3. 02. Nov. - Onedim diagnostics (cont.)

Revised: 0.5. Nov.

A concept to be added: if there are two methods competing, and they need sample sizes n resp. n', n'/n is called the relative efficiency.

What is the relative efficiency of a QQ-plot in comparison to a histogram, used as a test for uniform distribution at level 5%?

More on QQ-plots:

Ref. Ross Ihaka: Statistics 787 Topic in Computational Data Analysis and Graphics, Auckland 2007

`https://www.stat.auckland.ac.nz/~ihaka/787/lectures-quantiles2.pdf`

R code: `https://www.stat.auckland.ac.nz/~ihaka/787/quantiles.R`

3.1. **Bounds for PP- and QQ-plots.** Using $\sqrt{n}\sup|F_n - F| \leq \kappa$ gives bounds. In original PP-space, this is a uniform tubes. If we go to QQ-space using the quantile functions for transformation, the tube will transform. In particular for the normal distribution, the flat parts in the tail give steep tails in the quantile function, and the bounds expand.

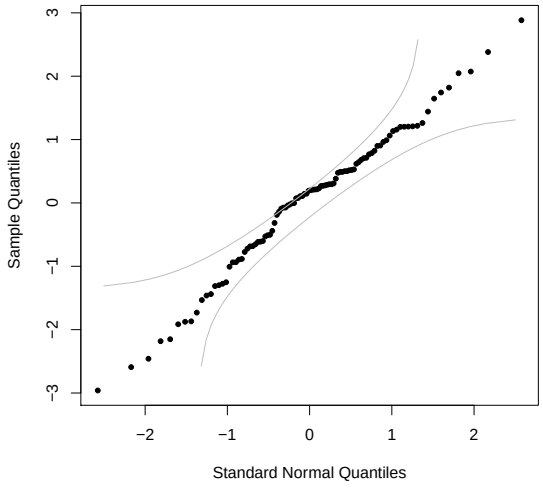

Figure 4. QQ bounds

Since the expression is symmetric in F, F_n, we can as well get bounds starting with F_n Using $\kappa = F_{KS}^{-1}(1-\alpha)$ to define bounds around F_n, we get asymptotic confidence bounds. Since the bound is uniform, these are simultaneous confidence bounds.

Unfortunately, default *qqnorm*() as provided in R does not show these bounds. They are provided in the add-on package *extRemes*. You have to download and install this package first from the R-repository using `install.packages("extRemes")`. To use it, you have to activate it in a session with `library(extRemes)`.

With `library(extRemes)`, the behaviour of `qqnorm` changes to include asymptotic 95% confidence bands. If the normality assumption is compatible with the data, some straight line may fit within the bounds.

Input

```
library(extRemes)
qqnorm(x)
qqline(x)
```

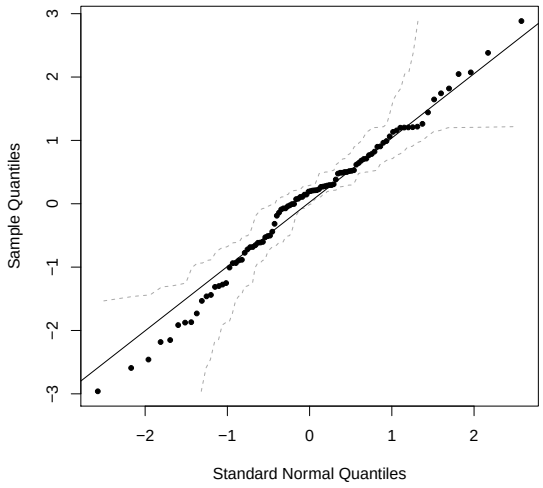

FIGURE 5. QQ with asymptotic 95% bounds

Exercise 5. *(Cont.) This figure 6 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).*

Exercise 6. *(Cont.) 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.*
Squaring is another common transformation, for example when area is reported were diameter is in effect. How can you differ square transformed data for exponential transformation? Can you tell the difference by histograms? By QQ-plots?

The Kolmogorov-Smirnov test is available as `ks.test`. The distribution function of the test statistic is hidden in the internals of R.

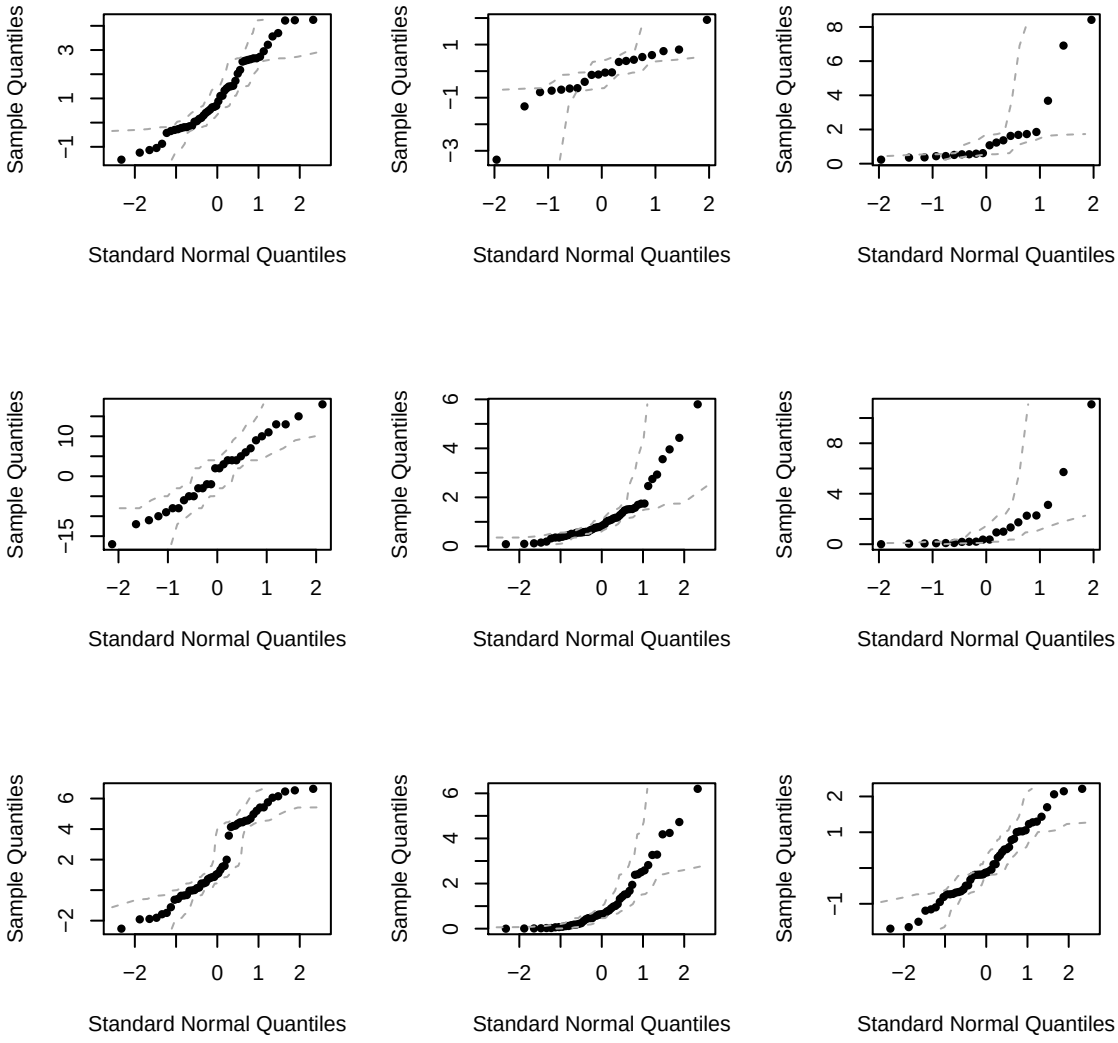

FIGURE 6. QQ-Plot as Diagnostics

Exercise 7. *Using* `help(ks.test)` *you get information on how to invoke the function* `ks.test`*. Which results do you expect if you test the following vectors for a uniform distribution?*

```
(1:100)/100
runif(100)
sin(1:100)
rnorm(100)
```

Perform these tests and discuss the results.
For the test, scale the values so that they fall into the interval $[0, 1]$*, or use a uniform distribution on an interval that is adapted to the data.*

3.2. **Project task.** *This is a first draft. Please contact me for extensions and corrections.*

Exercise 8. **** Project Task ****

`OddOneOut()` *provides a training framework for diagnostic plots.*
`OddOneOut()` *is available in* `library(Sintro)`*. You have to install the library using* `install.packages("sintro",repos="http://r-forge.r-project.org")` *and load it in your session using* `library(sintro)`*. The calling conventions are shown using* `help(OddOneOut)`
In a grid of `nrows, nrcols`*, all but one panels are shown as defined by* `goodplot`*, and one random panel if from* `badplot`*. The task is to identify the bad plot by clicking on it.*
Use `OddOneOut()` *as a starting point to design a training for histograms and for QQ-plots.*
You have to provide two documents: a manual with instructions for the trainee, and a manual for the trainer.
Please record your personal required size to use (a) histograms (b) QQ-plots to test for normality with an error level not exceeding 5%.

References

R session info:

- R version 3.3.2 (2016-10-31), `x86_64-apple-darwin13.4.0`
- Locale: `en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: car 2.1-3, distillery 1.0-2, Lmoments 1.2-3, sintro 0.1-5
- Loaded via a namespace (and not attached): extRemes 2.0-8, grid 3.3.2, lattice 0.20-34, lme4 1.1-12, MASS 7.3-45, Matrix 1.2-7.1, MatrixModels 0.4-1, mgcv 1.8-15, minqa 1.2.4, nlme 3.1-128, nloptr 1.0.4, nnet 7.3-12, parallel 3.3.2, pbkrtest 0.4-6, quantreg 5.29, Rcpp 0.12.5, SparseM 1.72, splines 3.3.2, tools 3.3.2

LaTeX information:

textwidth: 6.00612in linewidth:6.00612in

textheight: 9.21922in

CVS/Svn repository information:

`$Source: /u/math/j40/cvsroot/lectures/src/insider/profile/Rnw/profile.Rnw,v $`

`$Revision: 1.1 $`

`$Date: 2013/05/20 20:24:04 $`

`$name:  $`

`$Author: j40 $`

E-mail address: `g.sawitzki@uni-heidelberg.de`

Günther Sawitzki
StatLab Heidelberg
Im Neuenheimer Feld 294
D 69120 Heidelberg