STATISTICAL DATA ANALYSIS: EXERCISES

GÜNTHER SAWITZKI

Contents

1. Setup

These are instructions to set up the graphical system and random number generations helpful for print manuscript preparation. You can ignore this section for now.

Input

```
save.RNGseed <- 87149 #.Random.seed
save.RNGkind <- RNGkind()
# save.RNGseed
save.RNGkind
```

Output

```
[1] "Mersenne-Twister" "Inversion"
```

Input

```
set.seed(save.RNGseed, save.RNGkind[1])
```

Input

```
# install.packages("sintro",repos="http://r-forge.r-project.org",type="source")
library(sintro)
```

Please add your name and student identification.

Date: November 5, 2016.
Key words and phrases. data analysis, distribution diagnostics, parallel coordinates.
This waste book is a companion to "G. Sawitzki: Statistical Data Analysis"
Typeset, with minor revisions: November 5, 2016 from cvs *Revision* : 1.1
gs@statlab.uni-heidelberg.de .

2. 26. Oct. - Onedim diagnostic

Revised: 5.Nov.

So far, in this course we used the (F_n, F) for the distribution function. This approximately gives a diagonal, if F_n is approximately F, as is the case if F_n is the empirical distribution function based on a sample from F.

For the exercises, we use the QQ-plot, the quantile-quantile plot (Q_n, Q). The quantile function is roughly the inverse of the distribution function. Again, this approximately gives a diagonal, if Q_n is approximately Q.

In both plots, a transposed image may be shown (F, F_n)

You can do simulation and generate histograms and distributions in any statistical software. R is recommended.

In R, for a standard distribution $xxxx$, $rxxxx()$ gives random numbers, $pxxxx$ gives the distributions function, $qxxxx$ the quantile function.

For the special case of the normal distribution, the QQ-plot is provided as $qqnorm()$. $qqline()$ adds a line which passes through the quantiles, by default the first and third quartiles. For example,

Input

```
x <-  rnorm(10)
qqnorm(x)
qqline(x)
```

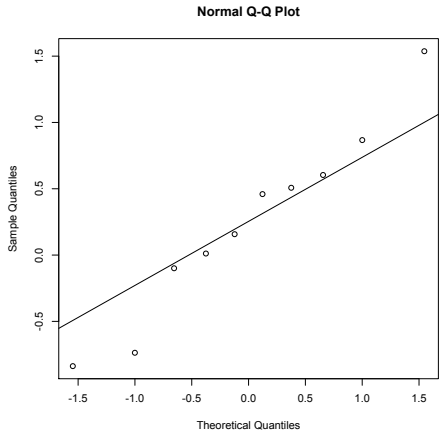

Figure 1. Normal qq-plot

Exercise 1. *Use the distribution function plot with data from a uniform distribution. What is the minimum sample size n so that it can be used as a test for uniformity with an error level not exceeding5%?*

Input

```
par(mfrow=c(4,5)) # set the graphics for multiple displays
n <- 10
for  (i in 1:20) plot( ecdf( runif(n) ),
        main="", xlab="" ) # repeated plots, no main title, no x label
```

Note: 5% is used as a level to keep the time you will need restricted. Using 5% as a limit means that if you use this as a routine procedure, your are accepting one wrong result in 20 as an average.

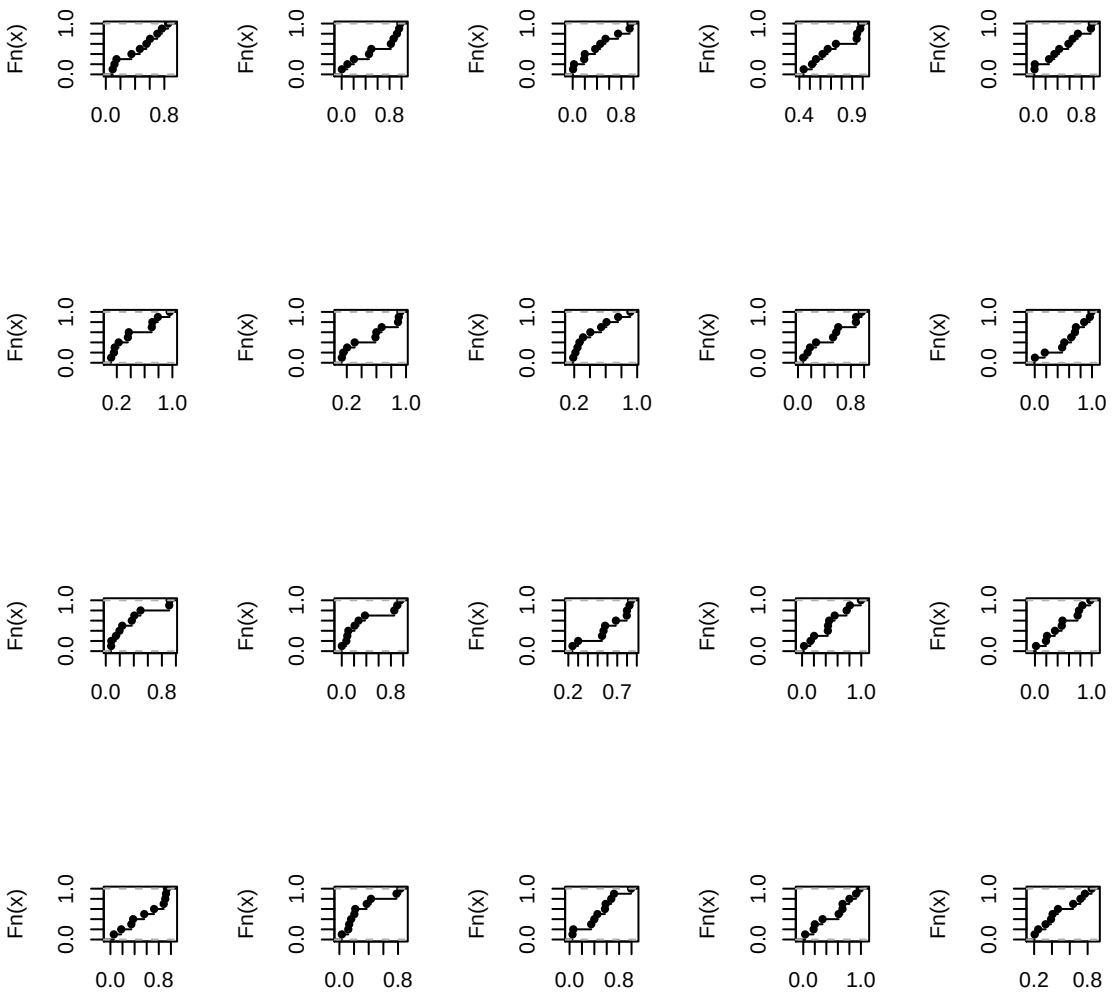

FIGURE 2. Empirical distribution functions

Exercise 2. *Section 4.3 of* `http://www.statlab.uni-heidelberg.de/studinfo/sda/onedim.pdf` *gives a short list of common problems to check for. Which of these can you identify in histograms? In QQ-plots?*
Just give three examples, such as effect xxx can be detected by yyy. It will show up as zzz in a QQ-plot.

Exercise 3. *Sometimes data can be transformed to normality. A very important case is that of logarithmic scales, that is you have underlying data y from a normal distribution, but they are measured on a logarithmic scale as* $x = exp(y)$. *Use the normal quantile plot on a sample* $x = exp(rnorm(n))$ *an try to describe what you could use as a diagnosis. Compare to the normal quantile plot for* $ln(x)$.

Exercise 4. *This is a short collection of normal QQ-plots from different data. Inspect these plots an decide whether you would accept them as from a normal population. If not, try to formulate which feature of the plot does not fit into a normal picture. (Pick 3 from these plots).*

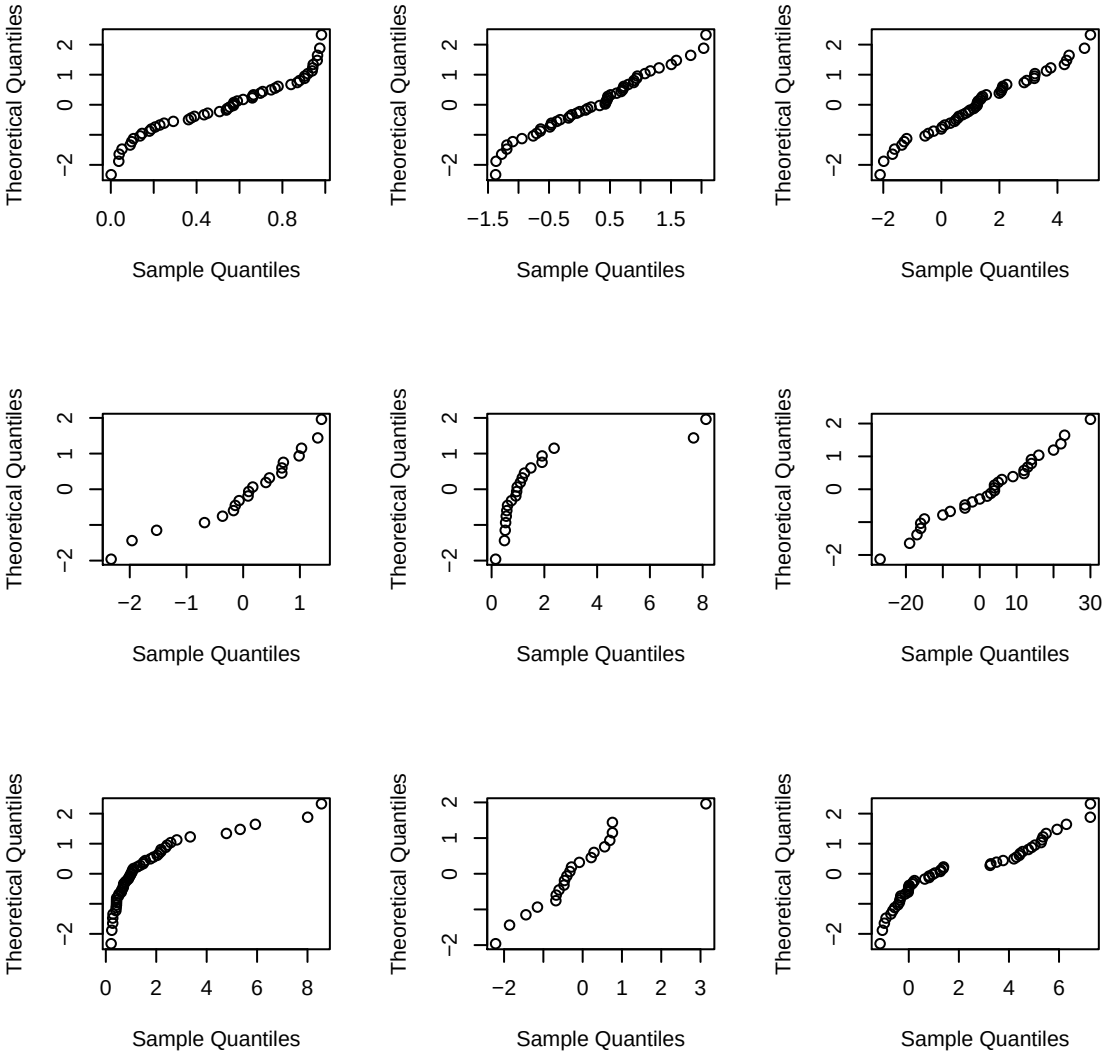

FIGURE 3. QQ-Plot as Diagnostics