Mock Exam - Solution

Applied Statistics 2019, IT University of Copenhagen

1 Probability Theory (8pt)

(a) A natural choice for the sample space is

$$\Omega = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(2,4), \\ (3,1),(3,2),(3,3),(3,4),(4,1),(4,2),(4,3),(4,4)\}$$

(b) The set of outcomes are

- $A = \{(2,4),(3,3),(3,4),(4,2),(4,3),(4,4)\}$
- $B = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$
- $C = \{(1,1),(2,2),(3,3),(4,4),\}$

(c) Since the dice is fair and the roles are independent, all outcomes in the sample space are equally likely. Therefore we get the probability by taking the number of outcomes in the event and divide it by the size of the sample space: $P(A) = P(B) = \frac{6}{16} = \frac{3}{8}$ and $P(C) = \frac{4}{16} = \frac{1}{4}$.

(d) $P(A|C)$ is the probability of the sum of the two dice being greater than or equal to 6 given that the values of the first and second roll are the same. Using the definition of conditional probability it is given by $P(A|C) = \frac{P(A\cap C)}{P(C)} = \frac{2/16}{4/16} = \frac{1}{2}$.

(e) We can consider this experiment to be 10 Bernoulli trials, where "success" corresponds to rolling a 2 and "failure" corresponds to not rolling a 2. The probability of success is 1/4. Now X represents the number of successes in 10 Bernoulli trials and accordingly X has a $Bin(10, 1/4)$ distribution. Therefore the probability of rolling 2 exactly 5 times is

$$P(X=5) = \binom{10}{5} \cdot (1/4)^5 \cdot (3/4)^5 = 0.0583992$$

2 Hypothesis Testing

(a) Our null hypothesis is that the expected grade μ is 7, that is $H_0 : \mu = \mu_0$ for $\mu_0 = 7$. The alternative hypothesis is that the expected grade is not 7, that is $H_1 : \mu \neq \mu_0$.

(b) We can use a one-sample t-test and since we are considering a large sample, our test statistics $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ can be assumed to approximately follow a standard normal distribution.

(c) The value of the test statics is $t = \frac{\bar{x}_{681} - \mu_0}{s_{681}/\sqrt{681}} = \frac{7.15-7}{3.07/\sqrt{681}} = 1.2750$. Since we know that test statistic follows a $N(0,1)$ distribution, the critical values at level $\alpha = 0.05$ are $z_{1-\alpha/2} =$ `qnorm(0.025)` $= -1.959964$ and $z_{\alpha/2} =$ `qnorm(0.975)` $= 1.959964$. As $z_{1-\alpha/2} < t < z_{\alpha/2}$ is not in the critical region, we cannot reject the null hypothesis at significance level $\alpha = 0.05$. So we cannot conclude that the expected grade for the course is different from 7.

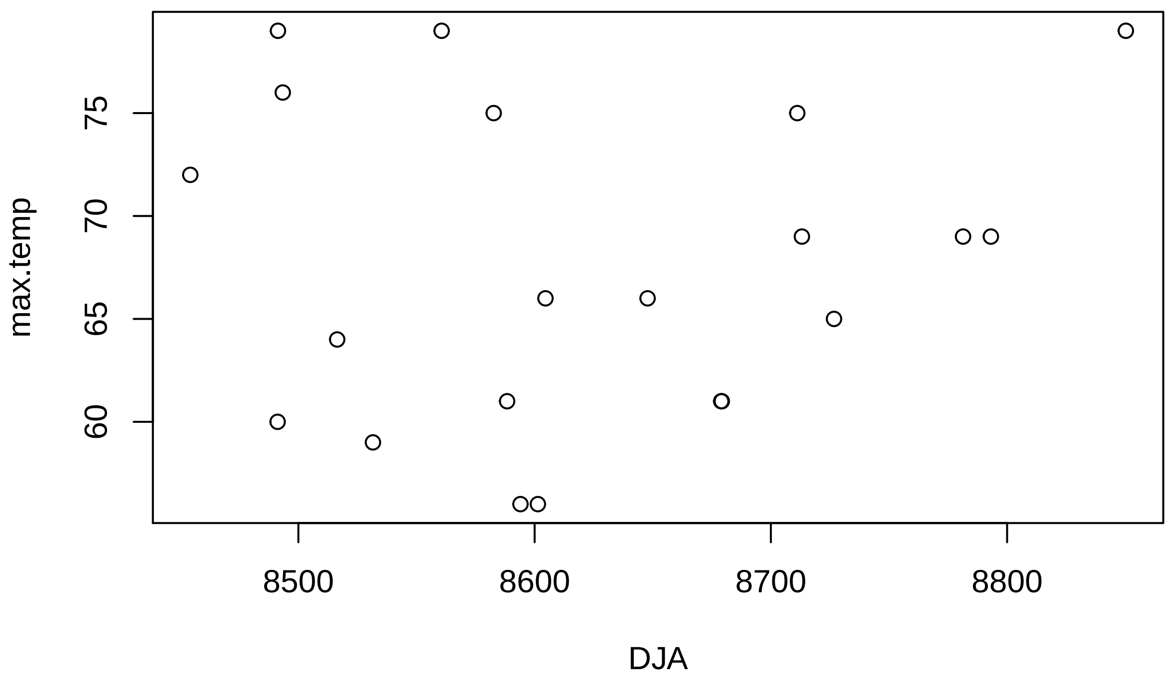
(d) Since we have a large sample, the one-tailed p-value $P(T \geq t) = P(T \geq 1.2750)$ can be approximated by $P(Z \geq 1.2750)$, where Z has an $N(0,1)$ distribution. We find the p-value to be $P(Z \geq 1.2750) =$ `1-pnorm(1.2750)` $= 0.1011$. As we in 1/10 cases expect to see a test statics at least this extreme, the evidence is very weak. That is also why we reject the null hypothesis.

3 Small R Questions

3a Correlation

First we plot the data:

```
plot(max.temp ~ DJA, data=maydow)
```



From the plot, it does not look like there are any correlation between the Dow Jones index and the temperature. To investigate further, we compute the correlation coefficient.

```
cor(maydow$DJA,maydow$max.temp)
```

```
## [1] 0.08626861
```

The correlation coefficient is very close to zero, so the variables are practically uncorrelated.

3b Simple statistics

First we compute the minimum, maximum, median, and standard deviation of those mothers' age:

```
sub = subset(babies, age !=99)
min(sub$age)
```

```
## [1] 15
```

```
max(sub$age)
```

```
## [1] 45
```

```
mean(sub$age)
```

```
## [1] 27.25527
```

```
median(sub$age)
```

```
## [1] 26
```

```
sd(sub$age)
```

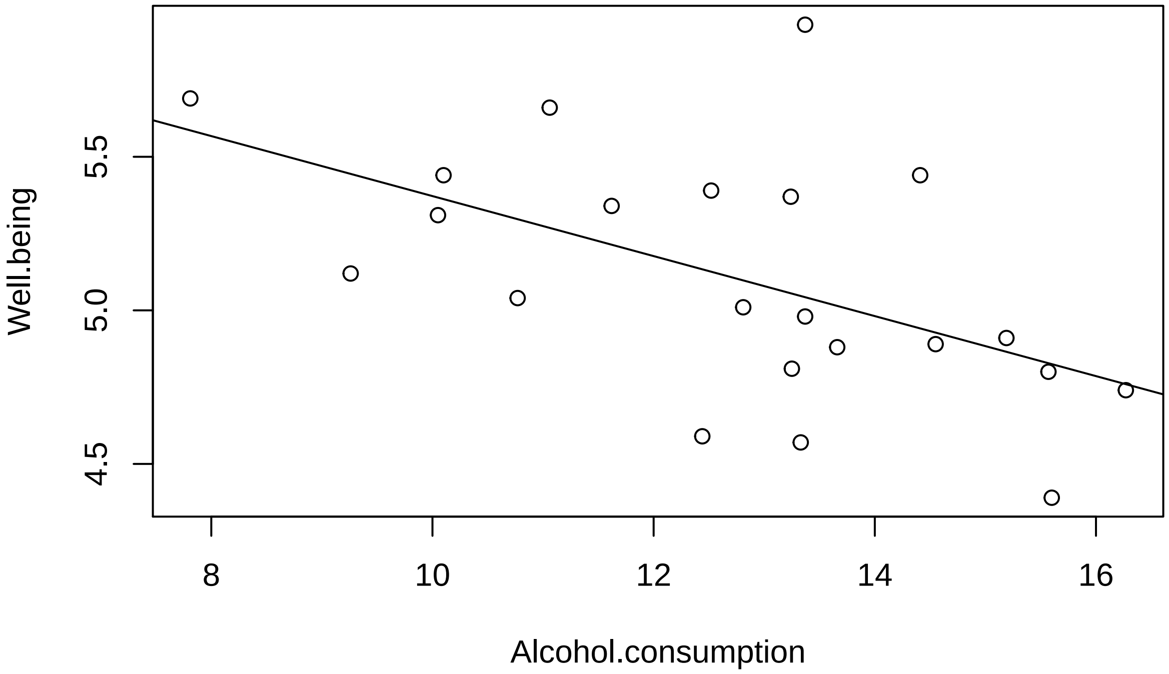
```
## [1] 5.781405
```

The youngest known mother in the study were 15 years old and oldest known was 45 years old. The mean age is 26 years, while the median is 27 years, which indicates that the age distribution is not symmetric but leaned to the right.

3c Linear Regression

First we fit a linear model and plot the model together with the data.

```
plot(Well.being ~ Alcohol.consumption, data=wellbeing)
model = lm(Well.being ~ Alcohol.consumption, data=wellbeing)
abline(model)
```



Then we estimate how many percentage the well-being is increased, in the reported units, if the alcohol consumption is decreased from 16 to 8 units:

```
prediction <- predict(model, newdata = data.frame(Alcohol.consumption=c(16,8)))
percentage <- (prediction[2]-prediction[1])/prediction[1]
percentage
```

```
##         2
## 0.1632916
```

The well being is improved about 16% when the alcohol consumption is decreased from 16 units to 8.

3d Confidence intervals

We'll first extract the 5-year-olds using the command

```
yr5 <- subset(kid.weights, subset = 5*12 <= age & age < 6*12)
```

There is no direct command in R for getting confidence intervals. Instead, if one makes the t.test one gets those as a by product together with other information. To get only the confidence interval we may use the `confint` wrapper after calling the t-test with the 90% confidence level as follows

```
confint(t.test(yr5$weight, conf.level = 0.90))
```

```
## (42.75, 47.09) with 90 percent confidence
```

that is the confidence interval we were after.

4 Viking Lottery

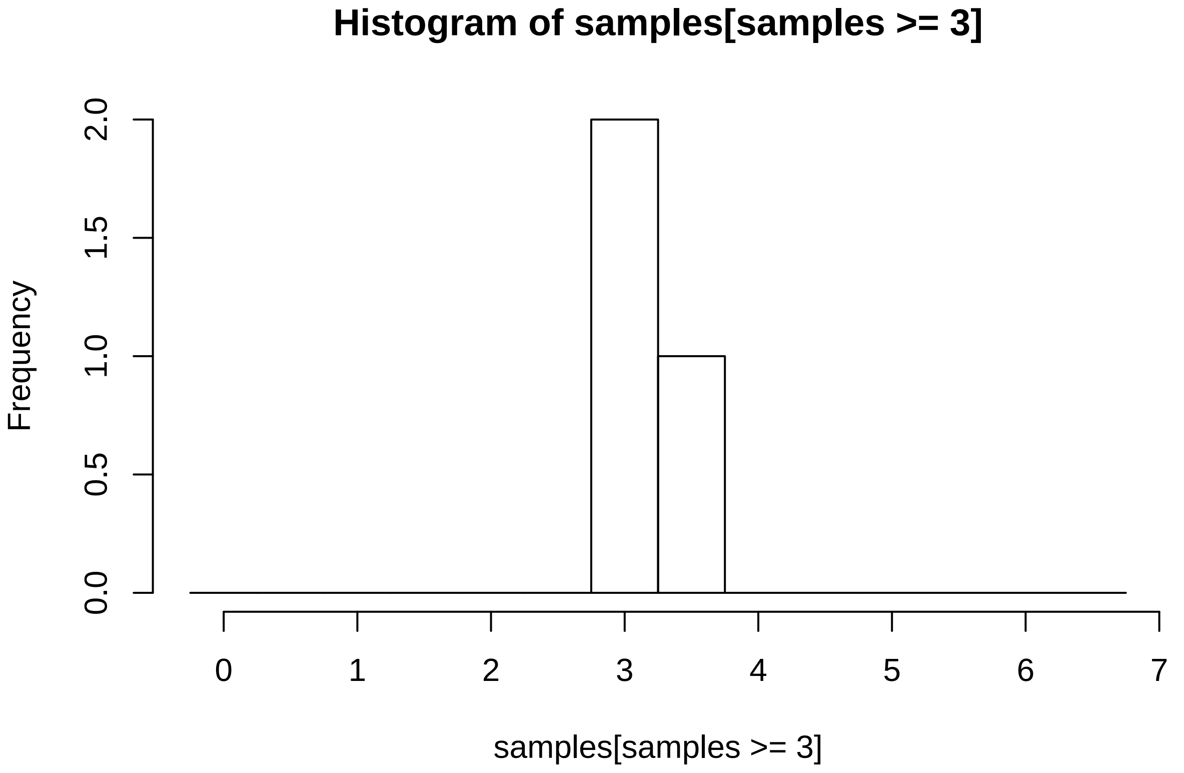
Let us make a program that simulates the lottery.

```
n=100
our.normal.row = sample(1:48,6,replace=FALSE)
our.viking.number = sample(1:8,1,replace=FALSE)

samples = replicate(n,{
  sample.row <- sample(1:48,6,replace=FALSE)
  sample.viking <- sample(1:8,1,replace=FALSE)

  tab <- rep(0,48)
  vtab <- rep(0,8)
  tab[our.normal.row]=1
  vtab[our.viking.number]=1
  tab[sample.row] <- tab[sample.row]+1
  vtab[sample.viking] <- vtab[sample.viking]+1
  result <- length(tab[tab==2])+0.5*length(vtab[vtab==2])
  result
})
```

In the code, we'll first generate our luck row in random, of course, that could be also generated manually. The results, indicating the number of right numbers, will be stored in the variable samples over the `n=100` trials. Inside the `replicate` loop we generate two hit tables, `tab` for the normal numbers and `vtab` for the viking number. The hit table is initialised to zeros in the beginning of each trial, after which those elements corresponding to our luck row and luck Viking number will be incremented by one. Then we draw the lotto row and the Viking number and increment the hit tables. Those elements that have a double hit will indicate a match, so we test for how many of those double hits we got; in addition, the Viking number is modelled as a half hit in the result.

We first visualise the result by making a histogram. We are only interested in cases where we got more than 3 right so we restrict the histogram to exclude the worse results.

```
h <- hist(samples[samples>=3], (0:14)/2-0.25)
```



From the histogram, we can already see that we did not get many right. Let us then compute the total profit which is the total amount of wins from which the cost of the playing (80 EUR) is subtracted. To compute the total about of wins, we can access the histogram values taking its counts and multiply elementwise by the corresponding wins given in the table, and finally summing over the entries.

```
profit = sum(h$counts*c(0,0,0,0,0,0,4,8,25,50,1000,10000,250000,1000000)) - n*0.8
profit
```

```
## [1] -64
```

The profit of the game is the above (in EUR).