STA 312f22 Practice Questions

1. A strange three-sided die has $Pr(1) = 1/6$, $Pr(2) = 2/6$, and $Pr(3) = 3/6$. If you roll this die four times, what is the probability of getting one 1, one 2 and two 3s? The answer is a number. **Circle your answer**.

$$\binom{4}{1\ 1\ 2}\frac{1}{6}\frac{2}{6}\left(\frac{3}{6}\right)^2 = \frac{4!}{1!\,1!\,2!}\cdot\frac{1}{6}\cdot\frac{2}{6}\cdot\frac{3}{6}\cdot\frac{3}{6}$$

$$= 12\cdot\frac{6\cdot 3}{6\cdot 6\cdot 6\cdot 6} = \frac{1}{6} \approx 0.167$$

2. For $i = 1, \ldots, n$, let $\mathbf{X}_i = (X_{i,1}, X_{i,2}, X_{i,3})$ be independent multinomial $M(1, (\pi_1, \pi_2, \pi_3))$ random vectors.

   (a) Consider $H_0 : \pi_2 = 2\pi_3$. Starting with the likelihood on the formula sheet, derive the maximum likelihood estimator under this restriction. Your answer is a vector of *three* quantities, each a function of n_1, n_2 and n_3 (and maybe n, depending on how you write it.) Show your work and **Circle your final answer**.

$$L_0 = (1-\hat{\pi}_2-\hat{\pi}_3)^{n_1}\hat{\pi}_2^{n_2}\hat{\pi}_3^{n_3} = (1-2\hat{\pi}_3-\hat{\pi}_3)^{n_1}(2\hat{\pi}_3)^{n_2}\hat{\pi}_3^{n_3}$$

$$= (1-3\hat{\pi}_3)^{n_1}2^{n_2}\hat{\pi}_3^{n_2}\hat{\pi}_3^{n_3} = (1-3\hat{\pi}_3)^{n_1}2^{n_2}\hat{\pi}_3^{(n_2+n_3)}$$

$$\frac{\partial \log L_0}{\partial \hat{\pi}_3} = \frac{\partial}{\partial \hat{\pi}_3}\left(n_1\log(1-3\hat{\pi}_3) + n_2\log 2 + (n_2+n_3)\log\hat{\pi}_3\right)$$

$$= \frac{-3n_1}{1-3\hat{\pi}_3} + 0 + \frac{n_2+n_3}{\hat{\pi}_3} \stackrel{\text{set}}{=} 0$$

$$\Rightarrow \frac{3n_1}{1-3\hat{\pi}_3} = \frac{n_2+n_3}{\hat{\pi}_3} \Rightarrow 3n_1\hat{\pi}_3 = n_2+n_3 - 3(n_2+n_3)\hat{\pi}_3$$

$$\Rightarrow 3n_1\hat{\pi}_3 + 3(n_2+n_3)\hat{\pi}_3 = n_2+n_3$$

$$\Rightarrow 3\hat{\pi}_3(n_1+n_2+n_3) = n_2+n_3 \Rightarrow \hat{\pi}_3 3n = n_2+n_3$$

$$\Rightarrow \hat{\pi}_3 = \frac{n_2+n_3}{3n} \qquad \hat{\pi}_2 = 2\hat{\pi}_3 = \frac{2(n_2+n_3)}{3n}$$

$$\hat{\pi}_1 = 1-\hat{\pi}_2-\hat{\pi}_3 = \frac{3n}{3n} - \frac{2(n_2+n_3)}{3n} - \frac{n_2+n_3}{3n}$$

$$= 1 - \frac{n_2+n_3}{n} = \frac{n_1}{n}$$

9 2 cont

So $\hat{\pi}_1 = \frac{n_1}{n}$, $\hat{\pi}_2 = \frac{2(n_2+n_3)}{3n}$, $\hat{\pi}_3 = \frac{n_2+n_3}{3n}$

(b) Suppose we sample 300 adults, give each of them three unlabeled cups of tea to taste, and ask them to indicate their preference. We find that 104 prefer tea Type A, 87 prefer B and 109 prefer C. Give the restricted maximum likelihood estimate. This is the numerical version of your answer to 2a. The answer is a vector of three numbers. **Circle your answer.** $n_1 = 104,\ n_2 = 87,\ n_3 = 109$

$$\hat{\pi}_1 = \frac{n_1}{n} = \frac{104}{300} \qquad \hat{\pi}_{12} = \frac{2(n_2+n_3)}{3n} = \frac{2\times 196}{900} \qquad \hat{\pi}_3 = \frac{n_2+n_3}{3n} = \frac{196}{900}$$

$$= (0.347, 0.436, 0.218)$$

3. For the following table, show that $P(Y = 1|X = 1) = P(Y = 1|X = 2)$ implies that the odds ratio $\theta = 1$.

	$Y = 1$	$Y = 2$
$X = 1$	π_{11}	π_{12}
$X = 2$	π_{21}	π_{22}

$$P(Y=1|X=1) = P(Y=1|X=2) \Rightarrow \frac{\pi_{11}}{\pi_{11}+\pi_{12}} = \frac{\pi_{21}}{\pi_{21}+\pi_{22}}$$

$$\Rightarrow \pi_{11}\pi_{21} + \pi_{11}\pi_{22} = \pi_{11}\pi_{21} + \pi_{12}\pi_{21}$$

$$\Rightarrow \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1$$

$$\| \theta$$

4. In a famous genetics experiment first described by Gregor Mendel, plants produce peas that are either smooth or wrinkled. For this experiment (a back-cross involving a single dominant/recessive gene pair), classical genetic theory tells us that the probability of observing a plant with smooth peas is 0.75. We breed and raise 50 plants according to the experimental protocol, and observe that 34 produce smooth peas.

   (a) What is a reasonable model for these data? $X_1, \ldots, X_n$ iid Bernoulli(π)

   (b) What is the null hypothesis, in symbols? $H_0: \pi = 0.75$

   (c) Would rejection of H_0 be evidence for the theory, or against it? Against

   (d) You will test the null hypothesis with a large-sample likelihood ratio test. What is the approximate distribution of the test statistic when H_0 is true? You don't have to show anything. Just write down the answer. Chi-squared ($df = 1$)

   (e) What is the critical value of the test statistic at $\alpha = 0.05$? The answer is a number. 3.841

   (f) Calculate the test statistic. Show some work. Your answer is a number. **Circle your answer.** $n_1 = 34$, $n_2 = 16$ $\hat{\mu}_1 = 50(0.75) = 37.5$ $\hat{\mu}_2 = 50(0.25) = 12.5$

$$G^2 = 2\sum_{j=1}^{c} n_j \log \frac{n_j}{\hat{\mu}_j} = 2\left(34 \log \frac{34}{37.5} + 16 \log \frac{16}{12.5}\right)$$

$$= 2(-3.33 + 3.95) = 1.24$$

   (g) Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No. NO

   (h) Do the results of this experiment provide evidence against the theory? Answer Yes or No. NO

   (i) Do the results of this experiment prove that the theory is correct? Answer Yes or No. NO

5. Information on a sample of drivers includes Age, Sex and how many traffic tickets (moving violations) they had the past 12 months. The number of tickets is modeled as a Poisson random variable with conditional mean λ, and

$$\log \lambda = \beta_0 + \beta_1 x + \beta_2 s,$$

where x is age, and s is a dummy variable that equals one for females and zero for males.

(a) Make a table with one row for females and one row for males. Make one column for the dummy variable s, and a second, wider column for the expected number of tickets.

	s	λ
F	1	$e^{\beta_0} e^{\beta_1 x} e^{\beta_2}$
M	0	$e^{\beta_0} e^{\beta_1 x}$

(b) In terms of β quantities, what is the expected number of traffic tickets for a 20 year old woman?

$$e^{\beta_0 + \beta_2 + 20\beta_1}$$

(c) Suppose that for any age, the expected number of traffic tickets is twice as great for men as for women. What is β_2? Show some work. **Circle your answer.**

$$e^{\beta_0 + \beta_1 x} = 2 e^{\beta_0 + \beta_1 x} e^{\beta_2}$$

$$\Rightarrow e^{\beta_2} = \frac{1}{2} \Rightarrow \beta_2 = \log \frac{1}{2}$$

or $\beta_2 = -\log 2$

6. This question is based on the following printout from the Birth Weight study. Remember, `lwt` refers to mother's weight.

```
> library(MASS)
> head(birthwt)
   low age lwt race smoke ptl ht ui ftv  bwt
85   0  19 182    2     0   0  0  1   0 2523
86   0  33 155    3     0   0  0  0   3 2551
87   0  20 105    1     1   0  0  0   1 2557
> attach(birthwt)
> race=factor(race,labels=c("White","Black","Other"))
> contrasts(race)
      Black Other
White     0     0
Black     1     0
Other     0     1
> mod1 = glm(low ~  lwt + smoke + race, family=binomial)
> summary(mod1)

Call:
glm(formula = low ~ lwt + smoke + race, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5278  -0.9053  -0.5863   1.2878   2.0364

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.10922    0.88211  -0.124  0.90146
lwt         -0.01326    0.00631  -2.101  0.03562 *
smoke        1.06001    0.37832   2.802  0.00508 **
raceBlack    1.29009    0.51087   2.525  0.01156 *
raceOther    0.97052    0.41224   2.354  0.01856 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 215.01  on 184  degrees of freedom
AIC: 225.01

Number of Fisher Scoring iterations: 4

> anova(mod1,test='LRT')
Analysis of Deviance Table

Model: binomial, link: logit

Terms added sequentially (first to last)
```

```
      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    188     234.67
lwt    1   5.9813       187     228.69 0.014458 *
smoke  1   4.3500       186     224.34 0.037009 *
race   2   9.3260       184     215.01 0.009438 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) For any race and any mother's weight, the odds of a low birth weight baby are estimated to be _____ as great for a mother who smokes during pregnancy. The answer is a number.

$e^{1.06} = 2.886$

(b) Give a 95% confidence interval for that last odds ratio. Your answer is a pair of numbers, a lower confidence limit and an upper confidence limit. Show your work. **Circle your answer.**

95% CI for $\beta_2 = (1.06 - 1.96(0.378),\ 1.06 + 1.96(0.378))$

$= (0.319, 1.801)$ So 95% CI for e^{β_2} is

$(e^{0.319}, e^{1.801}) = (1.376, 6.056)$

(c) Estimate the probability of a low birth weight baby for a 130 pound, White, non-smoking mother. The answer is a number. **Circle your answer.**

$x'\hat{\beta} = -0.109 + 130(-0.013) = -1.799$

$\hat{\pi} = \dfrac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}} = \dfrac{0.165}{1.165} \approx 0.14$

(d) Controlling for mother's weight and smoking status, the estimated odds of a low birth weight baby are _____ as great for a Black mother, compared to an Other mother.

$e^{1.29}/e^{0.971} = e^{0.319} = 1.376$ (1.4 is OKAY)

(e) Controlling for mother's weight and smoking status, do White and Black mothers differ in their odds of having a low birth weight baby?

i. Give the test statistic (Z or χ^2). The answer is a number.

$Z = 2.525$

ii. What is the critical value of the test statistic at $\alpha = 0.05$? The answer is a number.

1.96

iii. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

Yes

iv. In plain, non-statistical language, what do you conclude?

Controlling (correcting) for mother's weight and smoking status, Black mothers are more likely to have a low birth weight baby than white mothers.

7. Using the Math data, we investigate choice of university Calculus course as a function of High School Calculus mark and sex. The questions come after the printout.

```
> # Choice of university course based on High School data, sex and first language
> # install.packages("mlogit", dependencies=TRUE) # Only need to do this once
> library(mlogit) # Load the package every time
> datta = math[,c(1,5,9)] # Just course, hscalc and sex
> datta = na.omit(datta)
> summary(datta); attach(datta)
      course        hscalc       sex
 Catch-up: 20   Min.   : 50.00   F:193
 Elite   : 28   1st Qu.: 67.00   M:186
 Mainstrm:331   Median : 77.00
                Mean   : 76.09
                3rd Qu.: 86.00
                Max.   :100.00

> # Make Mainstream the reference category for course by changing alphabetical order.
> n = length(course); Course = character(n)
> Course[course=='Mainstrm'] = '1_Mainstrm'
> Course[course=='Elite'] = '2_Elite'
> Course[course=='Catch-up'] = '3_Catch-up'
> Course = factor(Course); table(Course)
Course
1_Mainstrm    2_Elite 3_Catch-up
       331         28         20
> datta$course = Course # Put the fixed-up version back in the data frame
>
> # Make an mlogit data frame in long format
> long = mlogit.data(datta,shape="wide",choice="course")
>
> # Fit full model
> full = mlogit(course ~ 0 | hscalc + sex, data=long)
> summary(full)

Call:
mlogit(formula = course ~ 0 | hscalc + sex, data = long, method = "nr",
    print.level = 0)

Frequencies of alternatives:
1_Mainstrm    2_Elite 3_Catch-up
  0.873351   0.073879   0.052770

nr method
6 iterations, 0h:0m:0s
g'(-H)^-1g = 1.94E-07
gradient close to zero

Coefficients :
                         Estimate Std. Error t-value  Pr(>|t|)
2_Elite:(intercept)    -6.277453   1.573941 -3.9884 6.653e-05 ***
3_Catch-up:(intercept)  4.213750   1.472793  2.8611  0.004222 **
2_Elite:hscalc          0.036789   0.018833  1.9535  0.050762 .
3_Catch-up:hscalc      -0.107888   0.023086 -4.6732 2.965e-06 ***
2_Elite:sexM            1.411205   0.475800  2.9660  0.003017 **
3_Catch-up:sexM         0.733976   0.497212  1.4762  0.139895
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -153.11
McFadden R^2:  0.13307
Likelihood ratio test : chisq = 47.003 (p.value = 1.5226e-09)
>
> # Restricted models
> NoCalculus = mlogit(course ~ 0 | sex, data=long)
> summary(NoCalculus)

Call:
mlogit(formula = course ~ 0 | sex, data = long, method = "nr",
    print.level = 0)

Frequencies of alternatives:
1_Mainstrm    2_Elite 3_Catch-up
  0.873351   0.073879   0.052770

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 3.08E-08
gradient close to zero

Coefficients :
                          Estimate Std. Error t-value  Pr(>|t|)
2_Elite:(intercept)     -3.39563    0.41503 -8.1816  2.22e-16 ***
3_Catch-up:(intercept) -3.10794    0.36137 -8.6005 < 2.2e-16 ***
2_Elite:sexM             1.46279    0.47359  3.0887   0.00201 **
3_Catch-up:sexM          0.56897    0.46957  1.2117   0.22564
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -170.31
McFadden R^2:  0.035674
Likelihood ratio test : chisq = 12.601 (p.value = 0.0018356)

> anova(NoCalculus, full)
Error in UseMethod("anova") :
  no applicable method for 'anova' applied to an object of class "mlogit"

> NoSex      = mlogit(course ~ 0 | hscalc, data=long)
> summary(NoSex)

Call:
mlogit(formula = course ~ 0 | hscalc, data = long, method = "nr",
    print.level = 0)

Frequencies of alternatives:
1_Mainstrm    2_Elite 3_Catch-up
  0.873351   0.073879   0.052770

nr method
6 iterations, 0h:0m:0s
g'(-H)^-1g = 1.56E-07
gradient close to zero
```

```
Coefficients :
                          Estimate Std. Error t-value  Pr(>|t|)
2_Elite:(intercept)     -5.641500   1.514151 -3.7258 0.0001947 ***
3_Catch-up:(intercept)  4.472142   1.453862  3.0760 0.0020977 **
2_Elite:hscalc           0.040052   0.018460  2.1696 0.0300373 *
3_Catch-up:hscalc       -0.106000   0.022873 -4.6343 3.581e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -159.34
McFadden R^2:  0.097752
Likelihood ratio test : chisq = 34.528 (p.value = 3.1797e-08)
```

(a) The tests summary output for the full model include two tests (excluding tests for the intercepts) that are statistically significant. In plain, non-statistical language and mentioning *no numbers*, give the conclusions from these two tests. You have more room than you need.

Controlling for sex, students with higher HS calculus marks are less likely to choose the catchup course over the mainstream course.

Controlling for HS calculus mark, male students are more likely to choose the elite course over the mainstream course

(b) We seek a *single* test of the relationship between sex and choice of university Calculus course, controlling for mark in High School Calculus.

i. Write the numerical value of the test statistic in the space below. The answer is a number. If you need to calculate this number from material on the printout, show a little work. **Circle the number.**

There are 2 ways

(1) Subtract the χ^2 tests of model vs null model

$47.003 - 34.528 = (12.475)$

(2) $-2(\log \text{lik}_{\text{restricted}} - \log \text{lik}_{\text{full}})$

$= -2(-159.34 - -153.11)$

$= -2(-6.23) = (12.46)$ rounding error

ii. What is the critical value? The answer is a number from the formula sheet.

χ^2 with $df = 2$ is 5.991

iii. Do you reject the null hypothesis? Answer Yes or No.

Yes

iv. Is there evidence that sex is related to choice of Calculus course, controlling for High School performance? Just answer Yes or No.

Yes