

## October 2008 Midterm solutions

Q1. a. We begin by calculating the relevant denominators for obtaining rates. We need to account for the number of cars and the number of miles driven by the cars. For the first segment we have 200,000 cars (average) per week, driving 2.5 miles, over 10 weeks, so  $T_1 = 200,000 \times 2.5 \times 10 = 5,000,000$  miles driven over the segment. By analogy with the concept of person-years we could call these vehicle-miles. A similar calculation for segment 2 yields  $T_2 = 9,000,000$  vehicle-miles resulting in rates of

$$\hat{\lambda}_1 = 111 \div T_1 = 2.22 \times 10^{-5} \quad \text{and} \quad \hat{\lambda}_2 = 156 \div T_2 = 1.68 \times 10^{-5}$$

Thus the estimated rates are 2.22 and 1.68 accidents per 100,000 vehicle-miles for the segments, respectively.

b. The estimated ratio of rates (we'll consider segment 1 versus segment 2 for convenience, as that yields a ratio  $> 1$ ) is  $RR = 2.22/1.68 = 1.32$ . To obtain a confidence interval we begin with the confidence interval for the logarithm of the ratio of Poisson means,  $\theta$ . The point estimate is

$$\hat{\theta} = \log(111/156) = -.3077,$$

$$\text{with standard error} = \sqrt{\frac{1}{111} + \frac{1}{156}} = .1242$$

leading to a confidence interval of

$$-.3077 \pm 1.96 \times .1242 = (-0.5528, -0.06270).$$

Exponentiating this interval yields  $(0.5754, 0.9392)$ . To

convert this to rates, we divide by  $T_1/T_2$  (or multiply by  $T_2/T_1$ ) to get  $(1.04, 1.69)$ .

Note: These calculations can be conducted using Poisson regression as follows:

```
vMiles <- 10*c(200000*2.5,300000*3.0)
accidents <- c(111,151)
segment <- factor(1:2)
pfit <- glm(accidents ~ offset(log(vMiles)) + segment, family=poisson)

> coefTable(pfit,transf = function(x) exp(-x) )
      Estimate Std. Error  Pr(>|z|)      2.5 %      97.5 %
(Intercept) 45045.045045 0.09491578 0.00000000 37398.533001 54254.964575
segment2    1.323179 0.12502608 0.02510203  1.035610  1.690600
```

I've used the `coefTable` function that's on the web-page (updated Oct. 16). I've backtransformed with  $\exp(-x)$  rather than  $\exp(x)$  because otherwise I would get the relative ratio for 2 vs. 1 and I want 1 vs. 2.

Q2. We are provided with frequencies of the six possible outcomes from throwing a die. If the die were fair, the probabilities for each outcome would be  $1/6$ , so based on 120 throws, we would have expected frequencies of 20 for in each case. We can apply a Pearson's goodness of fit test and calculate

$$\chi^2 = \frac{\sum_{i=1}^6 (n_i - e_i)^2}{e_i} = \frac{(15 - 20)^2}{20} + \frac{(19 - 20)^2}{20} + \dots = 18.8$$

We have  $n=6$  frequencies with no parameters estimates, so the degrees of freedom for referring to  $\chi^2$  tables is  $n-1 = 5$ .

We calculate the p-value =  $\text{Prob}(X_{(5)}^2 > 18.8) < .005$  to the accuracy provided in the table. Thus we have strong evidence that the die is "loaded".

b. The observed frequency of "6" is 36, so the estimated probability is  $36/120 = .30$ . The estimated probability of two "6"'s is  $.30 \times .30 = .09$ . There are two ways to proceed to get a confidence interval. The easiest way is to start from the confidence interval for  $p = \text{Prob}(\text{getting "6"})$ . Since  $\hat{p} = .30$ , we calculate

$$.30 \pm 1.96 \times \sqrt{\frac{.3 \times .7}{120}} = (0.218, 0.382)$$

To obtain the interval for  $p^2$  we just square this, yielding a 95% confidence interval of  $(0.048, 0.146)$

The second (and likely less accurate) approach would be to

apply the delta method. since  $\frac{d p^2}{d p} = 2 \times p$  yields the

estimated standard error for  $\hat{p}^2 = 2 \times \hat{p} \times \sqrt{\frac{.3 \times .7}{120}} = .0251$

which when plugged into the generic asymptotic formula yields a confidence interval of  $(0.041, 0.139)$ . If the dice were fair the probability of getting double sixes would be

$\frac{1}{6^2} = 0.028$ . This value is not contained in the interval, so

by this confidence interval indicates that the data is not consistent with a fair die.

Q3. For each of the independent observations,  $Y_i$ , we have

$$\text{that } \text{Prob}(Y_i = y_i) = \frac{(\theta t_i)^{y_i} \times e^{-\theta t_i}}{y_i!}$$

so the log likelihood is

$$\log \text{Lik}(\theta) = \sum_{i=1}^n \left[ y_i (\log(\theta) + \log(t_i)) - \theta t_i - \log(y_i!) \right]$$

and

$$\frac{d \log \text{Lik}(\theta)}{d \theta} = \sum_{i=1}^n \left[ \frac{y_i}{\theta} - t_i \right] = \frac{\sum_{i=1}^n y_i}{\theta} - \sum_{i=1}^n t_i$$

setting this equal to 0 yields that  $\hat{\theta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n t_i}$

b. The second derivative of the log likelihood is

$$\frac{d^2 \log \text{Lik}(\theta)}{d\theta^2} = -\frac{\sum_{i=1}^n y_i}{\theta^2}.$$

Taking the expected value of the negative of this yields the Fisher information,

$$I(\theta) = \frac{\sum_{i=1}^n \theta t_i}{\theta^2} = \frac{\sum_{i=1}^n t_i}{\theta}$$

Plugging in the maximum likelihood estimate provides an estimate for the variance of

$$I(\hat{\theta}) = \frac{\sum_{i=1}^n t_i}{\hat{\theta}} = \frac{\left(\sum_{i=1}^n t_i\right)^2}{\sum_{i=1}^n y_i}$$

which provides an estimated standard error of

$$\sqrt{\left(I(\hat{\theta})\right)^{-1}} = \frac{\sqrt{\sum_{i=1}^n y_i}}{\sum_{i=1}^n t_i}$$