Data provided: Formula Sheet, incorporating Critical Values Tables

MAS361

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2020–21**

Medical Statistics

This is an open book exam.

Answer ***both*** *questions.*

You can work on the exam during the 24 hour period starting at 10am (GMT), and you must submit your work within 2 hours 30 minutes of accessing the exam paper or by the end of the 24 hour period (whichever is earlier). ***Late submission will not be considered without extenuating circumstances.***

Unless it is explicitly stated otherwise, it is intended that calculations are performed by hand (possibly with the aid of a calculator). To gain full marks, you will need to show your working.

By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged, and that no unfair means have been used.

Total marks available: 30.

1 (i) In a study aimed at assessing the effectiveness of a new drug in lowering high systolic blood pressure, the following trial was conducted. Eighty patients with high systolic blood pressure were recruited and split at random into two groups of size 40. Blood pressure meaurements were taken from each patient before the study began and after 1 week of treatment. In Group T the subjects were treated with the new drug and in Group P a placebo version. The data available for patient i ($i = 1, ..., 80$) is:

- G_i their Group indicator (coded as 1 for Group T and 0 for Group P)
- B_i their systolic blood pressure Before the treatment
- A_i their systolic blood pressure After the treatment
- $D_i = B_i - A_i$ the Difference in their systolic blood pressure over the trial.

Four possible analyses and their results are outlined below. You may assume that the tests specified have been correctly conducted and reported and that all necessary underlying assumptions are valid. It is observed that $\bar{D}_T > \bar{D}_P > 0$ where $\bar{D}_j$ is the sample mean of the Differences for Group j ($j = T, P$).

- **Analysis 1**
Two sided two sample t test using the A_i values for the two Groups is non significant.
- **Analysis 2**
Two sided two sample t test using the B_i values for the two Groups is non significant.
- **Analysis 3**
Two sided two sample t test using the D_i values for the two Groups is significant.
- **Analysis 4**
A regression analysis is performed using the D_i values as the response, with G_i and B_i values as the explanatory variables, as well as an intercept term μ, i.e. with regression function $E(D_i) = \mu + \gamma G_i + \beta B_i$. Separate tests of null hypotheses $H_0 : \gamma = 0$ and $H_0 : \beta = 0$ are both rejected, with estimated coefficients $\hat{\gamma} > 0$ and $\hat{\beta} > 0$.

In each case, explain what the analysis shows and whether (with reasons) it contributes usefully to meeting the aims of the study. What do you conclude from the study overall?
(9 marks)

(ii) As part of a court case considering whether a particular industrial solvent was implicated in development of the lung condition emphysema, the following data were collected. Eighty present and former male manual employees of a large company were identified; 30 of whom had been diagnosed with emphysema and 50 (who had worked there for a similar length of time) who had not. The men were interviewed and their employment records examined in order to determine whether they had been required to work with the solvent on a regular basis. The dataset was tabulated as follows:

	Have emphysema	Do not have emphysema
Used solvent regularly	12	10
Did not use solvent regularly	18	40

What can you conclude about the solvent as a risk factor for emphysema? ***(6 marks)***

2 (i) A clinical trial was conducted on 16 patients with chronic myelogenous leukemia, a cancer of the white blood cells, randomized into two groups of treatment and followed up. The first group of patients (8 patients) was given Drug A while the second group (8 patients) was given Drug B. The outcome of interest was mortality. Five patients were lost to follow-up; these censored observations are denoted by asterisks(*). The data below show each patient's time until death in months.

Patient	Drug A Time (months)	Drug B Time (months)
1	11.60	2.10*
2	9.00	1.60
3	1.30*	4.40
4	21.00	2.80
5	7.80	16.90
6	14.70*	9.40
7	1.20	9.40
8	21.90*	3.20*
Total	88.5	49.8

Obtain by hand the Kaplan-Meier (KM) estimate of the survivor function for Drug B.

(4 marks)

2 (continued)

(ii) A soft-tissue sarcoma is a malignant tumour, a type of cancer, that develops in soft tissue. A clinical trial involved the drugs Pazopanib, Gemcitabine, and Docetaxel. The purpose of the trial was to test the effectiveness of a combination of Gemcitabine and Pazopanib (treatment A) compared with a combination of Gemcitabine and Docetaxel (treatment B) in participants with soft tissue sarcoma. Patients are all in remission at the beginning of the trial and effectiveness is measured by time to relapse. Subjects were followed up for 2 years for a potential relapse. Patients still in remission after the end of study were considered right-censored. The data are stored in an R dataframe with the following coding:

Coding:

Sex:	0 = male ; 1 = female
Treatment:	0 = Gemcitabine+Pazopanib; 1= Gemcitabine+Docetaxel
Age:	age of patient centred on 40 years (i.e. 45yrs old becomes 5 after centring)
Time:	time until relapse (months)
Status:	0 = censored ; 1 = relapse

An analysis was implemented in R producing the following output:

```
Model 1
> # Survival object
> SARC.sv <- Surv(Time, Status)
> SARC.fit1 <- survreg(SARC.sv ~Sex+Age+Treatment, dist="exponential")
> summary(SARC.fit1)

Call:
survreg(formula = SARC.sv ~ Sex + Age + Treatment, dist = "exponential")
               Value Std. Error     z       p
(Intercept)  0.41743    0.12801  3.26 0.00111
Sex          0.28233    0.14564  1.94 0.05255
Age          0.01985    0.00552  3.60 0.00032
Treatment   -0.39474    0.14443 -2.73 0.00627

Scale fixed at 1

Exponential distribution
Loglik(model)= -271.4   Loglik(intercept only)= -286.3
Chisq= 29.82 on 3 degrees of freedom, p= 1.5e-06
Number of Newton-Raphson Iterations: 5
n= 250
```

(a) Describe the model estimated (Model 1) and write down the fitted model for the time to relapse, using appropriate notation. ***(2 marks)***

2 (continued)

(b) Based on the estimated output (Model 1), would you say that combination Gemcitabine and Pazopanib (treatment A) seems to be more effective than combination Gemcitabine and Docetaxel (treatment B)? ***(1 mark)***

(iii) An alternative model fitted (Model 2), gave the following results:

```
Model 2
> SARC.fit2 <- survreg(SARC.sv ~Sex+Age+Treatment, dist="weibull")
> summary(SARC.fit2)

Call:
survreg(formula = SARC.sv ~ Sex + Age + Treatment, dist = "weibull")
               Value Std. Error     z       p
(Intercept)  0.41383    0.13065  3.17 0.00154
Sex          0.28189    0.14812  1.90 0.05703
Age          0.01996    0.00562  3.55 0.00038
Treatment   -0.39420    0.14695 -2.68 0.00731
Log(scale)   0.01750    0.05457  0.32 0.74849

Scale= 1.02

Weibull distribution
Loglik(model)= -271.4   Loglik(intercept only)= -284.9
Chisq= 27.02 on 3 degrees of freedom, p= 5.8e-06
Number of Newton-Raphson Iterations: 5
n= 250
```

(a) Write down the survivor function of Model 2. ***(1 mark)***

(b) Explain how Model 1 and Model 2 are related. Compare the estimates of Model 1 and Model 2 explaining what drives potential differences or similarities. ***(2 marks)***

2 (continued)

(iv) Further analysis was undertaken, giving the following results (Model 3):

```
Model 3
> SARC.fit3 <- coxph(SARC.sv ~Sex+Age+Treatment)
> summary(SARC.fit3)
Call:
coxph(formula = SARC.sv ~ Sex + Age + Treatment)

  n= 250, number of events= 196

               coef exp(coef)  se(coef)     z  Pr(>|z|)
Sex       -0.253851  0.775807  0.147546 -1.720 0.085343
Age       -0.019307  0.980878  0.005552 -3.477 0.000506
Treatment  0.373432  1.452712  0.147294  2.535 0.011236
```

(a) Write down the fitted model (Model 3) in terms of the baseline hazard function $h_0(t)$. Explain whether this is a parametric, semi-parametric or non-parametric model. ***(2 marks)***

(b) Using the R output for Model 3, describe the effect of the treatment on time to relapse. ***(1 mark)***

(c) Based on the model fitted (Model 3), estimate the hazard ratio comparing two females, 35 and 45 years old respectively, both under treatment B (Gemcitabine+Docetaxel). ***(2 marks)***

End of Question Paper

STANDARD FORMULAE FOR MEDICAL STATISTICS (INCLUDING TABLES OF CRITICAL VALUES)

1 Clinical Trials Formulae

Two Sample t-Test — Separate variances form $r = min(n_1, n_2)$

$$t_r \;=\; |\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}|$$

Two Sample t-Test — Pooled variance form $r = n_1 + n_2 - 2$

$$t_r \;=\; |\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}(\frac{1}{n_1} + \frac{1}{n_2})}}|$$

Sample Size Calculations — Two sample test for proportions NB number in each group

$$n \simeq \frac{\theta_2(1-\theta_2) + \theta_1(1-\theta_1)}{(\theta_2 - \theta_1)^2}[\Phi^{-1}(\beta) + \Phi^{-1}(\alpha/2)]^2$$

Sample Size Calculations — Two sample test for means NB number in each group

$$n \simeq \frac{2\sigma^2}{(\mu_2 - \mu_1)^2}[\Phi^{-1}(\beta) + \Phi^{-1}(\alpha/2)]^2$$

Standard Error for Natural Logarithm of Relative Risk

$$s.e[(log_e(RR)] = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

Standard Error for Natural Logarithm of Odds Ratio

$$s.e[(log_e(OR)] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

2 Survival Analysis Formulae

Exponential Distributions — MLE of rate λ with censoring The mle

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} = \frac{\Delta}{\mathcal{T}} \qquad var(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{\sum_{i=1}^{n} \delta_i}.$$

For any (differentiable, monotonic) function $g(\cdot)$,

$$var(g(\hat{\lambda})) \approx \left[\{g'(\lambda)\}^2 var(\lambda)\right]_{\lambda=\hat{\lambda}}.$$

so e.g.

$$var(\frac{1}{\hat{\lambda}}) = var(\hat{\mu}) \approx \frac{\hat{\mu}^2}{\sum_{i=1}^{n} \delta_i}$$

Exponential Distributions — MLE test

$$W = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\frac{\hat{\lambda}_1^2}{\Delta_1} + \frac{\hat{\lambda}_2^2}{\Delta_2}}} \approx N(0,1).$$

Exponential Distributions — LRT test

$$2\left\{\Delta_1 \log\frac{\Delta_1}{\mathcal{T}_1} + \Delta_2 \log\frac{\Delta_2}{\mathcal{T}_2} - (\Delta_1+\Delta_2)\log\frac{\Delta_1+\Delta_2}{\mathcal{T}_1+\mathcal{T}_2}\right\} \approx \chi_1^2$$

Log-rank Statistic

$$LR = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \sim \chi_1^2$$

3 Tables of Percentage Points (also known as Quantiles or Critical Values) for Three Standard Distributions

The tables contain values of quantiles q such that $P[X \leq q] = p$ for various probabilities p when X has the specified distribution (which may depend on particular degrees of freedom ν). In these tables, p has been expressed as a percentage rather than a decimal. The relevant R commands for generating the q are also shown. For the $N(0,1)$ distribution, the tabulated function is also known as the Φ^{-1} function.

STANDARD NORMAL DISTRIBUTION PERCENTAGE POINTS
`qnorm`(p) where p is percentage, e.g. for 95%, $p = 0.95$

	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
`qnorm`	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

CHI-SQUARED PERCENTAGE POINTS
`qchisq`(p, ν) where p is percentage, e.g. for 95%, $p = 0.95$

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588

STUDENT'S t PERCENTAGE POINTS

`qt(`p, ν`)` where p is percentage, e.g. for 95%, $p = 0.95$

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090