$\frac{7}{10}$

Identification Code 24

1. The objective of this problem is to demonstrate that *the moments may not always determine a probability distribution.* Consider a random variable X having the standard lognormal distribution with density given by

$$f(x) = \begin{cases} \frac{1}{x\sqrt{2\pi}} \exp\{-\frac{1}{2}(\log x)^2\} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) ✓ *(3 points)* Show that the k-th raw moment of X, corresponding to the above density function $f(x)$ is given by

$$E(X^k) = \exp\{\frac{1}{2}k^2\}, \quad \text{for } k = 1, 2 \ldots$$

(b) *(3 points)* Consider the function given by

$$g(y) = \begin{cases} f(y)(1 + \sin(2\pi \log(y)) & \text{if } y > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $f(y)$ is as defined above. Show that $g(y)$ is a probability density function.

(c) *(4 points)* Let Y be a random variable with density function $g(y)$ as defined in part (b). Show that

$$E(Y^k) = E(X^k), \quad \text{for } k = 1, 2 \ldots$$

Problem 1

(a). $M_X(t) = E(e^{tX})$

$= \int_0^\infty e^{tx} \frac{1}{x\sqrt{2\pi}} \exp\{-\frac{1}{2}(\log x)^2\} dx$

$= \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{1}{x} \exp\{tx - \frac{1}{2}(\log(x))^2\} dx$

~~$E(X^k) = \int_0^\infty x^k \frac{1}{x\sqrt{2\pi}} \exp(-\frac{1}{2}(\log x)^2) dx$~~

~~$= \int_0^\infty \frac{1}{\sqrt{2\pi}} x^{k-1} \exp(-\frac{1}{2}(\log x)^2) dx$~~

~~$= \int_0^\infty \frac{1}{\sqrt{2\pi}} x^{k+1} \exp(-\frac{1}{2}(\log x)^2 +$~~

$M_X(t)$ is not finite

$E(X^k) = \frac{\partial^k}{\partial t^k} M_X(t) |_{t=0}$

$= \frac{\partial^{k-1}}{\partial t^{k-1}} \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{x}{x} \exp\{tx - \frac{1}{2}(\log(x))^2\} dx |_{t=0}$

$= \frac{1}{\sqrt{2\pi}} \int_0^\infty x^{k-1} \exp\{tx - \frac{1}{2}(\log(x))^2\} dx |_{t=0}$

$\boxed{=} \exp(\frac{1}{2}x^2)$

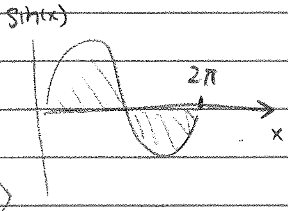
(b). $\int_0^\infty g(y) dy$

$= \int_0^\infty \frac{1}{y\sqrt{2\pi}} \exp\{-\frac{1}{2}(\log y)^2\} (1 + \sin(2\pi \log(y))) dy$

$= \int_0^\infty \frac{1}{y\sqrt{2\pi}} \exp\{-\frac{1}{2}(\log y)^2\} dy + \int_0^\infty \frac{1}{y\sqrt{2\pi}} \exp\{-\frac{1}{2}\log(y)^2\} \sin(2\pi \log(y)) dy$

$= 1 \; (+0)$

$= 1$



Thus $g(y)$ is a probability function, as it sums to 1 and is 0 everywhere else outside its support.

(this since $f(y) = 0$ when $y < 0$).

(Thus $g(y) = 0$ when $y < 0$)

This is because recall sin(·) is symmetric to 0, thus the sum (integral) due to the symetry, equals to 0.

Thus. $= E(\sin(2\pi \log(y)))$ where $y \sim$ lognormal

$= 0$

why $g(y) \geq 0$?

-1

$= \frac{1}{\sqrt{2\pi}} \int_0^\infty x^{k-1} \exp\{-\frac{1}{2}(\log(x))^2\} dx$ at $t=0$

$= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{uk} \exp\{-\frac{1}{2}u^2\} du$ let $u = \log(x)$ $du = dx\frac{1}{x}$

$= E(e^{uk})$ where $u \sim N(0,1)$

$= e^{\frac{1}{2}k^2} \longleftrightarrow$ MGF of $N(0,1)$

Problem 1 continuous

(c). Showing $E(Y^k) = E(X^k)$

it equivalent to show $E(Y^k) = e^{(\frac{1}{2}k^2)}$

$$M_Y(t) = E(e^{tY})$$
$$= \int_0^\infty e^{ty} f(y)(1+\sin(2\pi\log(y)))\,dy$$
$$= \int_0^\infty e^{ty} f(y)\,dy + \int_0^\infty e^{ty}\underset{f(y)}{\wedge}\sin(2\pi\log(y))\,dy$$

$$E(Y^k) = \frac{\partial^k}{\partial t^k} M_Y(t)\,\Big|_{t=0}$$
$$= e^{\frac{1}{2}k^2} + \int_0^\infty y^k \overset{f(y)}{\perp}\sin(2\pi\log(y))\,dy$$
$$= e^{\frac{1}{2}k^2} + \int_{-\infty}^\infty e^{ku}\underset{f(u)}{\wedge}\sin(2\pi u)\,du$$
$$= e^{\frac{1}{2}k^2} + 0$$
$$= e^{\frac{1}{2}k^2}$$

let $u = \log y$

$du = \frac{1}{y}dy$

This may again from the symmetric property of $\sin(\cdot)$, or by using the geometric expansion of $\sin(\cdot)$ such that shows the integral is 0. However, due to the limit time and I couldn't recall the formular for $\sin(\cdot)$ expansion, I cannot perform this.

$\sin 2\pi u = \sin 2\pi(u-k)$

$\forall k \in \mathbb{N}$.

-2

This is from

$$\frac{\partial^k}{\partial t^k}\int_0^\infty e^{ty}\sin(2\pi\log(y))f(y)\,dy\Big|_{t=0}$$
$$= \int_0^\infty e^{ty}\, y^k \sin(2\pi\log(y))f(y)\,dy\,\Big|_{t=0}$$
$$= \int_0^\infty y^k\sin(2\pi\log(y))f(y)\,dy$$
$$= E(y^k\sin(2\pi\log(y)))$$

MGF is not finite in this case!

Or this question can be done by first transform y to log(y), then do the computation, then the answer maybe easier to get.

9

2. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed (iid) random variables with common probability density function (pdf)

$$f_X(x|\theta) = \begin{cases} \frac{2x}{\theta} \exp\left\{-\frac{x^2}{\theta}\right\} & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter.

(a) *(2 points)* Show that $T = \sum_{i=1}^{n} X_i^2$ is a complete sufficient statistic for θ.

(b) *(4 points)* Determine the uniform minimum variance unbiased estimator (UMVUE) of θ^2.

(c) *(4 points)* Determine the uniformly most powerful (UMP) rejection region for testing $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, where θ_0 is a known positive number. State your rejection region in terms of T and a quantile from a named probability distribution.

Problem 2

(a). $f(\underline{x}) = \prod_{i=1}^{n} \frac{2x_i}{\theta} \exp\{-\frac{x_i^2}{\theta}\} I(x_i > 0)$

$= \exp\{\sum_{i=1}^{n} \log(2x_i) - n\log(\theta) - \sum_{i=1}^{n} x_i^2/\theta\}$

$\in$ exponential family with $\eta(\theta) = \frac{-1}{\theta}$, $T(x) = \sum_{i=1}^{n} x_i^2$

Since $\theta > 0$, $\frac{-1}{\theta}$ spans $\mathbb{R}$, $\exists$ open set, thus $T(x)$ is sufficient & complete for $\frac{-1}{\theta}$

Since $\frac{-1}{\theta}$ and θ is one-to-one, then $T(x)$ is sufficient & complete for θ.

* can also use factorization theorem show $\sum x_i^2$ is minimal sufficient.

(b). $E(X^2) = \int_0^\infty \frac{2x^3}{\theta} \exp\{-\frac{x^2}{\theta}\} dx$ let $u = x^2$, $du = 2x\,dx$

$= \int_0^\infty \frac{u}{\theta} \exp\{-\frac{u}{\theta}\} du$ → Recall $x \sim \text{Gamma}(2, \frac{1}{\theta})$ scale

$= \theta.$

$f_X(x) = \frac{1}{\Gamma(2)\theta^2} x^1 \exp(-\frac{x}{\theta})$

first recognize, $Y = X^2 \sim \exp(\theta)$

Then $Var(Y) = \theta^2$

Thus, $\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_.)^2$ is an unbiased estimator of θ^2.

$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - \overline{X^2})^2$

Then by Rao-Blackwell & Lehmann Scheffe theorem.

UMVUE = E[unbiased estimator | sufficient, complete statistics]

$= E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - \overline{X^2})^2 \mid \sum_{i=1}^{n} X_i^2]$

$= E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i^4 - 2X_i^2\overline{X^2} + (\overline{X^2})^2 \mid \sum_{i=1}^{n} X_i^2]$

please simplify

Let $Y = X^2$; $X = \sqrt{Y}$ (one-to-one since $X > 0$) $J = \frac{\partial}{\partial y}\sqrt{y} = \frac{1}{2\sqrt{y}}$

$f_Y(y) = f_X(y) \cdot J$

$= \frac{2\sqrt{y}}{\theta} \cdot \exp\{-\frac{y}{\theta}\} \cdot \frac{1}{2\sqrt{y}}$

$= \frac{1}{\theta} \exp\{-\frac{y}{\theta}\}$

$\sim \exp(\theta)$ → scale, mean parameter

(c). Recall from part (b), we recongnize $Y = X^2 \sim \exp(\theta)$, then, a natural choice of statistics is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Likelihood-ratio: $\frac{L(\hat{\theta})}{L(\hat{\theta}_0)} = \frac{(\frac{1}{\hat{\theta}})^n \exp\{-\frac{\sum y_i}{\hat{\theta}}\}}{(\frac{1}{\hat{\theta}_0})^n \exp\{-\frac{\sum y_i}{\hat{\theta}_0}\}}$ where $\hat{\theta}_0$ is the MLE estimator of θ under H_0, $\hat{\theta}$ is the MLE estimator of θ under H_1

why

is monotonic in $\sum y_i = \sum x_i^2 = T(x)$, thus by Karlin Rubin's theorem, we reject when Ratio is large, that is when $\frac{1}{n}\sum x_i^2 = \hat{\theta}$ ~~< is~~ small (due to the H_0 direction).

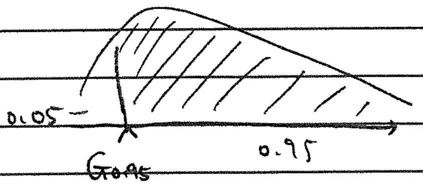
Since $\frac{1}{n}\sum x_i^2 \sim \text{Gamma}(n, \frac{\theta}{n})$, reject when

$\frac{1}{n}\sum x_i^2 < G_{(n, \frac{\theta}{n}), 0.95}$

↑

5th quantile for Gamma.

αth



This is a UMP test.

3. The Pima Indians Diabetes data contains information of 768 women from a population near Phoenix, Arizona, USA. The data contain medical records from Pima-Indian women at least 21 years of age. Below is reproduced the list of 5 attribute variables and the class variable (response):

   - 'glucose': plasma glucose concentration
   - 'pregnancy': number of times pregnant
   - 'bmi': body mass index (weight in kg/squared height in m^2)
   - 'pedigree': diabetes pedigree function
   - 'age': age in years
   - 'diabetes': class variable (0 or 1)

   The 'diabetes' variable is an indicator, with (1) indicating a positive test for diabetes between 1 and 5 years from the examination and (0) meaning a negative test.

   We first fit a logistic regression model with 'diabetes' as the response variable and all the above attribute variables as predictors. The R code is produced below

```
model =  glm(diabetes~glucose+pregnancy+bmi+pedigree+age,
             data = PimaIndiansDiabetes2, family = binomial(link = "logit"))
summary(model)
```

   and we have the following output:

```
Call:
glm(formula = diabetes ~ glucose + pregnancy + bmi + pedigree +
    age, family = binomial(link = "logit"), data = PimaIndiansDiabetes2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.673124   0.689980 -12.570  < 2e-16 ***
glucose      0.032893   0.003403   9.667  < 2e-16 ***
pregnancy    0.119458   0.031653   3.774 0.000161 ***
bmi          0.079550   0.013810   5.760  8.4e-09 ***
pedigree     0.891487   0.292239   3.051 0.002284 **
```

```
age            0.012230   0.009085   1.346 0.178247
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 732.51  on 762  degrees of freedom
AIC: 744.51

Number of Fisher Scoring iterations: 5
```

Then the code

```
anova(model)
```

gives the output here:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: diabetes

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        767     993.48
glucose    1  184.764       766     808.72 < 2.2e-16 ***
pregnancy  1   23.770       765     784.95 1.086e-06 ***
bmi        1   40.825       764     744.12 1.665e-10 ***
pedigree   1    9.819       763     734.31  0.001727 **
age        1    1.797       762     732.51  0.180067
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) *(3 points)* Write down the fitted model formula.

(b) *(4 points)* Conduct a hypothesis test at the 5% significance level that the coefficients for the two predictors ‘pedigree’ and ‘age’ are both zero. Give the numerical value of the test statistic and specify the rejection region.

(c) *(3 points)* We refit the model with the predictor ‘age’ being dropped and the summary output is below:

```
Call:
glm(formula = diabetes ~ glucose + pregnancy + bmi + pedigree,
    family = binomial(link = "logit"), data = PimaIndiansDiabetes2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.415851   0.656908 -12.811  < 2e-16 ***
glucose      0.033826   0.003345  10.112  < 2e-16 ***
pregnancy    0.141926   0.027105   5.236 1.64e-07 ***
bmi          0.078097   0.013771   5.671 1.42e-08 ***
pedigree     0.901294   0.291696   3.090    0.002 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 734.31  on 763  degrees of freedom
AIC: 744.31

Number of Fisher Scoring iterations: 5
```

Based on the fitted model, suppose that a woman in the study has a probability of 0.4 of having a positive diabetes test between 1 and 5 years from the examination. If her body mass index, i.e., the predictor ‘bmi’, is actually one unit less while the values of the other predictors remain the same, give the probability of this woman having a positive diabetes test.

2 24

Problem3 -0.5 for not giving values of betas

(b) First write out the model:

$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{glucose} + \beta_2 \text{bmi} + \beta_3 \text{pedigree} + \beta_4 \text{age} + \beta_5 \text{pegnancy} + \varepsilon$ not correct -0.5

β_0: intercept

$H_0: \beta_3 = \beta_4 = 0 \qquad H_1: \text{not } H_0$

Recall the LRT test under GLM:

$$T = \frac{D_{M_0} - D_M}{\varphi} \overset{H_0}{\sim} \chi^2_{p-q}$$ when sample size n=768 large

$D_M = 732.51 \quad \varphi = 1 \quad q = 762$

$D_{M_0} = 732.51 + 9.819 + 1.797 \quad p = 762+2$

$$T = \frac{9.819 + 1.797}{1}$$

$$= 11.616 \overset{H_0}{\sim} \chi^2_2$$

$$> \chi^2_{2,0.05} = 5.991$$

4

Reject H_0 when $T > \chi^2_{2,0.05} = 5.991$

In our case, we reject it, at least one of β_3, β_4 matters.

(a). Let $P_i = \text{Prob}(\text{Diabetiz} \mid \text{other variables})$:

$\log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 \text{glucose}_i + \beta_2 \text{Preg}_i + \beta_3 \text{bmi}_i + \beta_4 \text{pedi}_i + \beta_5 \text{age}_i + \varepsilon_i$

$i = 1, \ldots, 768$

(c). Origion $p = 0.4$

3

$\log(\frac{p}{1-p}) = \log(\frac{0.4}{0.6}) = -0.405$

With 1 unit decrease in bmi:

$\Rightarrow \log(\frac{p}{1-p})_{new} = -0.405 - 0.078097$

$= -0.48356$

$\Rightarrow$

$$P_{new} = \frac{1}{1+e^{0.48356}}$$

$= 0.381$

thus the true prob is 0.381.

6|10

4. Consider a study in which we count the number of defects in 1-square-inch specimens of fabrics. We have two different types of fabrics, A and B. Define the response variable y as the count of defects and a binary variable x that takes the values 0 or 1 when type is A or B, respectively. The data is shown below.

Specimen ID (i)	Type (x_i)	Count (y_i)
1	0	1
2	0	0
3	0	0
4	0	1
5	1	2
6	1	1
7	1	2
8	1	3

Consider the following two models:

Model I: $y \sim$ *Poisson*, two types have two different means, λ_A and λ_B

Model II: $y \sim$ *Poisson*, the mean is modeled as $\lambda = \beta_0 + \beta_1 x$.

Answer the following questions.

(a) *(2 points)* Write the mathematical forms of the two models described above.

(b) *(3 points)* Interpret the parameters in each model and find the relationship between (λ_A, λ_B) and (β_0, β_1).

(c) *(3 points)* Write the log-likelihood for model I. Using the data, find the Maximum Likelihood Estimator (MLE) of (λ_A, λ_B). Give numerical value of the MLE.

(d) *(2 points)* Using the data, find the MLE of (β_0, β_1) and give its numerical value.

Problem 4

(a) $\lambda = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, 8.$ *

$\uparrow$ model 2

$Y \sim \text{Poisson}(\lambda)$

$f_Y = \frac{\lambda^y e^{-\lambda}}{y!}$

model 1: mixed model depends on $I(x)$.

$\lambda = \lambda_A (1-X_i) + \lambda_B X_i$

$= \lambda_A (1-X_i) + \lambda_B (X_i) + \varepsilon_i$

$= \lambda_A + X_i(\lambda_B - \lambda_A) + \varepsilon_i$

-1

· * This should be something Identical to the poisson model in ST704, which it in GLM with link $\ln(\cdot)$. But in that case, $\ln(\lambda)$ is linear, since we aleady have λ linear, may be do not need to log again ?

(b). In model 2, β_0 is the expected mean when $X_i = 0$ (Type A)

β_1 is the expected change in mean when X_i change from type A to type B.

please see next page for part (a),(b) discussion.

In model 1, I'm fitting a linear model here using λ_A and λ_B, and by comparing with model 2, we see $\beta_0 \equiv \lambda_A$, $\beta_1 \equiv (\lambda_B - \lambda_A)$

-1

* However again, I'm not sure about the model, thus interpetation & relationships varies.

λ_A & λ_B ?

c), $L(y) = \prod_{i=1}^{4} \frac{\lambda_A^{y_i} e^{-\lambda_A}}{y_i!} \prod_{i=5}^{8} \frac{\lambda_B^{y_i} e^{-\lambda_B}}{y_i!}$

$\ell(y) = c + \sum_{i=1}^{4} y_i \log(\lambda_A) - 4\lambda_A + \sum_{i=5}^{8} y_i \log(\lambda_B) - 4\lambda_B$

c - constant invariant of λ_A, λ_B.

$\Rightarrow \frac{\partial \ell}{\partial \lambda_A} = \frac{\sum y_i}{\lambda_A} - 4 \overset{set}{=} 0$

$\Rightarrow \hat{\lambda}_A = \frac{\sum y_i}{4} = \bar{y} \quad i = 1,2,3,4$

$\frac{\partial^2 \ell}{\partial \lambda_A^2} = -\frac{\sum y_i}{\lambda_A^2} < 0$

$\frac{\partial^2 \ell}{\partial \lambda_B^2} = -\frac{\sum y_i}{\lambda_B^2} < 0$

concave.

$\Rightarrow \frac{\partial \ell}{\partial \lambda_B} = \frac{\sum y_i}{\lambda_B} - 4 = 0$

Thus maximizers.

$\hat{\lambda}_B = \bar{y} \quad i = 5,6,7,8$

$(\hat{\lambda}_A, \hat{\lambda}_B) = \left(\frac{2}{4}, \frac{2+1+2+3}{4}\right)$

$= \left(\frac{1}{2}, 2\right)$

(d). by definition $\lambda_i = \beta_0 + \beta_1 X$.

$\ell(y) = \sum_{i=1}^{8} y_i \log(\beta_0 + \beta_1 X_i) - 8(\beta_0 + \beta_1 X_i) + c$.

-2

$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{8} \frac{y_i}{\beta_0 + \beta_1 X_i} - 8 \overset{set}{=} 0$

$\beta_0 + \beta_1 \bar{X} = \bar{y}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X} = \bar{y} = \frac{5}{4}$

$\bar{X} = 0.5 \qquad \bar{y} = \frac{5}{4}$

$\frac{\partial \ell}{\partial \beta_1} = \sum^{8} \frac{y_i X_i}{\beta_0 + \beta_1 X_i} - \sum X_i \overset{set}{=} 0$

$\sum (X_i - \bar{X})^2 = 0.5^2 \cdot 8 = 2$

$\sum \beta_0 + \beta_1 X_i = \frac{\sum y_i X_i}{\sum X_i}$

$\sum (X_i - \bar{X})(y_i - \bar{y}) = \left(0.5 \cdot \left(1 - \frac{5}{4}\right) + 0.5\left(-\frac{5}{4}\right)\right) 2$

result as SLR $\longrightarrow$ $\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(y_i - \bar{y})}{\sum (X_i - \bar{X})^2}$

due to the time limit, I cannot recalculate these numbers, but the expessions should be sound.

$+ \left(0.5\left(2 - \frac{5}{4}\right)\right) 2 + 0.5\left(1 - \frac{5}{4}\right) + 0.5\left(3 - \frac{5}{4}\right)$

$= \left(-\frac{1}{8} - \frac{5}{8}\right) 2 + \left(\frac{3}{8} \cdot 2\right) - \frac{1}{8} + \frac{7}{8}$

$= \frac{0}{2}$

$= 0$

$= \frac{-6}{4} + \frac{3}{4} + \left(-\frac{1}{8}\right) + \frac{7}{8}$

$= \frac{-6}{4} + \frac{3}{4} + \frac{3}{4}$

$= 0$?

Problem4 continuous

(a)-(b)*. Given poisson distribution

$Y_i \sim \text{Poisson}(\lambda) \Rightarrow f_Y(y) = \frac{\lambda^y e^{-\lambda}}{y!} = P(Y=y)$. $\ln(P(Y=y)) = y\log(\lambda) - \lambda - \log(y!)$.

And under GLM, we normally write $g(E(y|X)) = \beta_0 + \beta_1 X \boxed{+\varepsilon}$ where $g(\cdot)$ link function, and the natural link function for poisson is $\log(\cdot)$ link, thus.

we normally have $\ln(\lambda) = \beta_0 + \beta_1 X \boxed{+\varepsilon}$

in this sense $\ln(\lambda_i) = \ln(\lambda_A)(1-X_i) + \ln(\lambda_B) X_i + \varepsilon_i$

$= \ln(\lambda_A) + (\ln(\frac{\lambda_B}{\lambda_A}) X_i + \varepsilon_i$ model 1

In this case, $\ln(\lambda_A)$ is the expected log-mean when $X=0 \Rightarrow$ Type A.

$\ln(\frac{\lambda_B}{\lambda_A})$ is the expected log-mean change when $X \uparrow 1 \Rightarrow$ Type A to Type B

Similarly to model 2, but since already have $\lambda = \beta_0 + \beta_1 X$,

we may use identity link, that is

$\lambda = \beta_0 + \beta_1 X + \varepsilon$

Interpretation then consists with part (b).

7/10

5. An education researcher receives a grant to test the effectiveness of a new textbook (treat=1) compared to the standard textbook (treat=0). She randomly chooses J schools in a large district to participate. For each school, exactly one class is chosen at random to participate in the trial, so we use a single index j to refer to the school/class. All students within a class use the same textbook. Finally, the treatment is assigned at random.

   Let Y_{ij} denote a testing outcome (difference in test scores, after treatment minus before treatment) for student i in classroom j, and $\text{treat}_j \in \{0, 1\}$ denotes treatment that is applied at the classroom level. The researcher assumes the following model:

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{treat}_j + v_{0j} + v_{1j} \cdot \text{treat}_j + \epsilon_{ij}$$

   where:

$$\begin{aligned} v_{0j} &\sim \mathcal{N}(0, \tau_0^2) \quad \text{(random intercept for classroom)} \\ v_{1j} &\sim \mathcal{N}(0, \tau_1^2) \quad \text{(random offset for treat=1)} \\ \text{Cov}(v_{0j}, v_{1j}) &= \tau_{01} \quad \text{(covariance between intercept and offset)} \\ \epsilon_{ij} &\sim \text{i.i.d. } \mathcal{N}(0, \sigma^2) \quad \text{(residual error)} \end{aligned}$$

   The β values are fixed effects, and we consider β_1 of primary interest.

   (a) *(2 points)* From the model, what is $\text{var}(Y_{ij})$ for a randomly selected student in a random classroom that has been assigned the new textbook (treat=1)?

   Below is partial R lme4 output from an analysis of the researcher's data, with 100 schools/classrooms (50 using the old text, 50 new), and classroom sizes ranging from 10 to 30 (average class size ~20).

```
> model <- lmer(y ~ treat + (1 + treat | class_id), data = df)
> summary(model)
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ treat + (1 + treat | class_id)
   Data: df
```

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 class_id (Intercept) 21.60    4.647
          treat       17.18    4.145    0.01
 Residual             15.85    3.982
Number of obs: 2066, groups:  class_id, 100

Fixed effects:
            Estimate Std. Error t value
(Intercept)  49.7791     0.6838  72.798
treat         2.4224     1.1132   2.176
```

(b) *(2 points)* For the null hypothesis that the treatment has no effect versus a two-sided alternative, would you reject the null at the 0.05 significance level? Provide your reasoning. (You may assume that the effective degrees of freedom is large enough that a standard normal approximation may be used in place of a t distribution).

(c) *(2 points)* Suppose the researcher took the same data and simply averaged scores within each class, and used these average scores to perform an unequal variance two sample t-test of the 50 classes with the old text vs. the 50 with the new text.. Would this be a valid approach, in the sense of controlling type I error? Explain.

(d) *(4 points)* Another researcher wants to replicate these results in a different city, where the average classroom size is 100, but they can afford to conduct the trial in only 50 classrooms (25 with the new text, 25 old). Using the output above, which study (first or second) do you think will have greater statistical power? Explain.

[HINT: Following part (c), think of observing a single class mean $\bar{Y}_j$ for each class, with $J/2$ classes per treatment group, and assume all classes are the same size n. Let $\bar{\bar{Y}}_{\text{treat}=k}$ denote the grand mean within treatment group k. The power of the t-test will be mainly determined by $\text{var}(\bar{\bar{Y}}_{\text{treat}=1} - \bar{\bar{Y}}_{\text{treat}=0})$, with small values providing greater power. From the output, you should be able to formulate $\text{var}(\bar{\bar{Y}}_{\text{treat}=1} - \bar{\bar{Y}}_{\text{treat}=0})$ in terms of J, n, and plugging in other estimates above, and you may assume $\tau_{01} = 0$. The first study has $J = 100$, $n = 20$, and the second study has $J = 50$, $n = 100$. Which study has smaller $\text{var}(\bar{\bar{Y}}_{\text{treat}=1} - \bar{\bar{Y}}_{\text{treat}=0})$?]

Problem5

(a). $Var(Y_{ij})$

$= Var(\beta_0 + \beta_1 \cdot treat_j + V_{0j} + V_{ij} \cdot trt_j + \varepsilon_{ij})$

$= Var(V_{0j} + V_{ij} + \varepsilon_{ij})$ at trt $= 1$, drop fixed effects.

$= \sigma^2 + Var(V_{0j} + V_{ij})$ since ε_{ij} should be indep from V_{0j}, V_{ij}

$= \sigma^2 + Var(V_{0j}) + Var(V_{ij}) + 2Cov(V_{0j}, V_{ij})$

$= \sigma^2 + \tau_0^2 + \tau_1^2 + 2\tau_{01}$ ✓

if want numerical results → $= 15.85 + 21.6 + 17.18 + 0.02 = 54.65$

(b). $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

$Z = \frac{\hat{\beta_1}}{SE(\hat{\beta_1})} = \frac{2.4224}{1.1132} = 2.176 > Z_{0.975} = 1.96$ ✓

Thus, we reject H_0 at 0.05. ✓

(c). The approach itself is valid, that we are doing a 2 sample t-test. however, under the unequal variance test, the df need to be adjusted using satterthwaite approximation. If the df is not been adjusted by the approx. method, the Type I error will be unpredicted. ✓ good

To comparing the Type I error of doing such test with our current test, we recall the type I error is =

$Prob(\text{reject } H_0 \mid H_0 \text{ is correct}) = \alpha$

Thus when SE is larger, we tend to fail to reject H_0.

When comparing the variance of mean, the overall variance tend to be smaller than fitting a full model (as $var(\bar{X}) = \frac{1}{n} var(X)$)

Thus are more likely to reject H_0, thus α may increase, not so optimal.

Also, testing on average losses information (indeed). Actually, not much loss

(d). now we have less J but more n in 2nd study:

for $J = 100$, $n = 20$,

$Var(\bar{Y}_{..trt} - \bar{Y}_{..control})$

$= Var(\bar{U}_{0.} + \bar{U}_{1.} + \bar{\varepsilon}_{..} - \bar{U}_{0.} - \bar{\varepsilon}_{..})$

$= Var(\bar{U}_{1.})$

$= \frac{1}{100} \tau_1^2$

≈ 0.1718

3

for $J = 50$, $n = 100$

similarly,

$var(\bar{Y}_{..trt} - \bar{Y}_{..control})$

$= \frac{1}{50} \tau_1^2$

$= 0.3436$

Thus, first study have smaller variance. thus first test has more power.

More complicated tradeoff between J and n