Exam 2010
Statistics for Artificial Intelligence and Informatics

Date: November 11 2010
Time: 18.30 - 21.30
Place: Tentamenhal 01, Blauwborgje 4
Code: WISTAKI-07

Rules to follow:

- This is closed book exam.
- The usage of an (graphical) electronic calculator is admitted.
- Always give an argument underpinning your answer!
- Do not forget to fill in your name and student number.
- Use four decimals precision, unless otherwise indicated.
- Take significance level 0.05 throughout.
- Choose one out of questions 7 and 8!
- We wish you success with the completion of the exam.

START OF EXAM

1. Drug Treatment. A certain drug treatment cures 80% of the children with a certain rare disease. During a year in a hospital twenty otherwise unrelated children are treated. Recall that the binomial density is defined as

$$P(X = x|\pi, n) = \begin{cases} \binom{n}{x}\pi^x(1-\pi)^{n-x} & \text{for } x = 0, 1, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

   (a) How many children do you expect to be cured?
   (b) What is the probability that all are cured?
   (c) What is the probability that twelve are cured?
   (d) What is the probability that more than seventeen are cured?

2. Methodological questions.

   (a) Give a descriptive and a testing approach to investigate whether a vector of scores is normally distributed?
   (b) How would you test that the correlation coefficient between two variables equals .30 in case both are not normally distributed?
   (c) Give an important reason to use model diagnostics.

3. Questions on using R.

(a) Briefly describe the purpose of the functions `lm`, `rm`, `setwd`?

(b) Give three built-in-functions to investigate properties of an object that is new to you.

4. Questions on output from R

(a) Explain what the code below does. Which object from the data contains the measurements and which the experimental factor? What is the null hypothesis of the test? Give the conclusion from the test.

```
> library(ISwR)
> data(red.cell.folate)
> kruskal.test(folate ~ ventilation, data=red.cell.folate)

        Kruskal-Wallis rank sum test

data:  folate by ventilation
Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234
```

(b) Explain what the code below does, give the statistic, its meaning as well as the null-hypothesis and conclusion from the test.

```
> y <- c(30,29,44,86)
> smoke <- gl(2,1,4,labels=c("smoke","no smoke"))
> birthweight <- gl(2,2,labels=c("low","normal"))
> ov <- xtabs(y ~ birthweight + smoke)
> fisher.test(ov)
        Fisher's Exact Test for Count Data
data:  ov
p-value = 0.03618
95 percent confidence interval:
 1.028780 3.964904
sample estimates: odds ratio
  2.014137
```

(c) Explain in detail what the code below does. Specifically, give the null-hypothesis and your conclusion from the statistical test involved.

```
library(ggm); data(marks)
data <- marks; p <- ncol(data); n <- nrow(data); nboot <-1000
eigenvalues <- array(dim=c(nboot,p))
for (i in 1:nboot){
   dat.star <- data[sample(1:n,replace=TRUE),]
   eigenvalues[i,] <- eigen(cor(dat.star))$values}

> for (j in 1:p) print(quantile(eigenvalues[,j], c(0.025,0.975)))
    2.5%     97.5%
2.726191 3.588401
```

```
0.5256718 1.0275732
0.3522736 0.6162052
0.2593002 0.4702665
0.1575396 0.3106478
```

5. Soporific drugs. Classical data in Table 1 show the effect of two soporific drugs (increase in hours of sleep compared to control) on 20 persons in two groups (Cushny & Peebles, 1905).

Table 1: Effect of two soporific drugs.

	Increase in hours of sleep									
	1	2	3	4	5	6	7	8	9	10
Group 1	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
Group 2	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4

The data are stored in R as follows.

```
> str(sleep)
'data.frame':   20 obs. of  2 variables:
 $ extra: num  0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2 ...
 $ group: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

(a) A preliminary test gives the following results.

```
> shapiro.test(sleep[group=="1",1])
data:  sleep[group == "1", 1]
W = 0.9258, p-value = 0.4079
> shapiro.test(sleep[group=="2",1])
data:  sleep[group == "2", 1]
W = 0.9193, p-value = 0.3511
```

Give the null hypothesis and your conclusion.

(b) Yet, another preliminary test gives the following output.

```
> var.test(extra ~ group, data = sleep)

data:  extra by group
F = 0.7983, num df = 9, denom df = 9, p-value = 0.7427
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.198297 3.214123
sample estimates:
ratio of variances
         0.7983426
```

Formulate the null hypothesis, the value of the test statistic, its distribution under H_0, and your conclusion.

(c) A t-test is conducted on the data with the following result.

```
> t.test(extra ~ group, data = sleep)

        Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.0794
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

Formulate the null hypothesis and give the value of the test statistic, its distribution under H_0, and your conclusion.

6. Wages. A regression analysis was conducted on data containing weekly wages for US male workers and years of education sampled from the Current Population Survey in 1988. (Bierens & Ginther, 2001). The following output is generated.

```
> library(faraway)
> data(uswages)
> summary(lm(wage ~ educ, data=uswages))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  109.754     44.616    2.46   0.0140 *
educ          38.011      3.317   11.46   <2e-16 ***

Residual standard error: 445.5 on 1998 degrees of freedom
Multiple R-squared: 0.06167,     Adjusted R-squared: 0.0612
F-statistic: 131.3 on 1 and 1998 DF,  p-value: < 2.2e-16
```

(a) What is the dependent and what the independent variable?

(b) How many valid observations are there?

(c) Give the equation of the line which is best fitting to the data according to the least squares criterion.

(d) Explain from the output how the p-values are computed.

(e) What is the null hypothesis and your conclusion with respect to testing the parameters?

(f) What is the fit of the model to the data? What is your evaluation on this?

(g) Which null hypothesis is tested by the F-statistic?

Choose one out of the last two questions.

7. Give a brief, but precise, answer to the following.

   (a) Briefly describe the different types of t-tests.

   (b) Describe a non-parametric test of the difference between two data vectors.

   (c) Explain in general terms how maximum likelihood estimation is defined.

8. ML estimation. The Poisson density with intensity parameter λ is defined as

$$P(X = x|\lambda) = \begin{cases} \frac{\lambda^x}{x!}e^{-\lambda} & \text{for } x \in \mathbb{Z}_+ \\ 0 & \text{otherwise} \end{cases}$$

   The following measurements were found 3 5 3 5.

   (a) Formulate the likelihood of the measurements.

   (b) Maximize the log likelihood by setting its derivative to zero.

   (c) Solve the first order equations to find the maximum likelihood estimate $\widehat{\lambda}$.

END OF EXAM