محاسبات آماری پیشرفته

ترم اول سال تحصیلی ۹۳

جلسه ششم: نمایش گرافیکی داده‌ها

حسین باغیشنی

دانشگاه شاهرود

۱۶ آبان ۱۳۹۳

مقدمه

بحث این جلسه خیلی گسترده است و نمی‌توان همه چیز را پوشاند.

مقدمه

بحث این جلسه خیلی گسترده است و نمی‌توان همه چیز را پوشاند.

در واقع همراه با هر روش آماری، نمایش‌های گرافیکی برای بررسی وضعیت موجود، تحلیل نیکویی مدل در نظر گرفته شده، پیدا کردن برخی مشکلات، بهتر کردن مدل قبلی و خیلی موارد دیگر، به کار می‌روند.

مقدمه

بحث این جلسه خیلی گسترده است و نمی‌توان همه چیز را پوشاند.

در واقع همراه با هر روش آماری، نمایش‌های گرافیکی برای بررسی وضعیت موجود، تحلیل نیکویی مدل در نظر گرفته شده، پیدا کردن برخی مشکلات، بهتر کردن مدل قبلی و خیلی موارد دیگر، به کار می‌روند.

نمایش گرافیکی، بیشتر مربوط به تحلیل اکتشافی داده‌ها ($ExploratoryDataAnalysis$) (EDA) و نمودارهای آماری است.

مقدمه

بحث این جلسه خیلی گسترده است و نمی‌توان همه چیز را پوشاند.

در واقع همراه با هر روش آماری، نمایش‌های گرافیکی برای بررسی وضعیت موجود، تحلیل نیکویی مدل در نظر گرفته شده، پیدا کردن برخی مشکلات، بهتر کردن مدل قبلی و خیلی موارد دیگر، به کار می‌روند.

نمایش گرافیکی، بیشتر مربوط به تحلیل اکتشافی داده‌ها ($ExploratoryDataAnalysis$) (EDA) و نمودارهای آماری است.

اصطلاح اکتشافی در مقابل تاییدی است که روش‌های تاییدی را می‌توان به آزمون فرضیه‌ها منتسب کرد.

مقدمه

بحث این جلسه خیلی گسترده است و نمی‌توان همه چیز را پوشاند.

در واقع همراه با هر روش آماری، نمایش‌های گرافیکی برای بررسی وضعیت موجود، تحلیل نیکویی مدل در نظر گرفته شده، پیدا کردن برخی مشکلات، بهتر کردن مدل قبلی و خیلی موارد دیگر، به کار می‌روند.

نمایش گرافیکی، بیشتر مربوط به تحلیل اکتشافی داده‌ها ($ExploratoryDataAnalysis$) (EDA) و نمودارهای آماری است.

اصطلاح اکتشافی در مقابل تاییدی است که روش‌های تاییدی را می‌توان به آزمون فرضیه‌ها منتسب کرد.

در واقع توکی، معتقد بود قبل از آزمون فرضیه، بهتر است تحلیل اکتشافی بر روی داده‌ها انجام شود تا سوالات مناسب برای پرسیدن و روش‌های مناسب برای پاسخ به آن‌ها را بیابیم.

مقدمه

بحث این جلسه خیلی گسترده است و نمی‌توان همه چیز را پوشاند.

در واقع همراه با هر روش آماری، نمایش‌های گرافیکی برای بررسی وضعیت موجود، تحلیل نیکویی مدل در نظر گرفته شده، پیدا کردن برخی مشکلات، بهتر کردن مدل قبلی و خیلی موارد دیگر، به کار می‌روند.

نمایش گرافیکی، بیشتر مربوط به تحلیل اکتشافی داده‌ها ($ExploratoryDataAnalysis$) (EDA) و نمودارهای آماری است.

اصطلاح اکتشافی در مقابل تاییدی است که روش‌های تاییدی را می‌توان به آزمون فرضیه‌ها منتسب کرد.

در واقع توکی، معتقد بود قبل از آزمون فرضیه، بهتر است تحلیل اکتشافی بر روی داده‌ها انجام شود تا سوالات مناسب برای پرسیدن و روش‌های مناسب برای پاسخ به آن‌ها را بیابیم.

در این جلسه، قسمتی از موارد مهم و شاید کمی جذاب‌تر را (از دیدگاه مدرس) مطرح می‌کنیم.

نمودارهای چندک ـ چندک

یکی از مواردی که در مورد داده‌ها مایلیم بدانیم، این است که آیا آن‌ها از توزیع نرمال پیروی می‌کنند؟ آیا از یک توزیع شناخته‌شده پیروی می‌کنند؟

نمودارهای چندک ـ چندک

یکی از مواردی که در مورد داده‌ها مایلیم بدانیم، این است که آیا آن‌ها از توزیع نرمال پیروی می‌کنند؟ آیا از یک توزیع شناخته‌شده پیروی می‌کنند؟

در بعضی از موارد، واضح است که داده‌ها مثلا از نرمال پیروی نمی‌کنند. زیرا به عنوان مثال چگالی داده‌ها دو مدی است. اما مواردی هم وجود دارند که چنین آگاهی واضح و روشن نیست.

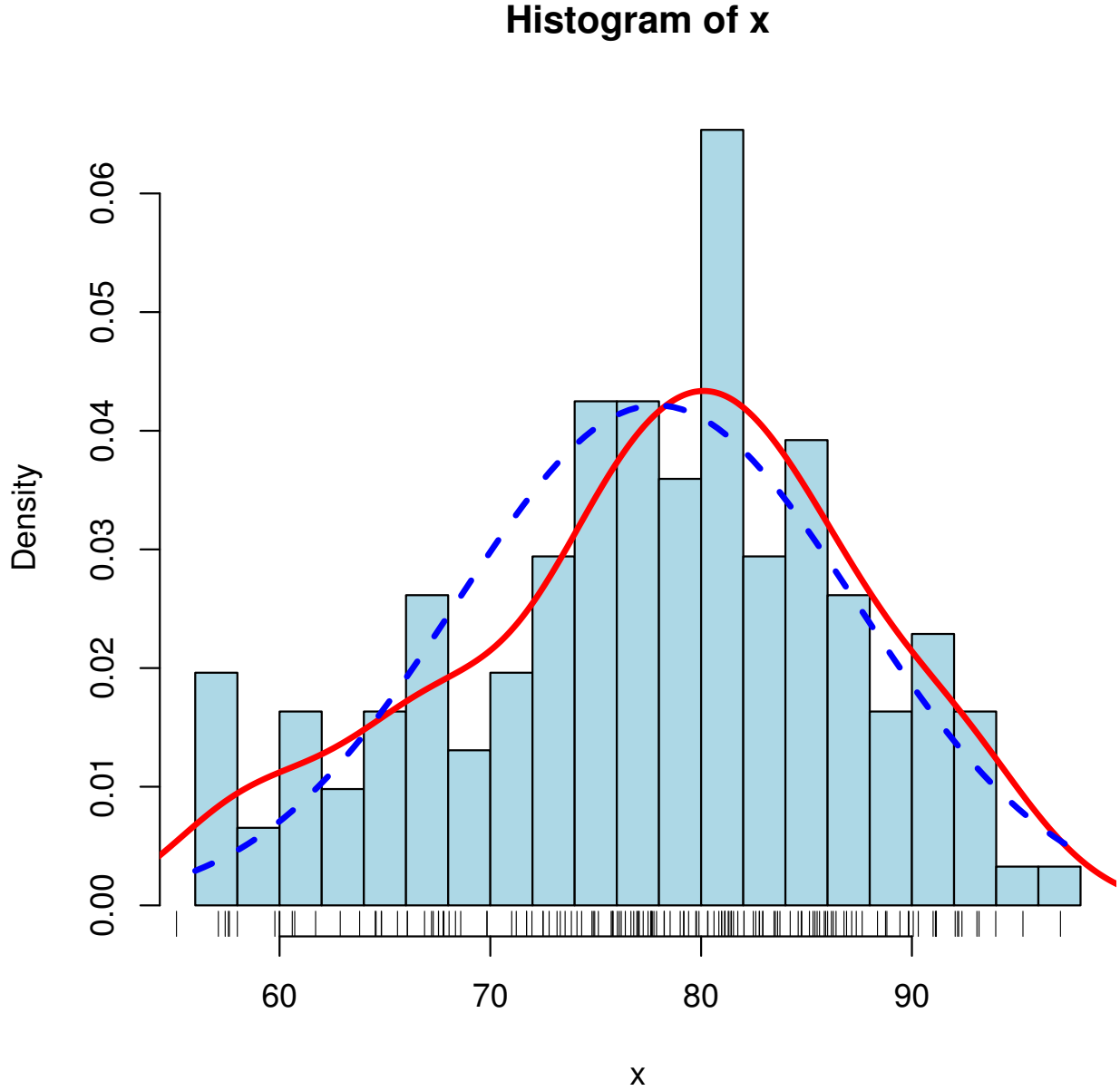
نمودارهای چندک ـ چندک

یکی از مواردی که در مورد داده‌ها مایلیم بدانیم، این است که آیا آن‌ها از توزیع نرمال پیروی می‌کنند؟ آیا از یک توزیع شناخته‌شده پیروی می‌کنند؟

در بعضی از موارد، واضح است که داده‌ها مثلا از نرمال پیروی نمی‌کنند. زیرا به عنوان مثال چگالی داده‌ها دو مدی است. اما مواردی هم وجود دارند که چنین آگاهی واضح و روشن نیست.

یک راه برای بررسی این موضوع، مقایسه چگالی برآوردشده با چگالی توزیع نرمال است.

```
data(airquality)
x <- airquality[,4]
hist(x, probability=TRUE, breaks=20, col="light blue")
rug(jitter(x, 5))
points(density(x), type='l', lwd=3, col='red')
f <- function(t) {
  dnorm(t, mean=mean(x), sd=sd(x) )
}
curve(f, add=T, col="blue", lwd=3, lty=2)
```


Histogram of x
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
60
70
80
90
x

راه حل دوم

روش دیگر که یک روش بصری است، رسم چندک‌های نمونه‌ای در مقابل چندک‌های نرمال است. به این نمودار، چندک‌ـ‌چندک گویند.

راه حل دوم

روش دیگر که یک روش بصری است، رسم چندک‌های نمونه‌ای در مقابل چندک‌های نرمال است. به این نمودار، چندک‌ـ‌چندک گویند.

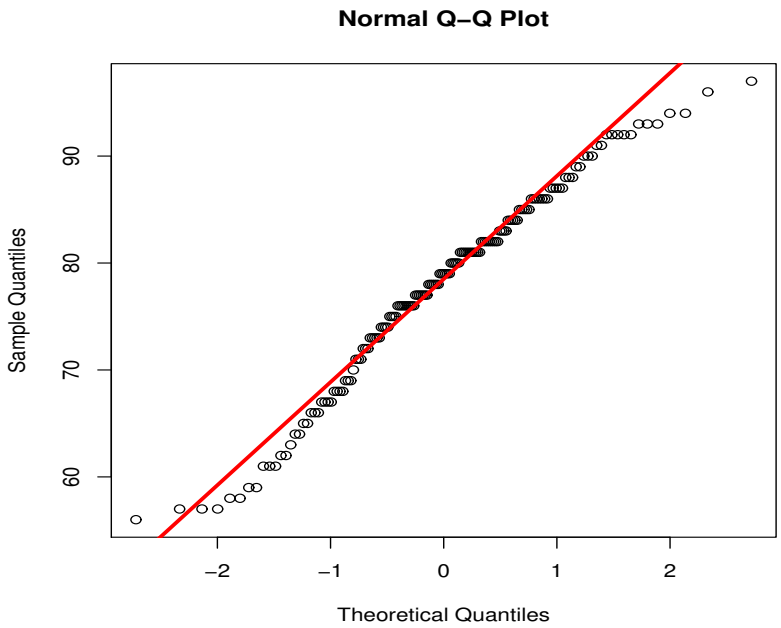
در نمودار زیر به نظر می‌رسد داده‌ها تقریبا نرمال هستند، اما می‌توان دید که داده‌ها گسسته‌اند!!

راه حل دوم

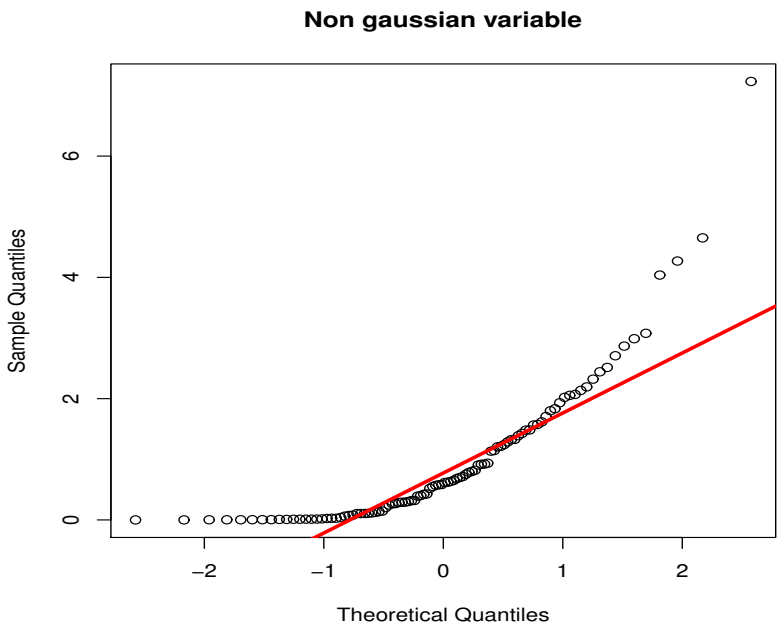
روش دیگر که یک روش بصری است، رسم چندک‌های نمونه‌ای در مقابل چندک‌های نرمال است. به این نمودار، چندک‌ـ‌چندک گویند.

در نمودار زیر به نظر می‌رسد داده‌ها تقریبا نرمال هستند، اما می‌توان دید که داده‌ها گسسته‌اند!!

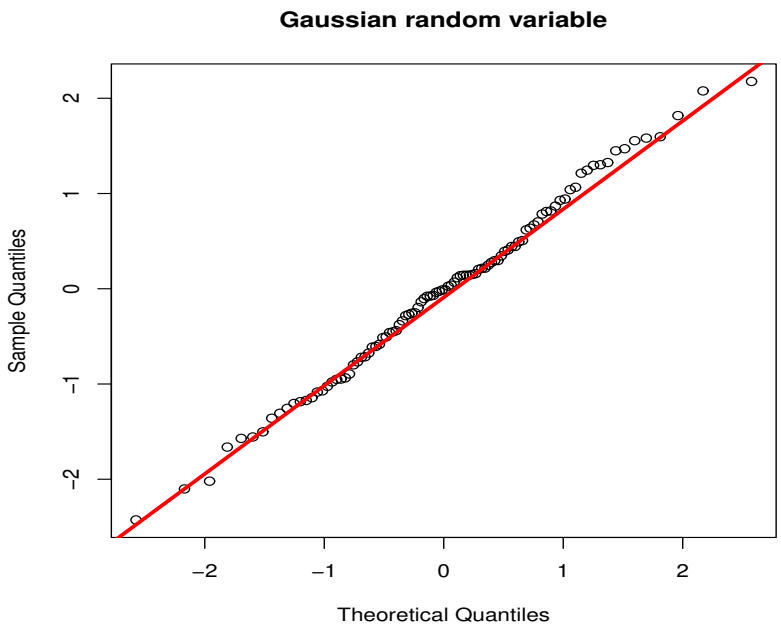
```
x <- airquality[,4]
qqnorm(x)
qqline(x, col="red", lwd=3)
```



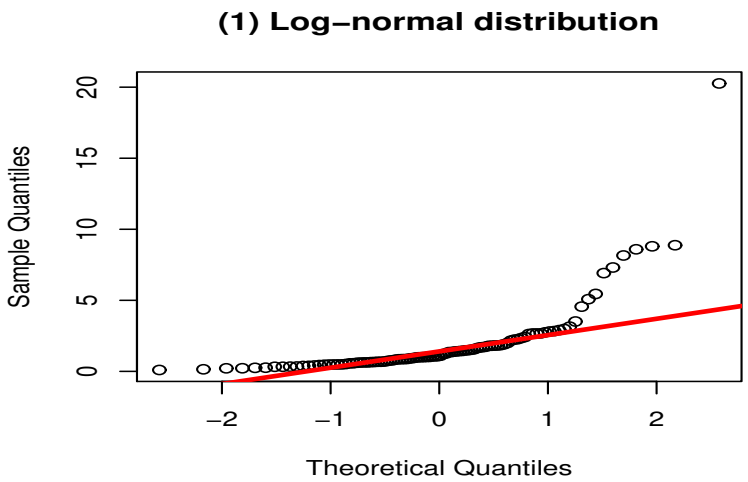
```
y <- rnorm(100)^2
qqnorm(y, main="Non gaussian variable")
qqline(y, col="red", lwd=3)
```

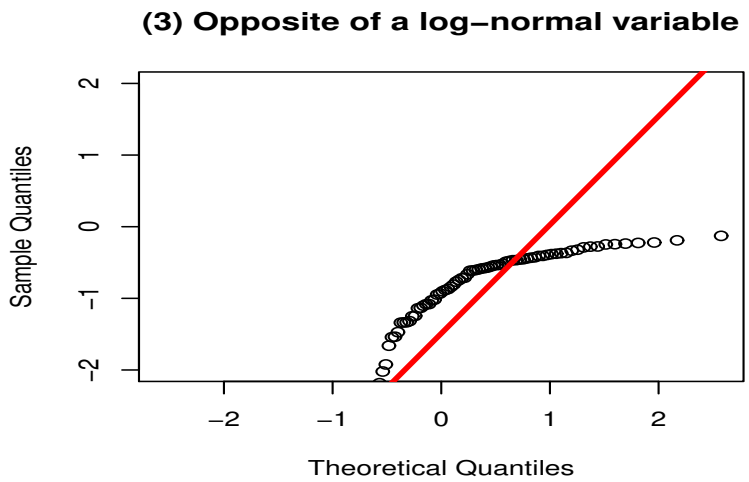


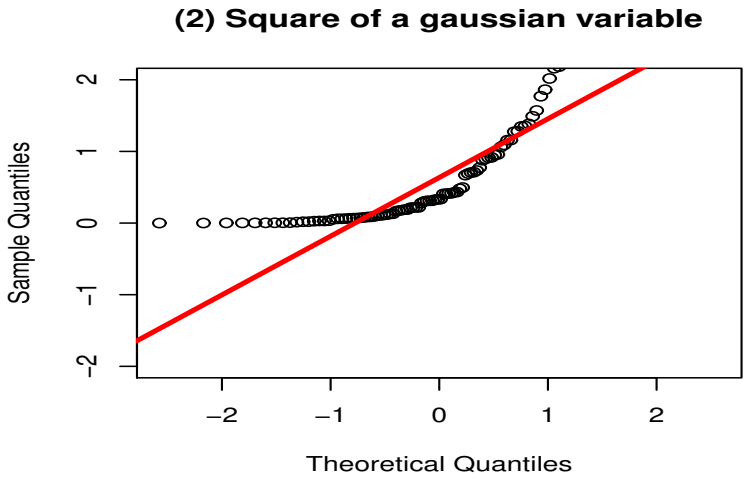

```
y <- rnorm(100)
qqnorm(y, main="Gaussian random variable")
qqline(y, col="red", lwd=3)
```

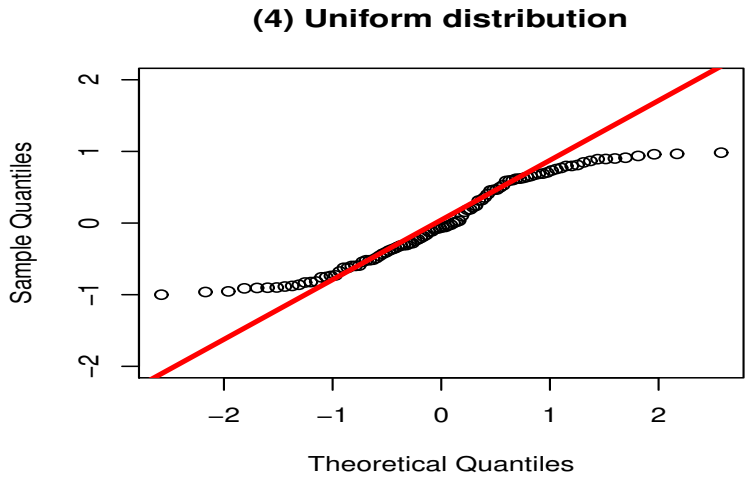


سایر مثال‌ها

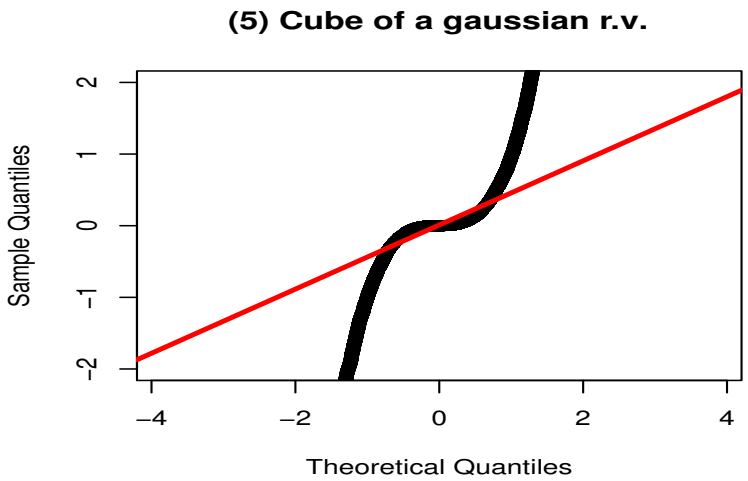

(1) Log−normal distribution
Sample Quantiles
Theoretical Quantiles
0
5
10
15
20
−2
−1
0
1
2

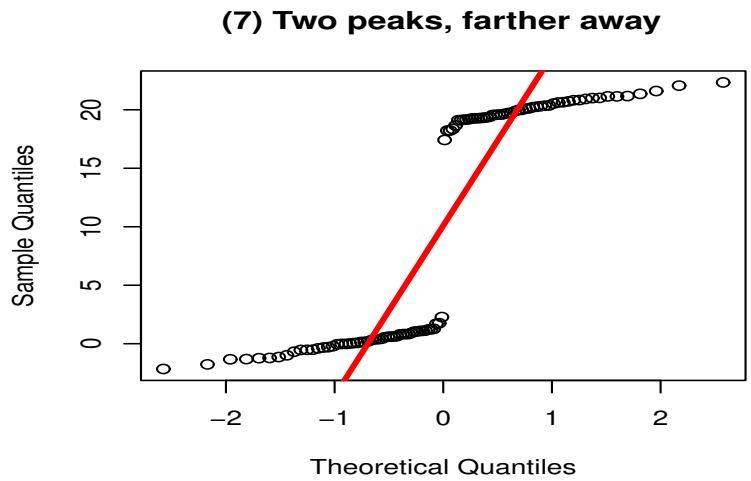

(3) Opposite of a log−normal variable
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

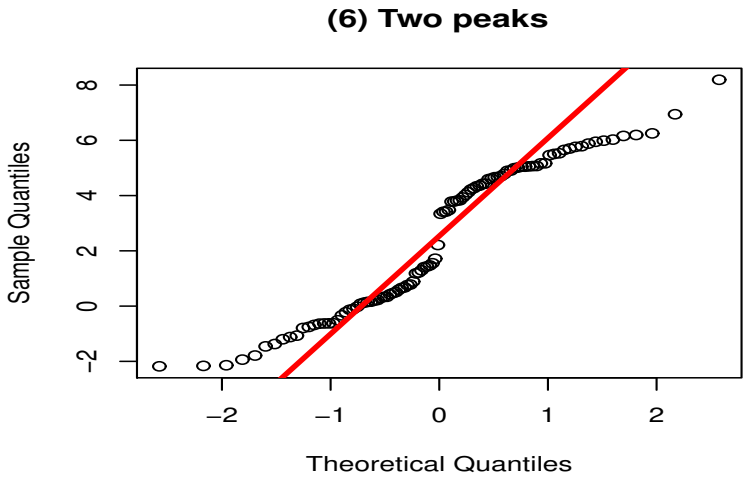

(2) Square of a gaussian variable
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

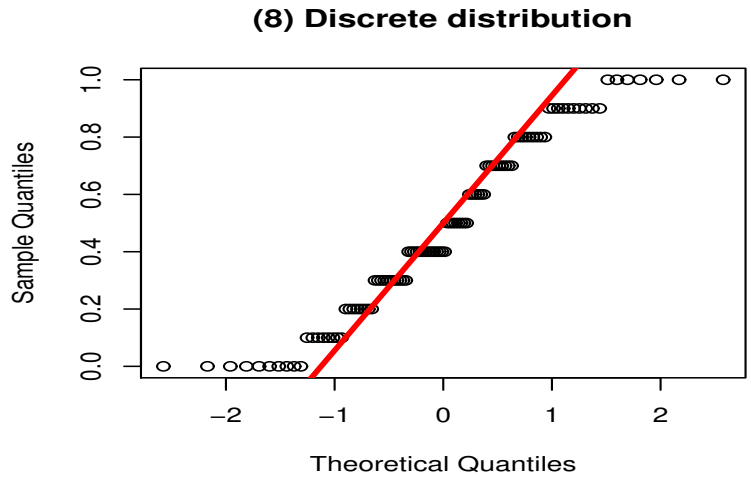

(4) Uniform distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

سایر مثال‌ها


(5) Cube of a gaussian r.v.
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
4
−2
−1
0
1
2

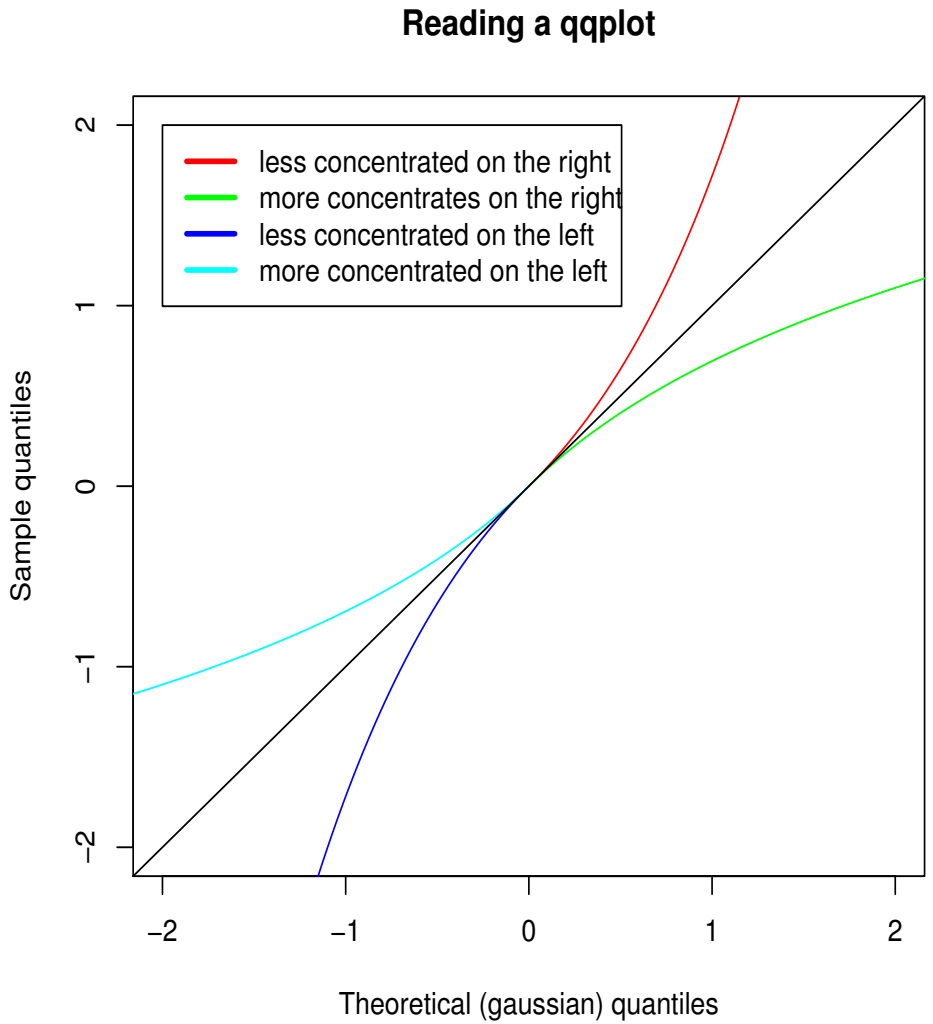

(7) Two peaks, farther away
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
0
5
10
15
20


(6) Two peaks
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−2
0
2
4
6
8

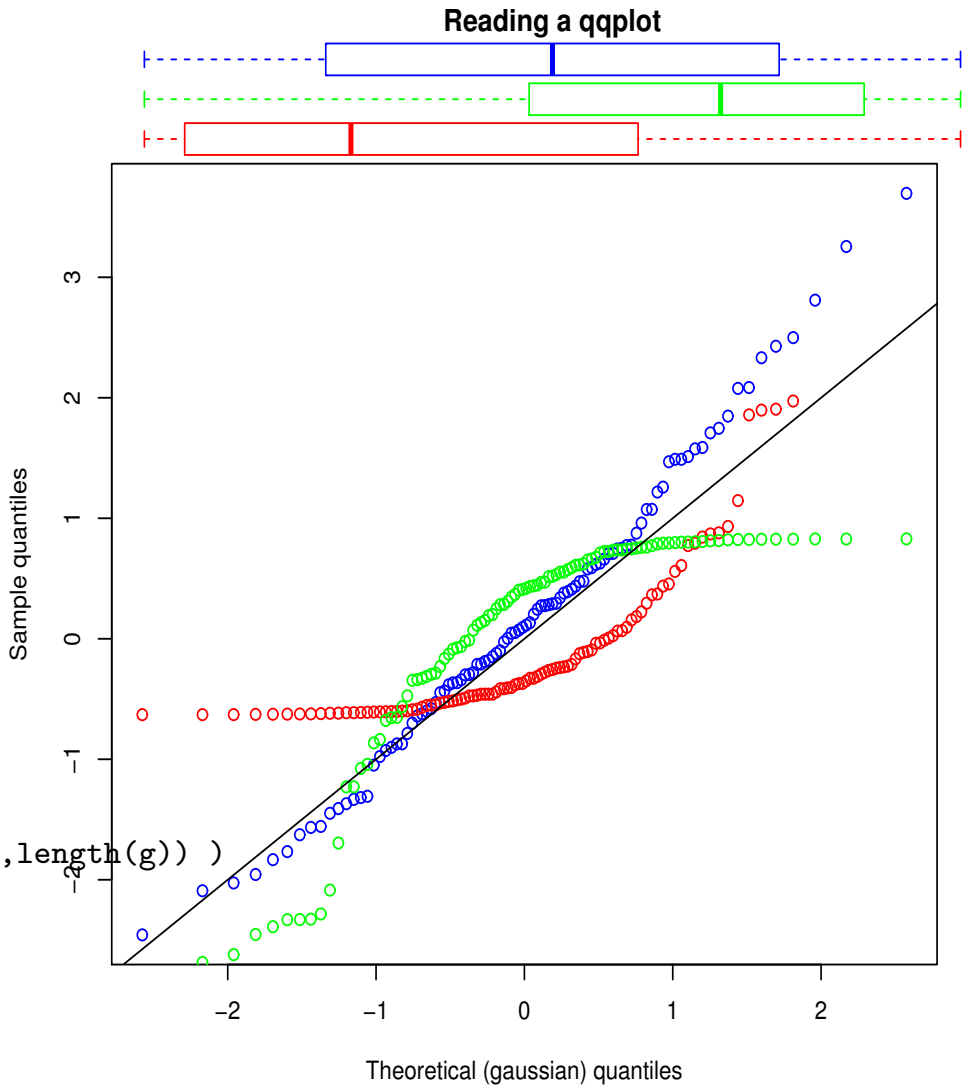

(8) Discrete distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0

دستورالعمل کلی

```
x <- seq(from=0, to=2, length=100)
y <- exp(x)-1
plot( y ~ x, type = 'l', col = 'red',
      xlim = c(-2,2), ylim = c(-2,2),
      xlab = "Theoretical (gaussian) quantiles",
      ylab = "Sample quantiles")
lines( x~y, type='l', col='green')
x <- -x
y <- -y
lines( y~x, type='l', col='blue', )
lines( x~y, type='l', col='cyan')
abline(0,1)
legend( -2, 2,
        c( "less concentrated on the right",
           "more concentrates on the right",
           "less concentrated on the left",
           "more concentrated on the left"
         ),
        lwd=3,
        col=c("red", "green", "blue", "cyan")
      )
title(main="Reading a qqplot")
```



```
op <- par()
layout( matrix( c(2,2,1,1), 2, 2, byrow=T ),
        c(1,1), c(1,6),
      )
# The plot
n <- 100
y <- rnorm(n)
x <- qnorm(ppoints(n))[order(order(y))]
par(mar=c(5.1,4.1,0,2.1))
plot( y ~ x, col = "blue",
      xlab = "Theoretical (gaussian) quantiles",
      ylab = "Sample quantiles" )
y1 <- scale( rnorm(n)^2 )
x <- qnorm(ppoints(n))[order(order(y1))]
lines(y1~x, type="p", col="red")
y2 <- scale( -rnorm(n)^2 )
x <- qnorm(ppoints(n))[order(order(y2))]
lines(y2~x, type="p", col="green")
abline(0,1)
# The legend
par(bty='n', ann=F)
g <- seq(0,1, length=10)
e <- g^2
f <- sqrt(g)
h <- c( rep(1,length(e)), rep(2,length(f)), rep(3,length(g)) )
par(mar=c(0,4.1,1,0))
boxplot( c(e,f,g) ~ h, horizontal=T,
         border=c("red", "green", "blue"),
         col="white", # Something prettier?
         xaxt='n',
         yaxt='n',
         )
title(main="Reading a qqplot")
par(op)
```



چندک_چندک برای سایر توزیع‌ها

می‌توان برای سایر توزیع‌ها از همین ایده نمودار چندک_چندک نرمال استفاده کرد

```
qq <- function (y, ylim, quantiles=qnorm,
    main = "Q-Q Plot", xlab = "Theoretical Quantiles",
    ylab = "Sample Quantiles", plot.it = TRUE, ...)
{
  y <- y[!is.na(y)]
  if (0 == (n <- length(y)))
    stop("y is empty")
  if (missing(ylim))
    ylim <- range(y)
  x <- quantiles(ppoints(n))[order(order(y))]
  if (plot.it)
    plot(x, y, main = main, xlab = xlab,
    ylab = ylab, ylim = ylim, ...)
  # From qqline
  y <- quantile(y, c(0.25, 0.75))
  x <- quantiles(c(0.25, 0.75))
  slope <- diff(y)/diff(x)
  int <- y[1] - slope * x[1]
  abline(int, slope, ...)
  invisible(list(x = x, y = y))
}
y <- runif(100)
qq(y, quantiles=qunif)
```