

*Prof. Dr. Vladimir Shikhman*  
*Professur für Wirtschaftsmathematik*  
*Technische Universität Chemnitz*

*Übungsleiter: David Müller*  
*david.mueller@mathematik.tu-chemnitz.de*

## **Mathematical Foundation of Big Data Analytics (SS 2019)**

### **Classification II**

**Ex. 1** You are working at the controlling department of a big company. The marketing department wants to increase the sales of a product. In order to do so, they plan to advertise the excellent quality of the product. Therefore the controlling has to assure a constant quality. Due to time constraints, it is not possible to check every single copy of the product, so the decision has to be made out of sampling.

The product has two possible conditions. It is either intact or defect. Hence the product's condition can be described by a random variable:

$$\mathbb{P}(\theta = def) = \pi, \quad \mathbb{P}(\theta = int) = 1 - \pi.$$

- a) You install a test device and draw an independent sample of size  $n$ . Your device recognizes a defect product with probability  $p \in [0, 1]$  and the functionality of an intact one with probability  $q \in [0, 1]$ . Denote the random variable for the number of defect messages  $0 \leq k \leq n$  by  $X$ . Give the formulas for the conditional probabilities of exactly  $k$  defect messages?
- b) Write the problem of cleaning out the product as a classification problem.
- c) Based on b), derive a formula, in dependence of  $n$ , for which number  $k$  you start to clean out.

**Ex. 2** Finally, the production department has been able to provide some concrete numbers. According to them, in average 5% of the products are defect. Your test device recognizes a defect product with 90% and an intact with 70% probability.

- a) How many defect messages have to appear, in order to clean the product out, if  $n = 10$ ?
- b) Suppose you have access to a huge sample. Calculate the percentage for cleaning the product out.