# TMA4255 Applied Statistics
# Exercise 9

## Problem 1

One wants to monitor an electric signal (in dB) that characterizes a particular electric component.

| Sample | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| 1 | 15.5 | 16.5 | 14.4 | 15.0 | 16.6 |
| 2 | 16.1 | 15.2 | 14.9 | 14.6 | 16.3 |
| 3 | 14.9 | 16.9 | 16.2 | 17.2 | 16.8 |
| 4 | 16.6 | 16.0 | 16.6 | 16.8 | 16.9 |
| 5 | 15.7 | 16.2 | 16.1 | 16.5 | 15.9 |
| 6 | 15.9 | 15.2 | 13.7 | 15.9 | 16.0 |
| 7 | 16.1 | 15.5 | 17.5 | 14.6 | 16.9 |
| 8 | 16.1 | 15.0 | 16.9 | 16.1 | 16.8 |
| 9 | 10.9 | 16.6 | 12.9 | 16.8 | 15.6 |
| 10 | 15.0 | 13.0 | 15.7 | 15.8 | 17.1 |
| 11 | 16.8 | 15.6 | 14.9 | 14.0 | 16.3 |

Rational subgroups of size $n = 5$ are used.

Use these data as training data and set up $\bar{X} - S$ and $S$-charts for the data. Comment

## Problem 2

The number of defective transistors out of 1000 inspected has been registered during one month. Use the given data as training data and set up a $p$-chart. Comment.

| Day | Number of defective transistors | Day | Number of defective transistors |
|---|---|---|---|
| 1 | 7 | 16 | 10 |
| 2 | 5 | 17 | 9 |
| 3 | 11 | 18 | 12 |
| 4 | 13 | 19 | 14 |
| 5 | 9 | 20 | 12 |
| 6 | 12 | 21 | 13 |
| 7 | 10 | 22 | 7 |
| 8 | 10 | 23 | 9 |
| 9 | 6 | 24 | 12 |
| 10 | 14 | 25 | 8 |
| 11 | 9 | 26 | 14 |
| 12 | 13 | 27 | 12 |
| 13 | 8 | 28 | 12 |
| 14 | 11 | 29 | 11 |
| 15 | 12 | 30 | 13 |

## Problem 3

During teeming of metal bolts the number of pores in a certain area is registered.

| Bolt | Number of pores | Bolt | Number of pores |
|---|---|---|---|
| 1 | 9 | 14 | 3 |
| 2 | 15 | 15 | 7 |
| 3 | 11 | 16 | 2 |
| 4 | 8 | 17 | 3 |
| 5 | 17 | 18 | 3 |
| 6 | 11 | 19 | 6 |
| 7 | 5 | 20 | 2 |
| 8 | 11 | 21 | 7 |
| 9 | 13 | 22 | 9 |
| 10 | 7 | 23 | 1 |
| 11 | 10 | 24 | 5 |
| 12 | 12 | 25 | 8 |
| 13 | 4 | 26 | 8 |

Use these data as training data and set up a $c$-chart. Comment.

## Problem 4: Goodness of fit test

This problem requires both theoretical work and use of software.

Some hand dryers will blow hot air for a time $t_0$ once you turn them on and then they are automatically turned off. If the hands are not dry during the time $t_0$, the machine has to be turned on again.

We assume that the time $T$ (in seconds) that a randomly selected individual needs to get the hands dry is gamma distributed $(2, \theta)$ with probability density function

$$f(t) = \frac{t}{\theta^2} e^{-\frac{t}{\theta}} \text{ for } t > 0 \tag{1}$$

where $\theta = E(T)$ is an unknown parameter. Alternatively you may use that for $T \sim \text{gamma}(2, \theta)$ it is also known that $Z = \frac{2T}{\theta} \sim \chi^2_4$ (chi-squared-distributed with 4 degrees of freedom). (So you either use the gamma or the chi-square distribution in calculating probabilites and expected values. Both are available in MiniTab and in R, and you need to perform this part of the calculations with software.)

Assume that $n$ individuals have used this kind of hand dryer and that their drying times $T_1, ..., T_n$ are independently and identically distributed with the probability density function given in 1. The standard estimator (maximum likelihood) for $\theta$ is then

$$\bar{T} = \frac{1}{n} \sum_{i=1}^{n} T_i \tag{2}$$

Assume that $n = 60$ and $\sum_{i=1}^{60} T_i = 1800$, so that $\bar{T} = 1800/60 = 30$. (Remark: you need to use this estimator for the unknown $\theta$ in your calculation.)

One wants to test if $T$ has the probability density function 1. Use the data and the separation in intervals as given in the table below. Choose the null hypothesis and alternative hypothesis yourself.

| Interval no. | Intervall | Number of observations in the interval ($X_j$) |
|---|---|---|
| 1 | $< 0, 25]$ | 12 |
| 2 | $< 25, 40]$ | 7 |
| 3 | $< 40, 55]$ | 11 |
| 4 | $< 55, 70]$ | 12 |
| 5 | $< 70, 90]$ | 8 |
| 6 | $< 90, \infty]$ | 10 |

## Problem 5: Test for independence

This problem can be solved using software alone.

The health authorities are interested in knowing whether the consumption of skimmed, semi-skimmed and homogenized milk is the same among young people as among the older ones. A survey has been conducted where 500 adults were asked about their age and what kind of milk they drink. The results were as follows:

| | Skimmed milk | Semi-skimmed milk | Homogenized milk | Total |
|---|---|---|---|---|
| Age $< 45$ | 138 | 83 | 64 | 285 |
| Age $> 45$ | 64 | 67 | 84 | 215 |
| Total | 202 | 150 | 148 | 500 |

Formulate the situation as a hypothesis testing problem. Perform the test using a 5% significance level.

## Problem 6: Homogeniety

Exam August 2012, Problem 1: Use of cannabis among youths in Norway. See the course WWW-page. This is done by hand (no software at the exam).

## Minitab

Problem 1:

Stat→Control Charts →Variables Charts for Subgroups → Xbar-S

- Obs. from subgroups are in one row across columns: C2-C6

Problem 2:

Stat→Control Charts → Attributes Charts → P

Variable: C2

- Subgroup size: 1000

Problem 3: Stat→Control Charts → Attributes Charts→ C

Variable: C2

**Problem 4)** Put the interval limits in one column and the number of observations in another column and perform operations on these columns to first find the probability of observing a time within the interval (based on gamma$(2, \theta)$ or Chi-squared 4), then the expected number of events in the intervals, and finally the $\chi^2$ test statistic and the corresponding $p$-value.

**Problem 5)** Stat $\rightarrow$ Tables $\rightarrow$ Chi-Square Test.

**R:**

We may use the R-library `qcc`. See the help pages for details. The `qcc` function takes a data set and specification of which plot to be made as arguments.

```
install.packages("qcc")
library(qcc)
?qcc

# Problem 1
data1 <- read.table("http://www.math.ntnu.no/~mettela/TMA4255/Data/data10_1.txt")
data1
cc1XS <- qcc(data=data1,type="xbar",std.dev="UWAVE-SD")
cc1S <- qcc(data=data1,type="S")
# leaving out observation at row 9
cc1Sm9<- qcc(data=data1[-9,],type="S")
cc1XSm9 <- qcc(data=data1[-9,],type="xbar",std.dev="UWAVE-SD")

# Problem 2
data2 <-  scan("http://www.math.ntnu.no/~mettela/TMA4255/Data/data10_2.txt")
data2
# binomial data with n=1000
cc2p <- qcc(data2,type="p",sizes=1000)

# Problem 3
data3 <- scan("http://www.math.ntnu.no/~mettela/TMA4255/Data/data10_3.txt")
data3
# binomial data with n=1000
cc3c <- qcc(data3,type="c")
# violation for obs 5
cc3cm5 <- qcc(data3[-5],type="c")
```

**Problem 4)** First read in the interval limits, and use the gamma or chisq-distribution to calculate probabilites for each interval. For all probability distributions *pdist* gives cumulative probabilites, where *dist* is replaced with e.g. `gamma` or `chisq`. Then use $P(T \leq U) - P(T \leq L)$ (U=upper, L=lower) to find the probability to be inside the interval. Then multiply with $n = 60$ and we have the expected cell counts under the null hypothesis. Then calculate the chi-squared goodness of fit test statistic and compare with the correct chi-squared distribution. The degree of freedom is then "number of cells - 1 - number of parameters estimated", and we have estimated one parameter.

```
thetaest<-1800/60
lowerL <- c(0,25,40,55,70,90)
# probabilites for each interval can be calculated using the gamma for T
# first using gamma to find the prob of being lower than equal to the lower
# interval limit
intprobL <- pgamma(lowerL,shape=2,rate=1/thetaest)
# for comparison, also chisq 4 for 2*T/theta
intprobL2 <- pchisq(lowerL*2/thetaest,4)
# check: this gives the same - not use intprobL2 further - just compare

# prob of being lower than upper limit, shuffle and add 1
intprobU <- c(intprobL[-1],1)
intprob <- intprobU-intprobL
# expected in each interval, multiply intprob with n=60
e <- intprob*60
# entering the observed values
o <- c( 12,7,11,12,8,10)
# (o-e)^2/e and sum
testobscell <- (o-e)^2/e
testobs <- sum(testobscell)
# compare with chisq distribution with 6-1-1=4 degrees of freedom
# since 6 cells, all sum to 1 (minus 1), and estimated theta (minus 1).
# p-value
pchisq(testobs,4,lower.tail=FALSE)
# critical value, just for comparison - how large need to be for rejection
qchisq(0.05,4,lower.tail=FALSE)
#9.487729
```

**Problem 5)** First read the data into a matrix, one row for each sample. Use the function `chisq.test` which takes the 2 by 3 matrix of data as input. Choose the `correct=FALSE` option to skip the continuity correction.

```
milktab <- rbind(c(138,83,64),c(64,67,84))
chisq.test(milktab,correct=FALSE)
```