Semester 2 Final Exam, 2015

School of Mathematics and Statistics

MAST90058 Elements of Statistics

Writing time: 3 hours

Reading time: 15 minutes

This is NOT an open book exam

This paper consists of 6 pages (including this page)

Authorised materials:

- Hand-held scientific calculators (not CAS or graphics) may be used.
- Students may use one double-sided A4 sheet of handwritten notes.

Instructions to Students

- You may remove this question paper at the conclusion of the examination
- This examination contains 9 questions.
- All questions may be attempted. The total number of marks available is 70.

Instructions to Invigilators

- Students may remove this question paper at the conclusion of the examination
- All graphics or CAS calculators should be confiscated.
- Students may use one double-sided A4 sheet of handwritten notes.

Blank page (ignored in page numbering)

Question 1 (17 marks) Let X denote the proportion of allotted study time that a randomly selected student spends on a particular subject. Let $X_1, \ldots, X_n$ be a random sample on X assuming to have the density:

$$f(x;\theta) = (1+\theta)x^{\theta}, \quad 0 < x < 1,$$

and 0 otherwise, where $\theta > -1$. A sample of students yields the following data:

```
0.72  0.29 0.70 0.25 0.86 0.17 0.47 0.53
```

(a) Write the log-likelihood function.

(b) Determine the maximum likelihood estimator of θ.

(c) Give the Crámer-Rao lower bound of unbiased estimators of θ.

(d) Determine a sufficient statistic for θ.

(e) Determine the maximum likelihood estimate of θ and give an approximate 95% confidence interval for θ. Some R output that may help.

```
>  z <- c(0.95,0.975,0.99,0.995)
> qnorm(z)
[1] 1.644854 1.959964 2.326348 2.575829
```

(f) Derive an estimator for θ using the method of moments.

Question 2 (10 marks) A theoretical model suggests that the survival time (in weeks) of a randomly selected cell from a particular culture of cells follows an exponential distribution $Exp(\lambda)$ with pdf $f(x;\lambda) = \lambda \exp\{-\lambda x\}$, $\lambda > 0$, $x \geq 0$.

(a) Derive a $(1-\alpha)100\%$ confidence interval for the expected (true average) cell lifetime. [*Hint*: Use the fact that if $X_1, \ldots, X_n$ are i.i.d. from $Exp(\lambda)$, then $2\lambda \sum_{i=1}^{n} X_i$ follows a chi-square distribution with $2n$ degrees of freedom]

(b) A sample of 11 cells gave $\sum_i x_i = 10.66$. Find the endpoints for a 90% confidence interval for the mean.

(c) Another sample of 40 cells gave $\overline{x} = 1.02$. The following table gives frequencies corresponding to 3 intervals:

[0, 1)	[1, 2)	[2, ∞)
26	8	6

Test the hypothesis that the exponential model is adequate for these data using the 0.05 significance level of significance.

Some useful R output:

```
> t=c(0.005,0.01,0.05,0.1, 0.90, 0.95, 0.95, 0.99)
> qchisq(t, 2)
[1] 0.01 0.02 0.10 0.21 4.61 5.99 5.99 9.21
> qchisq(t, 1)
[1] 0.00 0.00 0.00 0.02 2.71 3.84 3.84 6.63
```

```
> qchisq(t, 7)
[1]  0.99  1.24  2.17  2.83 12.02 14.07 14.07 18.48
> qchisq(t, 9)
[1]  1.73  2.09  3.33  4.17 14.68 16.92 16.92 21.67
> qnorm(t, 9)
[1]  6.42  6.67  7.36  7.72 10.28 10.64 10.64 11.33
> qchisq(t, 20)
[1]  7.43  8.26 10.85 12.44 28.41 31.41 31.41 37.57
```

Question 3 (7 marks) A manufacturer has developed a new type of bicycle frame which will be sold with a 2-year warranty. To see whether this is economically feasible, 20 prototype frames are subjected to an accelerated life experiment to simulate 2 years of use. The proposed warranty will be modified only if fewer than 90% of such frames would survive the 2-year period.

(a) Let p be the true proportion of frames that survive. Find a rejection region for testing of the null hypothesis $H_0 : p = 0.9$ against $H_1 : p < 0.9$ at the 0.05 level of significance. Since the sample is relatively small, avoid normal approximations.

(b) It is found that 14 frames survive the accelerated life test. Do these data suggest that the true proportion of for the frames that survive is smaller than 90%?

(c) Find the probability of a Type II error assuming that the true proportion of frames surviving the accelerated life experiment is 0.8. Briefly comment on the accuracy of this test. How could you improve this result?

You may use the following R output:

```
> round(pbinom(10:20, size=20, prob=0.8), digit=2)
 [1] 0.00 0.01 0.03 0.09 0.20 0.37 0.59 0.79 0.93 0.99 1.00
> round(pbinom(10:20, size=20, prob=0.9), digit=2)
 [1] 0.00 0.00 0.00 0.00 0.01 0.04 0.13 0.32 0.61 0.88 1.00
```

Question 4 (5 marks) Let Y be a single observation from a Binomial(10, θ) distribution, where θ is the unknown probability of success. For θ, assume the $Beta(\alpha, \beta)$ prior distribution with pdf

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \beta > 0, \quad 0 < \theta < 1,$$

where $\Gamma(\cdot)$ denotes the Gamma function.

(a) Find the posterior distribution of θ.

(b) A sample gave $y = 3$. Find the Bayesian point estimate under the loss function $[w(y) - \theta]^2$ when the prior hyper-parameters are $\alpha = \beta = 1$. [Hint: Recall that for the $X \sim Beta(a, b)$ distribution, $E(X) = a/(a + b)$.

Question 5 (6 marks) An experiment investigates the performance of four different brands of light bulbs. Five bulbs of each brand were tested and the time until failure (min) was observed. The partial ANOVA table for the data is given below. Fill in the missing entries, state the relevant hypotheses, and carry out an appropriate statistical test using the $\alpha = 0.05$ level of significance. What can you conclude about the four brands based on your test?

Source	df	Sum of Squares	Mean Square	F
Brand				
Error			14,713	
Total		310,500		

You may use the following R output:

```
> qf(c(0.95), 3, 16)
[1] 3.238872
> qf(c(0.975), 3, 16)
[1] 4.076823
> qf(c(0.975), 4, 12)
[1] 4.121209
> qf(c(0.95), 4, 12)
[1] 3.259167
```

Question 6 (3 marks) A study on the ability to walk in a straight line reported the following data on cadence (strides per second) for a sample of 10 healthy men:

```
0.95  0.85 0.92 0.95 0.93 0.86 1.00 0.92 0.85 0.81
```

Find an approximate 90% confidence interval for the 25th percentile, $\pi_{0.25}$. You may use:

```
> pbinom(0:7, size=10, prob=0.25)
 [1] 0.056 0.244 0.526 0.776 0.922 0.980 0.996 1.000
```

Question 7 (5 marks) The following observations are pH values of synovial fluid which lubricates joints and tendons taken from the knwees of individuals suffering from arthritis. Asuuming that the true average pH for nonarthritic individuals is 7.1, use a Wilcoxon signed rank test at level 0.05 to see whether the data indicates a difference between average pH values for arthritic and nonarthritic individuals.

```
5.9  7.3  5.2  6.8  7.1  5.8  5.9  6.6  6.4  6.0
```

(a) Using the Wilcoxon statistic, W, define a critical region that has an approximate significance level of $\alpha = 0.05$.
Recall that under H_0 the sum of the signed ranks has an approximate normal distribution with mean zero and variance $n(n+1)(2n+1)/6$ and $z_{0.1} = 1.282$, $z_{0.05} = 1.645$, $z_{0.025} = 1.96$, and $z_{0.01} = 2.326$.

(b) Compute the test statistic and give your conclusion.

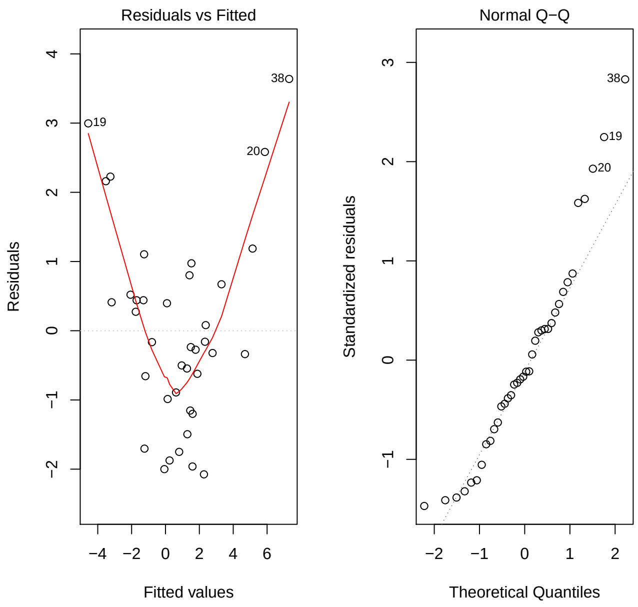
Question 8 (6 marks) A candidate single nucleotide polymorphism was thought to be associated with drug response in a group of epilepsy patients. To test this a sample of 100 patients that did not respond to the drug and 100 that did were examined and the values of the SNP (coded as 0,1 and 2) were recorded as follows:

	SNP		
	0	1	2
Non-responder	19	24	47
Responder	15	43	42

At the 5% level of significance, is there evidence that the SNP is related to drug response?

```
> z=c(0.95,0.975,0.99,0.995)
> qchisq(z,1)
[1] 3.841459 5.023886 6.634897 7.879439
> qchisq(z,2)
[1]  5.991465  7.377759  9.210340 10.596635
> qchisq(z,3)
[1]  7.814728  9.348404 11.344867 12.838156
```

Question 9 (11 marks) A geneticist thought that the gene expression on two genes may be related and sampled pairs (x_i, y_i), $i = 1, \ldots, 38$ of observations from 38 individuals. He then conducted a regression of one of the gene expressions (y) on the other (x). Some relevant R output is:



```
> m.1 <- lm(y~x)
> summary(m.1)
Call:
lm(formula = y ~ x)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.0749 -0.9628 -0.2004  0.6336  3.6382

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3896     0.2553  -1.526    0.136
x             1.2358     0.1131  10.927 5.52e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05

Residual standard error: 1.438 on 36 degrees of freedom
Multiple R-squared:  0.7683,    Adjusted R-squared:  0.7619
F-statistic: 119.4 on 1 and 36 DF,  p-value: 5.524e-13

> confint(m.1)
                 2.5 %    97.5 %
(Intercept) -0.9074457 0.1282999
x            1.0064281 1.4651772
> new.data=data.frame(x=c(-1,0,1))
> predict(m.1,newdata=new.data,interval="confidence")
         fit        lwr        upr
1 -1.6253755 -2.2715080 -0.9792431
2 -0.3895729 -0.9074457  0.1282999
3  0.8462298  0.3728138  1.3196458
```

Consider the model $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

(a) Give point estimates of β_0, β_1 and σ.

(b) Give 95% confidence intervals for β_0 and β_1.

(c) Test $H_0 : \beta_1 = 0$. Give the value of the test statistic, the degrees of freedom and the associated p-value.

(d) What should he conclude at the 5% level of significance?

(e) What further models would it be appropriate to fit to these data?

End of Exam