Total: 100 points

Note: In any question, if you are using R, all R codes and R outputs must be included in your answers. Submit the R markdown file along with your answers (with knitted pdf as well). Provide comments for R programs and TAs can easily follow your codes.

Q. 1 (7 points) For known $k \geq 2$ show that the negative binomial distribution with probability mass function,

$$f(y,k,\mu) = \binom{y+k-1}{y}\left(1-\frac{k}{\mu+k}\right)^{y}\left(\frac{k}{\mu+k}\right)^{k}, \quad y = 0,1,2,\ldots$$

belongs to the exponential family of distributions. Find the natural parameter for this distribution.

Q. 2 (12 points) An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the numbers of imperfections are 8, 5, 6, 6, 5, 4, 7, 2, 3, 4. Treatment B applied to 10 other wafers has 9, 10, 8, 14, 8, 13, 11, 4, 7, 6 imperfections. Treat the counts as independent Poisson random variables having means μ_A and μ_B.

(a) (3 pts) Fit the model $log(\mu) = \alpha + \beta x$, where $x = 1$ for treatment B and $x = 0$ for treatment A.

(b)(3 pts) Show that $\beta = log(\mu_B/\mu_A)$ and interpret your estimate of obtained in part (a) above.

(c) (3 pts) Test $H_0 : \mu_A = \mu_B$ against two sided alternative $H_1 : \mu_A \neq \mu_B$ using Wald test.

(d) (3 pts) Construct a Wald type 95% confidence interval for μ_B/μ_A.

Q. 3 (23 points) For the horseshoecrabs data in Table 3.2, fit the Poisson regression model to use weight to predict the number of satellites. The dataset "horseshoecrabs.dat" is posted at Quercus.

(a) (3 pts) Using x = weight and Y = number of satellites, fit a Poisson loglinear model. Report the prediction equation.

(b) (5 pts) Use $\hat{\beta}$ to describe the weight effect. Construct and interpret a 95% Wald confidence intervals for β and another for the multiplicative effect of a 1kg increase on Y.

(c) (5 pts) Conduct a likelihood-ratio test about the weight effect. Interpret.

(d) (5 pts) Plot the standardized deviance residuals against the fitted values. Do you see any outliers?

(e) (5 pts) Obtain the deviance of the model. Can one use it to conduct a Goodness-of-Fit test? Why or why not?

Q. 4 (18 points) Continue the previous problem. This time we want to fit a negative binomial loglinear model that allows overdispersion.

(a) (5 pts) Count the number of female crabs with weights in the intervals < 1.5, 1.5-1.7, 1.7-1.9, . . . , 3.1-3.3, > 3.3 kg. Find the sample mean and sample variance for Y = number of satellites of female crabs in each Weight category. Display the results in a table as follows

Weight (kg)	No. Cases	Sample Mean	Sample Variance
< 1.5	5	0.80	3.20
1.5-1.7	11	0.82	1.56
1.7-1.9	16	1.25	7.93
1.9-2.1	?	?	?
⋮	⋮	⋮	⋮
3.1-3.3	?	?	?
> 3.3	?	?	?

Do you see any evidence of overdispersion?

(b) (3 pts) Allow overdispersion by fitting the negative binomial loglinear model. Report the prediction equation and the estimate of the dispersion parameter θ and its SE.

(c) (5 pts) Construct and interpret a 95% Wald confidence interval for $exp(\beta)$ based on the negative binomial loglinear model. Compare it with the one in (b) in the previous problem, and explain why the interval is wider with the negative binomial model.

(d) (5 pts) Plot the standardized deviance residuals against the fitted values. Do you see any outlier(s)?

Q. 5 (15 points)

The table below refers to ratings of agricultural extension agents in North Carolina. In each of 5 districts, agents were classified by their race and by whether they got a merit pay increase. The data can be load in R using the R codes below

Race	Blacks		Whites	
Merit Pay	Yes	No	Yes	No
District				
NC	24	9	47	12
NE	10	3	45	8
NW	5	4	57	9
SE	16	7	54	10
SW	7	4	59	12

```
PayYes = c(24,10,5,16,7,47,45, 57,54,59)
PayNo = c(9,3,4,7,4,12,8,9,10,12)
District = rep(c("NC", "NE", "NW", "SE", "SW"),2)
Race = c(rep("Blacks",5), rep("Whites",5))
```

```
data.frame(Race, District, PayYes, PayNo)
```

(a) (5 pts) Fit a logistic regression model with the main effects of race and district but no interaction. Show how you test whether the merit pay increase is independent of race, conditional on the district using a Wald test and a LR test about a parameter in the logistic regression model you just fit.

(b) (5 pts) Estimate the common odds ratio between Merit Pay and Race based on the logistic model in the previous part, and report the 95% Wald confidence interval and the 95% likelihood ratio confidence interval for the common odds ratio. Interpret the confidence intervals.

(c) (5 pts) Test whether the association between merit pay decision and race is homogeneous across the five districts using a likelihood ratio test. Please report the test statistic, degrees of freedom, P-value, and then make conclusion.

Q. 6 (25 points) We are using the same MBTI dataset. We fit a model using the four scales as predictors of whether a subject drinks alcohol frequently. Answer the following questions.

(a) (5 pts) Conduct a model goodness-of-fit test, and interpret. If you were to simplify the model by removing a predictor, which would you remove? Why?

(b) (10 pts) Report AIC values for the model with the four main effects and the six interaction terms; the model with only the four main effect; the model with no predictors. According to this criterion, which model is preferred? Explain the rationale for using AIC.

(c) (10 pts) Using the MBTI data set, use model-builing methods to select a model for this alcohol response variable.