Mock Exam Solution

Applied Statistics 2022, IT University of Copenhagen

11th May 2022

1. Probability Theory (6 pts)

(a)

$$\Omega = \{H,T\} \times \{H,T\} \times \{H,T\} \times \{H,T\}.$$

(b)

$$A = \{HHHT, HHTH, HTHH, THHH\},$$
$$B = \{HHHH, TTTT\}.$$

(c) We get either exactly one tails or the coin comes always the same side up.

(d) The events A and B are disjoint so

$$P(C) = 1 - P(A) - P(B) = 1 - 4 \cdot \frac{1}{16} + 2 \cdot \frac{1}{16} = \frac{5}{8}.$$

2. Expectation, Variance, Discrete Distributions (6 pts)

(a) $E[Z] = E[X]E[Y]$ due to independence, where $E[X] = E[Y] = \sum_{a=1}^{6} a/6 = \frac{7}{2}$. Hence $E[Z] = \frac{49}{4}$.

(b) Let us see the possible outcomes for the product:

```
z=outer(c(1,2,3,4,5,6),c(1,2,3,4,5,6),'*')
z
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    2    3    4    5    6
## [2,]    2    4    6    8   10   12
## [3,]    3    6    9   12   15   18
## [4,]    4    8   12   16   20   24
## [5,]    5   10   15   20   25   30
## [6,]    6   12   18   24   30   36
```

We can thus conclude that the probability mass function for Z is

$$p(c) = P(Z = c) = \begin{cases} \frac{1}{36} & \text{when } c = 1, 9, 16, 25, 36 \\ \frac{1}{18} & \text{when } c = 2, 3, 5, 8, 10, 15, 18, 20, 24, 30 \\ \frac{1}{12} & \text{when } c = 4 \\ \frac{1}{9} & \text{when } c = 6, 12 \\ 0 & \text{otherwise.} \end{cases}.$$

(c) $\text{Var}[Z] = E[(Z - E[Z])^2] = E[Z^2] - E[Z]^2$. The numerical value is hence obtained as

```
VarZ = mean(as.vector(z*z)) - mean(as.vector(z))^2
VarZ
```

```
## [1] 79.96528
```

So, the variance of the product is approximately 79.97.

3. Maximum likelihood (4pts)

(a) The likelihood of the parameters is

$$L(\alpha, \beta) = \prod_{i=1}^{n} f(x_i; \alpha, \beta)$$

where probability density of the data is

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \leq x \leq \beta \\ 0 & \text{otherwise.} \end{cases}$$

(b) We first observe that L is always non-negative. We can also see that if we set α so that any observation would be smaller than it, the likelihood function will be zero. Similarly if any of the observations is larger than β the likelihood will be likewise zero. The likelihood function hence yields the largest value when the interval is the smallest as long as these two conditions are met. Therefore, we conclude that the maximum likelihood estimate for the parameters is $\hat{\alpha} = \min_i x_i$ and $\hat{\beta} = \max_i x_i$.

4. Small R Problems (8 pts)

(a) We get several simple statistics by calling the summary of the data as follows.

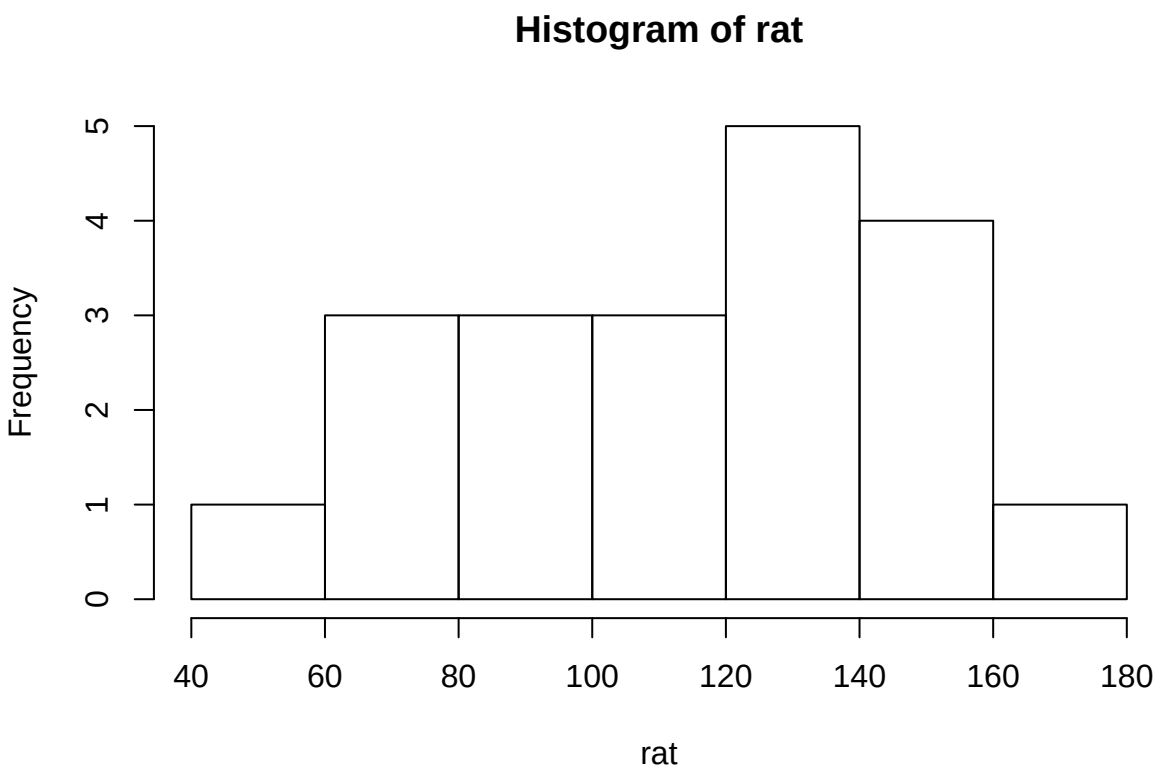
```
summary(firstchi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   20.00   23.00   23.98   26.00   42.00
```

The mean is different (larger) from the median that indicates that the distribution is leaning to the right.

(b) To visualise the dataset we use histogram.

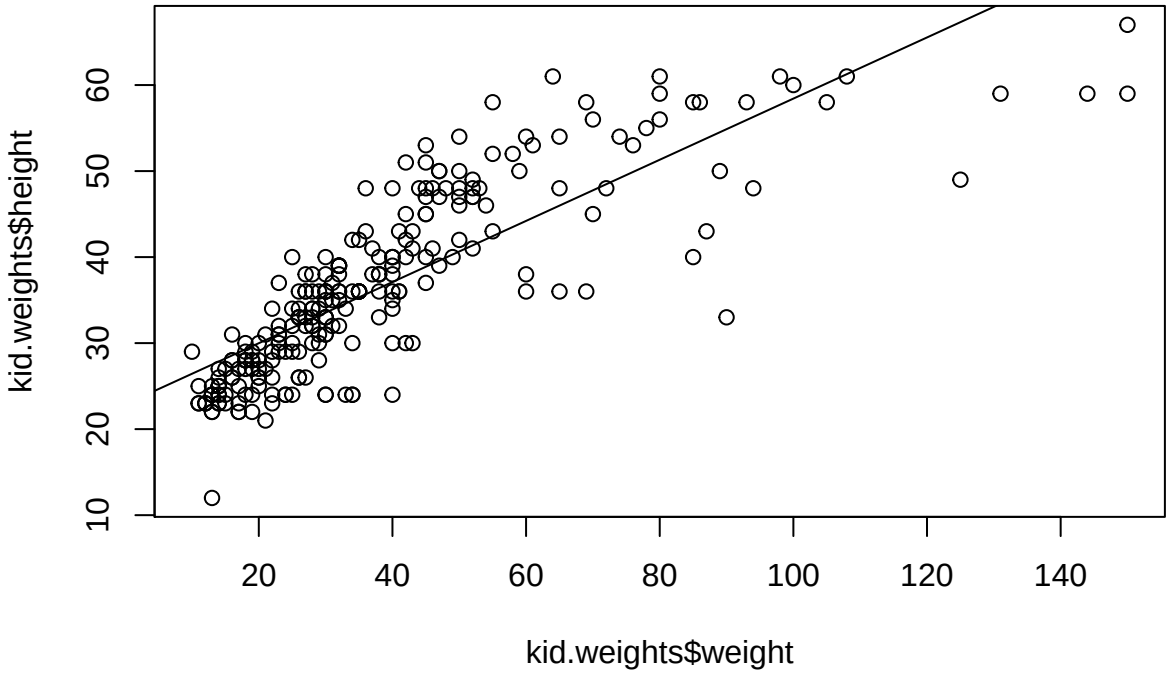
```
hist(rat)
```



From the visual shape and the small number of observations we conclude that the normal distribution model seems appropriate for the survival times.

(c) Let us make the scatter plot and add the regression line to the plot.

```
plot(kid.weights$weight,kid.weights$height)
abline(lm(kid.weights$height ~ kid.weights$weight))
```



The weight and height seem to have a linear relationship, apart from obese people that make the curve to lean to the right.

(d) The null hypothesis is

H_0 : The methods are equally good.

The alternative hypothesis is

H_1 : My method is better than the state-of-the art.

Due to very small number of data points, and the distribution of the data a normal model for them seems appropriate. Since the variances seem different, we select the two sample t-test with unequal variances.

```
ref = c(0.908,0.915,0.908,0.905,0.904);
my = c(0.910,0.914,0.909,0.914,0.910);

var(ref)
```

```
## [1] 1.85e-05
```

```
var(my)
```

```
## [1] 5.8e-06
```

```
t.test(ref,my,var.equal=FALSE, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  ref and my
## t = -1.5423, df = 6.2836, p-value = 0.08587
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf 0.0008494366
```

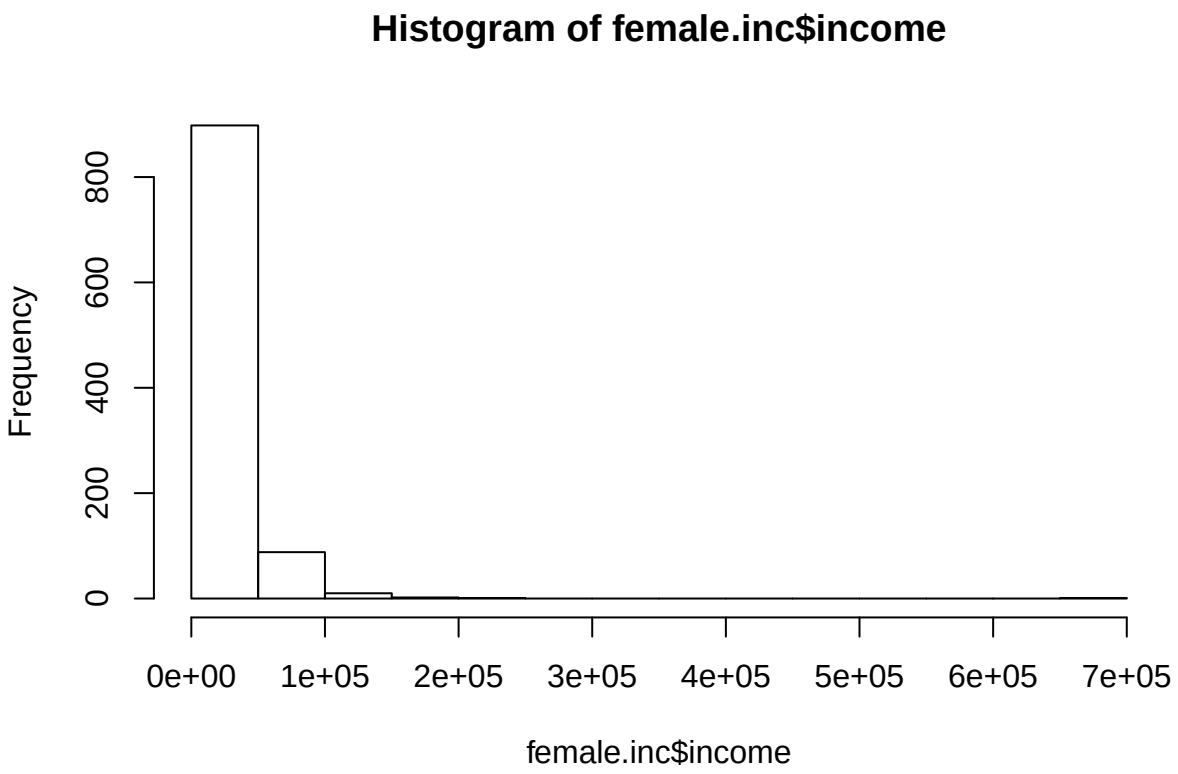
```
## sample estimates:
## mean of x mean of y
##    0.9080    0.9114
```

Since the obtained p-value $0.086 > 0.05$, we conclude that there is not enough statistical evidence to reject the null hypothesis.

5. Bootstrapping and confidence intervals (8 pts)

Let us first investigate how the data is distributed by looking at the histogram.

```
hist(female.inc$income)
```



Bootstrapping is the right strategy for determining the confidence intervals since the data is clearly non-normal.

(b)

```
mean.income <- mean(female.inc$income)
n <- length(female.inc$income)
studentized.means <- c()
for(i in 1:10000) {
  income.star <- sample(female.inc$income,n,replace=TRUE)
  studentized.mean<-(mean(income.star)-mean.income)/(sd(income.star)/sqrt(n))
  studentized.means <- c(studentized.means,studentized.mean)
}

c.lu=quantile(studentized.means,probs=c(0.025,0.975))
interval = c(mean.income-c.lu[2]*sd(female.inc$income)/sqrt(n),
             mean.income-c.lu[1]*sd(female.inc$income)/sqrt(n) )
```

(c) The mean is

```
mean.income
```

```
## [1] 21071.61
```

and the bootstrapped confidence interval is

```
unname(interval)
```

```
## [1] 19466.13 23518.52
```

Visualisation of the distribution of the bootstrapped studentised means is below.

```
densityplot(studentized.means)
```