Australian National University

RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS

Second Semester Final Exam 2012

Survival Models / Biostatistics

(STAT3032/7042/8003)

Duration: 15 minutes reading time and 3 hours writing time

Permitted materials: Calculators, lecture notes, dictionary

You must attempt to answer all questions.
All questions are to be completed in the script book provided.

Question 1 (10 marks = 5 +5)

a) Under the assumption of a uniform distribution of deaths (i.e. ${}_t p_x \mu_{x+t}$ is a constant for $0 \le t \le 1$), show that ${}_{t-s}q_{x+s} = \dfrac{(t-s)q_x}{1-s\times q_x}$ $0 \le s \le t \le 1.$

Solution:

We know that ${}_t p_x = {}_s p_x \, {}_{t-s} p_{x+s}$ and ${}_{t-s} p_{x+s} = 1 - {}_{t-s} q_{x+s}$. Using these results we can then show the following:

$$= 1 - \frac{{}_t p_x}{{}_s p_x} = 1 - \frac{1 - {}_t q_x}{1 - {}_s q_x} = 1 - \frac{1 - tq_x}{1 - sq_x} = \frac{(t-s)q_x}{1 - sq_x}.$$

b) Show that $\dfrac{d}{dx} {}_t p_x = (\mu_x - \mu_{x+t}) {}_t p_x.$

Solution:

$$\log {}_t p_x = \log(l_{x+t}) - \log(l_x)$$
$$\Rightarrow \frac{1}{{}_t p_x} \frac{d}{dx} {}_t p_x = \frac{1}{l_{x+t}} \frac{d}{dx} l_{x+t} - \frac{1}{l_x} \frac{d}{dx} l_x = (\mu_x - \mu_{x+t}) {}_t p_x.$$

Question 2 (10 marks = 2+2+2+2+2)

In this question you will be looking at the breastfeeding data that was discussed in class. As a reminder, this dataset contains information on the duration of breastfeeding for over 900 mothers as well as information on a number of covariates. To investigate the effects of these covariates a Cox regression model was fitted. The following is output from a Cox regression analysis conducted in R:

```
cox.mod<-
coxph(Surv(duration,delta)~as.factor(race)+as.factor(poverty)+as.factor(smoke)+as.f
actor(alcohol)+agemth+I(agemth^2)+ybirth+yschool+pc3mth,data=bfeed)

> summary(cox.mod)

Call:

coxph(formula = Surv(duration, delta) ~ as.factor(race) + as.factor(poverty) +

    as.factor(smoke) + as.factor(alcohol) + agemth + I(agemth^2) +

    ybirth + yschool + pc3mth, data = bfeed)


  n= 927, number of events= 892


                        coef exp(coef)  se(coef)      z Pr(>|z|)
```

```
as.factor(race)2     0.187823  1.206620  0.105469  1.781 0.074940 .
as.factor(race)3     0.296366  1.344962  0.097186  3.049 0.002293 **
as.factor(poverty)1 -0.226792  0.797087  0.094469 -2.401 0.016363 *
as.factor(smoke)1    0.246100  1.279027  0.079599  3.092 0.001990 **
as.factor(alcohol)1  0.167771  1.182666  0.123253  1.361 0.173454
agemth              -0.173193  0.840975  0.186168 -0.930 0.352212
I(agemth^2)          0.003615  1.003621  0.004251  0.850 0.395118
ybirth               0.079672  1.082932  0.020505  3.886 0.000102 ***
yschool             -0.056739  0.944841  0.023167 -2.449 0.014320 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Concordance= 0.577  (se = 0.012 )
Rsquare= 0.05   (max possible= 1 )
Likelihood ratio test= 47.1  on 10 df,   p=9.051e-07
Wald test            = 47.41  on 10 df,   p=7.956e-07
Score (logrank) test = 47.48  on 10 df,   p=7.728e-07
```

The variables *race*, *poverty*, *smoke*, and *alcohol* are categorical variables representing the race of the mother, whether the mother is living in poverty, whether the mother smokes, and whether the mother consumes alcohol. The variables *agemth*, *ybirth*, and *yschool* are continuous variables representing the age of the mother, the year of birth of the child and the numbers of years that the mother attended school. Based on the above output answer the following questions:

a) Provide a 95% confidence interval for the multiplicative change in the hazard for a 1 year increase in the variable *yschool*, everything else held constant.

$$\exp(-0.056739 \pm 2 \times 0.023167)$$

b) Conduct a test of the overall significance of the fitted Cox model, that is, test the null hypothesis all the parameters in the model (the $\beta's$) are zero versus the alternative that at least one of the parameters in the model is non-zero. Conduct this test at the 5% level. You must state both the value of the test-statistic and its associated p-value.

This test can be conducted using either of the following tests from the R output

```
Likelihood ratio test= 47.1  on 10 df,   p=9.051e-07
Wald test            = 47.41  on 10 df,   p=7.956e-07
Score (logrank) test = 47.48  on 10 df,   p=7.728e-07
```

Clearly the null hypothesis is rejected at the 5% level.

c) Provide an estimate of the quantity $\frac{1}{\beta_8}$, where β_8 is the parameter corresponding to *yschool*. You must provide a standard error for your estimate.

Need to use the delta method here. Our estimate is $\frac{1}{-0.056739}$ and the standard error is: $(0.023167)^2 \times (\frac{1}{-0.056739})^2$

d) After the above analysis had been conducted you are told that the values of *duration* were mistakenly reported using units of days rather than months. For example, a duration of 4 months was accidentally recorded as a duration of 4 days etc. How would this information change the parameter estimates reported above? You must provide a reason for your answer.

Estimates would not change. All that matters is the order of the deaths.

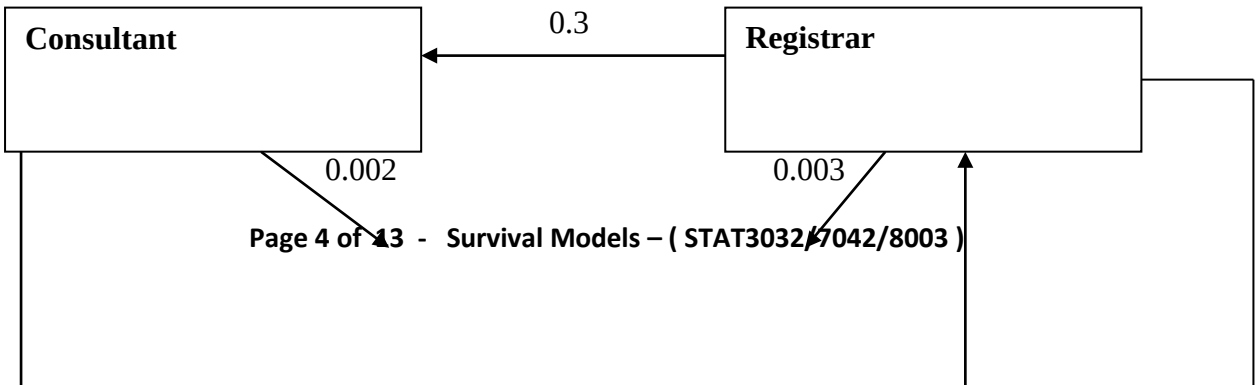
e) In class we saw that when we produced Kaplan-Meier survival curves for the exact data analyzed above, broken down by the variable *poverty* that we were unable to conclude (using a log-rank test) that the two survival curves were different (i.e. the curve corresponding to mothers living in poverty and the curve corresponding to mothers not living in poverty). However, for the above fitted Cox model we note that poverty appears to be a significant variable. Briefly describe reasons for this apparent contradiction.

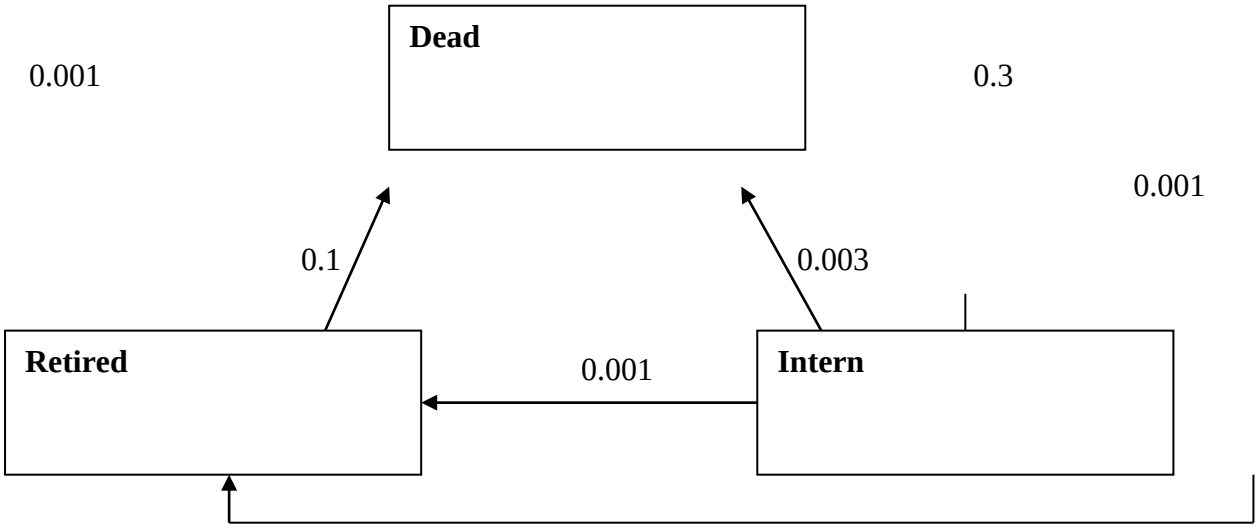
There are a number of reasons: (1) models are based on different assumptions and (2) cox model adjusted for many other covariates.

Question 3 (10 marks = 2+2+2+2+2)

A multi-state Markov model has been used to describe the transition of medical doctors between the various levels of their career. The levels of career are *consultant*, *registrar* and *intern*. In total the Markov model has five states.

The transition intensities given in the figure below were estimated by observing 1000 years of waiting time in each of the three states consultant, registrar and intern and 10 years for the retired state. For example, the estimated transition intensity between registrar and retired is 0.001.





Based on the above information and figure answer the following questions:

(a) How many of registrars were observed to move to become consultants during the period of observation?

300

(b) Calculate the probability that a new registrar will still be a registrar in 3 years' time.

$$\exp\left(-3(0.304)\right) = 0.40172.$$

(c) Calculate a 90% confidence interval for the probability calculated in (b).

$$\hat{\lambda}_{AF} = 0.3 \quad Var\left(\hat{\lambda}_{AF}\right) = \frac{300}{1000^2} \quad \hat{\lambda}_{AD} = 0.003 \quad Var\left(\hat{\lambda}_{AD}\right) = \frac{3}{1000^2} \quad \hat{\lambda}_{AE} = 0.001 \quad Var\left(\hat{\lambda}_{AE}\right) = \frac{1}{1000^2}$$

$$Var\left(\hat{\lambda}_{AF} + \hat{\lambda}_{AD} + \hat{\lambda}_{AE}\right) = \frac{304}{1000^2}$$

Let $Y = \hat{\lambda}_{AF} + \hat{\lambda}_{AD} + \hat{\lambda}_{AE}$ and using the delta method,

$$Var\left(e^{-3Y}\right) \approx \left(-3e^{-3Y}\right)^2 .Var\left(Y\right) = 9\left(e^{-0.912}\right)^2 \frac{304}{1000^2} = 0.00044$$

Hence an approximate 95% CI is

$$0.40172 \pm 2\sqrt{0.00044}.$$

(d) Jennifer has been an intern for exactly two years. Calculate the 90th percentile (ie the upper 10th percentile) of the probability distribution of the amount of remaining time that Jennifer will spend as an intern.

Force of exit = 0.304.

$$S_T(t) = e^{-0.304t} = 0.1 \rightarrow t = -\frac{1}{0.304}\ln(0.1) = 7.58 \text{ years.}$$

(e) Explain briefly why the Markov assumption may not be appropriate in this model of the career progression of doctors. Use two examples to illustrate your answer.

(1) More time spend in a state gives rise to increasing probability of death.
(2) Over-time increased chance of moving from the intern state.

Question 4 (10 marks = 2 + 2 + 2 + 2)

A set of crude mortality rates were graduated using a particular smoothing technique. The smoothing technique that was used required 5 parameters to be estimated. The table below provides details of the graduation:

Age	Crude Rate	Graduated Rate	Population	Deaths	Expected Deaths	Variance	Standardised Deviation
20	0.028884	0.02287	1004	29	22.96	22.44	1.2748
21	0.01626	0.02525	984	16	24.85	24.22	-1.7975
22	0.040123	0.02778	972	39	27	26.25	2.3416
23	0.018386	0.03049	979	18	29.85	28.94	-2.2027
24	0.04375	0.03339	960	42	32.05	30.98	1.7867
25	0.022293	0.03648	942	21	34.36	33.11	-2.3225
26	0.052295	0.03978	937	49	37.27	35.79	1.9601
27	0.031083	0.04332	933	29	40.42	38.67	-1.8361
28	0.061159	0.04712	932	57	43.92	41.85	2.0226
29	0.039474	0.05122	912	36	46.71	44.32	-1.6092
30	0.060241	0.05566	913	55	50.82	47.99	0.6037
31	0.048206	0.06047	892	43	53.94	50.68	-1.5367
32	0.078409	0.0657	880	69	57.82	54.02	1.5217
33	0.067045	0.07139	880	59	62.82	58.34	-0.5006
34	0.087558	0.07759	868	76	67.35	62.12	1.0977
35	0.077816	0.08434	861	67	72.62	66.49	-0.6888
36	0.104094	0.09167	855	89	78.38	71.19	1.2589
37	0.092216	0.09963	835	77	83.19	74.9	-0.7153
38	0.112961	0.10824	841	95	91.03	81.18	0.4406
39	0.102871	0.11752	836	86	98.25	86.7	-1.3153
40	0.142327	0.12747	808	115	103	89.87	1.2663
					Total	**1070.04**	

(a) Perform the chi-square test to see whether check whether the graduated rates are appropriate. State the null and alternative hypotheses, test statistic, critical value and conclusion of the statistical test. Use a significance level of 5%.

H_0: The graduation is appropriate

H_1: The graduation is not appropriate

Test Statistic = 50.2672 which is chi-squared with 21-5 degrees of freedom.

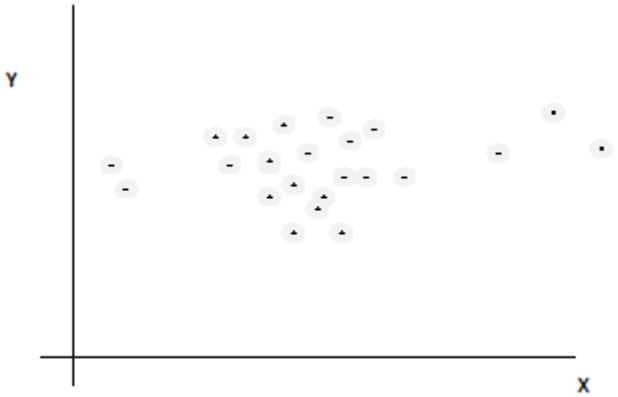
Critical Value at the 5% significance level is 26.29.

Conclusion is to reject the null.

(b) Calculate the test statistic for sign test. State the conclusion of the sign test using a significance level of 5%.

There are 11 positive deviations. Our test-statistic is 11-(21*0.5) / sqrt(21*0.5*0.5)=0.22. Clearly we fail to reject the null hypothesis.

(c) The figure below shows a plot of a particular set of data that is going to be smoothed using Kernel smoothing (in conjunction with the Normal Kernel).



Would you have any concerns smoothing the above data using kernel smoothing?

Low density of points in some areas.

(d) In class we discussed Kernel smoothing using both the triangle Kernel and also the rectangular kernel. State two advantage of using a triangular kernel over a rectangular kernel?

Question 5 (10 marks = 2 + 2 + 2 + 2+2)

Below the survival function for the lifetimes of a particular population is given. For this survival function the parameter φ represents the smallest time at which a death could occur.

$$s(x) = \begin{cases} 1 & \text{if } x \le \varphi \\ \exp[-\lambda(x-\varphi)^{\alpha}] & \text{if } x > \varphi \end{cases}.$$

Based on the above survival function answer the following questions:

(a) Find the density function for this population.

(b) Find the hazard function for this population.

(c) Provide an example for humans where this form of the survival function might be appropriate.

In answering the following two questions you are given that $\alpha = 1, \lambda = 0.0075$, and $\phi = 100$.

(d) Compute the mean survival time of this population.

(e) Compute the median survival time of this population.

Solution:

(a) $f(x) = \lambda\alpha(x-\varphi)^{\alpha-1}\exp[-\lambda(x-\varphi)]^{\alpha}$ and 0 if $x < \varphi$.

(b) $h(x) = \lambda\alpha(x-\varphi)^{\alpha-1}$ and 0 if $x < \varphi$.

(c) Length of time students take to complete and exam where the students cannot leave the exam room for at least one hour. Time in hospital after a particular operation and the hospital says patients must stay in hospital for at least 12 hours. Anything along these lines is okay.

(c) 233.33

(e) 192.42

Question 6 (8 marks = 2 + 2 + 2 + 2)

The output below is from a Kaplan-Meir survival curve fitted to the breastfeeding data of question 2.

```
> KM.est<-
survfit(Surv(duration,delta)~alcohol,data=bfeed,conf.type="plain")

> KM.est
```

```
Call: survfit(formula = Surv(duration, delta) ~ alcohol, data = bfeed,
    conf.type = "plain")

          records n.max n.start events median 0.95LCL 0.95UCL
alcohol=0     848   848     848    816     12       9      12
alcohol=1      79    79      79     76      8       6      12
> summary(KM.est)
Call: survfit(formula = Surv(duration, delta) ~ alcohol, data = bfeed,
    conf.type = "plain")

                alcohol=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    848      69  0.91863 0.00939      ???????      ???????
    2    778      65  0.84188 0.01253      0.81732      0.86645
    3    710      46  0.78734 0.01407      0.75977      0.81491
    4    661      61  0.71468 0.01554      0.68421      0.74514
    5    597      16  0.69553 0.01585      0.66447      0.72659
    6    578      53  0.63175 0.01664      0.59913      0.66436
    7    519      11  0.61836 0.01677      0.58549      0.65123
    8    505      68  0.53509 0.01729      0.50121      0.56898
    9    435       3  0.53140 0.01730      0.49750      0.56531
   10    431      16  0.51168 0.01734      0.47768      0.54567
   11    414       2  0.50921 0.01735      0.47520      0.54321
   12    412      69  0.42393 0.01722      0.39018      0.45767
   13    343       3  0.42022 0.01720      0.38651      0.45392
   14    338       6  0.41276 0.01716      0.37913      0.44639
   15    331       5  0.40652 0.01713      0.37296      0.44009
   16    326      49  0.34542 0.01663      0.31283      0.37801
   17    277       1  0.34417 0.01661      0.31161      0.37674
   18    276       9  0.33295 0.01649      0.30063      0.36527
```

20	265	29	0.29651	0.01601	0.26513	0.32790
21	236	1	0.29526	0.01599	0.26391	0.32660
22	235	2	0.29274	0.01596	0.26147	0.32402
24	232	58	0.21956	0.01458	0.19099	0.24813
25	174	1	0.21830	0.01455	0.18979	0.24681
26	173	5	0.21199	0.01440	0.18377	0.24021
28	168	18	0.18927	0.01381	0.16220	0.21635
30	150	2	0.18675	0.01375	0.15981	0.21369
32	148	25	0.15521	0.01279	0.13014	0.18027
34	123	1	0.15394	0.01275	0.12896	0.17893
36	122	16	0.13375	0.01203	0.11017	0.15734
38	106	2	0.13123	0.01194	0.10783	0.15463
40	104	16	0.11104	0.01112	0.08925	0.13283
42	88	3	0.10726	0.01095	0.08579	0.12872
44	85	10	0.09464	0.01036	0.07432	0.11495
46	75	2	0.09211	0.01024	0.07204	0.11218
48	73	29	0.05552	0.00812	0.03961	0.07143
50	44	2	0.05300	0.00794	0.03743	0.06857
52	42	16	0.03281	0.00632	0.02042	0.04520
56	26	5	0.02650	0.00570	0.01533	0.03767
60	21	5	0.02019	0.00499	0.01040	0.02998
64	16	2	0.01767	0.00468	0.00850	0.02683
68	14	1	0.01640	0.00451	0.00756	0.02524
72	13	4	0.01136	0.00376	0.00398	0.01873
76	9	1	0.01009	0.00355	0.00314	0.01705
80	8	1	0.00883	0.00332	0.00232	0.01535
96	7	5	0.00252	0.00178	0.00000	0.00602
120	2	1	0.00126	0.00126	0.00000	0.00373
192	1	1	0.00000	NaN	NaN	NaN

```
                alcohol=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     79       8   0.8987  0.0339      0.83221       0.9653
    2     70       6   0.8217  0.0432      0.73701       0.9064
    3     64       3   0.7832  0.0466      0.69193       0.8744
    4     61       9   0.6676  0.0533      0.56318       0.7721
    5     52       3   0.6291  0.0547      0.52198       0.7362
    6     49       3   0.5906  0.0557      0.48152       0.6997
    7     46       4   0.5392  0.0564      0.42864       0.6498
    8     42       4   0.4879  0.0566      0.37696       0.5988
   10     38       3   0.4494  0.0563      0.33897       0.5598
   12     35       6   0.3723  0.0548      0.26502       0.4796
   13     29       2   0.3467  0.0539      0.24101       0.4523
   14     27       1   0.3338  0.0534      0.22913       0.4385
   16     26       5   0.2696  0.0503      0.17110       0.3681
   18     19       1   0.2554  0.0496      0.15825       0.3526
   20     18       1   0.2412  0.0488      0.14556       0.3369
   21     17       1   0.2270  0.0480      0.13304       0.3211
   24     16       4   0.1703  0.0436      0.08489       0.2557
   28     12       1   0.1561  0.0422      0.07341       0.2388
   32     11       4   0.0993  0.0351      0.03050       0.1682
   36      7       1   0.0851  0.0328      0.02077       0.1495
   40      6       2   0.0568  0.0273      0.00316       0.1104
   48      4       3   0.0142  0.0141      0.00000       0.0418
  104      1       1   0.0000     NaN          NaN          NaN

> survdiff(Surv(duration,delta)~alcohol,data=bfeed,rho=0)
Call:
survdiff(formula = Surv(duration, delta) ~ alcohol, data = bfeed,
    rho = 0)
```

```
             N Observed Expected (O-E)^2/E (O-E)^2/V
alcohol=0 848      816    826.3     0.129      2.01
alcohol=1  79       76     65.7     1.628      2.01

 Chisq= 2  on 1 degrees of freedom, p= 0.156
```

Based on the above output answer the following questions:

(a) Compute the two missing values labeled ??????? in the above output.

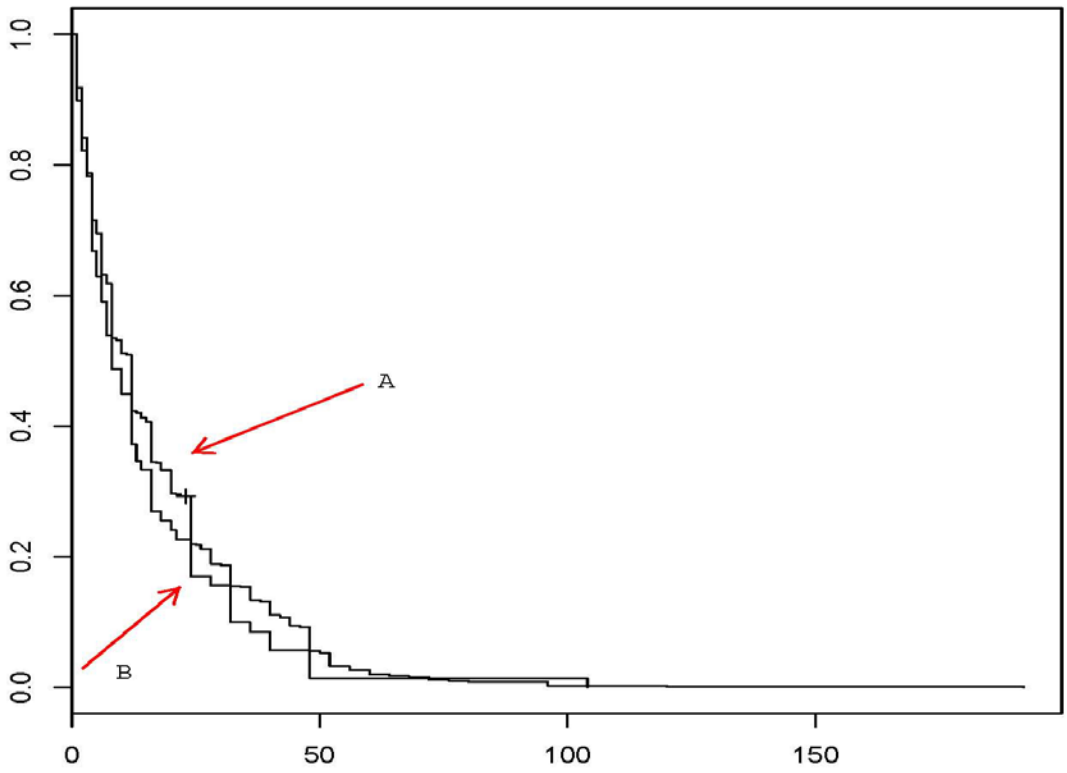
(0.90023,0.93703)

(b) At the 5% level of significance are you able to conclude that the survival curves based on different alcohol status are different?

P value for Chisq test is 0.156, we fail to reject null. We conclude that the survival curves based on different alcohol status are not different.

(c) The R command plot(KM.est) was used to produce the plot below [ideally we would want to label these plot!]. Which of the two curves (curve “A” or curve “B”) corresponds to the alcohol=1?

Curve B because alcohol=1 has less records than alcohol=0. We have reasons to believe alcohol=1 should have a lower surviving probability.



(e) Provide a rough estimate of the mean duration of breastfeeding for a mother who consumes alcohol. Please provide clear justification for the method you use to get your rough estimate.

Calculate the area under the curve. Details see mid-term exam 2018.