایش می کنیم : (n = 205)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	SD
13	19	24	25.22	30	49	0	6.542142

فرض صفر : متوسط شاخص MPG شهر 25 است. $H_0: \mu = 25$

فرض متقابل : متوسط شاخص MPG شهر بیشتر یا کمتر از 25 است. $H_A: \mu \neq 25$

$$\bar{X} \sim N(\mu = 25\, , \frac{S}{\sqrt{n}} = \frac{6.542}{\sqrt{205}} \approx 0.47)$$

→ test statistic: Z = $\frac{25.22-25}{0.47}$ =0.468

→ p-value = $P(Z > 0.468) + P(Z < -0.468) = 0.683$

از آنجا که $\text{p-value} > 0.05$ فرض صفر را نمی توان رد کرد.

و)

i)

$$y = \theta^T x + \varepsilon \qquad (11)$$

$$f_{(y_i|x_i;\theta)} = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} exp\left(-\frac{(y_i - \theta^T x_i)^2}{2{\sigma_\varepsilon}^2}\right)$$

$$LL(\theta) = \sum_{i=1} ln\left(\frac{1}{\sqrt{2\pi}\sigma_\varepsilon} exp\left(-\frac{(y_i - \theta^T x_i)^2}{2{\sigma_\varepsilon}^2}\right)\right)$$

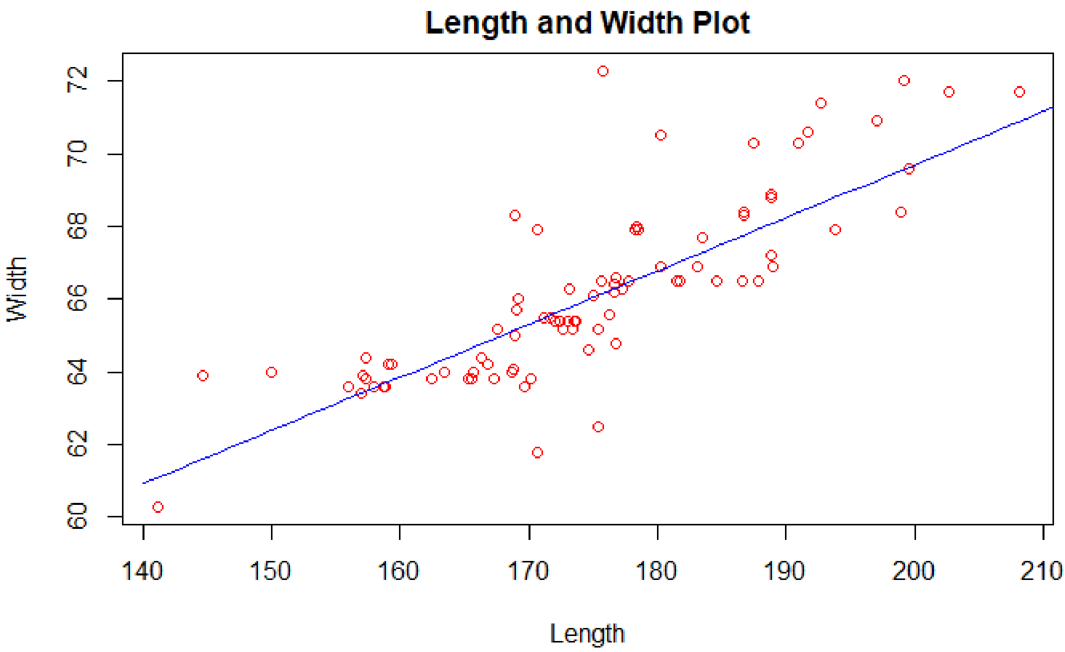
$$= n \times \ln(\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}) - \frac{1}{2{\sigma_\varepsilon}^2}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$

$$\frac{\partial LL}{\partial \theta} = 0 \quad \rightarrow \quad \theta = (X^T X)^{-1} X^T Y \qquad (12)$$

ii) با توجه به رابطه 13 و تخمین LLS[1] و نتایج شکل 30 رابطه خطی رسم شده میان طول و پهنا خودرو ها برقرار است.

$$\hat{Y}_{LLS} = a + bX \qquad (13)$$

If we have $b = \rho\frac{\sigma_y}{\sigma_x}$ and $a = E(Y) - bE(X)$



شکل 30 : نمودار پراکندگی طول و عرض ماشین همراه تقریب رگرسیون خطی

[1] Linear least squares

فایل های جانبی

به همراه این گزارش یک پوشه به نام codes در فایل zip تحویل داده شده ارائه میگردد که حاوی اطلاعات زیر است :

EPSProject_Problem1 to 5_810198375.ipybn : فایل ژوپیتر کد های پایتون سوالات 1 تا 5 پروژه

EPSProject_Problem6_810198375.ipybn : فایل ژوپیتر کد های R سوال 6 پروژه