به نام خدا

پروژه پایانی درس آمار و احتمال مهندسی

نیمسال دوم سال تحصیلی ۹۸- ۹۷

نگارنده : سهند خوشدل

استاد : دکتر ابوالقاسمی دهاقانی

شماره دانشجویی : ۸۱۰۱۹۶۶۰۷

مقدمه :

برای انجام تمرین های این پروژه از دیتاست bike و از دیتا های قسمت train استفاده شده است .

برای import کردن داده ها از فایل excel به R از تابع read.csv اسفاده شده که فایل با فرمت csv. را به این صورت می خواند که آرگومان اول (file) نشان دهنده آدرس فایل excel مربوطه است و دو متغیر دیگر (header,sep) را هم به ترتیب برابر با TRUE و "," قرار می دهیم.

سوال ۱)

شرح syntax و الگوریتم :

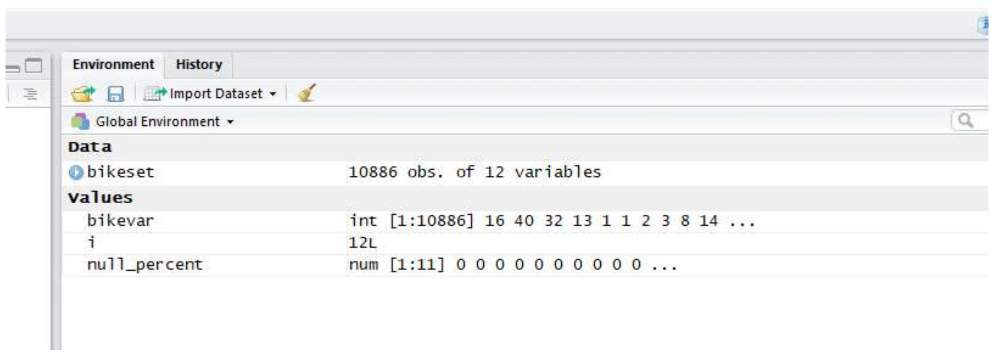
در این سوال از ما خواسته شده تا اگر در داده های مربوط به هر یک از متغیر ها ، جایی عدم وجود داده مشاهده شد اعلام کنیم و درصد نسبت تعداد داده های تهی (null) را به کل داده ها نیز بیان کنیم .

برای این کار از تابع (is.na) می توان استفاده کرد به این شکل که اگر به آن برداری پاس بدهیم به ما در خروجب برداری ارز مقادیر منطقی TRUE یا FALSE بر می گرداند به اسن صورت که در صورتی که آن عنصر از بردار فاقد مقدار (NULL) باشد به ما TRUE برمیگرداند و بالعکس.

حال اگر روی برداری که این تابع به عوان خروجی به ما می دهد جمع ببندیم (sum) در این صورت تعداد داده های فاقد محتوا (null) بدست می آید و با تقسیم کردن این تعداد به کل تعداد و ضرب کردن آن در ۱۰۰ ، درصد نسبت مورد نظر برای هر متغیر (ستون داده) بدست خواهد آمد.
برای ذخیره این درصد برای هر متغیر و پیمایش رئی ستون های مختلف (محاسبه این مقدار برای متغیر های مختلف) یک بردار برای ذخیره سازی با نام null percent تعریف می کنیم و روی عناصر آن حرکت می کنیم و عملیات مذکور را تکرار می کنیم .(البته با توجه به اینکه ستون اول داده ها به صورت تاریخ است بهتر است پیمایش روی ستون دوم تا آخر انجام شود)

مشاهده و ثبت داده :

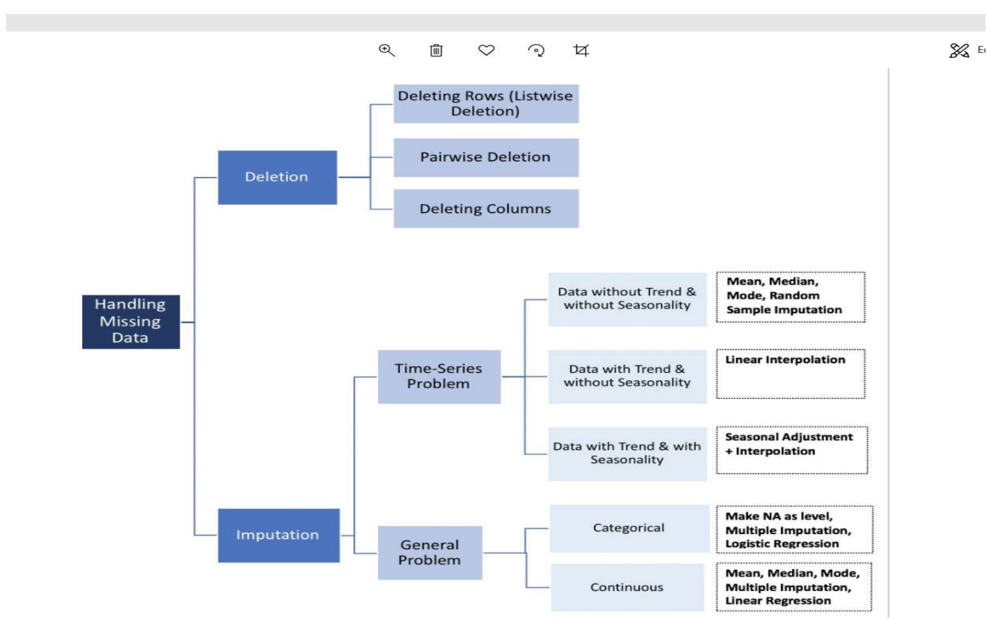
بدین ترتیب این درصد ها در برداری ۱۱ عنصره ذخیره شده و نمایش داده می شود. (شکل ۱-۱)
مشاهده می شود که مشاهده می شود هیچ فقدانی برای متغیر های دیتا ست train وجود ندارد.



شکل ۱-۱

راه حل :

راه های متعددی برای حل مشکل فقدان داده (data miss) وجود دارد که در نمودار زیر (شکل ۲-۱) دسته بندی شده اند



(شکل ۲-۱)

با توجه به شکل بالا روش کلی برای کنترل فقدان داده وجود دارد :

۱) حذف خانه های فاقد داده / ۲) روش های جبران سازی

۱) روش های حذف داده خود به سه روش تقسیم می شود :

۱- حذف سطری(row deletion)

۲ – حذف ستونی (column deletion)

۳ - حذف pairwise یا جفت حفت

۲) روش های جبران سازی نیز به چند دسته تقسیم می شوند که می توان به روش های درونیابی خطی (linear interpolation) ، رگرسیون خطی (linear regression) ، جبران سازی با نمونه برداری تصادفی (random sample imputation) ، روش های مبتنی بر (میانه ، مد و میانگین) و روش های جبران سازی چندگانه و logistic regression اشاره کرد .

همچنین قابل ذکر است که انتخاب روش مناسب برای کنترل داده های ناپدید شده به نسبت داده های ناپدید شده ی مربوط به یک متغیر ، دلیل ناپدید شدن داده مربوطه ، ارزش داده مربوطه و ... بستگی دارد .

در زبان برنامه نویسی R در برخی توابع flag ای به نام na.rm به عنوان آرگومان ورودی موجود است که در صورت TRUE بودن آن ، پیش از انجام عملیات توسط تابع ، داده های ناپدید شده از دیتا ست حذف می شوند.

سوال ۲)

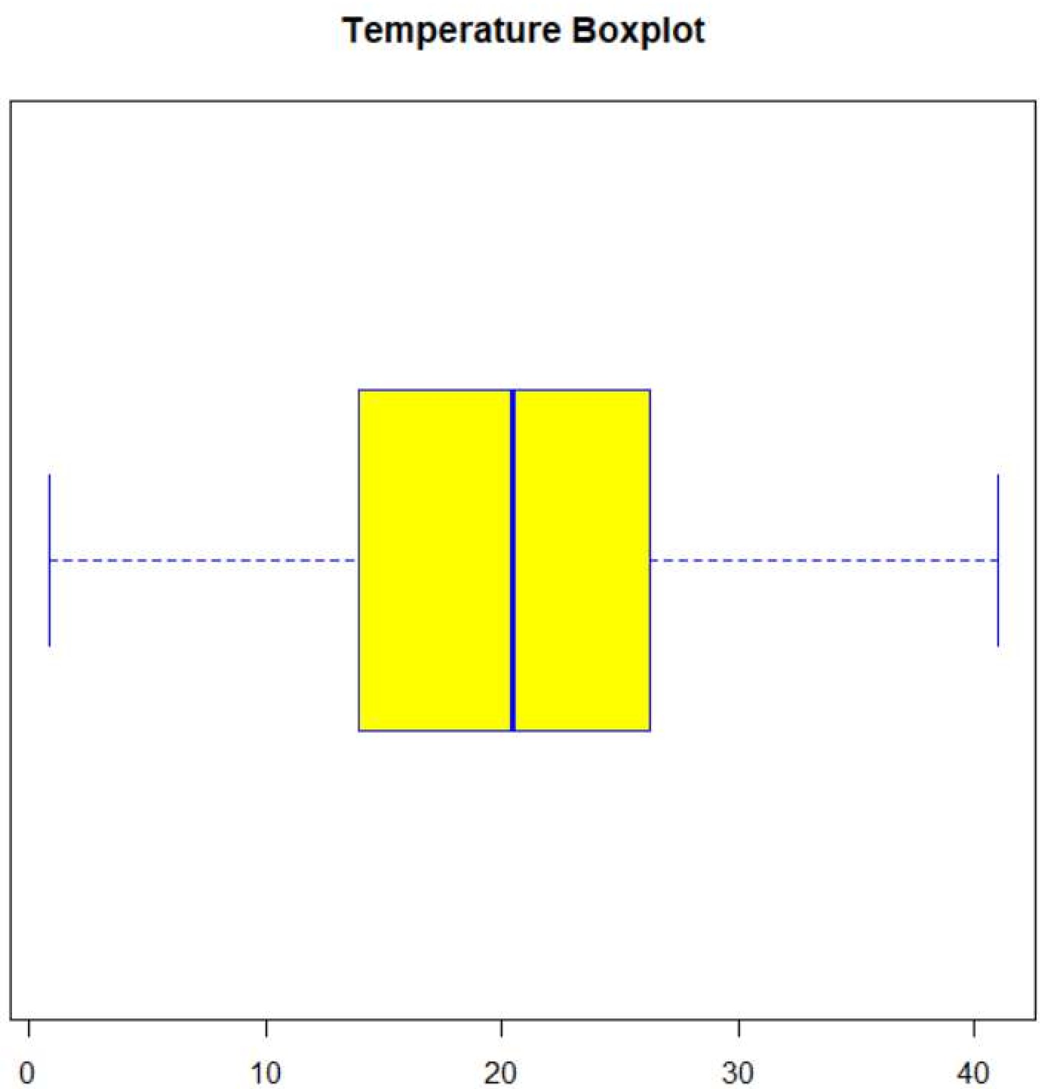
بخش اول : رسم نمودار جعبه ای

شرح syntax و الگوریتم :

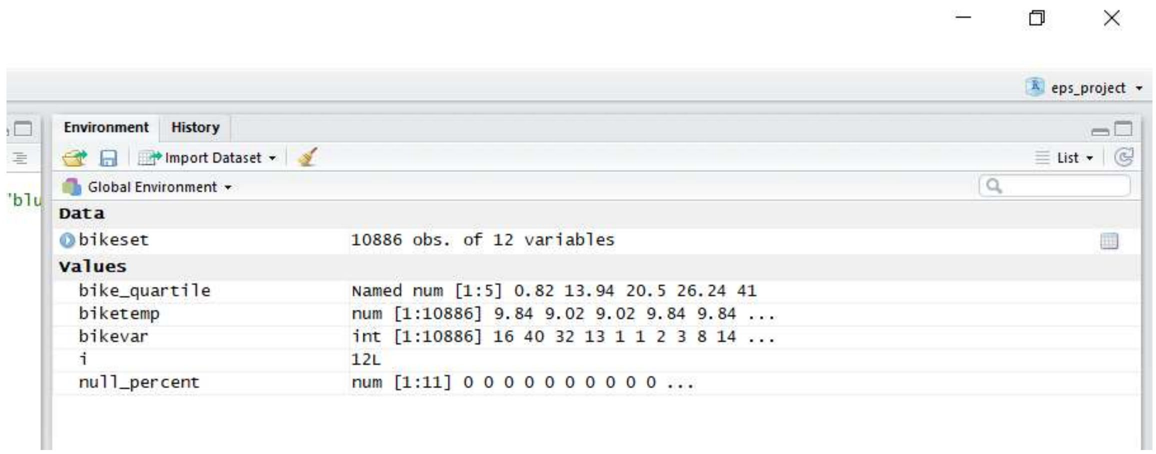
برای رسم نمودار جعبه ای در R از دستور boxplot استفاده میکنیم . آرگومان های ورودی این دستور (متغیر دلخواه برای رسم نمودار ، عنوان نمودار ، رنگ نمودار و رنک حاشیه نمودار) می باشد که آن ها را وارد می کنیم .

مشاهده و ثبت داده :

در این سوال متغیر دما را به عنوان متغیر دلخواه انتخاب کرده ام و نمودار جعبه ای آن را رسم کرده ام (شکل ۲-۱) . مقادیر میانه ، چارک اول ، چارک سوم در جدول زیر (شکل ۲-۲) گزارش شده اند.



(شکل ۱-۲)



شاخص	چارک اول (Q1) (25th percentile)	چارک دوم (Q2) (Median)	چارک سوم (Q3) (75th percentile)	چارک چهارم (Q4) (Maximum)
مقدار	13.94	20.5	26.24	41

(شکل ۲-۲)

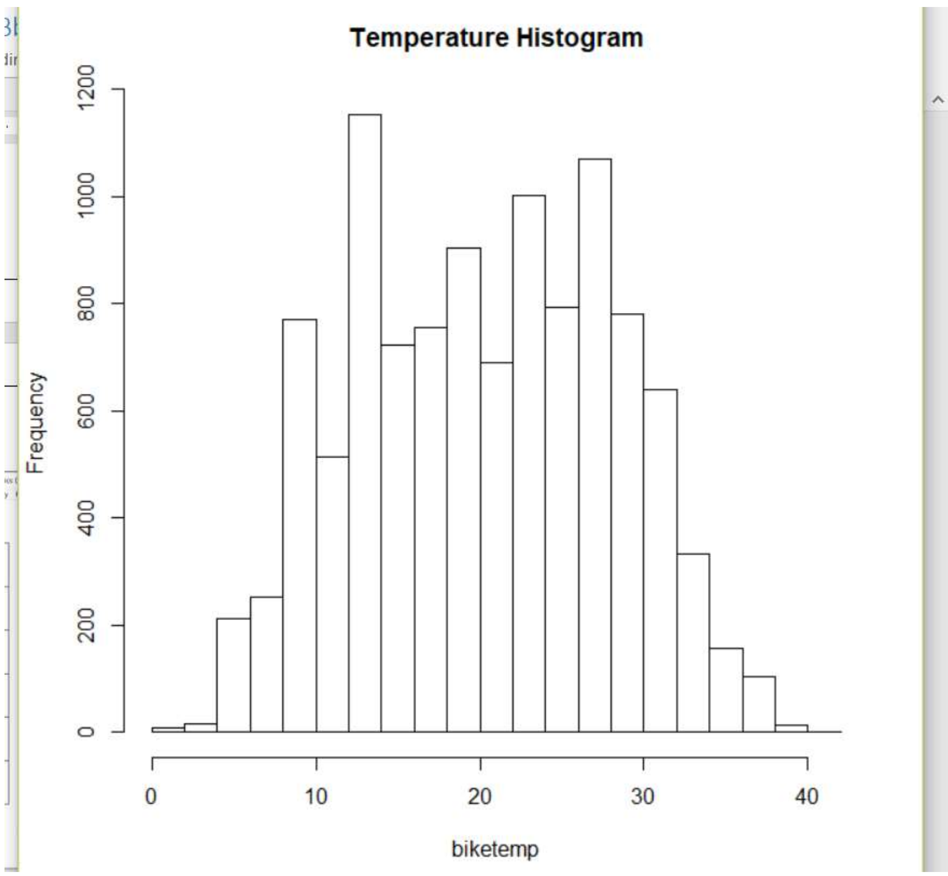
بخش دوم : رسم histogram (نمودار ستونی)

شرح syntax و الگوریتم :

برای رسم histogram از دستور hist استفاده می کنیم که آرگومان های ورودی آن شامل متغیر انتخاب شده ، عنوان histogram ، رنگ ستون ها و همچنین رنگ حاشیه ها می باشد .

مشاهده و ثبت داده :

هیستوگرام رسم شده در شکل (۳-۲) قابل مشاهده است . محور عمودی در این نوع نمودار نشان دهنده فراوانی متناظر با مجموع فراوانی های بازه ی مربوطه است و محور افقی هم شامل بازه های گوناگون از مقادیر قابل اختیار برای متغیر انتخاب شده می باشد.



(شکل ۳-۲)

بخش سوم : رسم تابع توزیع تجمعی (CDF) :

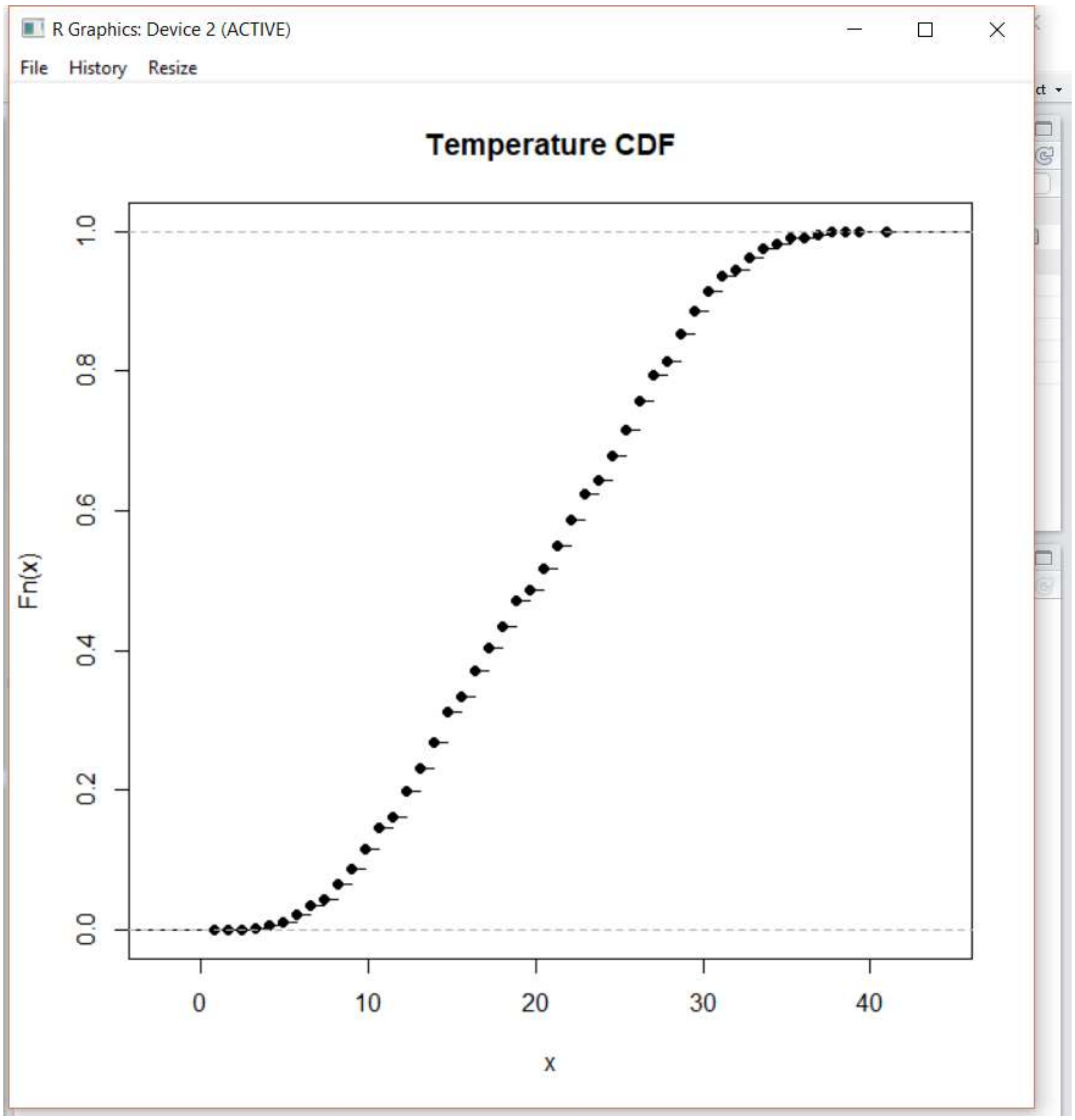
شرح syntax و الگوریتم :

ابتدا برای محاسبه تابع توزیع تجمعی از تابع ecdf استفاده می کنیم . آرگومان ورودی این تابع همان متغیر انتخاب شده برای رسم cdf است . سپس این تابع (خروجی این تابع) را به عنوان اولین آرگومان ورودی به تابع

plot پاس می دهیم تا نمودار مربط به آن را رسم کند . (آرگومان های دیگر تابع Plot می توانند شامل عنوان ، رنگ و .. نیز باشند .

مشاهده و ثبت داده :

تابع cdf رسم شده برای متغیر دما در شکل (۴-۲) قابل مشاهده است .



(شکل ۴-۲)

سوال ۳)

در این سوال نیز از متغیر دما برای انجام خواسته های سوال استفاده کرده ام .

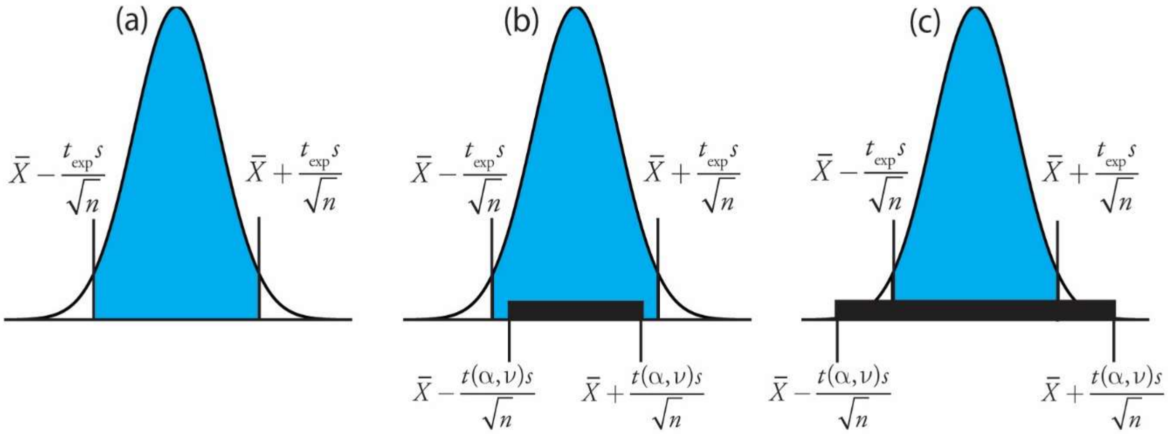
شرح syntax و الگوریتم :

- ابتدا برای بخش اول سوال به کمک تابع sample یک نمونه صدتایی از داده های متغیر دما می گیریم .
- سپس برای بخش دوم سوال میانگین و واریانس اسن نمونه صدتایی را به ترتیب به کمک دستور های mean و var محاسبه می کنیم و در دو متغیر دیگر ذخیره می کنیم . سپس با جذرگرفتن از واریانس محاسبه شده انحاف معیا مربوط به این نمونه صدتایی نیز بدست می آید.
- در بخش سوم سوال برای مقایسه توزیع بدست آمده از این نمونه صدتایی با توزیع نرمال از دستور qqnormاستفاده می کنیم و بردار (لیست) داده ی حاصل از نمونه گیری را به عنوان ورودی به این تابع می دهیم .
- در بخش چهارم باید بازه اطمینانی ۹۵ درصدی برای میانگین نمونه ای بدست آمده پیدا کنیم . این به معنای آن است که بازه محاسبه شده با احتمال ۹۵ درصد شامل میانگین ثابت و دقیق جامعه داده های ما خواهد بود. (باید دقت کنیم که ابن بازه است که متغیر است و با توجه به تخمین گر M (میانگین نمونه ای) آن را بدست آورده ایم و تخمینی است از حدود استقرار میانگین واقعی جامعه و میانگین جامعه پارامتری ثابت است که با سرشماری از جامعه می توان به مقدار دقیق آن رسید . لذا بهتر است از عبارت ((بازه با احتمال ... شامل متغیر خواهد بود)) استفاده کنیم)
- رابطه مربوط به محاسبه بازه اطمینان با سطح اطمینان دلخواه ($1-\alpha$) در صورت استفاده از واریانس نمونه ای در فقدان واریانس جامعه در شکل زیر (۱-۳) آمده است :

$$[\bar{x} + t_{1-\frac{\alpha}{2},n-1}\ (\frac{s}{\sqrt{n}})\ ,\ \bar{x} - t_{1-\frac{\alpha}{2},n-1}\ (\frac{s}{\sqrt{n}})$$

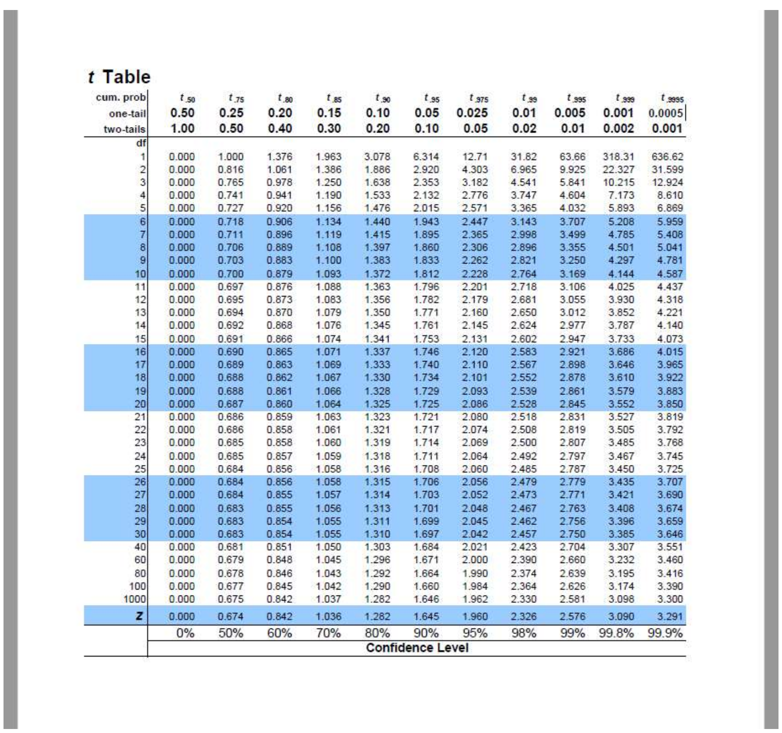
(شکل ۱-۳)

- همچنین در شکل زیر(شکل ۲-۳) دیاگرام متناظر با بازه اطمینان نمایش داده شده است .



شکل (۲-۳)

- همانطور که مشاهده می شود برای پیدا کردن بازه اطمینان به تابع توزیع تجمعی وارون مربوط به توزیع استاندارد student t نیاز داریم تا مقدار t را در نقطه ای که cdf آن برابر با 1-a/2 است بدست بیاوریم . هم می توان با اسفاده از تِوری(با استفاده از جدول cdf تابع توزیع t-student)مقدار متناظر با 1-a/2 که در اینجا برابر با 0.975 است و همچنین ۹۹ درجه آزادی با توجه به n=100 ، را بصورت دستی (t = 1.984) از طریق جدول زیر (شکل (۳-۳) بدست آورد و هم داخل کد R با استفاده از تابع CDF وارون توزیع t –student بدست می آوریم که د ادامه می بینیم .



t Table

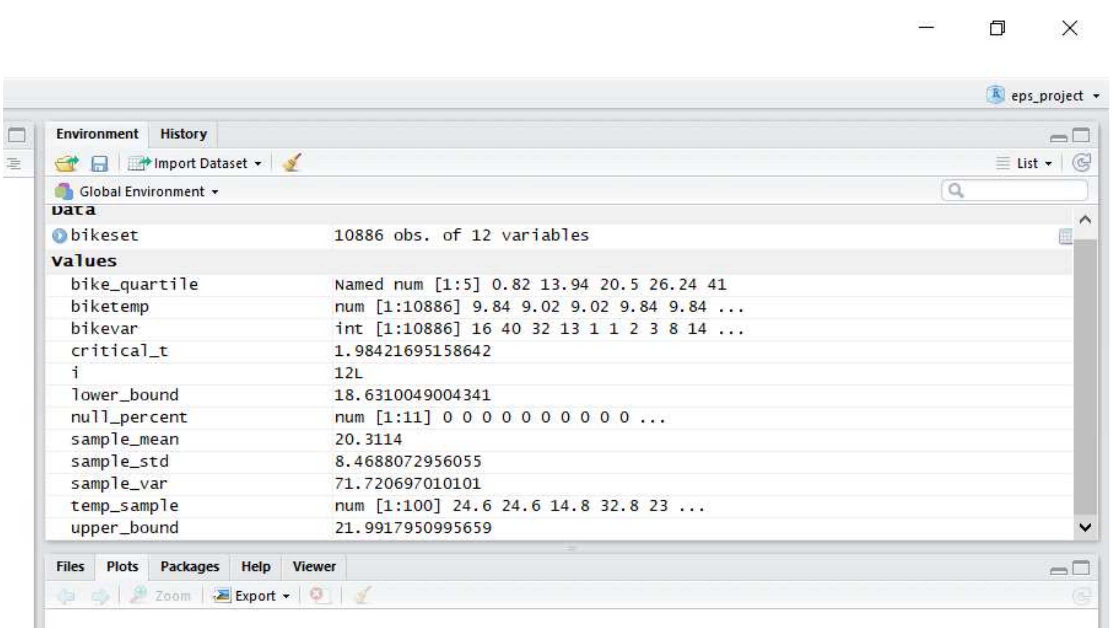
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

شکل (۳-۳)

- در بخش پنجم برای طراحی آزمون فرض دو طرفه ابتدا فرضی را مبنی بر یک ادعا مطرح می کنیم (مثلا ادعا می کنیم میانگین دما برابر با ۲۵ درجه است . سپس با توجه به میانگن نمونه ای اندازه گیری شده P-value مورد نظر را محاسبه می کنیم(مقدار p –value برابر است با مجموع احتمال رخداد میانگین نمونه ای یا بیشتر از آن و رخداد مکمل میانگین نمونه ای نسبت به میانگین فرض شده و ضعیف تر از آن) اگر مقدار محاسبه شده برای p-value کمتر از 0.05 باشد می توان فرض صفر را رد کرد .
برای محاسبه p-value به کمک قضیه حد مرکزی توزیع میانگین نمونه ای استاندارد شده را معادل با توزیع t-student) یا با کمی اغماض توزیع normal) می گیریم و مقدار احتمال متناظر با عبارتی بر حسب cdf توزیع t یا توزیع normal خواهد بود . رابطه ی مذکور در شکل (۴-۳) نشان داده شده است .

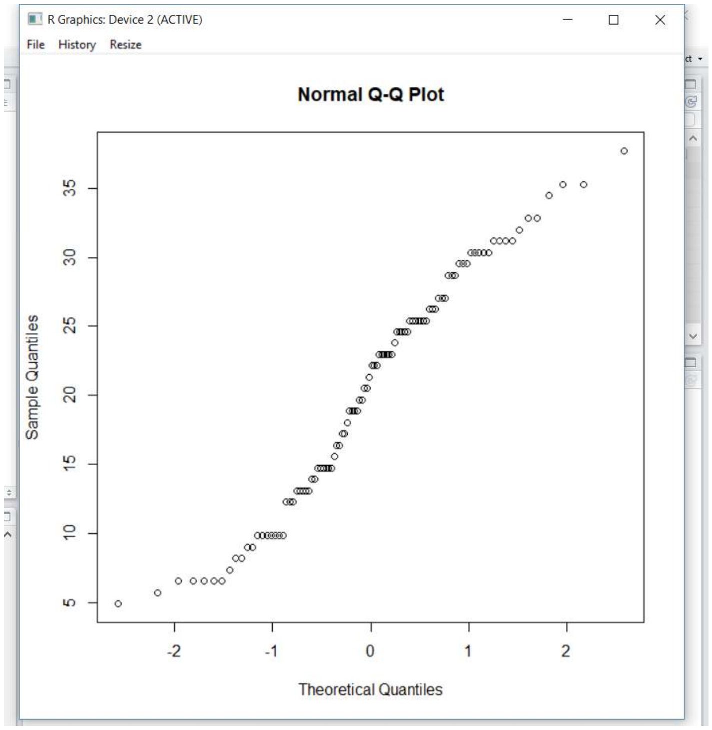
$$p - value \ =$$
$$P\{sample\ mean\ or\ more\ extreme\ observation\} + P\{sample\ mean\ complement\ (origin : H0)\}$$
$$= 2P\{sample\ mean\ or\ more\ extreme\ observation\}$$

مشاهده و ثبت داده :



(شکل ۴-۳)

قسمت سوم : نمودار qqnorm (مقایسه ی توزیع دما با نمونه گیری با سایز ۱۰۰ با توزیع نرمال):



(شکل ۵-۳)

تحلیل نتیجه :

با توجه به حطی بودن رابطه ی توزیع نرکال با توزیع دما لذا با ۱۰۰ واحد نمونه گیری توزیع هر آمارگان نمونه ای به توزیع نرمال میل می کند(قضیه حد مرکزی) . لذا رابطه آن ها برحسب هم به صورت تقریبا خطی خواهد بود .

قسمت چهارم :

در آوردن مقدار p-value متناظر با داده ها برای تخمین میانگین جامعه :

پارامتر	n	s	$\bar{x}$	1 - α	1 - $\alpha/2$	t(1 - $\alpha/$ 2)	Confidence interval
مقدار	100	8.469	20.32	0.95	0.975	1.984	[18.631,21.992]

تحلیل نتیجه :

با احتمال ۹۵ درصد بازه فوق شامل میانگین اصلی جامعه می باشد .

قسمت پنجم :

حال اگر فرض صفر (H0) را آن بگذاریم ک میانگین دما ۲۰ درجه بگیریم با توجه به بازه اطمینان ۹۵ درصدی برای آزمون فرض ، فرض ضفر رد می شود .

اما با توجه به خواسته سوال p-value را نیز محاسبه می کنیم .

اجرای تست فرض : با فرض h(0) = 20 به محاسبه مقدار برای p-value بر اساس فرمول زیر(ناشی از استاندارد سازی و قضیه حد مرکزی) می پردازیم :

$$p - value(bilateral) = \{2 * (1 - CDF((\bar{X} - \mu)/s))\}$$

$$p - value(unilateral) = \{(1 - CDF((\bar{X} - \mu)/s))\}$$

(μ - در عبارت بالا همان فرض صفر است.)

برای قسمت ششم:

همان عملیات قسمت قبل را یکطرفه انجام می دهیم و مقدار p-value تنها برابر با احتمال رخداد میانگین نمونه ای مشاهده شده یا بیشتر خواهد بود .

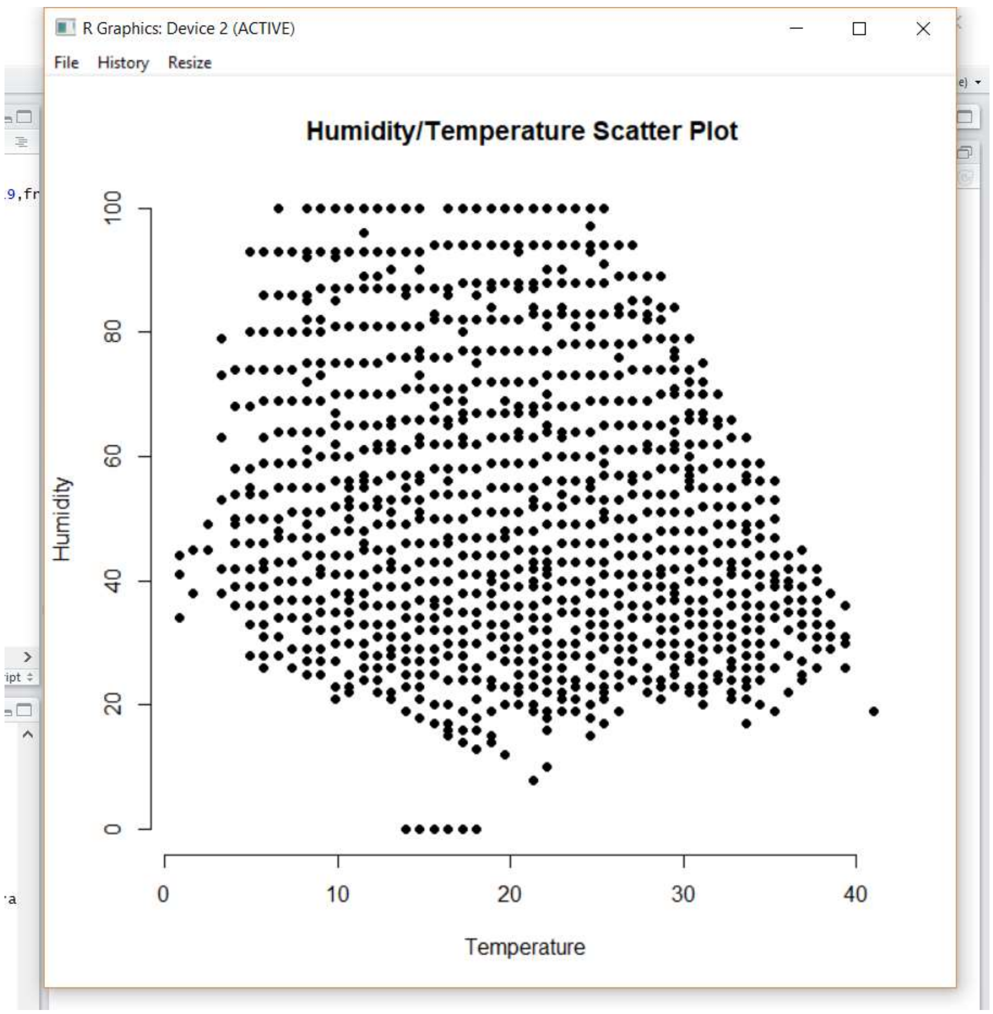
	$\bar{x}$	s	$H_0 = \mu$	CDF	p-value
Unilateral Hypothesis Test	20.32	8.469	20	0.515	0.485
Bilateral Hypothesis Test	20.32	8.469	20	0.515	0.97

تحلیل نتیجه :

با توجه به اینکه میزان P-value محاسبه شده در هر دو آزمون فرض یکطرفه و دوطرفه بیشتر از p-value بحرانی (0.05) می باشد لذا نمی توانیم فرض صفر را رد کنیم . پس تخمین دمای ۲۰ درجه برای میانگین دما در داده های ما تخمین مناسبی است و نزدیک بودن p-valueدر حالت دو طرفه به مقدار ۱ نیز این موضوع را تایید می کند .

سوال ۴)

در این سوال دو متغیر دلخواه انتخابی دما و رطوبت می باشند . ابتدا خواسته شده تا scatterplot بین توزیع این دو متغیر را نشان دهیم .(شکل ۱-۴)



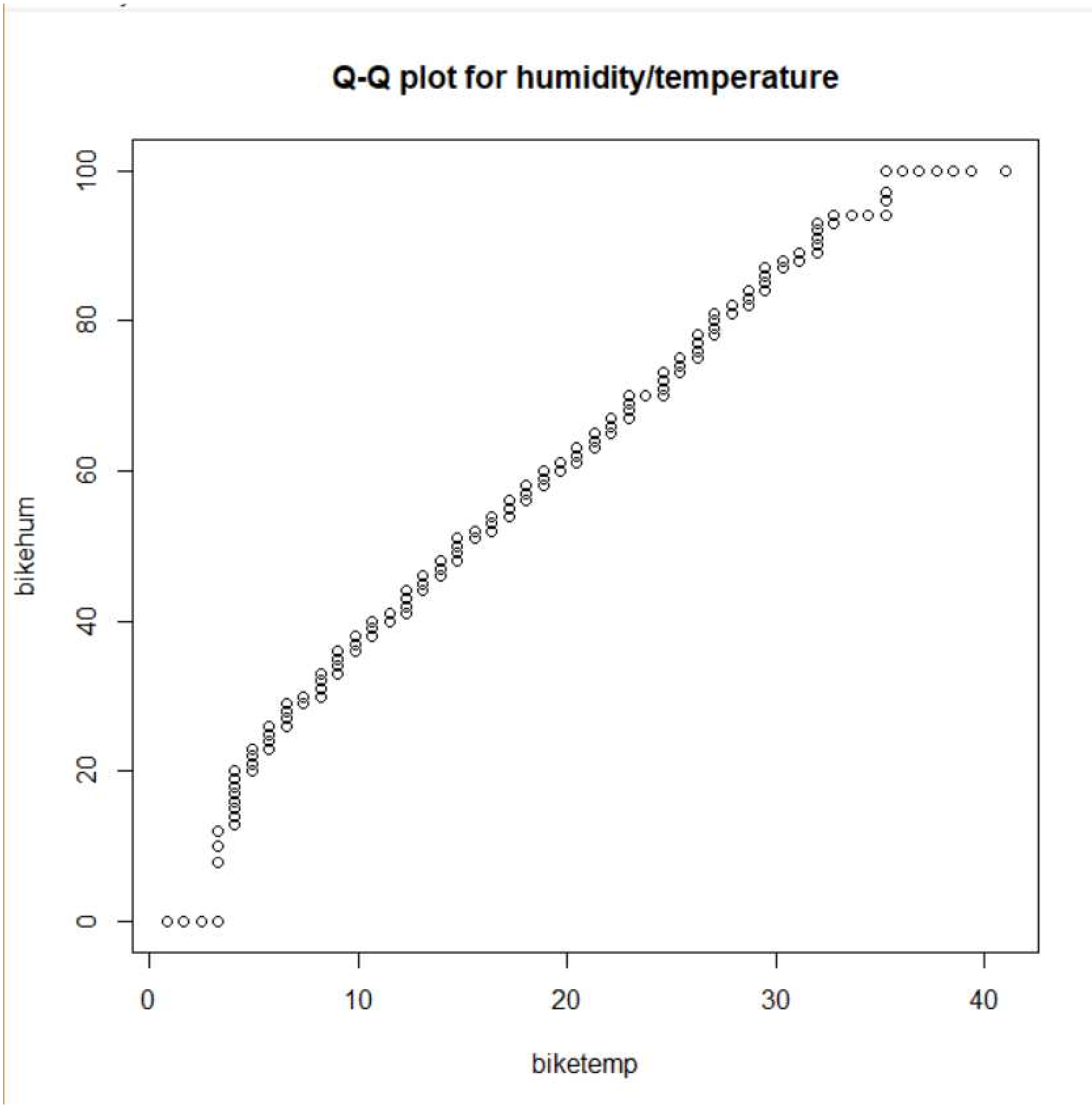
(شکل ۱-۴)

$r_{spearman} = -0.0468$ **/** $r_{pearson} = -0.0649$

تحلیل نتایج :

شکل scatter-plot به نوعی توصیف کننده تابع توزیع توام دو متغیر است و می توان آن را مدلی توصیفی برای نمایش رویه تابع چگالی توام دو متغیر دانست به این شکل که تراکم نقاط معیاری از میزان $f_{xy}(x,y)$ می باشد و هر چه تراکم نقاط در قسمتی از صفحه حاصل از دو متغیر بیشتر باشد در واقع ارتفاع رویه مربوطه بیشتر است و احتمال ظهور جفت مقادیر حول آن قسمت ها برای دو متغیر بیشتر است .

با توجه به نمودار مشاهده شده و البته ضریب همبستگی بدست آمده می توان گفت این دو داده یعنی دما و رطوبت همبستگی چندانی با هم ندارند(اگر به سراغ بدست آوردن توابع توزیع حاشیه ای با برش زدن و انتگرال گیری روی یک محور برویم نیز توابع حاشیه ای هر متغیر تا حدود بسیار بالایی مستقل از مقدار اختیار شده توسط متغیر دیگر خواهد بود) (اما همان همبستگی اندک آن ها در جهت منفی است)


Q-Q plot for humidity/temperature
bikehum
0
20
40
60
80
100
0
10
20
30
40
biketemp

(شکل ۲-۴)

تحلیل نتایج :

نمودار Q-Q plot را برای مشاهده ی میزان شباهت بین توزیع دو متغیر مختلف استفاده می کنند . یکی از مزیت های این نمودار عدم نیاز به هم سایز بودن نمونه های دو متغیر است . هر چه این نمودار به خط y=x بیشتر تمایل پیدا کند شکل توزیع های مورد نظر به هم بیشتر شبیه اند . اگر یکی از توزیع های نرمال باشد این دستور همان Q-Qnorm خواهد بود .

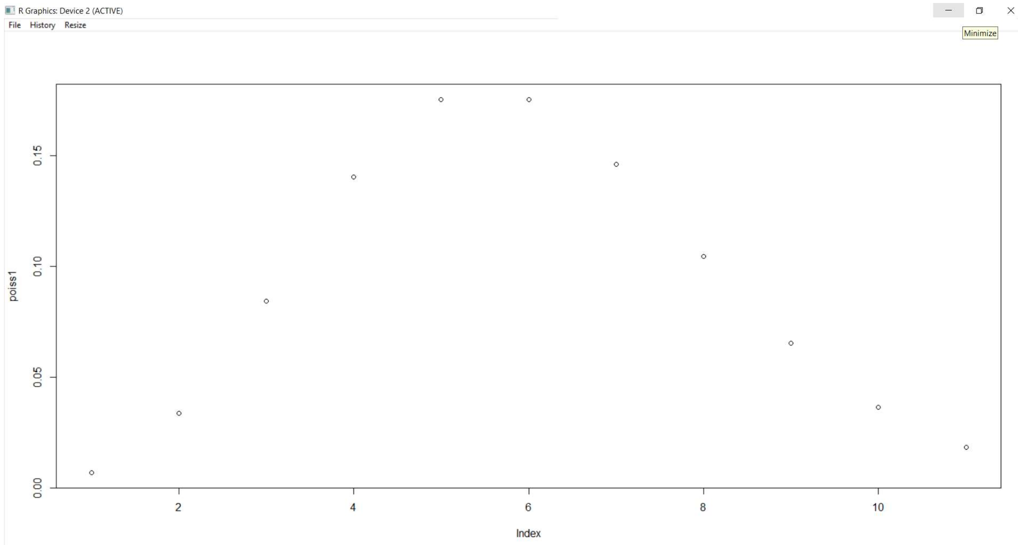
با توجه به این که متغیر های رطوبت و دما انحراف کمی از این خط دارند لذا شکل توزیع این دو متغیر تا حدود زیادی به هم شباهت دارد

سوال ۵)

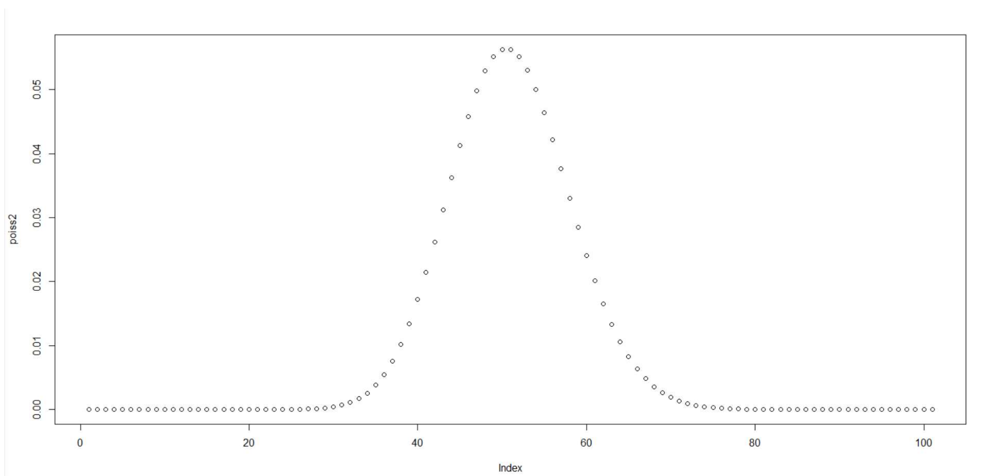
ابتدا به کمک تابع dpoiss سه توزیع پواسن با پارامتر های ۵ و۵۰ و۵۰۰ را رسم میکنیم

این توزیع حدودی را برای نمایش میگیرد که بهتر است بازه ای متقارن حول میانگین (پارامتر lambda) به آن بدهیم .

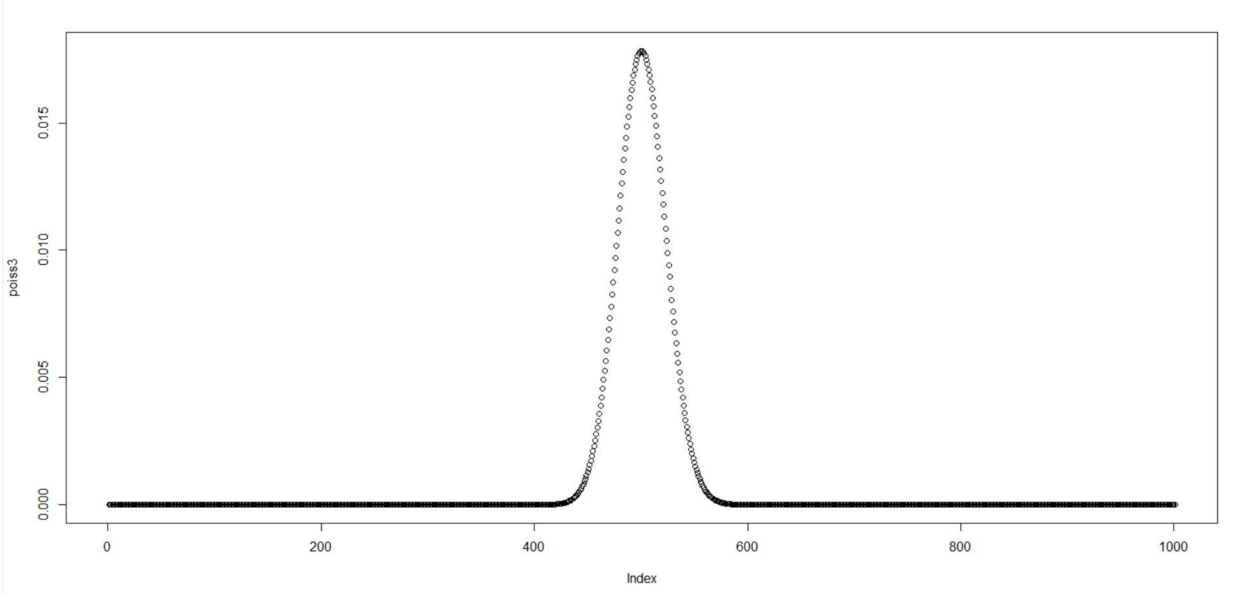
الف) نمودار توزیع پواسن با پارامتر lambda = 5:



ب) نمودار توزیع پواسن با پارامتر lambda = 50:

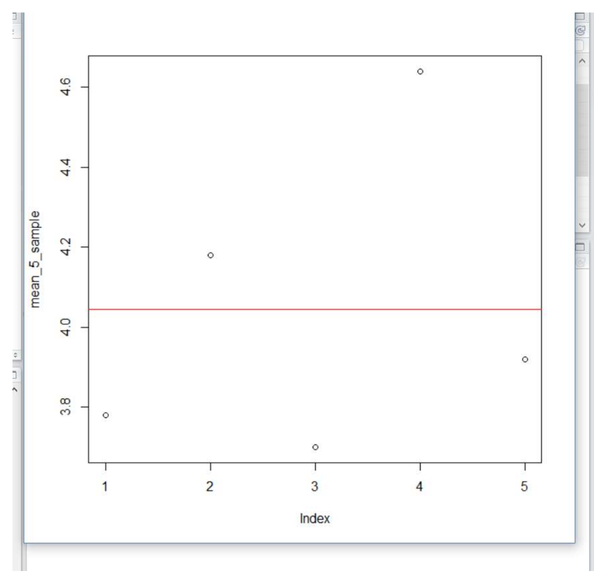


ج) نمودار توزیع پواسن با پارامتر lambda = 500:

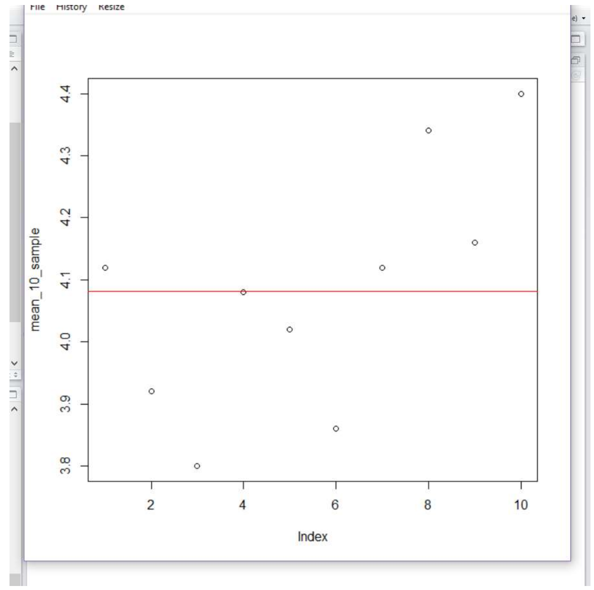


سپس شروع به نمونه برداری از یک توزیع پواسن با پارامتر **(lambda = 4)** می کنیم به این صورت که ابتدا در حالت ((د)) ، ۵ نمونه ۵۰ تایی ، در حالت ((ه)) ، ۱۰ نمونه ۵۰ تایی ، در حالت ((و)) ، ۵۰۰ نمونه ۵۰ تایی ، در حالت ((ز)) ، ۵۰۰۰ نمونه ۵۰ تایی را بررسی می کنیم .

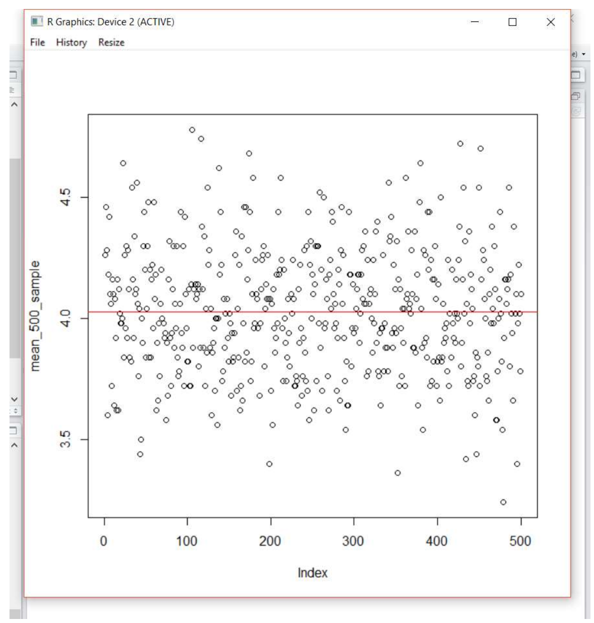
د) نمودار توزیع آمارگان میانگین نمونه ای در حالت ۵ تایی :



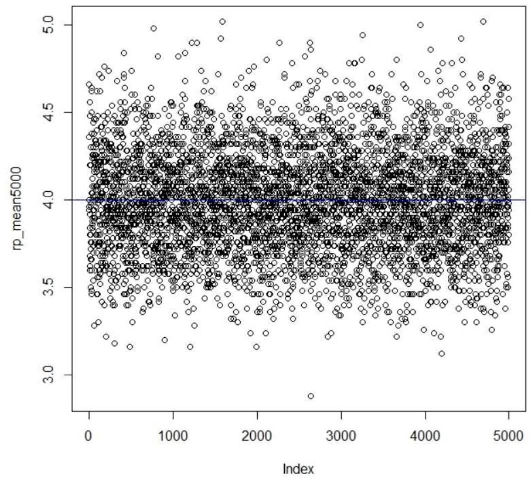
ه) نمودار توزیع آمارگان میانگین نمونه ای در حالت ۱۰ تایی :



و) نمودار توزیع آمارگان میانگین نمونه ای در حالت ۵۰۰ تایی :



ز) نمودار توزیع آمارگان میانگین نمونه ای در حالت ۵ تایی :



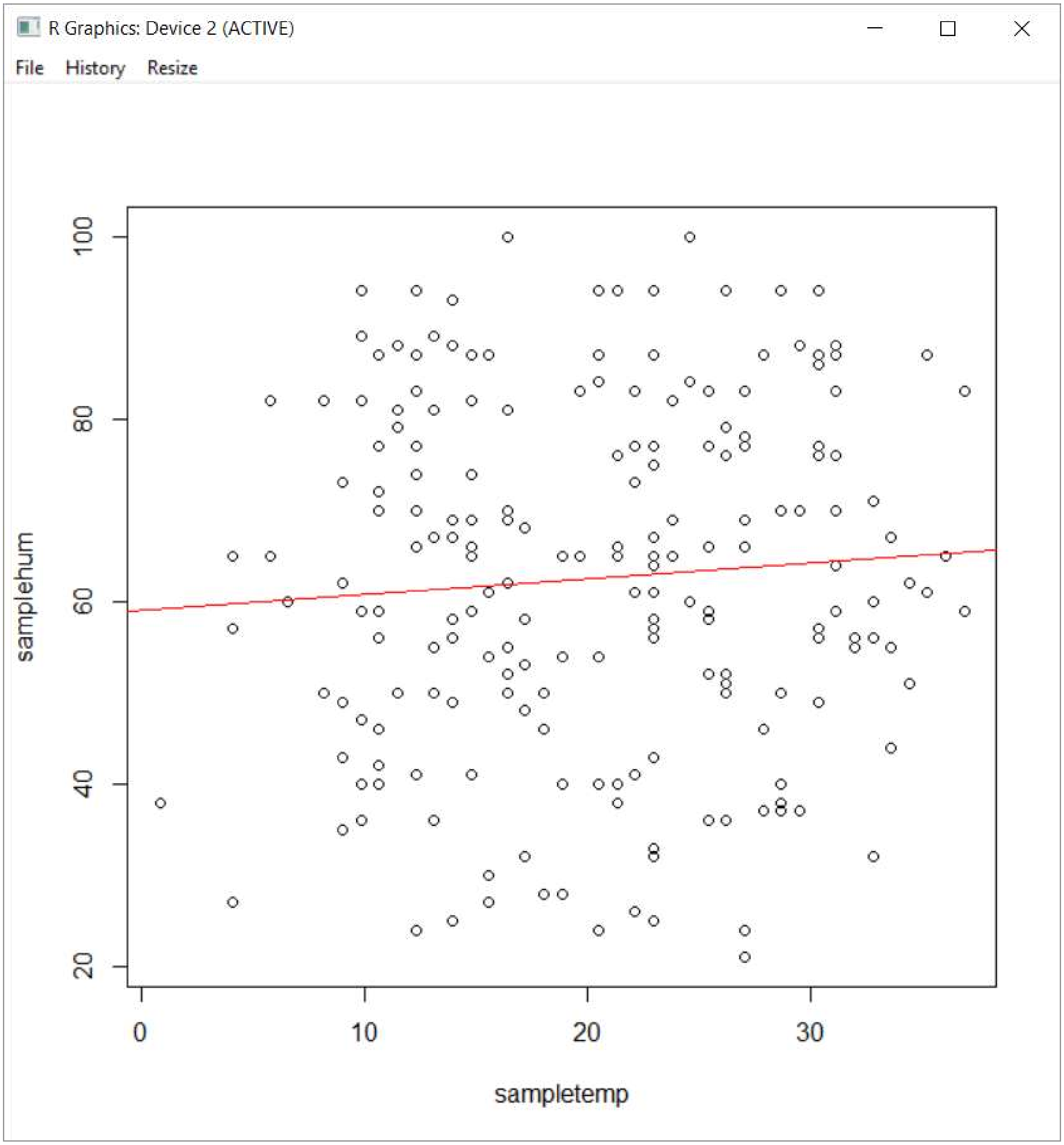
تحلیل نتیجه :

با توجه به قضیه حد مرکزی با افزایش تعداد نمونه های تصادفی با سایز ۵۰ باید توزیع آمارگان میانگین نمونه این به توزیع نرمال با میانگین جامعه (که با توجه به تبعیت جامعه از توزیع پواسن برابر با lambda

این توزیع که برابر با 4 است و هماهنطور که مشاهده می شود هم خطوط متناظر با میانگین توزیع میانگین های نمونه ای در هر یک از نمودار ها که رنگی رسم شده است با افزایش این تعداد به ۴ میل کردند که قضیه حد مکزی را تایید می کند .

سوال ۶)

با استفاده از دستور های sample ابتدا نمونه برداری از متغیر های رطوبت و دما را انجام می دهیم ، سپس برای بدست آوردن بهترین تخمین خطی برای ارتباط بین دو متغیر مذکور (رگرسیون خطی) از دستورات plot ,abline,lm استفاده می کنیم . دستور lm توزیع متناظر با نمونه های تصادفی دو متغیر را می گیرد و شیب خط رگرسیون و عرض از مبدا آن را تحویل می دهد . سپس خروجی lmرا به abline پاس می دهیم تا خط رگرسیون کشیده شود و همچنین با دستور plot اقدام به رسم این نمودار نقطه ای این دو متغیر بر حسب هم میکنیم .


R Graphics: Device 2 (ACTIVE)
File History Resize
100
80
60
40
20
samplehum
0
10
20
30
sampletemp

منابع :

1) https://stackoverflow.com/questions/2613420/handling-missing-incomplete-data-in-r-is-there-function-to-mask-but-not-remove

2) https://www.researchgate.net/figure/Student-distribution-and-its-confidence-interval_fig3_232637324

3) www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf