Remember, I want your name only on the back of the last page of your answers. If you have answers written on that page, please put your name on a clean sheet of paper.

R and SAS code and output for problem 1 will be found after the questions. Unless specifically stated, all R models use the default contrasts (contr.treatment for lm and contr.helmert for lmer).

1. Further west from Iowa, it is quite common to find road signs with gunshot holes in them. The following data were collected on one of my trips West. As I drove past a sign along the road, I noted whether or not it had been shot. Here are data from 2 states. I counted signs for a set number of miles, but the number of miles was not the same in the two states. The total number of signs for each state is random.

State	shot	not	total	miles
Iowa	2	48	50	10
Colorado	20	20	40	20

   (a) 3 pts. These data an example of which sampling model for contingency tables?

   (b) 5 pts. Test whether the probability that a sign has been shot is the same in both states. Report your test statistic and give its distribution under the null hypothesis. Or, if none of the provided code includes an appropriate test, say "can't be done" and briefly describe what information you need / test you want to do.

   (c) 5 pts. Test whether the number of shot-at signs per mile is the same in both states. Report your test statistic and give its distribution under the null hypothesis. Or, if none of the provided code includes an appropriate test, say "can't be done" and briefly describe what information you need / test you want to do.

   Further west (Utah and Nevada), I recorded data separately for each mile. Analyses of these data are labeled "Data set 2".

   (e) 5 pts. You want to compare the mean number of shot-at signs per mile in the two states (Utah and Nevada). Write a model that allows you to do this. Define your subscripts and make sure you specify the distribution of the response variable.

   (f) 5 pts. Define μ_i as the mean number of shot-at signs per mile in state i. Estimate the logarithm of the ratio μ_{Utah}/μ_{Nevada}. If none of the provided code includes an appropriate estimate, say "can't be done" and briefly describe what information you need to construct an appropriate estimate or s.e.

   (g) 5 pts. What is the standard error of the log transformed ratio in part 1f?

2. Combines are large, expensive machines that cut down a crop (mostly corn or soybeans in IA) when it is ready to harvest and separate the grain (the good stuff that makes money for the farmer) from other unwanted plant parts (stems, leaves, cobs, husks). The grain goes into the hopper, a storage area in the combine; the unwanted stuff goes out the back. A few years ago, I was involved in the design and analysis of studies in the Ag. Engineering department.

Researchers there developed two new designs for the part of a combine called the thresher. It is surprisingly hard to design a good thresher. A bad thresher pulverizes the grain and sends it out the back of the combine, instead of into the hopper. They wanted to know if one of their new designs was more efficient (more grain into the hopper). The efficiency of the thresher could depend on the speed (miles per hour) of the combine and the slope of the land (the combine was expected to be more efficient on flat land than when the land was sloping).

Efficiency data (kg corn per hectare harvested) were collected using the follwing design. The one experimental combine available to the researchers was taken to a field with one of three (3) average slopes: steep, shallow, or flat. The combine was fitted with one of the three (3) thresher designs and harvested one-third of the field. For simplicity, I'll call the 1/3'rd field a field-part. Thresher designs were randomly assigned to "field-parts". While harvesting a field-part, the combine was driven at one of four (4) speeds. Hence, there are 12 "field-bits" in each field corresponding to one combination of thresher design and speed. Combine speed was randomly assigned to a "field-bit" within a "field-part". The amount of grain collected in the hopper was recorded for each "field-bit". You can consider slope to be randomly assigned to field, althrough that is an observational factor and not strictly randomly assigned. This design was repeated on a total of 12 fields, 4 for each slope. Summary: 12 fields, 36 "field-parts" and 144 "field-bits".

(a) 5 pts. How many different sizes of experimental units (including observational factors treated like experimental units) are there in this study? What are those experimental units?

(b) 5 pts. What treatments levels are randomly assigned (or observationally "assigned") to each size of experimental unit?

(c) 5 pts. Write out the non-zero columns of the $\boldsymbol{Z}$ matrix for the following four observations. Only include the columns with a non-zero value for one or more of these four observations. Do not include any column that is all zeros for these four observations.

Slope	Field	Design	Speed
steep	1	A	5 mph
steep	1	A	12 mph
flat	5	A	5 mph
flat	5	B	5 mpgh

(d) 10 pts. Write out the skeleton ANOVA table (sources of variability and associated d.f.) for this study.

3. This question is based around the question "does the average dietary intake of Vitamin A change with age?". Daily intake of Vitamin A is measured by asking participants to record (in a food diary) the foods they eat and the portion size each day. Researchers convert this information to a daily intake of Vitamin A. In this study, subjects were recruited in 3 age groups (teen, young adult, and middle aged). There were 10 subjects in each age group. Each subject filled out the food dairy for 10 days randomly selected by the researchers. For each

subject, 5 of the days were week days and 5 of the days were weekend days. Summary: 30 subjects, 300 observations.

One possible model that accounts for variation among subjects and variation among days within a subject is:

$$\begin{aligned} Y_{ijkl} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_{ik} + \varepsilon_{ijkl} \\ \gamma_{ik} &\sim N(0, \sigma_u^2) \\ \varepsilon_{ijkl} &\sim N(0, \sigma_e^2), \end{aligned}$$

where i indicates age group, j the type of day (week or weekend), k the subject, and l the observation.

(a) 5 pts. Write out the skeleton ANOVA table indicating sources of variation and degrees of freedom for this study.

(b) 5 pts. Which sources of variation should be considered fixed and which should be considered random? Briefly explain your choices.

(c) 5 pts. Derive the Expected Mean Square for "Age group" (α_i in the model).
You may find the following helpful:

$$\begin{aligned} MS_{trt} &= \frac{nm}{t-1}\sum_{i=1}^{t}(\overline{y}_{i\ldots} - \overline{y}_{\ldots\cdot})^2 \\ E\sum_{i=1}^{t}(\overline{\gamma}_{i\ldots} - \overline{\gamma}_{\ldots\cdot})^2 &= \frac{t-1}{n}\sigma_u^2 \\ E\sum_{i=1}^{t}\sum_{k=1}^{n}(\overline{\gamma}_{i.k.} - \overline{\gamma}_{i\ldots})^2 &= (t-1)\sigma_u^2 \\ E\sum_{i=1}^{t}(\overline{\varepsilon}_{i\ldots} - \overline{\varepsilon}_{\ldots\cdot})^2 &= \frac{t-1}{mn}\sigma_e^2 \\ ,E\sum_{i=1}^{t}\sum_{k=1}^{n}(\overline{\varepsilon}_{i.k.} - \overline{\varepsilon}_{i\ldots})^2 &= \frac{t-1}{m}\sigma_e^2 \\ , \end{aligned}$$

where t is the number of groups, n is number of subjects per group, and m is the number of observations per subject. Hence, $m/2$ is the number of observations per subject and type of day.

(d) 5 pts. For subject-matter reasons, the linear combination of age group means given by:

$$-1.0\,\overline{y}_{1\ldots} - 0.3\,\overline{y}_{2\ldots} + 1.3\,\overline{y}_{3\ldots}$$

is of considerable interest to the researchers. What is the variance of this quantity? Since you don't have data, give your answer as an equation in terms of appropriate parameters.

For example, if I had asked about $\overline{y}_{11\cdot\cdot} - \overline{y}_{12\cdot\cdot}$, my answer would be $\frac{2\sigma_e^2}{nm/2}$ or $\frac{2\sigma_e^2}{150}$.

(e) 5 pts. The variance in part 3d can be estimated by what function of Mean Squares? Continuing the example, my answer would be $\frac{2\,MSE}{tnm/2}$ or $\frac{2\,MSE}{150}$

4. There is a second study of Vitamin A intake. Unlike the first study, this study makes no distinction between week and weekend days. The original plan was for each subject to record their diet on 7 days randomly chosen by the investigators. Unfortunately, some subjects forgot. In this study, there are 15 subjects per age group. The number of observations per subject is tabulated here:

# days:	1	2	3	4	5	6	7
# subjects:	5	1	2	8	10	8	11

These data were analyzed with the model

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + u_{ij} + \varepsilon_{ijk} \\ u_{ij} &\sim N(0, \sigma_u^2) \\ e_{ijk} &\sim N(0, \sigma_e^2), \end{aligned}$$

where i indicates the age group, j indicates the subject within the age group, and k indicates the observation within subject and age group.

The ANOVA table for this model, with type I (sequential) SS and (most of the) expected mean squares, is:

Source	df	SS	E MS
Age group	2	1,916.9	$\sigma_e^2 + 5.7089\sigma_u^2 + Q(t)$
Subject(age)	42	57,298.0	$\sigma_e^2 + 5.3776\sigma_u^2$
Error	198	4,650.6	

(a) 2 pts. What is the Expected Mean Square for the Error line (currently omitted from the ANOVA table)?

Note: If you do not remember this, see me to get the answer so you can complete the rest of the parts.

(b) 5 pts. A colleague suggests discarding the data from the 5 subjects with only 1 day's food dairy. His reason is "those observations don't provide any information about variability between subjects or days". Do you agree with this claim? Briefly explain why or why not.

Note: Don't worry about subject matter issues such as "these folks are so forgetful that we can't trust the data that was recorded.".

(c) 5 pts. Estimate σ_u^2.

(d) 5 pts. There are about 80 observations for each of the 3 age groups. Test whether the circa 80 observations within an age group are independent. Report your test statistic and state its distribution under the null hypothesis.

(e) 5 pts. What is the value of the denominator Mean Square for an F test of no differences among age groups?

(f) 5 pts. What are the degrees of freedom associated with the MS in part 4e

And because I didn't want to make this too long, you get 15 points for free.

Make sure you write your name on the back of the last page of your answers

R code for problem 1

```
# R code for problem 1

IA.CO <- matrix(c(2,20,48,20),ncol=2, dimnames=list(c('IA','CO'),c('shot','not')))
IA.CO

chisq.test(IA.CO,correct=F)

IA.CO2 <- data.frame(count=as.vector(IA.CO),
  status=factor(c('shot','shot','not','not')),
  state=factor(c('IA','CO','IA','CO')))

IA.CO2

ia.co1 <- glm(count~status+state,data=IA.CO2,
  family=poisson)
summary(ia.co1)
anova(ia.co1, test='Chi')

miles <- c(10,20)
IA.CO3 <- data.frame(count=IA.CO[,1],
  state=factor(c('IA','CO')),
  miles=miles, logmiles=log(miles))

IA.CO3

ia.co2 <- glm(count~state,data=IA.CO3,family=poisson)
summary(ia.co2)

ia.co3 <- glm(count~state,data=IA.CO3,family=poisson,
  offset=miles)
summary(ia.co3)

ia.co4 <- glm(count~state,data=IA.CO3,family=poisson,
  offset=logmiles)
summary(ia.co4)

# ------------------------------------
# Data set 2

UT.NV <- read.csv('UTNV.csv',as.is=T)
UT.NV$state.f <- factor(UT.NV$state)
```

```
UT.NV$logmiles <- log(UT.NV$miles)

UT.NV1 <- glm(count~state.f,data=UT.NV, family=poisson)
summary(UT.NV1)

UT.NV2 <- glm(count~state.f,offset=miles,data=UT.NV, family=poisson)
summary(UT.NV2)

UT.NV3 <- glm(count~state.f,offset=logmiles,data=UT.NV, family=poisson)
summary(UT.NV3)
```

R output for problem 1

```
>
> IA.CO <- matrix(c(2,20,48,20),ncol=2, dimnames=list(c('IA','CO'),c('shot','not')))
> IA.CO
   shot not
IA    2  48
CO   20  20
>
> chisq.test(IA.CO,correct=F)

        Pearson's Chi-squared test

data:  IA.CO
X-squared = 25.4599, df = 1, p-value = 4.517e-07

>
> IA.CO2 <- data.frame(count=as.vector(IA.CO),
+   status=factor(c('shot','shot','not','not')),
+   state=factor(c('IA','CO','IA','CO')))
>
> IA.CO2
  count status state
1     2   shot    IA
2    20   shot    CO
3    48    not    IA
4    20    not    CO
>
> ia.co1 <- glm(count~status+state,data=IA.CO2,family=poisson)
> summary(ia.co1)
```

```
Call:
glm(formula = count ~ status + state, family = poisson, data = IA.CO2)

Deviance Residuals:
     1       2       3       4
-3.634   2.860   1.595  -1.983

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.4086     0.1691  20.157  < 2e-16 ***
statusshot   -1.1285     0.2453  -4.601 4.21e-06 ***
stateIA       0.2231     0.2121   1.052    0.293
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 53.634  on 3  degrees of freedom
Residual deviance: 27.861  on 1  degrees of freedom
AIC: 51.871

Number of Fisher Scoring iterations: 5

> anova(ia.co1, test='Chi')
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                       3     53.634
status  1  24.6597         2     28.974 6.84e-07 ***
state   1   1.1134         1     27.861   0.2913
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> miles <- c(10,20)
> IA.CO3 <- data.frame(count=IA.CO[,1],
+   state=factor(c('IA','CO')),
```

```
+   miles=miles, logmiles=log(miles))
>
> IA.CO3
   count state miles logmiles
IA     2    IA    10 2.302585
CO    20    CO    20 2.995732
>
> ia.co2 <- glm(count~state,data=IA.CO3,family=poisson)
> summary(ia.co2)

Call:
glm(formula = count ~ state, family = poisson, data = IA.CO3)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9957     0.2236  13.397   <2e-16 ***
stateIA      -2.3026     0.7416  -3.105   0.0019 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.7094e+01  on 1  degrees of freedom
Residual deviance: 1.7764e-15  on 0  degrees of freedom
AIC: 11.456

Number of Fisher Scoring iterations: 3

>
> ia.co3 <- glm(count~state,data=IA.CO3,family=poisson, offset=miles)
> summary(ia.co3)

Call:
glm(formula = count ~ state, family = poisson, data = IA.CO3,
    offset = miles)

Deviance Residuals:
[1]  0  0

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -17.0043     0.2236  -76.05   <2e-16 ***
stateIA       7.6974     0.7416   10.38   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  2.6598e+01  on 1  degrees of freedom
Residual deviance: -3.5527e-15  on 0  degrees of freedom
AIC: 11.456

Number of Fisher Scoring iterations: 3

>
> ia.co4 <- glm(count~state,data=IA.CO3,family=poisson, offset=logmiles)
> summary(ia.co4)

Call:
glm(formula = count ~ state, family = poisson, data = IA.CO3,
    offset = logmiles)

Deviance Residuals:
[1]  0  0

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.468e-16  2.236e-01    0.00     1.00
stateIA     -1.609e+00  7.416e-01   -2.17     0.03 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7.2091e+00  on 1  degrees of freedom
Residual deviance: 1.7764e-15  on 0  degrees of freedom
AIC: 11.456

Number of Fisher Scoring iterations: 3

>
> # ------------------------------------
> # Data set 2
```

```
>
> UT.NV <- read.csv('UTNV.csv',as.is=T)
> UT.NV$state.f <- factor(UT.NV$state)
> UT.NV$logmiles <- log(UT.NV$miles)
>
> UT.NV1 <- glm(count~state.f,data=UT.NV, family=poisson)
> summary(UT.NV1)

Call:
glm(formula = count ~ state.f, family = poisson, data = UT.NV)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2361  -1.2021  -0.6679   0.8718   3.5673

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6292     0.1400   11.63   <2e-16 ***
state.fUT    -0.7129     0.2441   -2.92   0.0035 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 59.522  on 19  degrees of freedom
Residual deviance: 50.445  on 18  degrees of freedom
AIC: 107.77

Number of Fisher Scoring iterations: 5

>
> UT.NV2 <- glm(count~state.f,offset=miles,data=UT.NV, family=poisson)
> summary(UT.NV2)

Call:
glm(formula = count ~ state.f, family = poisson, data = UT.NV,
    offset = miles)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8996  -0.7378   0.2660   1.8453   3.0003

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.1600     0.1400 -22.567   <2e-16 ***
state.fUT     0.5580     0.2441   2.286   0.0223 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 73.715  on 19  degrees of freedom
Residual deviance: 68.859  on 18  degrees of freedom
AIC: 126.18

Number of Fisher Scoring iterations: 6

>
> UT.NV3 <- glm(count~state.f,offset=logmiles,data=UT.NV, family=poisson)
> summary(UT.NV3)

Call:
glm(formula = count ~ state.f, family = poisson, data = UT.NV,
    offset = logmiles)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.42675  -1.33740   0.08447   0.86533   1.92067

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1476     0.1400   1.054    0.292
state.fUT    -0.0198     0.2442  -0.081    0.935

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 33.509  on 19  degrees of freedom
Residual deviance: 33.503  on 18  degrees of freedom
AIC: 90.828

Number of Fisher Scoring iterations: 5
```