UNIVERSITEIT VAN DIE VRYSTAAT
UNIVERSITY OF THE FREE STATE

STSM 3714

WISKUNDIGE STATISTIEK & AKTUARIËLE WETENSKAP/
MATHEMATICAL STATISTICS & ACTUARIAL SCIENCE

Test 3 — 29 May 2018

MEMORANDUM

TYD/TIME: 120 Minutes PUNTE/MARKS: 44

INSTRUCTIONS:

- Answer all questions in a single Word document but convert the final version to .pdf before submitting (backup the Word document and all working files separately).
- Label questions clearly, as it is done on this question paper. Also provide your details in the document header.
- Show all code, calculations, formulas and reasoning.
- Use a significance level of 5% for all hypothesis tests in Question 1.
- No USB drives or data transfer devices may be used during the course of the test.
- No attempt at communication of any kind is allowed with other students. No email or messaging service on any device may be accessed.

Question 1

For this question, load the warpbreaks data from R [command: 'data(warpbreaks)']. The data compares two types of wool by comparing the number of breaks under different loads. Thus, the goal is to find which wool type breaks less.

(a) Test whether the breaks could follow an approximate Normal distribution if you consider ONLY the data for wool B under low tension. [4]

```
data(warpbreaks)
attach(warpbreaks)
names(warpbreaks)
shapiro.test(breaks[(wool=='B') & (tension=='L')])
```

✔

```
 Shapiro-Wilk normality test
data:  breaks[(wool == "B") & (tension ==
"L")]
W = 0.94899, p-value = 0.679
```

✔

Since the p-value > 0.05 we fail to reject H_0 ✔ and continue to assume an approximate Normal distribution for wool B under low tension. ✔

(b) Test whether wool A and wool B could have the same variance, considering all loads. [5]

$$H_0 : \ \sigma_A^2 = \sigma_B^2 \ vs \ H_1 : \ \sigma_A^2 \neq \sigma_B^2 \ \checkmark$$

```
var.test(breaks~wool)
```

✔

```
 F test to compare two variances
data:  breaks by wool
F = 2.9046, num df = 26, denom df = 26,
p-value = 0.008461
alternative hypothesis:  true ratio of
variances is not equal to 1
95 percent confidence interval:
1.323695 6.373561
sample estimates:
ratio of variances
2.904591
```

✔

Since the p-value < 0.05 we reject H_0 ✔ and conclude that the variance in breaking of the two types of wool is probably not the same. ✔

(c) Test whether wool A and wool B could have the same mean, considering all loads, and taking the result of the previous question into account. [5]

$$H_0 : \ \mu_A^2 = \mu_B^2 \ vs \ H_1 : \ \mu_A^2 \neq \mu_B^2 \ \checkmark$$

```
t.test(breaks~wool)
```

✔

```
 Welch Two Sample t-test
data:  breaks by wool
t = 1.6335, df = 42.006, p-value = 0.1098
alternative hypothesis:  true difference in
means is not equal to 0
95 percent confidence interval:
-1.360096 12.915652
sample estimates:
mean in group A mean in group B
31.03704     25.25926
```

✔

Since the p-value > 0.05 we fail to reject H_0 ✔ and continue to assume the number of breaks are the same on average. ✔

Total for Question 1: 14

Question 2

At an archery tournament you are asked to model the distances between 9 arrows on a target and the bullseye of the target. Assuming Normal variation in all dimensions you know you can model the distances using a Rayleigh distribution, with density $f(x|\tau) = \tau x e^{-0.5\tau x^2}$, and CDF $F(x|\tau) = 1 - e^{-0.5\tau x^2}$. τ is the precision parameter and represents the archer's skill. Potentially this approach could be used to more accurately differentiate archers, but for now you only wish to model one sample given in 'rayleighsample.csv'. Distances are in inches.

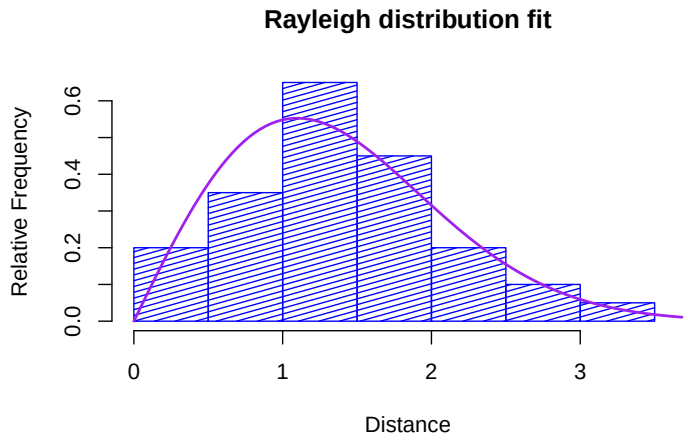
(a) Derive the log likelihood and show (in detail) that the ML estimator for τ is $\frac{2n}{\sum X_i^2}$. [4]

$$\begin{aligned}
\ell &= \sum_{i=1}^{n} \ln f(x_i|\tau) \\
&= \sum_{i=1}^{n} \left(\ln \tau + \ln x_i - 0.5\tau x_i^2\right) \checkmark \\
&= n \ln \tau + \sum_{i=1}^{n} \ln x_i - 0.5\tau \sum_{i=1}^{n} x_i^2 \checkmark \\
\therefore \frac{d\ell}{d\tau} &= \frac{n}{\tau} - 0.5 \sum_{i=1}^{n} x_i^2 \checkmark \\
\therefore \hat{\tau} &= \frac{2n}{\sum X_i^2} \checkmark \qquad \text{[By setting derivative to zero.]}
\end{aligned}$$

(b) Draw a histogram of the given sample and overlay the ML estimated density. You may use relative frequencies, or match the scale another way. [4]

```
d <- read.csv(file.choose(),row.names=1)
sam <- d$x
class(sam)
(n <- length(sam))
tauest <- 2*n/sum(sam^2)
hist(sam,6,freq=FALSE,col='blue',density=10,angle=30,main='Rayleigh distribution fit'
    ,xlab='Distance',ylab='Relative Frequency',xlim=c(0,1.2*max(sam)))
curve(tauest*x*exp(-0.5*tauest*(x^2)),0,1.2*max(sam),add=TRUE,lwd=2,col='purple')
```

✔



✔✔✔

(c) Derive a generalised likelihood ratio test for the hypothesis $H_0 : \tau = \tau_0$. [Hint: $Q = (\tau \sum_{i=1}^{n} X_i^2) \sim \chi^2_{2n}$] [7]

$$\ln \lambda = \ell(\tau_0) - \ell(\hat{\tau}) \checkmark$$

$$= \left[n \ln \tau_0 + \sum_{i=1}^{n} \ln x_i - 0.5\tau_0 \sum_{i=1}^{n} x_i^2 \right] - \left[n \ln \hat{\tau} + \sum_{i=1}^{n} \ln x_i - 0.5\hat{\tau} \sum_{i=1}^{n} x_i^2 \right]$$

$$= n\left[\ln \tau_0 - \ln \hat{\tau} + 1\right] - 0.5\tau_0 \sum_{i=1}^{n} x_i^2 \checkmark$$

$$\therefore \lambda = n^{-n} e^n (0.5Q)^n \exp(-0.5Q), \; Q = \left(\tau_0 \sum_{i=1}^{n} X_i^2 \right) \checkmark\checkmark\checkmark$$

$\therefore \lambda \leq c \Rightarrow$ Reject H_0 when $Q \leq k_1$ or $Q \geq k_2$, $k_1 < k_2$ ✔

Under H_0, $Q \sim \chi^2_{2n}$

$\therefore$ Reject H_0 when $\sum_{i=1}^{n} X_i^2 \leq \frac{1}{\tau_0} \chi^2_{2n;\frac{\alpha}{2}}$ or $\sum_{i=1}^{n} X_i^2 \geq \frac{1}{\tau_0} \chi^2_{2n;1-\frac{\alpha}{2}}$ ✔

(d) Use the pivotal quantity $Q = (\tau \sum_{i=1}^{n} X_i^2) \sim \chi^2_{2n}$ to construct a 90% symmetric interval for τ based on the observed sample, then use it to test $H_0 : \tau = 0.5$. [6]

$$P\left[\frac{\chi^2_{2n;\frac{\alpha}{2}}}{\sum_{i=1}^n X_i^2} \leq \tau \leq \frac{\chi^2_{2n;1-\frac{\alpha}{2}}}{\sum_{i=1}^n X_i^2}\right] = 1-\alpha$$

$$\Rightarrow P\left[0.6269 \leq \tau \leq 1.0576\right] = 0.9 \checkmark\checkmark\checkmark$$

```
(SS  <- sum(sam^2))
qchisq(c(0.05,0.95),2*n)/SS
```

✔

Since the interval does not contain 0.5 ✔, we reject H_0 and conclude that $\tau \neq 0.5$ ✔ with 90% confidence.

(e) Create a function that calculates the Anderson-Darling statistic for a sample assumed to come from a Rayleigh distribution. It should automatically calculate and use the ML estimator. Show that it works by applying it to your sample. It must give a result of about 0.25. [5]

```
ADstat.rayleigh <- function(sam) {
n <- length(sam)
ML <- 2*n/sum(sam^2)
x <- 1 - exp(-ML*(sort(sam)^2)/2)
A <- -n - sum(((2*(1:n)-1)/n)*(log(x)+log(1-rev(x))))
return(A)
}
(basestat <- ADstat.rayleigh(sam))
```

✔✔✔✔

Result checked ✔

(f) Test the null hypothesis that the sample originates from a Rayleigh model using the parametric bootstrap approach. Use 9999 random samples generated with the 'rrayleigh' function supplied by the lecturer. [4]

```
n <- length(sam)
acc <- 9999
stats <- rep(0,acc)
for (i in 1:acc) {
  newx <- rrayleigh(n,tauest)
  stats[i] <- ADstat.rayleigh(newx)
}
(pvalue <- mean(stats>basestat))
```

✔✔✔

The p-value should be close to 90%. ✔

Total for Question 2: 30

Total half marks on memo = 88 vs. 88 = Double total margin points (=44).