**Chapter 4 – Analyzing a count response**

Chapters 1 and 2 examined a binomial response (a count) arising from a FIXED number of trials. Thus, W, a binomial random variable, could be 0, 1, …, n, where n is the number of trials.

Count responses also arise from other mechanisms that have nothing to do with binary trials. Examples include:

* The number of arrests for a city per year
* The number of credit cards an individual owns
* The number of people arriving at an airport on a given day
* The number of cars stopped at the 33rd and Holdrege streets intersection in Lincoln
* The number of people standing in line at Starbucks.

For these settings, a Poisson distribution can be used to model the count responses.

You may also think of particular explanatory variables that could affect these count responses. For example, the number of credit cards an individual owns could be dependent on income level, gender, where they live, … .

You may also want to control for other items that are not necessarily explanatory variables. If the purpose of the number of arrests example was to compare cities in the United States, we would want to account for city sizes. For example, it would not make sense to compare Omaha directly to New York City without taking into account their vastly different sizes.

Next, we begin with the basics of how we model counts through using a Poisson probability distribution.

**Section 4.1 – Poisson model for count data**

Poisson probability mass function:



for y = 0, 1, 2, …, where

Y is a random variable

y denotes the possible outcomes of Y

μ is a parameter that is greater than 0

I will sometimes write Y ~ Po(μ) as shorthand notation to mean that Y has a Poisson distribution with parameter μ.

Y will denote the number of occurrences of an event.

Properties

Below are characteristics of a Poisson distribution and associated items of interest:

* One can show the mean and variance of Y to be:

E(Y) = μ and Var(Y) = μ

Having the mean and variance BOTH equal to μ is convenient but also limiting. Often, the actual variability in application can be GREATER than μ. This is referred to as overdispersion. We will discuss how to handle this in a future chapter.

* If Y1, …, Yn are independent with distribution Po(μ), then . If each Yk had a different mean μk, we would have .
* Likelihood function:



* MLE for μ is , i.e., the sample mean
* The estimated variance for  is



Wald confidence interval for μ

What is the interval?

Can this interval have a lower bound less than 0? If so, how can we fix it?

Find the Wald interval for log(μ) first! Transform the interval back for μ:



One can show through a delta-method approximation that .

Hypothesis test for μ

H0: μ = μ0 vs. Ha: μ ≠ μ0

Wald test statistic: 

Score test statistic: 

How would you decide on rejection of the null hypothesis?

Score confidence interval for μ



How was this interval derived?

Which interval do you think is better: Wald or score? Please see an example in the book for true confidence level plots.

How else could you find a confidence interval for μ?

Example: 33rd and Holdrege (Stoplight.R, StoplightAdditional.R, Stoplight.csv)

The intersection at 33rd and Holdrege Streets is a typical north-south/east-west, 4-way intersection.



Below is a picture taken from a location northeast of the intersection:



Approximately 150 feet north of the intersection is a fire station located on the west side of the street. A back-up of vehicles at the stoplight waiting to go south could block the fire station's driveway, which would prevent emergency vehicles from exiting the station.

To examine this more closely, I took a sample of 40 consecutive stoplight cycles from 3:25PM to 4:05PM on a non-holiday weekday and recorded the number of vehicles stopped at the stoplight going. Below is part of the data:

> stoplight <- read.csv(file = "C:\\data\\stoplight.csv")

> head(stoplight)

 Observation vehicles

1 1 4

2 2 6

3 3 1

4 4 2

5 5 3

6 6 3

Note that there were no vehicles remaining in the intersection for more than one stoplight cycle. Why is this important to know?

Is a Poisson distribution appropriate for this data?

> # Summary statistics

> mean(stoplight$vehicles)

[1] 3.875

> var(stoplight$vehicles)

[1] 4.317308

> # Frequencies

> table(stoplight$vehicles)

0 1 2 3 4 5 6 7 8

1 5 7 3 8 7 5 2 2

> rel.freq <- table(stoplight$vehicles) / length(stoplight$vehicles)

> rel.freq2 <- c(rel.freq, rep(0, times = 7))

> #Poisson calculations

> y <- 0:15

> prob <- round(dpois(x = y, lambda = mean(stoplight$vehicles)), 4)

> data.frame(y, prob, rel.freq = rel.freq2)

 y prob rel.freq

1 0 0.0208 0.025

2 1 0.0804 0.125

3 2 0.1558 0.175

4 3 0.2013 0.075

5 4 0.1950 0.200

6 5 0.1511 0.175

7 6 0.0976 0.125

8 7 0.0540 0.050

9 8 0.0262 0.050

10 9 0.0113 0.000

11 10 0.0044 0.000

12 11 0.0015 0.000

13 12 0.0005 0.000

14 13 0.0001 0.000

15 14 0.0000 0.000

16 15 0.0000 0.000

> plot(x = y-0.1, y = prob, type = "h", ylab = "Probability", xlab = "Number of vehicles", lwd = 2, xaxt = "n")

> axis(side = 1, at = 0:15)

> lines(x = y+0.1, y = rel.freq2, type = "h", lwd = 2, lty = "solid", col = "red")

> abline(h = 0)

> legend(x = 9, y = 0.15, legend = c("Poisson", "Observed"), lty = c("solid", "solid"), lwd = c(2,2), col = c(“black”, “red”), bty = “n”)



Confidence interval calculations:

> alpha <- 0.05

> n <- length(stoplight$vehicles)

> mu.hat <- mean(stoplight$vehicles)

> #Wald

> mu.hat + qnorm(p = c(alpha/2, 1-alpha/2))\*sqrt(mu.hat/n)

[1] 3.264966 4.485034

> #Wald using a log() transformation

> exp(log(mu.hat) + qnorm(p = c(alpha/2, 1-alpha/2)) \* sqrt(1/(mu.hat\*n)))

[1] 3.310561 4.535674

> #Score

> (mu.hat + qnorm(p = c(alpha/2, 1-alpha/2)^2)/(2\*n)) + qnorm(p = c(alpha/2, 1-alpha/2)) \* sqrt((mu.hat + qnorm(p = 1-alpha/2)^2/(4\*n))/n)

[1] 3.311097 4.534939

> #Usual t-distribution based interval

> t.test(x = stoplight$vehicles, conf.level = 0.95)

 One Sample t-test

data: stoplight$vehicles

t = 11.7949, df = 39, p-value = 1.955e-14

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

 3.210483 4.539517

sample estimates:

mean of x

 3.875

Suppose all vehicles are 14 feet long with a distance between cars of 4 feet when stopped at the intersection. This suggests that 9 vehicles (150/18 = 8.3) or more waiting at the intersection could at least partially block the fire station's driveway. Using the estimated Poisson distribution, estimate the probability this will happen.

> prob9 <- 1 - ppois(q = 8, lambda = mu.hat)

> prob9

[1] 0.01786982

This is the probability for one stoplight cycle. What about this event occurring at least once over the 60 stoplight cycles that would occur over one hour?

Let W be a binomial random variable with n = 60 and π = 0.01786982. Then P(W ≥ 1) = 1 – P(W = 0) = 0.66.

> # Not in book’s program

> 1 - dbinom(x = 0, size = 60, prob = prob9)

[1] 0.661044

Questions:

* What do you think will happen to this probability during rush-hour?
* How could we take into account different time periods of the day?
* This is an estimated probability. How could we find a confidence interval for it? The interval is (0.3305, 0.9916). See my additional program for details.

Other intervals for μ exist. Please see the book for the likelihood ratio and exact confidence intervals.