رسم نمودار

(R)

تدوین: مرکز تحلیل آماری خوارزمی

***

www.kharazmi-statistics.ir

مرکز آماری خوارزمی

مقدمه:

نمودار از جمله خروجی ها در نرم افزار است که در تحلیل ها کمک زیادی به محقق می کند. یکی از مزیت های نمودار درک بسیاری از اطلاعات داده ها و مشاهدات در یک نگاه است. یافتن توزیع مشاهدات، وجود داده ی پرت و چگونگی پراکندگی داده ها از جمله ی موارد آن است.

در این بخش به بررسی چند نمودار پایه ای می پردازیم. مطالبی که در این متن بیان شده است عبارتند از:

انتخاب نمودار مناسب

نمودار میله ای

نمودار نقطه ای

نمودار دایره ای

نمودار هسیتگرام

نمودار جعبه ای

نمودار شاخه و برگ

نمودار های پراکندگی

نمودار QQ

دستورات تکمیلی برای رسم نمودار در نرم افزار R

انتخاب مناسب یک نمودار

اولین عامل در انتخاب نمودار مناسب، نوع داده هاست. نمودارهای میله ای، نقطه ای و دایره ای مقادیر هر داده را نشان می دهند. نمودارهای هیستگرام، جعبه ای و QQ توزیع ها را نشان می دهند و نمودارهای پراکندگی مقادیر جفتی را رسم می کنند.

عامل دوم، مخاطب است. اگر نمودار را برای خودتان یا یک تحصیلکرده ی آماری رسم می کنید، پس با یک محیط با تجربه و کارآزموده سروکار دارید. برای مثال نمودارهای جعبه ای و QQ توضیحات بیشتری را نسبت به هیستوگرام می طلبد و ممکن است برای عموم مناسب نباشد. همچنین اطلاع از درک افراد و چگونگی استنباط آن ها از محیط های آماری نیز اهمیت بسزایی دارد.

در نگاه اول به یک نمودار، می توان اطلاعات کمّی آن را استخراج کرده و در اصطلاح آن را رمز گشایی می کنیم. باید در انتقال و فهماندن اطلاعات از راه هایی استفاده کنیم که به راحتی قابل برداشت باشد، نه اینکه نمودارها در رساندن پیام ها اختلاف داشته باشند.

برای مثال میله های فراوانی در نمودار میله ای به سادگی قابل تشخیص است زیرا موقعیت و طول میله ها به راحتی دیده می شود.مساحت مستطیل های فراوانی نیز به استنباط درست، استحکام می بخشد.

اما اینکه در نگاه اول به طول و مساحت جزئیات نمودار دقت می کنیم ما را در تنگنا قرار می دهد. ما باید به طور معمول رسم نمودار میله ای را از صفر شروع کنیم. بنابراین موقعیت، طول و مساحت اطلاعات مشابه را انتقال می دهد. اگر اعدادی را نمایش می دهیم که در آن صفر منظور نشده، آنگاه نمودار نقطه ای برای رسم، انتخاب بهتری است. در نمودار نقطه ای، موقعیت نقطه هایی است که از سوی بیننده قابل برداشت است.

از تأمل در شیوه های مشاهده ی نمودار به این نتیجه می رسیم که انتخاب نمودارهای دایره ای برای نمایش داده ها از ارزش کمتری برخوردار است. برای مشاهده ی اندازه ی قطعه های دایره لازم است زاویه و مساحت آن ها را اندازه بگیریم که زیاد هم در آن تبحّر نداریم.

و بالاخره رنگ نمودار که می تواند در تشخیص گروه ها و طبقه بندی ها بسیار مفید باشد. مجموعه ی RcolorBrewer درR مقداری از گونه های رنگ(پالت) را در خود جای داده است. چند پالت رنگ نشان دهنده ی گروه های پشت سر هم از پایین تا بالا، برخی هم نشان دهنده ی گروه هایی از مقادیر خنثی هستند. و شاید آنها که به طور کامل کمّی هستند و برای این انتخاب شدند که تفاوت آن ها برای افراد کورنگ ها، به راحتی قابل مشاهده باشد.

نمودار میله ای

ابتدایی ترین نوع نمودار مجموعه ی تک عضوی از اعداد را نشان می دهد. مانند نمودارهای میله ای و نقطه ای که به ترتیب طول یک میله ای مکان و یک نقطه را به عدد مورد نظر نسبت می دهد. برای رسم نمودار میله ای

داده های پیش فرض در نرم افزار R را فراخوانی می کنیم. در این سری داده ی USPersonalExpenditure فراخوانی شده است. برای آگاهی از چگونگی فراخوانی داده های پیش فرض در نرم افزار به فایل "**فراخوانی داده های پیش فرض در R**" بخش آموزش نرم افزار در پایگاه تحلیل آماری خوارزمی مراجعه نمایید.

داده های فراخوانی شده متوسط هزینه های شخصی در آمریکا را در بردارد. مشتمل بر ۵ متغیر هزینه های غذا و دخانیات، هزینه های خانوار، هزینه های دارو و سلامت، مراقبت های شخصی، آموزش های عمومی را در سال های ۱۹۴۰ تا ۱۹۶۰ بیان کرده است.

برای رسم نمودار میله ای این متغیرها فرمان زیر را بنویسید.

```
> barplot(USPersonalExpenditure , beside=TRUE, legend=TRUE, ylim=c(0,90),
ylab="cost", main="US Personal Expenditure")
```

beside=TRUE سبب می شود که مقادیر هر ستون کنار هم رسم شوند.

legend=TRUE راهنمای تصویر را در بالا سمت راست آن اضافه می کند.

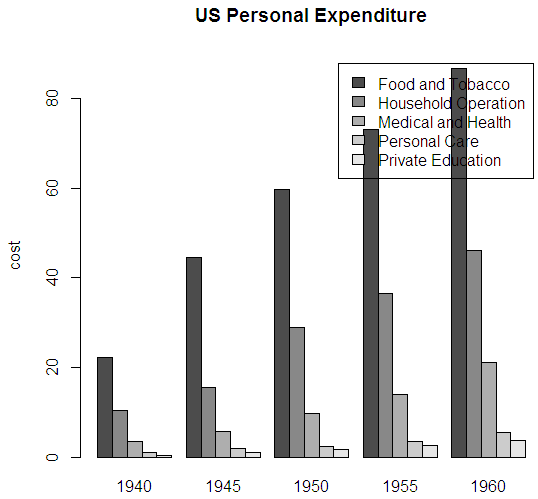
ylim=c(0,90) باعث می شود که بازه ی محور عمودی یا y نمودار را برای کل داده مشخص کند.

ylab="cost" عنوان محور عمودی یا y نمودار را اضافه می کند.

main="US Personal Expenditure" عنوان اصلی نمودار را به آن اضافه می کند.

```
> USPersonalExpenditure
                      1940   1945  1950 1955  1960
Food and Tobacco    22.200 44.500 59.60 73.2 86.80
Household Operation 10.500 15.500 29.00 36.5 46.20
Medical and Health   3.530  5.760  9.71 14.0 21.10
Personal Care        1.040  1.980  2.45  3.4  5.40
Private Education    0.341  0.974  1.80  2.6  3.64
> barplot(USPersonalExpenditure , beside=TRUE, legend=TRUE, ylim=c(0,90), ylab="Brow", main="Brownle")
```

خروجی نمودار در نرم افزار:



همان طور که مشاهده می شود نمودار متناسب با فرامینی نوشته شده در نرم افزار رسم و نمایش داده شده است. همچنین می توان رنگ ستون ها را متناسب را نیز با توجه به دستور رنگ انتخاب کرد. که در ادامه برای نمودارهای دیگر بیان شده است.

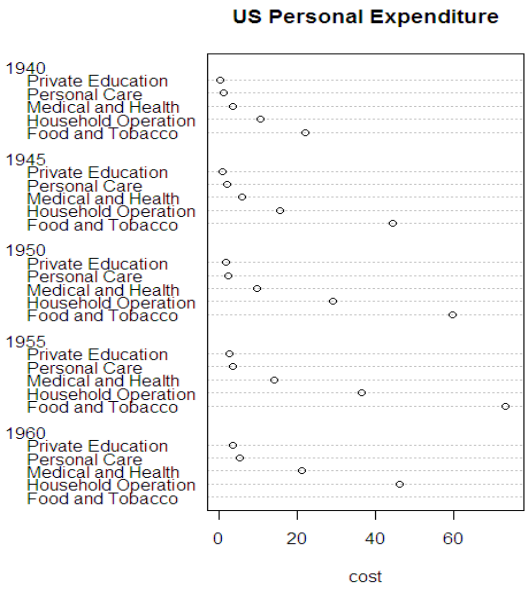
نمودار نقطه ای

از دیگر نمودارهای تک عضوی از اعداد، نمودار نقطه ای است. برای رسم این نمودار نیز از داده ها در بخش قبل استفاده می کنیم. برای رسم نمودار نقطه ای فرمان زیر را در صفحه ی فرامین نرم افزار بنویسید.

```
> dotchart(USPersonalExpenditure,xlim=c(0,75),xlab="cost",main="US Personal Expenditure")
```

محدوده ی x را از صفر تا ۷۵ تنظیم می کنیم که صفر را نیز در بردارد.

```
> dotchart(USPersonalExpenditure,xlim=c(0,75),xlab="cost",main="US Personal Expenditure")
```



نمودار دایره ای

نمودارهای دایره ای برداری از اعداد را به وسیله ی شکستن یک دایره به قسمت های مختلف و نسبت داده هر قسمت به یک عدد نشان میدهد.

برای این قسمت سری داده ای را معرفی می کنیم. نمرات ارزشیابی کارکنان یک شرکت از ۱۰۰ نمره به شرح زیر است.

احمد رضایی	۹۴
مهران بابایی	۸۹
شهین عسگری	۹۸
مریم احمدی	۸۰
لادن هاشمی	۹۰
ساسان پارسایی	۷۰

برای ورود اسامی افراد از فونت لاتین استفاده کنید چراکه نرم افزار R حروف فارسی را نمی شناسد.

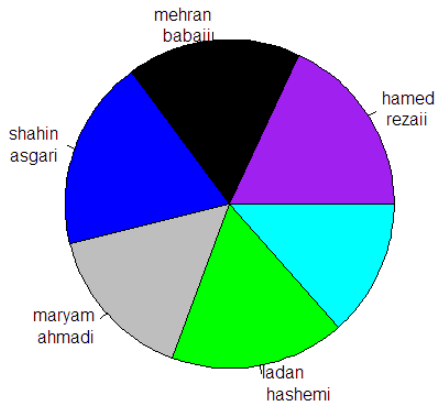
برای ورود داده های مربوطه فرامین زیر را در نرم افزار بنویسید.

```
level<-c(94,89,98,80,90,70)

labels<-c("hamed\n rezaii","mehran\n babaii","shahin\n asgari","maryam\n ahmadi","ladan\n hashemi","","sasan\n parsaii")

pie(level,labels,col=c("purple","black","blue","gray","green","cyan"))
```

```
> level<-c(94,89,98,80,90,70)
> labels<-c("hamed\n rezaii","mehran\n babaii","shahin\n asgari","maryam\n ahmadi","ladan\n hashemi","","sasan\n parsaii")
> pie(level,labels,col=c("purple","black","blue","gray","green","cyan"))
```



نمودار دایره ای بیشتر در نشریات غیر تخصصی کاربرد دارد.

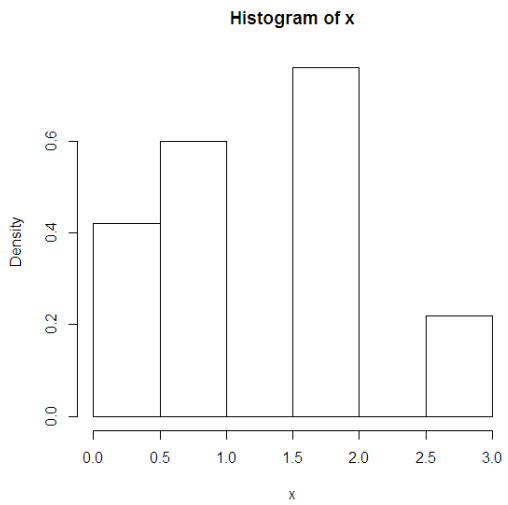
نمودار هیستوگرام

هیستوگرام نوع خاصی از نمودار میله ای که در آن از مستطیل ها برای نشان دادن فراوانی توزیع یک مجموعه از اعداد به کار می روند. هر مستطیل میزان مقادیر x را در هر کدام از پایه هایش نشان می دهد و معمولا تمام مستطیل ها پهنای یکسانی دارند و ارتفاع هر مستطیل به عدد مشاهدات مربوط به بازه ی مورد نظر بستگی دارد. اگر مستطیل ها پهنای متفاوتی داشته باشند آنگاه مساحت آن مستطیل متناسب با مقادیر آن است. در این روش ارتفاع نشان دهنده ی چگالی است.(فراوانی در هر واحد x)

در نرم افزار R دستور hist(x,…) روش اصلی برای رسم هیستگرام است. در اینجا x یک بردار شامل مشاهدات عددی و پارامترهای گزینش … برای کنترل جزئیات شکل است.

برای رسم نمودار هیستوگرام تعداد اعداد تصادفی با توزیع دو جمله ای در نظر گرفته شده است. فرمان ها زیر را در صفحه ی فرامین نرم افزار تایپ نمایید.

```
> n=5;p=.25
> x<-rbinom(100,n,p)
> hist(x,probability=TRUE)
```



اگر n مقدار از x داشته باشیم، نرم افزار R به طور خودکار بازه ی x را به $log_2(n)+1$ بازه ی تقریبی تقسیم می کند تا به هر کدام از بازه ها مقادیر فراوانی را اختصاص دهد. برای مثال داده های ما از ۱۰۰ اندازه تشکیل شده است. از آنجایی که

$$100 > 2^6 = 64$$

$$100 < 2^7 = 128$$

$$6 < log_2(100) < 7$$

مشاهده می شود که R باید از ۷ یا ۸ مستطیل فراوانی استفاده کند. در حقیقت از ۱۱ مستطیل استفاده کرده، زیرا R تلاش کرده است تا بازه را در اعداد رند بشکند که در اینجا این مقدار ۰/۵ است .

قاعده ی گفته شده در بالا، (که به استراگس معروف است) همواره قابل قبول نیست، خصوصا وقتی برای مقادیر بزرگ n، مستطیل های فراوانی کمی را اختصاص می دهد. تحقیقات حال حاضر پیشنهاد می کند که تعداد مستطیل های فراوانی باید به نسبت $n^{1/3}$ افزایش پیدا کند.

نمودار جعبه ای

نمودار جعبه ای یک روش جایگزین برای هیستوگرام است که نگاه کلی و اجمالی بر مشخصات اصلی مجموعه ی داده ها به ما می دهد و از جعبه های مستطیلی به همراه خطوطی که از دوطرف آن بیرون زده تشکیل می شود. این نمودار برای مختصرسازی داده ها به صورت فشرده و نمایش سریع در حالت چولگی داده ها می باشد. جعبه نشان دهنده ی موقعیت و تجمع قیمت مرکزی داده هاست. زمانی که وسعت خطوط(دم ها) کمک می کند تا بتوانیم در مورد بخش عمده ی داده ها نظر بدهیم. در اجرای برخی از برنامه ها، داده های پرت (مشاهداتی که از سایر داده ها تفاوت فاحشی دارد) نیز درنقاط جداگانه مشخص می شوند.

ساختار اساسی جعبه در نمودار جعبه ای بصورت زیر است:

۱) یک خط افقی که در میان جعبه به عنوان میانه کشیده می شود.
۲) داده ها به دو نیمه تقسیم می شوند که هر کدام میانه را دربر دارد.
۳) چارک اول و سوم به عنوان میانه های دو بخش مجزا محاسبه شده و خطوط افقی در هر کدام از این مقادیر کشیده شده است. آنگاه توسط یک جعبه ی مستطیلی به هم بسته شده اند.

این نمودار براساس خلاصه سازی ۵ معیار عددی است و ساده ترین کاربرد را دارد. نمودار جعبه ای یک کادر همراه با خطوطی در محور پایینی (Q_1) میانگین، محور بالایی (Q_3) و خطوط ریزی دارد که به min وmax گسترش می یابد.و برای نمایش چولگی مناسب است.

بنابراین جعبه ی نمودار نمایانگر موقعیت چارک میانی داده هاست. (IQR) که در واقع تفاوت میان چارک اول و سوم است. از IQR برای اندازه گیری میزان نوسانات در قسمت میانی داده ها استفاده می شود. زیرا ۵۰٪ داده ها دورن جعبه قرار خواهد گرفت. دم پایینی، محدوده ی بین انتهای پایین جعبه تا کوچکترین داده ای که از ۱/۵IQRکوچکتر نباشد را شامل می شود. مشابه آن دم بالایی نیز محدوده ی بین انتهای بالای جعبه تا

بزرگترین داده ای که از ۱/۵IQR بزرگتر نباشد را شامل می شود. فلسفه ی این تعاریف این است که وقتی داده ها از توزیع نرمال یا توزیع های همانند آن رسم می شوند ۹۹٪ مشاهدات در بین دم ها خواهند بود.

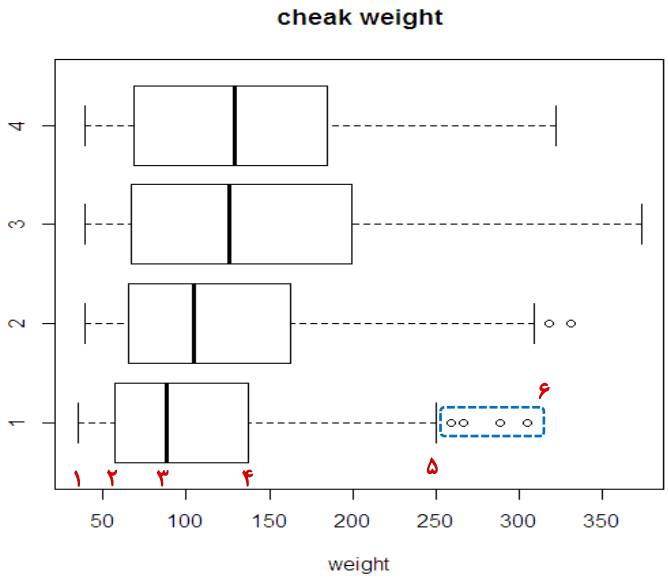
نمودار های جعبه ای برای مقایسه توزیع داده هایی در دو یا چند دسته، با چندین مشاهدات عددی (۱۰ یا بیشتر) در هر دسته به کار می روند.

برای رسم نمودار داده های ChickWeight را از سری داده های پیش فرض در نرم افزار انتخاب می کنیم. Nv در این سری داده قصد داریم نمودار را براساس متغیر weight را با توجه به متغیر Diet رسم کنیم. جهت رسم نمودار جعبه فرامین زیر را در صفحه فرمان نرم افزار بنویسید.

```
> boxplot(weight~Diet , data=ChickWeight, xlab="weight", main="cheak weight",
horizontal=TRUE)
```

```
> boxplot(weight~Diet , data=ChickWeight, xlab="weight", main="cheak weight", horizontal=TRUE)
```

- weight~Diet= انتخاب دو متغیر برای رسم. از آنجا مه در این مثال متغیر Diet دارای ۴ حالت است در نتیجه در خروجی نرم افزار ۴ نمودار جعبه برای هر یک از مشاهدات رسم شده است.
- data=ChickWeight= نام سری داده ای که در رسم نمودار از مشاهدات انها استفاده شده است.
- xlab="weight" عنوان محور x
- main="cheak weight" عنوان اصلی نمودار
- horizontal=TRUE= انتخاب رسم افقی نمودار. در صورت عدم تایپ این فرمان نمودار به صورت عمودی رسم می شود.



معرفی هر یک از قسمت های نمودار:

۱. کمینه
۲. چارک اول
۳. چارک دوم(میانه)
۴. چارک سوم
۵. بیشینه
۶. داده های پرت

نمودار شاخه و برگ

اگر مجموعه داده ها نسبتا کوچک باشد، نمودار شاخه و برگ برای مشاهده شکل توزیع و مقادیر بسیار مفید است. تعداد شماره ها در سمت چپ میله ی شاخه است و تعداد اعداد، سمت راست است. شما آنها را کنار یکدیگر قرار می دهید تا نمودار رسم شود. فرض کنید که جدول امتیازات بازی بستکبال را دارید و نمرات زیر را در هر بازی برای بازکنان در هر دو تیم در دست دارید.

۱، ۱۳، ۲۳، ۵۴، ۱۲، ۳۴، ۵۴، ۱۲، ۳۴، ۲۱، ۴۵، ۶۵، ۴۲، ۴۵، ۳۲، ۱، ۲، ۳۴، ۵، ۵، ۴، ۳، ۲، ۳۴، ۵۴، ۳۲

```
> y<-c(1,13,23,54,12,34,54,12,34,21,45,65,43,45,32,1,2,34,5,5,4,3,2,34,54,32)
> stem(y)
```

```
> y<-c(1,13,23,54,12,34,54,12,34,21,45,65,43,45,32,1,2,34,5,5,4,3,2,34,54,32)
> stem(y)

  The decimal point is 1 digit(s) to the right of the |

  0 | 11223455
  1 | 223
  2 | 13
  3 | 224444
  4 | 355
  5 | 444
  6 | 5
```

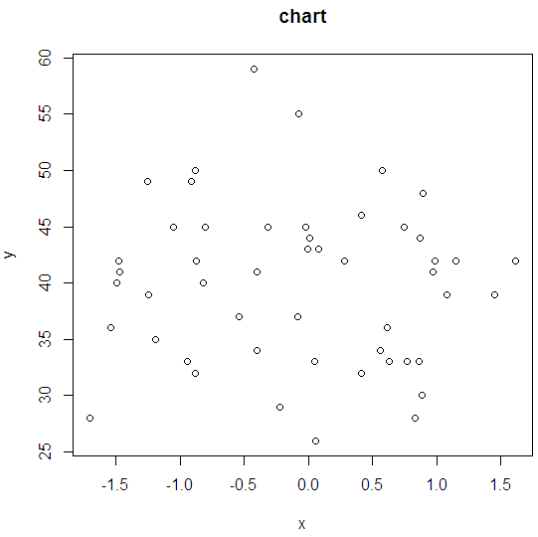
نمودار پراکندگی

هنگامی که از آمار استفاده می کنیم، اغلب مسائل از ارتباط بین اندازه های مختلف نشات می گیرد. برای بررسی این موضوع یکی از مهم ترین و رایج ترین نمودارها، نمودار پراکندگی است که نقاط (x_i, y_i) i =۱،۲،... را با علامت نقطه یا نشان های دیگر مشخص می کند. و ارتباط بین مقادیر x_i و y_i را نشان می دهد. در R نمودارهای پراکندگی (و بسیاری از نمودارها) از دستور plot() رسم می شوند. استفاده ی کلی از دستور plot(x,y,...) انجام می پذیرد که x و y بردارهای عددی با طول داده ی رسم شده هستند. آرگومان های بسیار دیگری در کنار انواع مختلف plot، برای داده های غیر عددی وجود دارد.

یک آرگومان مهم و قابل ذکر، type است. در R در حالت خودکار type=”p” است و نمودار پراکندگی را رسم می کند. نمودارهای خطی (که در آن بوسیله ی خطوط هر بازه، نقاط (x_i, y_i) از ابتدا تا انتها به هم وصل می شود) با آرگومان type=”l” رسم می شوند. انواع دیگر نیز وجود دارند که از جمله type=”n” است که چیزی را رسم نمی کند. این دستور تنها زمانی کاربرد دارد که بخواهیم دستورات دیگر را روی آن اجرا کنیم. برای

رسم نمودار از دو سری اعداد تصادفی از توزیع نرمال استاندارد و توزیع پواسن با میانگین ۴۰ در نظر می گیریم. فرامین زیر را در نرم افزار بنویسید.

```
> x<-rnorm(50)
> y<-rpois(50,40)
> plot(x,y,main="chart")
```



نمودار QQ

نمودارهای چندک-چندک(یا به عبارتی QQ) نوعی از نمودارهای پراکندگی هستند که برای مقایسه ی دو گروه از توزیع ها و یا مقایسه ی یک نمونه با توزیع مربوط به آن به کار می رود. در این حالت که در آن دو گروه به اندازه ی برابر جود دارند، نمودار های QQ ابتدا از ترتیبی کردن مشاهدات هر گروه به دست می آید:

$$X[1] \leq \cdots \leq X[n]$$

و

$$Y[1] \leq \cdots \leq Y[n]$$

سپس یک نمودار پراکندگی از (X[i],Y[i]) i=1,2,..,n رسم می شود.

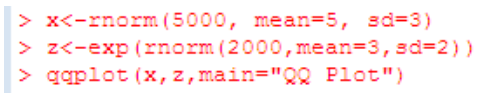
در حالتی که حجم گروه ها متفاوت است باید چند طرح ساده انجام شود تا دو گروه با هم هم حجم شوند. برای این کار R، حجم گروه بزرگتر را به حجم گروه کوچکتر کاهش می دهد، اما مقدار ماکسیمم و مینیمم آن را حفظ کرده و به طور مساوی چندک های بین آنها را انتخاب می کند. به طور مثال اگر ۵ مشاهده از X و ۲۰ مشاهده از Y داشته باشیم مقادیر X در مقابل مقادیر مینیمم، چارک اول، میانه، چارک سوم و ماکسیمم از Y رسم می شوند.

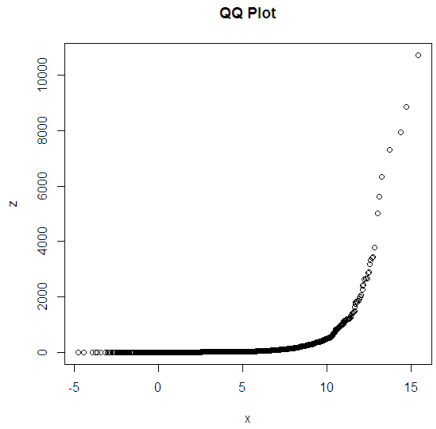
زمانی که یک نمونه ی تک عضوی در مقابل توزیع مربوط به آن رسم می شود، چندک های نظیر برای تنظیم آن در نظر گرفته می شود. R به طور خودکار چندک های نظیر روی محور X و داده ها را روی محور Y قرار می دهد. اما برخی دیگر از نرم افزار ها عمل عکس را انجام می دهند برای جلوگیری از اُریبی، چندک ها نسبت به احتمال $(i-\frac{1}{2})/n$ انتخاب و به صورت مساوی در میان صفر و ۱ پخش می شوند.

وقتی X و Y هم توزیع باشند، نقاط در نمودار QQ نزدیک به خط $y=x$ قرار می گیرد. اگر یک توزیع تبدیل خطی توزیع دیگری باشد، خط مستقیم متفاوتی را خواهیم دید. از طرف دیگر اگر دو توزیع یکسان نباشد، الگوی منظمی را خواهیم دید.

برای رسم نمودار QQ دو متغیر با عنوان x وy با توزیع نرمال با میانگین ۵، انحراف معیار ۳ و توزیع نمایی است.

```
> x<-rnorm(5000, mean=5, sd=3))
> z<-exp(rnorm(2000,mean=3,sd=2))
> qqplot(x,z,main="QQ Plot")
```





دستورات تکمیلی برای رسم نمودار در نرم افزار R

چند دستور دیگر برای اضافه کردن مولفه ها به ناحیه ی طرح در نمودار وجود دارد:

points(x,y,…)

lines(x,y,…) : بازه ی رسم خط را مشخص می کند

text(x,y,labels,…) : متن را به نمودار اضافه می کند.

abline(a,b,…) : خط y=a+bx را به نمودار اضافه می کند.

abline(h=y,…) : یک خط افقی را به نمودار اضافه می کند.

abline(v=x,…) یک خط عمودی را به نمودار اضافه می کند.

polygon(x,y,…) : یک چند ضلعی را به نمودار اضافه می کند.

segments(x0,y0,x0,x1,…) : خطوط بین بازه ها را رسم می کند.

arrows(x0,y0,x1,y1,…) : علامت فلش را رسم می کند.

symbols(x,y,…) : دایره ها،مربع ها، درجه بندی ها و ... را رسم می کند.

leged(x,y,legend,…) : فهرست علائم را رسم می کند.

آرگومان هایی همچون رنگ، اندازه و ... نیز می تواند به هر کدام از این دستورات اضافه شود.

منبع:

- مبانی برنامه نویسی آماری با R، تالیف: جان براون، دانکن مرداک، ترجمه: دکتر منوچهر بابانژاد،مبین ملکشاه
- استفاده از R در آمار مقدماتی، مولف:JOHN VERZAN، ترجمه: صغری طاهرخانی، نجم الدین گنج خانلو، علی فصیحی