Assessing normality with normal Quintilian Quintilian plots

by Craig W. Slinkman

November 15, 2015

An important aspect of analyzing data is to determine if a variable is approximately normally distributed. It is advantageous if the data is normally distributed because there are a wide number of statistical procedures that require the variables of interest be normally distributed.

# Basic theory

In statistics, a Q-Q plot[1] ("Q" stands for quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the (number of the) interval for the quantile.

If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line y = x. Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Quantile-Quantile (QQ) plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. QQ plots can be used to compare collections of data, or theoretical distributions. The use of QQ plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions.

A QQ plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples, but requires more skill to interpret. QQ plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary. QQ plots are also used to compare two theoretical distributions to each other. Since Q-Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

## Symetric distributions (<https://en.wikipedia.org/wiki/Kurtosis>)

In probability theory and statistics, kurtosis (from Greek: ????????????, kyrtos or kurtos, meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population. There are various interpretations of kurtosis, and of how particular measures should be interpreted; these are primarily tail weight, peakedness (width of peak), and lack of shoulders (distribution primarily peak and tails, not in between).

The standard measure of kurtosis, originating with Karl Pearson, is based on a scaled version of the fourth moment of the data or population. This number measures heavy tails, and not peakedness;hence, the "peakedness" definition is misleading. For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations.

The kurtosis of any univariate normal distribution is 3. It is common to compare the kurtosis of a distribution to this value. Distribution with kurtosis less than 3 are said to be **platykurtic**. An example of a platykurtic distribution is the uniform distribution, which does not have positive-valued tails. Distributions with kurtosis greater than 3 are said to be leptokurtic. An example of a leptokurtic distribution is the Laplace distribution, which has tails that asymptotically approach zero more slowly than a Gaussian. It is also common practice to use an adjusted version of Pearson's kurtosis, the excess kurtosis, which is the kurtosis minus 3, to provide the comparison to the normal distribution. Some authors useThe kurtosis of any univariate normal distribution is 3. It is common to compare the kurtosis of a distribution to this value. Distribution with kurtosis less than 3 are said to be **platykurtic**. An example of a platykurtic distribution is the uniform distribution, which does not have positive-valued tails. Distributions with kurtosis greater than 3 are said to be **leptokurtic**. An example of a leptokurtic distribution is the Laplace distribution, which has tails that asymptotically approach zero more slowly than a normal distributiion . It is also common practice to use an adjusted version of Pearson's kurtosis, the excess kurtosis, which is the kurtosis minus 3, to provide the comparison to the normal distribution. Some authors use "kurtosis" by itself to refer to the excess kurtosis. For the reason of clarity and generality, however, this articlepresentation follows the non-excess convention and explicitly indicates where excess kurtosis is meant. "kurtosis" by itself to refer to the excess kurtosis. For the reason of clarity and generality, however, this article follows the non-excess convention and explicitly indicates where excess kurtosis is meant.

## Platykurtic distribution

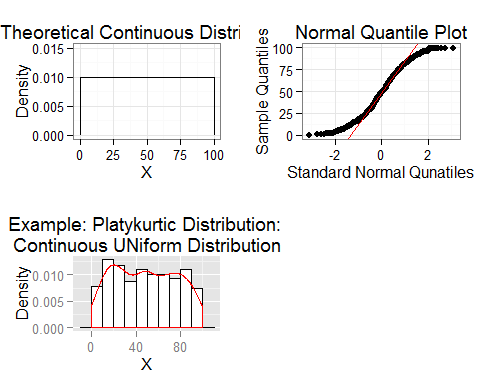
A platykurtic distribution is one in which most of the values share about the same frequency of occurrence. As a result, the curve is very flat, or plateau-like. A coin toss is an example of a particularity distribution because the .

In this document we will use a uniform random distribution with a mean of 50.00 and a standard deviation of 28.87. The code below shows how to plot the theoretical density diagram, generate a random sample of size 500 from a continuous uniform distribution with limits 0 and 100, and the normal QQ plot of the sample.

rm( list=ls() )  
set.seed( 11 )  
source( "ggQQ.r" )  
source( "multiplot.r ")  
  
require( ggplot2 )

## Loading required package: ggplot2

x <- c( 0, 100 )  
y <- c( 0.01, 0.01 )  
DF <- data.frame( x, y )  
p1 <-   
 ggplot( DF,  
 aes(X, y)) +  
 scale\_y\_continuous(limits = c( 0, 0.015)) +  
 annotate( "segment", x=0.00, y=0.01, xend=100.00, yend=0.01 ) +  
 annotate( "segment", x=0.00, y=0.00, xend=0.00, yend=0.01 ) +  
 annotate( "segment", x=100.00, y=0.00, xend=100.00, yend=0.01 ) +  
 xlab( "X" ) +  
 ylab( "Density" ) +  
 ggtitle("Theoretical Continuous Distribution" ) +  
 theme\_bw()  
  
x <- runif( 500, 0, 100 )  
DF <- data.frame( x )  
  
p2 <-   
 ggplot( DF,  
 aes( x=x )) +  
 geom\_histogram( aes( y=..density.. ),  
 binwidth=10,  
 color="black",  
 fill="white" ) +  
 geom\_density( color="red" ) +  
 xlab( "X" ) +  
 ylab( "Density" ) +  
 ggtitle( "Example: Platykurtic Distribution: \nContinuous UNiform Distribution" )  
  
p3 <- ggQQ( x )   
   
multiplot( p1, p2, p3, cols=2)



We that the normal QQ plot looks like a elongated **S**. The distribution has shorter tails than a normal distribution with the same standard deviations. As a result the central theorem works well for relatively small samples sizes. Empirical investigations shows that for sample sizes as low as 12 the resulting sampling distributions of the sample for uniform distributions is approximately normal distribution.

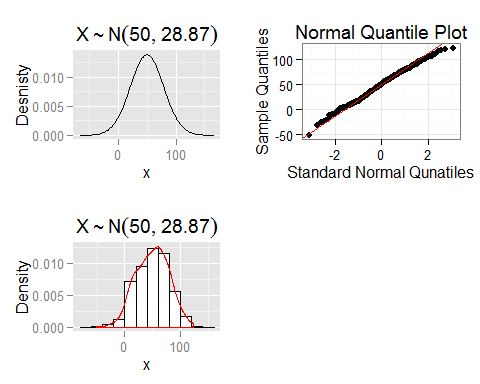
If you are working with platykurtic that are not uniform we expect that even small samples size 12 or larger will be approximately normally distributed. We can always bootstrap the sample to evaluate the sampling distribution for normality.

## The Normal distribution

It can be shown that if we sample from a normal distribution that the sample points on a normal QQ plot should be approximately linear. We know however that due to random sampling and other effects that there will be random variations around the theoretical expected line. Also the points at the end of the line will have the greater variability so we can discount them.

Below we who the normal distribution with a mean of 50 and a standard devation of 28.87 to match our continuous uniform distribution. As before, we draw the theoretical distribution, the histogram and the normal probability plots.

rm( list=ls())  
set.seed( 1123 )  
require( ggplot2)  
source("ggQQ.r")  
source( "multiplot.r ")  
  
mu <- 50  
sigma <- 28.87  
  
xl <- floor( mu - 4 \* sigma )  
xh <- ceiling( mu + 4 \* sigma )  
x <- c( xl, xh )  
DF <- data.frame( x )  
  
p1 <- ggplot( DF,   
 aes( x=x )) +  
 stat\_function( fun=dnorm, args=list( mean=mu, sd=sigma) ) +  
 ylab( "Desnisty" ) +  
 ggtitle( expression( X %~% N(50,28.87) ) )  
  
x <- rnorm( 500, mean=mu, sd=sigma )  
DF <- data.frame( x )  
  
p2 <-   
 ggplot( DF,  
 aes( x=x )) +  
 geom\_histogram( aes( y=..density.. ),  
 binwidth=20,  
 color="black",  
 fill="white" ) +  
 geom\_density( color="red" ) +  
 ylab( "Density" ) +  
 ggtitle( expression(X %~% N(50,28.87) ) )  
  
p3 <- ggQQ( x )  
  
multiplot( p1, p2, p3, cols=2)

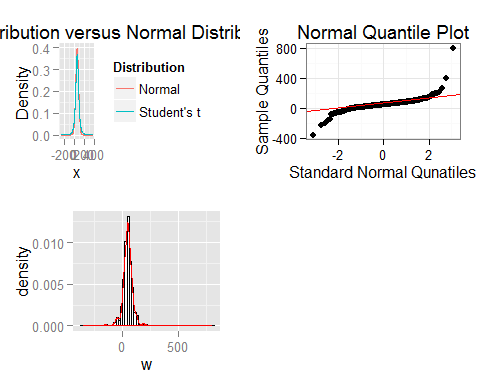


## Deptokurtic distributions

A distribution with positive excess kurtosis is called leptokurtic, or leptokurtotic. "Lepto-" means "slender". In terms of shape, a leptokurtic distribution has a more acute peak around the mean and fatter tails. Student's t distribution is leptohurtic when the degrees fo freedom are small.

Our simulated sample will be drean from a students t-distribution with three degrees of freedom. We will scale the distribution to have a mean of 50.00 and a standard deviation of 28.97. we will also include a plot comparing the dormal distribution to the t distribution.

rm( list=ls())  
require( ggplot2 )  
source( "ggQQ.r" )  
source( "multiplot.r")  
  
mu <- 50  
sigma <- 28.97  
xmin <- -270  
xmax <- 370  
x <- seq( xmin, xmax, 0.50 )  
fz <- dnorm( ( x-mu ) / sigma )  
ft <- dt( ( x-mu ) / sigma, df=3 )  
f <- c( fz, ft )  
n <- length( fz )  
Distribution <- c( rep( "Normal", n ),  
 rep( "Student's t", n ))  
x <- rep( x, 2)  
  
DF <- data.frame( x, f, Distribution )  
  
p1 <-  
 ggplot( DF,  
 aes( x=x, y=f, color=Distribution)) +  
 geom\_line() +  
 ylab( "Density ") +  
 ggtitle( "Leptokurtic Distribution versus Normal Distribution" )  
  
w <- mu + sigma \* rt( 500, df=3 )  
DF <- data.frame( w )  
  
p2 <-  
 ggplot( DF,  
 aes( x=w )) +  
 geom\_histogram( aes( y=..density.. ),  
 binwidth=20,  
 color="black",   
 fill="white" ) +  
 geom\_density( color="red" )  
  
p3 <- ggQQ( w )  
  
multiplot( p1, p2, p3, cols=2)

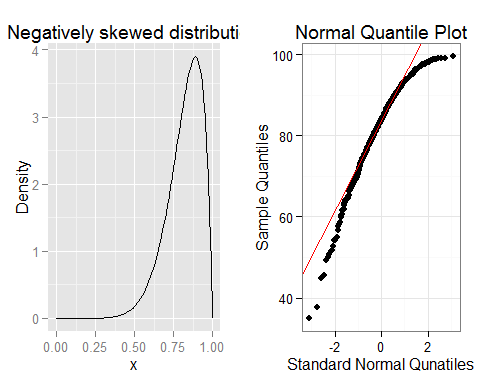


# Skewness

In this section we plot skewed data. We will have both distributions with negative and positive skewness.

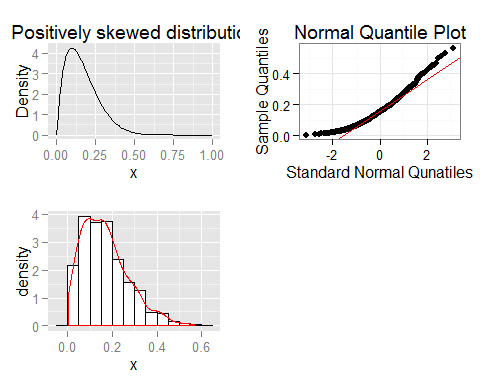
## Negative skewness

rm( list=ls())  
require( ggplot2 )  
source("ggQQ.r")  
source( "multiplot.r" )  
  
alpha <- 1  
beta <- 10  
  
x <- c(0,1)  
DF <- data.frame( x )  
  
p1 <-  
 ggplot( DF,  
 aes( x=x ))+  
 stat\_function( fun=dbeta,   
 args=list( shape1=9, shape=2 )) +  
 ylab( "Density" ) +  
 ggtitle( "Negatively skewed distribution" )  
  
x <- 100 \* rbeta( 500, shape1=9, shape2=2 )  
  
DF <- data.frame( x )  
  
p2 <- ggQQ( x )  
  
multiplot( p1, p2, cols=2 )



## Positive skewness

rm( list=ls())  
require( ggplot2 )  
source("ggQQ.r")  
source( "multiplot.r" )  
  
alpha <- 2  
beta <- 10   
  
x <- c(0,1)  
DF <- data.frame( x )  
  
p1 <-  
 ggplot( DF,  
 aes( x=x ))+  
 stat\_function( fun=dbeta,   
 args=list( shape1=alpha, shape2=beta )) +  
 ylab( "Density" ) +  
 ggtitle( "Positively skewed distribution" )  
  
x <- rbeta( 500, shape1= alpha, shape2=beta )  
  
DF <- data.frame( x )  
  
p2 <-   
 ggplot( DF,  
 aes( x= x)) +  
 geom\_histogram(   
 aes( y=..density.. ),  
 binwidth=0.05,  
 color="black",  
 fill="white") +  
 geom\_density( color="red" )   
  
p3 <- ggQQ( x )  
  
multiplot( p1, p2, p3, cols=2 )



# Practice

The rcode below provides a quick plot that will allow you to use graphics diagnose if a distribution is normally distributed. It will also allow to specify the departures from normality,

library( alr4 )

## Loading required package: car  
## Loading required package: effects  
##   
## Attaching package: 'effects'  
##   
## The following object is masked from 'package:car':  
##   
## Prestige

source( "diagnose.r")  
  
data( BGSgirls)  
head( BGSgirls)

## WT2 HT2 WT9 HT9 LG9 ST9 WT18 HT18 LG18 ST18 BMI18 Soma  
## 67 13.6 87.7 32.5 133.4 28.4 74 56.9 158.9 34.6 143 22.5 5.0  
## 68 11.3 90.0 27.8 134.8 26.9 65 49.9 166.0 33.8 117 18.1 4.0  
## 69 17.0 89.6 44.4 141.5 31.9 104 55.3 162.2 35.1 143 21.0 5.5  
## 70 13.2 90.3 40.5 137.1 31.8 79 65.9 167.8 39.3 148 23.4 5.5  
## 71 13.3 89.4 29.9 136.1 27.7 83 62.3 170.9 36.3 152 21.3 4.5  
## 72 11.3 85.5 22.8 130.6 23.4 60 47.4 164.9 31.8 126 17.4 3.0

weight <- BGSgirls$WT2 #Girls weight at 2 years  
  
diagnose( weight )

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.

