Problem 1 A random variable Y belongs to the univariate exponential family if the point mass (pmf) or probability density function (pdf) can be written on the form

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w)\right) \tag{1}$$

where θ is the natural (or canonical) parameter, b and c are functions, and the support of the pmf/pdf does not depend on θ. In the following you can assume that the dispersion parameter $\phi = 1$ and that the weight $w = 1$.

If $X \sim \text{bin}(n, p)$ and Y has the same distribution as X conditional on $X > 0$, then Y is zero-truncated binomial distributed with point mass function

$$f(y|n, p) = \frac{1}{1 - (1-p)^n} \binom{n}{y} p^y (1-p)^{n-y}$$

for $y = 1, 2, \ldots, n$.

a) Show that the zero-truncated binomial distribution belongs to the exponential family and identify θ and the function $b(\theta)$.

b) Using the connection between $E(Y)$ and $b(\theta)$, derive expressions for $E(Y)$ both in terms of n and θ or n and p.

Explain why $E(Y)$ should be asymptotically equal to np as n tends to ∞ and check that this is the case by verifying that

$$\lim_{n \to \infty} \frac{E(Y)}{np} = 1.$$

Problem 2

In this problem will we analyse traffic data on number of cyclists passing two different road points A and B in Trondheim at 56 different days in the months of May and June of 2022. The first 20 observations in the data frame are given below.

```
head(df,20)
```

```
##    daynr weekday location weekend  y
## 1      1     mon        A   FALSE 16
## 2      1     mon        B   FALSE 12
## 3      2     tue        A   FALSE 25
## 4      2     tue        B   FALSE 15
## 5      3     wed        A   FALSE 14
## 6      3     wed        B   FALSE 15
## 7      4     tur        A   FALSE 20
## 8      4     tur        B   FALSE  7
## 9      5     fri        A   FALSE 22
## 10     5     fri        B   FALSE  7
## 11     6     sat        A    TRUE 14
## 12     6     sat        B    TRUE  9
## 13     7     sun        A    TRUE  9
## 14     7     sun        B    TRUE  5
## 15     8     mon        A   FALSE 23
## 16     8     mon        B   FALSE  9
## 17     9     tue        A   FALSE 19
## 18     9     tue        B   FALSE 10
## 19    10     wed        A   FALSE 21
## 20    10     wed        B   FALSE 10
```

All variables are encoded as factors except `y` (the number of cyclists) which is a numeric variable.

We first fit a generalized linear model to the data in R as follows.

```
mod1 <- glm(y ~ location + weekday, poisson(link = "log"), data = df)
summary(mod1)
```

```
##
## Call:
## glm(formula = y ~ location + weekday, family = poisson(link = "log"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0356  -1.0687  -0.2070   0.7586   2.6667
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.97467    0.06730  44.202  < 2e-16 ***
## locationB   -0.65555    0.05294 -12.382  < 2e-16 ***
## weekdaytue   0.20448    0.08733   2.342   0.0192 *
## weekdaywed   0.14086    0.08861   1.590   0.1119
## weekdaytur   0.07291    0.09004   0.810   0.4181
## weekdayfri   0.10368    0.08938   1.160   0.2461
## weekdaysat  -0.57443    0.10800  -5.319 1.04e-07 ***
## weekdaysun  -0.62024    0.10961  -5.659 1.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 428.75  on 111  degrees of freedom
## Residual deviance: 135.04  on 104  degrees of freedom
## AIC: 638.26
##
## Number of Fisher Scoring iterations: 4
```

a) State the assumptions of this specific statistical model in suitable mathematical notation.

Based on this model, what is the expected number of cyclists passing road point B on a Monday?

Also give an interpretation of the coefficient estimate labelled `locationB` in the output.

Next we fit the following alternative model which includes the factor `weekend` (with levels `TRUE` and `FALSE`) instead of the factor `weekday` (which has 7 levels).

```
mod0 <- glm(y ~ location + weekend, poisson(link = "log"), data = df)
summary(mod0)


##
## Call:
```

```
## glm(formula = y ~ location + weekend, family = poisson(link = "log"),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.03707  -1.03971  -0.06234   0.77045   2.75113
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.08138    0.03290   93.65   <2e-16 ***
## locationB   -0.65555    0.05294  -12.38   <2e-16 ***
## weekendTRUE -0.70378    0.06762  -10.41   <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: 428.75  on 111  degrees of freedom
## Residual deviance: 141.31  on 109  degrees of freedom
## AIC: 634.52
##
## Number of Fisher Scoring iterations: 4
```

b) Explain what is meant by nested models and explain why `mod0` and `mod1` are nested.

Do a likelihood ratio test of `mod0` against `mod1` using a level of significance of 0.05.

c) Based on `mod0`, is there any indication of overdispersion in the data?

Discuss possible mechanism that may be responsible for generating overdispersion in the present data set.

Assuming a quasi-Poisson model where the true variance of the response is inflated by a factor φ, compute an estimate $\hat{\varphi}$ of this dispersion parameter and use this to compute an adjusted standard error of the estimated coefficient labelled `locationB` of model `mod0`.

Next we extend the GLM considered above by including a random intercept term using `daynr` (56 levels) as the grouping factor and fit the following GLMM:

```
mod2 <- glmer(y ~ location + weekend + (1|daynr), poisson(link = "log"), data = df)
summary(mod2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: y ~ location + weekend + (1 | daynr)
##    Data: df
##
##      AIC      BIC   logLik deviance df.resid
##    633.2    644.0   -312.6    625.2      108
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.51589 -0.91263 -0.01439  0.62878  2.72762
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  daynr  (Intercept) 0.01402  0.1184
## Number of obs: 112, groups:  daynr, 56
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.07461    0.03808  80.730   <2e-16 ***
## locationB   -0.65555    0.05290 -12.391   <2e-16 ***
## weekendTRUE -0.70458    0.07618  -9.249   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) loctnB
## locationB   -0.475
## weekendTRUE -0.380  0.000
```

d) Indexing clusters (different days) by $i = 1, 2, \ldots, 56$ and observations within clusters by $j = 1, 2$ in the usual way, and letting y_{ij} denote an observed response and $\mathbf{x}_{ij}$ the associated fixed effect covariate vector, state the assumptions of this specific statistical model in mathematical notation.

e) Given the model assumptions, what is the expectation of y_{ij} conditional on the random intercept for day number (cluster) i?

What is the conditional covariance between two observations y_{i1} and y_{i2} on the same day?

Derive an algebraic expression for the intraclass (or intraday) covariance between two observations y_{i1} and y_{i2} on the same day for given covariate vectors $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$. Hint: You will need to use properties of the log-normal distribution.

f) Does the data provide evidence that there is variation between days as modeled by `mod2`? Do a formal hypothesis test of this. This should include precise statements about the null and alternative hypothesis, your choice of test statistic and its (approximate) distribution. Compute the critical value of the test using a level of significance of $\alpha = 0.05$. The maximum log likelihood for the second GLM above (`mod0`) is -314.26.

g) Given your results in point c), e) and f), would you trust the estimated standard errors of the estimated regression coefficients of `mod0`? Why or why not?

h) Let $f(y_{ij}|\boldsymbol{\beta}, \gamma_i)$ denote the conditional point mass function of each observation conditional on the random intercept γ_i for day i, and let $f(\gamma_i|\tau^2)$ denote the density function of each random effect for the above GLMM. Derive an expression for the marginal likelihood $L(\boldsymbol{\beta}, \tau^2)$ of the above model expressed in terms of these two functions.

Name two methods by which this likelihood can be approximated or evaluated numerically.