Non-parametric tests and rehearsal

9.24 Is the Shapiro-Wilk test shapiro.test() resistant to outliers? Run it for 100, 1000, 4000 random numbers, generated by the standard normal distribution and add the number 5 to every vector with e.g. c(rnorm(100), 5). Does the presence of a single large outlier change the ability of the test to detect normality?

WW Synthesize 100 random number governed by the beta distribution rbeta() with the shape parameters (1, 1) and (2, 2).

- Have a look at the respective histograms.
- Now plot their empirical cumulative distribution functions into one plot using first plot.ecdf and then lines(ecdf()).
- Use the qbeta() command to read out the 0.05, 0.50, 0.95 percentiles of the respective beta distributions.

XX The Kruskal-Wallis test kruskal.test(x ~y, data = ...)) is an equivalent of the one-way test for non-normal data. Perform one on the data in the data set PlantGrowth, where weight is modeled by the factor group. Is there a significant difference in the means?

ECDF

YY The data set ChickWeight contains data on the weight of chicken held under different diets.

- Assign the data set to a variable cw and attach cw.
- Create four variables cw.i with i in 1 to 4, where you restrict the cw to the respective diet i.
- Plot all ecdfs in one plot (first plot.ecdf(...) then lines(ecdf(...)).The diets seem to have different effects on the chicken weight.
- Test this assumption with the Kruskal-Wallis test.

ZZ Download the data "data1.txt" from the courses homepage to your local drive. Read-in the data into a variable dat by the R command with the right file path

```
dat <- read.table("C:/...", sep = "\t", header = T)
```

Have a look at the data head and the structure of it (str(dat)) and attach it for convenience. We are interested in the effect of the factor set size on Hit RTs.

- Create four subsets of dat namely d.3, d.6, d.12, d.18 by restricting dat to the respective set size, to HIT and removing outliers (i.e. RT >= 850). You may do this as in the last sessions, by defining columns of interest, e.g.

ind3 <- (setsize == 3) & (message == “HIT”) & (RT <= 850)

and then setting d.3 accordingly.

- Now plot the denisities of the pairs d.3, d.6 and d.12, d.18 on the same plot using plot(density( )) to the RT data of each subset.
- After that plot the ECDFs of the RTs of d.3, d.6 on one plot and the ones of d.12, d.18 on another. Which plot suggest difference in distribution more clearly?

9.3 The data set stud.recs (UsingR) contains math and verbal SAT scores (SAT = Scholastic Assessment Test) for some students. Load the data by data(stud.recs, package = “UsingR”) and attach it. Assume naively that the two samples are independent, are the samples from the same population of scores?

First make a qqplot, a side-by-side boxplot and a plot of the ECDFs for the data sat.m, sat.v to see whether there is any merit to the question.

Now perform a Kolmogorov-Smirnof test ks.test() to see whether the ECDFs of sat.m and sat.v correspond. Does the test confirm the assumption from the visual inspection?

BS *Bootstrapping* is the process of sampling with replacement from the data of interest in order to infer on unknown parameters when sample sizes are to small.

The bycatch data set (UsingR) contains the number of albatrosses incidentally caught by squid fishers, measured by an observer program. Load the data and unfold it to the single measurements by

```
hauls <- with(bycatch, rep(no.albatross, no.hauls))
```

Assign n to the length of hauls. Take a look at the histogram of hauls. Now use the following code to perform the bootstrapping:

```
xbar <- c()
for(i in 1:1000){
    boot.samp <- sample(hauls, n, replace = T)
    xbar[i] <- mean(boot.samp)
}
```

What is the mean and the standard deviation of xbar? In which interval lies xbar in 90% of the time? Use the quantile() command on the percentiles 0.05 and 0.95 to address this question.