STOR 557: Fall 2023
Midterm Two: October 31, 2023

Open book in-class exam: time limit 75 minutes.

You are allowed to consult course notes and text (printed or e-read), homework assignments and any personal notes you have made during the course. Other outside materials are not permitted. Computers or ipads may be used *only* for the purpose of accessing pre-stored course notes; they are not to be used for computations during the exam. A hand-held calculator is permitted. Answers should preferably be written in a university examination book ("blue book"). You may consult the instructor if the wording is unclear or if you think there might be an error, but the instructor will not give hints how to solve the exam. The university Honor Code is in effect at all times.

The whole exam is worth 100 points (60 points for question 1, 40 for question 2). Statistical tables are provided.

1. Alcohol consumption is widely believed to be a causative factor in oesophageal cancer. A study of 975 hospital patients in France classified the patients according to whether or not they had oesophageal cancer, their alcohol consumption (high or low), and their age group. The raw data are as follows:

	Cancer		No Cancer	
Age Group	High Alc	Low Alc	High Alc	Low Alc
25–34	1	0	9	106
35–44	4	5	26	164
45–54	25	21	29	138
55–64	42	34	27	139
65–74	19	36	18	88
75+	5	8	0	31
Total	96	104	109	666

Note that the last row of the data is derived by simply adding up the numbers in the first six rows, and therefore represents the combined table of oesophageal cancer against alcohol consumption without taking account of age.

(a) Based on the last row of the table, the Pearson (continuity-corrected) chi-square statistic, for the null hypothesis that alcohol consumption and incidence of oesophageal cancer are independent, is 108.22. Is this a statistically significant result, and what do you conclude? **[7 points.]**

(b) The same calculation was repeated, separately, for each of the first six rows of the above table, with the resulting chi-square statistics of 2.19, 4.18, 24.1, 37.0, 5.36 and 9.9. State which of these values is statistically significant, and summarize your conclusion in words. **[7 points.]**

(c) Are the results in (a) and (b) an example of Simpson's paradox? What are the advantages and disadvantages of each approach? **[6 points.]**

(d) The data in the above table were combined into a dataframe "OC" with a numerical variable "Count" and factor variables "Cancer", "Alcohol" and "Agegp". The following code was run:

```
ct3=xtabs(Count~Cancer+Alcohol+Agegp,OC)
apply(ct3,3,function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
mantelhaen.test(ct3,exact=T)
```

The results were as follows:

```
       1        2        3        4        5        6
     Inf 5.046154 5.665025 6.359477 2.580247      Inf
Exact conditional test of independence in 2 x 2 x k tables
data:  ct3
S = 666, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 3.572140 7.758317
sample estimates:
common odds ratio
         5.250951
```

Explain what this code is doing and how you interpret the result. Why are two of the numbers in the second row stated as "Inf"? How does the "common odds ratio" of 5.250951 relate to the six numbers (including the two labelled "Inf") in the second row? **[15 points.]**

(e) A series of Poisson regression models was fitted to the data, as follows:

```
g1=glm(Count~Alcohol+Cancer+Agegp,family=poisson,OC)
g2=glm(Count~Alcohol*Cancer+Agegp,family=poisson,OC)
g3=glm(Count~Alcohol*Agegp+Agegp*Cancer,family=poisson,OC)
g4=glm(Count~(Alcohol+Cancer+Agegp)^2,family=poisson,OC)
g5=glm(Count~Alcohol*Cancer*Agegp,family=poisson,OC)
```

A summary of the results is given as follows:

Model	Residual Deviance	Residual DF
g1	224.2	16
g2	145.9	15
g3	90.56	6
g4	11.04	5
g5	0	0

Which do you conclude is the best of the five models? Give relevant calculations to support your conclusion. **[15 points.]**

(f) State in words what you conclude about the relationships among age, alcohol consumption and incidence of oesophageal cancer. **[10 points.]**

QUESTION TWO ON THE NEXT PAGE

2. Twelve hospital patients were given an experimental dietary regime. After each week, their levels of plasma ascorbic acid were measured. The results are illustrated graphically in Figure 1. Each broken-line curve represents one patient. It can be seen that most patients demonstrated a rise followed by a fall in their levels of ascorbic acid, but visually, there appears to be significant variability among patients.

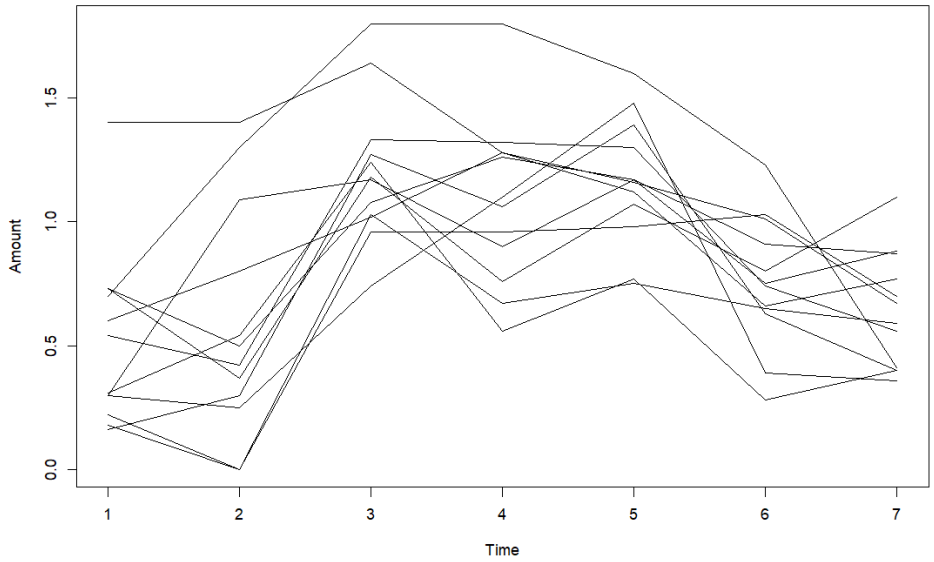

Figure 1: Weekly Variations in Plasma Ascorbic Acid for 12 Patients.

(a) The data was collected in a dataframe "Diet" with variables "Time" (factor), "Person" (factor) and "Amount" (numeric). A simple analysis of variance through the R function

```
a1=aov(Amount~Time+Person,Diet)
```

produced the following output

```
> summary(a1)
            Df Sum Sq Mean Sq F value   Pr(>F)
Time         6  6.165  1.0274  15.241 7.25e-11 ***
Person      11  3.682  0.3347   4.966 1.42e-05 ***
Residuals   66  4.449  0.0674
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Briefly interpret the results of this analysis. What is its main limitation? **[8 points.]**

For the next three parts, I want you to write brief R code for each of the following operations:

(b) Fit the same model as a mixed model with "Time" treated as a fixed effect and "Person" as a random effect. **[8 points.]**

(c) Within this model, write R code for *two* ways to test the significance of "Time". **[10 points.]**

(d) Within this model, write R code for one way to test the significance of the random effect of "Person". **[8 points.]**

(e) When the test of part (d) is applied to the model that includes both "Time" as a fixed effect and "Person" as a random effect, the p-value appears as "$<$ 2.2e-16" (which is 0 for all practical purposes). However when the same test is applied to the model from which "Time" has been omitted (so that the random effect due to "Person" is the only non-constant term in the model), the p-value is about 0.017. How would you interpret the discrepancy between these two results? **[6 points.]**

Sketch Solutions

Note: There is no single "right answer" to questions like these. Credit will be given for answers that pick up key points in the analysis, that may not necessarily agree with those here.

1. (a) The distribution of the Pearson statistic, if the null hypothesis of no association is true, is (asymptotically) χ^2_1, but the value 108.22 is way beyond the percentage points for that distribution (you didn't have the means to calculate this exactly in the exam, but you can read from the table that the 0.001-upper tail point is 10.83, and the observed value is far larger than that). Conclusion: there is evidence of an association between alcohol consumption and oesophageal cancer. Credit for correct statement of null hypothesis, correct discussion of χ^2_1 statistic, and final conclusion.

   (b) The p-value for the same null hypothesis in the 25–34 age group is > 0.05, and that for the 35-44 age group is between 0.01 and 0.05 (this is based on the relevant critical values of χ^2_1, which are 3.84 and 6.63). Thus the result for 25–34 is not statistically significant, and that for 35-44 is only mildly statistically significant. The other values are all statistically significant with p-values < 0.001 except for the 75+ age group. The conclusion is that there is strong evidence for an association in most age groups, though not as strong as when all age groups are combined. Credit for correctly identifying which age groups are significant and for correctly interpreting the χ^2_1 statistics. However you may also comment that the lack of statistical significance in the first two age groups is likely because of the low number of cancer cases overall, while that in the 75+ age group has far fewer individuals than some of the other groups. In other words, these may be cases where acceptance of the null hypothesis arises because of insufficient data and not because the null hypothsis is correct (you don't have to say all that, but I expect to see some reasoning along those lines).

   (c) This is not an example of Simpson's paradox because in all cases there is a positive association between the two variables so there is no reversal of the effect which is the characteristic of Simpson's paradox. Nevertheless you may note that both cancer rates (certainly) and alcohol consumption (possibly) vary with age group, so the main message of Simpson's paradox — the need to consider the subgroups separately — is valid for this example. You may also note that the fact that all the odds ratios are > 1 (see answer to part (d)) shows that the association is positive in every case — an odds ratio < 1 would imply a negative association.

   (d) This code is carrying out the Mantel-Haenszel test and the final conclusion (p-value $< 2.2e-16$) shows that the result is very highly statistically significant — in other words, we strongly reject the null hypothesis that there is no association between alcohol and cancer across the whole table. The odds ratio 5.250951 is the estimated combined odds ratio under the assumption that it is the same for all age groups. The six numbers in line 2 represent individual odds ratios that are specific to each group, and the two labelled "Inf" arise because in each case one of the cell counts is 0, which is in the denominator of the odds ratio calculation. Extra credit if you gave examples to show the odds ratios themselves (e.g. for the 45–54 age group, $OR = \frac{25\times138}{21\times29} = 5.665025$, as given).

   (e) g1 and g2 don't make sense because they imply that neither cancer nor alcohol is dependent on age, which we know to be false at least so far as cancer is concerned. g3 does

allow for an age dependence but not for any association between alcohol and cancer, but the high deviances imply that none of these models is supported by the data. g4 is plausible because it does allow for all the two-way associations but not for a three-way association. g5 allows for all the associations but it is not an attractive model because it just reproduces the data and has has no explanatory or predictive value (there are other ways you could express this, but the point is that there are reasons for not preferring this model). As far as a formal test is concerned, g4 shows a deviance of 11.04 with 5 degrees of freedom, and the 0.05 upper tail point for χ^2_5 is 11.07. So the test is borderline, but I would still find g4 the more attractive model (any answer advocating g4 or g5 will be accepted if you give appropriate arguments to support your choice).

(f) There are different ways to express this and some of your answer may have appeared under (e), but I would summarize the main point as follows: we need to take account of different age groups because of the confounding effect of cancer and age, but both within most of the age groups and when combined into one analysis, there is very strong evidence that alcohol use is associated with oesophageal cancer. If you're careful about how you express this, you could add that the statistical calculations don't necessarily imply that the association is causal.

2. (a) Both the Time effect and the Person (or Patient) effect are clearly significant with p-values well under 0.05. So the response does depend on time, and there are also clear differences among the patients. As for limitations: there is no test for interactions here (not possible with only one observation for each Time-Person combination); you could also note that treating both effects as fixed is itself a limitation, since the patients are clearly only a sample from a potentially much larger population, so a random effects model is more logical.

(b) Possible code to fit the mixed model is

```
library(lme4)
m1=lmer(Amount~Time+(1|Person),Diet)
summary(m1)
```

Equivalent code to fit the same model will be accepted.

(c) The point of this question was to highlight that there are two ways of doing it: the Kenward-Roger test (an approximate test, but not involving simulation) or the bootstrap (which does use simulation). Again, I would accept any code that produces both these tests, but something like:

```
library(pbkrtest)
m2=lmer(Amount~(1|Person),Diet)
KRmodcomp(m1,m2)
PBmodcomp(m1,m2)
```

(d) In this case we need a bootstrap procedure: again there are different ways to do it, but one possible code is

```
library(RLRsim)
exactRLRT(m1)
exactRLRT(m2)
```

though note that only the test for m1 is needed to answer the question in the form it was asked. (There are other ways of doing this, described on page 215 of the course text. Credit will also be given for one of those methods so long as it is carefully described, with explicit R code to implement.)

(e) Both results are statistically significant so there's no contradiction to resolve, but one way to explain the difference is this: omitting the Time effect is equivalent to assuming the response is constant across time points, and if we did assume that, then our estimate of the residual variability would increase whereas the inter-person variability is presumably about the same. However in that case, the residual variability could well be masking the effect of inter-person variability. (In fact the estimated variances for Person and Residual are 0.1954 and 0.2596 under m1, and 0.1636 and 0.3839 under m2, so the Residual variance is substantially higher under m2.)