ITX Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Microarrays Bioinformatics — WS 2017/18

Assignment No. 2 (5 points)

Hand out:	Thursday, November 9
Hand in due:	Thursday, November 16, 10:00
Tutorial date:	Tuesday, November 14, 10:00-11:30
Direct inquiries to:	fabian.moertter@student.uni-tuebingen.de

Theoretical Assignments

1. Descriptive Statistics (5p)

(a) For the discrete uniform distribution X on the integers $1, 2, \ldots, n$ calculate the expected value of X (=mean).

(b) Using `R`, calculate a normally distributed random vector of length 10. Using this vector, calculate the empirical cumulative distribution function (*ecdf*), the 0.75 quantile ($Q_{0.75}$) and the 0.95 quantile ($Q_{0.95}$) values by hand.

(c) Using your vector, how much does the mean differ from the median? What happens if you use a shorter/longer vector? Why?

(d) What is the IQR for normally distributed data $\mathcal{N}(\mu, \sigma^2)$?

2. Boxplots and their interpretation (5p)

In the lecture you learned about boxplots and how they can be used to visualize and interpret distributions.

(a) In the following there are two different boxplots showing two different distributions. Describe the two plots and try to find an interpretation, i.e. try to find out what type of distributions are shown in the two plots.

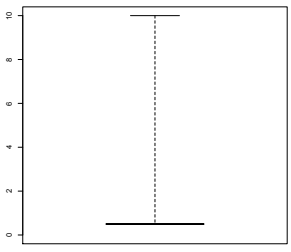

Abbildung 1: Boxplot 1

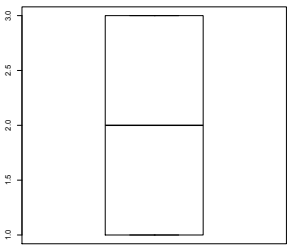

Abbildung 2: Boxplot 2

(b) Together with boxplots one can also draw the outliers of the data. In the lecture you learned about the definition of an outlier and the interquartile range (IQR) as a measure of statistical dispersion. In this context explain the two terms *true outlier* and *probable outlier* and why one uses $a = 3$ for a true outlier and $a = 1.5$ for a probable outlier.

Practical Assignments

3. Isothermality of an Oligo Design: handling strings in R

(a) Install the package "Biostrings" from Bioconductor. It contains methods for the handling of biological sequences. Other packages may also be used.

(b) Read in the sequences of the file `oligos2017.txt` from the material folder (the sequences are formatted in the `fasta` format).

(c) Implement a function that calculates the melting temperature using the Wallace method, and apply this method to the data. Are these oligos a good choice for a microarray? Are certain criteria missed by these oligos? Prepare a plot to answer this question.

(d) Use the following function to produce a sample of random generated sequences of length n:

```
1 library(stringi)
2 bases <- c("A","T","G","C")
3
4 randomOligos <- function(n) {
5   do.call(paste0, replicate(n, sample(bases, 1000, TRUE), FALSE))
6 }
```

i. Write a function that returns the average melting temperature (according to the Wallace method).

ii. Plot the average melting temperature with increasing length of your random generated oligos (up to 300 base pairs)

iii. Discuss your results and suggest an oligo length, which hybridisation temperature would also be appropriate to bake a pizza.

Please read the questions carefully. If there are any questions, you may ask them during the tutorial session or via e-mail to your tutor. You will usually get an answer in time, but late e-mails (e.g. on Thursday morning before class) might not be answered in time. Please upload your solutions in the Ilias system. Please pack your source code, the plots, as well as the theoretical part into **one single archive file (zip)**. Source code should compile correctly.