Week 7 Tests and CI's for Count Data: One Sample Poisson

STAT 111

Zhengfei Li (Alex) Instructed by Prof. Phil Everson

I would like to acknowledge that this handout is created with reference to Mathematical Statistics and Data Analysis by J. A. Rice (2007). I would also like to acknowledge the instructions from Prof. Everson.

1 Preliminaries

1.1 One Sample Poisson Distribution

$$X \sim Pois(\theta)$$
$$P(X = k) = \frac{\theta^k e^{-\theta}}{k!}$$

1.2 χ^2 Goodness of Fit Test

χ^2 Goodness of Fit Test is also called Pearson's chi-square test. We put a random sample of n test statistics into m bins, with count of statistics in each bin being x_i. The hypothesis is:

H_0 : X follows the suggested distribution.

H_A : X does not follow the suggested distribution (might be a wrong parameter or might be a totally different distribution).

and test statistic defined as:

$$\begin{aligned} X^2 &= \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \\ &= \sum_{i=1}^{n} \frac{[x_i - n * P_i(\hat{\theta})]^2}{nP_i(\hat{\theta})} \end{aligned}$$

Notice that we know the total count being n, which takes 1 away from degree of freedom. Also notice that $\hat{\theta}$ is estimated. Assume that we need to estimate k count of parameters, so our degree of freedom is

$$df = m - 1 - k$$

2 Problem 2 Setup

A classic example of Poisson data are the counts of deaths in the Prussian Calvary due to horse-kicks to the head. Data for 200 Corps-Years appear in the table below:

Deaths	Count	Proportion
0	109	0.545
1	65	0.325
2	22	0.110
3	3	0.015
4	1	0.005

Table 1: Deaths in the Prussian Calvary due to horse-kicks to the head

3 Problem 2a

3.1 First we estimate $\hat{\theta}$

Overall there were 122 such deaths in the 10 Army Corps over a 20 year period, for an estimated rate of

$$\hat{\theta} = \frac{0 \cdot 109 + 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1}{200} = 0.61$$

deaths per corp-year.

3.2 Then, we compute the expected counts

```
theta_hat = (0 * 109 + 1 * 65 + 2 * 22 + 3 * 3 + 4 * 1)/200
deaths = c("0", "1", "2", "3+")
cnt_obs = c(109, 65, 22, 4)
cnt_exp = c(0, 1, 2, 3)
cnt_exp = dpois(cnt_exp, lambda = theta_hat) * 200
cnt_exp[4] = 200 - cnt_exp[1] - cnt_exp[2] - cnt_exp[3]
df_death_chisq = data.frame(deaths, cnt_obs, cnt_exp)
df_death_chisq
```

```
  deaths cnt_obs    cnt_exp
1      0     109 108.670174
2      1      65  66.288806
3      2      22  20.218086
4     3+       4   4.822934
```

3.3 Then, we may carry out a Chi-square goodness of fit test

We are estimating θ, so $df = 4 - 1 - 1 = 2$.

```
chi_sq_stat = sum((df_death_chisq$cnt_obs - df_death_chisq$cnt_exp)^2/
                    df_death_chisq$cnt_exp)
df = 4 - 1 - 1
```

```
p_val <- 1 - pchisq(chi_sq_stat, df)
sprintf("chi-square stat %.2f; degree of freedom %d; p-value %.2f.",
        chi_sq_stat, df, p_val)
```

```
[1] "chi-square stat 0.32; degree of freedom 2; p-value 0.85."
```

This is very high, which means that the data is highly likely following poisson distribution.

3.4 (Optional) Poisson Dispersion Test

The Poisson Dispersion Test is a GLR test where the alternative hypothesis is that the data is Poisson but there are different rates.

$H_0 : x_i \sim Pois(\hat{\theta})$

$H_0 : x_i \sim Pois(\tilde{\theta}_i)$

Note that we will estimate $\hat{\theta} = \bar{x}$ and $\tilde{\theta} = x_i$

$$
\begin{aligned}
\Lambda &= \frac{\prod \hat{\theta}^{x_i} e^{-\hat{\theta}}/x_i!}{\prod \tilde{\theta}_i^{x_i} e^{-\tilde{\theta}_i}/x_i!} \\
&= \prod \left(\frac{\bar{x}}{x_i}\right)^{x_i} e^{x_i - \bar{x}} \\
-2\log(\Lambda) &= 2\sum x_i \log\left(\frac{x_i}{\bar{x}}\right)
\end{aligned}
$$

following chi-square distribution with degree of freedom is $df = n - 1$.

```
x <- c(rep(0, 109), rep(1, 65), rep(2, 22), rep(3,3), rep(4, 1))
pois_disp_stat = 2* sum(log((x/mean(x))^x))
df = 200 - 1
p_val = pchisq(pois_disp_stat, df)
sprintf("Test staistic %.2f with df %d yields p-value %.2f",
        pois_disp_stat, df, p_val)
```

```
[1] "Test staistic 212.47 with df 199 yields p-value 0.76"
```

Note that the chi-square distribution approximation for the the GLR statistic only works for large samples: the performance is poor even with sample size 100:

```
#### test Poisson dispersion test - not very good with lambda = 0.61
nsim=10000
teststat = rep(0,nsim)
n=200
lambda=100
```

```
for(i in 1:length(teststat)){
 x=rpois(n,lambda)
 xbar = mean(x)
 teststat[i] = 2*sum(log((x/xbar)^x))
 }

mean(teststat) # larger than what we'd expect for chi-square(199): 199
```

```
[1] 199.1848
```

4 Problem 2b

For $X_1, \cdots, X_n \overset{i.i.d.}{\sim} Pois(\theta)$, find the:

4.1 MLE

$$
\begin{aligned}
L(\theta) &= \frac{\theta^{\sum X_i} e^{-n\theta}}{\prod X_i!} \\
l(\theta) &= \sum X_i \cdot \ln(\theta) - n\theta = \sum \ln(X_i!) \\
\text{let } l'(\theta) &= \frac{\sum X_i}{\theta} - n = 0 \\
\hat{\theta}_{MLE} &= \frac{\sum X_i}{n} = \bar{X}
\end{aligned}
$$

4.2 Information

$$
\begin{aligned}
l'(\theta) &= \frac{\sum X_i}{\theta} - n \\
l''\theta &= -\frac{\sum X_i}{\theta^2}
\end{aligned}
$$

$$
\begin{aligned}
I(\theta) &= -E(l''(\theta)) \\
&= -E(-\frac{\sum X_i}{\theta^2}) \\
&= \frac{E(\sum X_i)}{\theta^2} \\
&= \frac{\sum E(X_i)}{\theta^2} \\
&= \frac{\sum \theta}{\theta^2} \\
&= \frac{n\theta}{\theta^2} \\
&= \frac{n}{\theta}
\end{aligned}
$$

4.3 Construct a large sample approximate 95% CI for θ

$$
\begin{aligned}
Var(\hat{\theta}_{MLE}) &= \frac{1}{I(\theta)} \\
&= \frac{\theta}{n} \\
&= \frac{\bar{X}}{n} \\
SE(\hat{\theta}_{MLE}) &= \sqrt{\frac{\bar{X}}{n}}
\end{aligned}
$$

Knowing that $\hat{\theta}_{MLE} = \frac{\sum X_i}{n} = \bar{X}$

Therefore, we may construct the 95% CI for θ:

$$
\left(\bar{X} - 1.96 \cdot \sqrt{\frac{\bar{X}}{n}}, \bar{X} + 1.96 \cdot \sqrt{\frac{\bar{X}}{n}}, \right)
$$

Using the data we have

```
n = 200
x_bar = (0 * 109 + 1 * 65 + 2 * 22 + 3 * 3 + 4 * 1)/200
se = sqrt(x_bar/n)
z = qnorm(0.975)
sprintf("The CI is (%.2f, %.2f)", x_bar - z*se, x_bar + z*se)
```

```
[1] "The CI is (0.50, 0.72)"
```

5 Problem 2c

The large-sample approximation considers $Y = \sum X_i \dot\sim N(n\theta, n\theta)$, so $0.95 \approx P(|Y - n\theta| < 2.0\sqrt{n\theta})$ Find the analog to the Binomial plus four CI by solving for the range of nθ values to make the inequality true.

$$
\begin{aligned}
|Y - n\theta| &< 2\sqrt{n\theta} \\
(Y - n\theta)^2 &< 4n\theta
\end{aligned}
$$

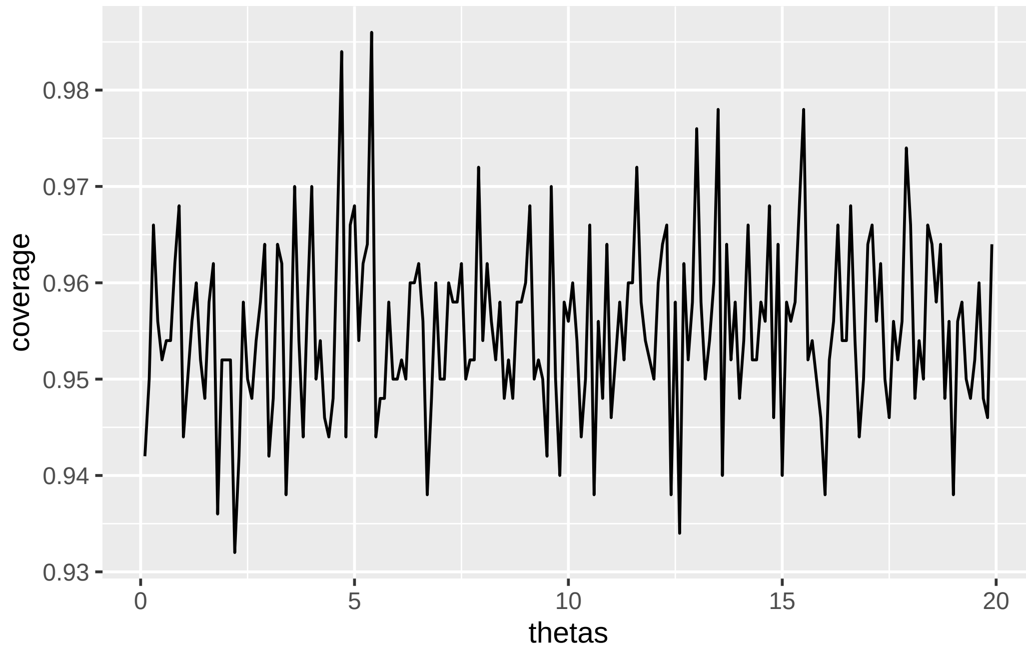
$$
\begin{aligned}
Y^2 - 2n\theta Y + n^2\theta^2 &< 4n\theta \\
n^2\theta^2 - (2nY + 4n)\theta + Y^2 &< 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{(2nY+4n) - \sqrt{(2nY+4n)^2 - 4n^2Y^2}}{2n^2} &< \theta < \frac{(2nY+4n) + \sqrt{(2nY+4n)^2 - 4n^2Y^2}}{2n^2} \\
\frac{(2nY+4n) - \sqrt{4n^2Y^2 + 16n^2Y + 16n^2 - 4n^2Y^2}}{2n^2} &< \theta < \frac{(2nY+4n) + \sqrt{4n^2Y^2 + 16n^2Y + 16n^2 - 4n^2Y^2}}{2n^2} \\
\frac{(2nY+4n) - \sqrt{16n^2Y + 16n^2}}{2n^2} &< \theta < \frac{(2nY+4n) + \sqrt{16n^2Y + 16n^2}}{2n^2} \\
\frac{Y + 2 - 2\sqrt{1+Y}}{n} &< \theta < \frac{Y + 2 + 2\sqrt{1+Y}}{n}
\end{aligned}
$$

5.1 Run simulations to test the coverage probabilities.

```
nreps = 500
n = 200
ntheta = 200
coverage = rep(NA, ntheta)
thetas = rep(NA, ntheta)
for (j in 1:ntheta-1){
  theta = j/10
  included = 0
  for (i in 1:nreps) {
    sum_x = sum(rpois(n, theta))
    lb = (sum_x + 2 - 2*sqrt(1 + sum_x))/(n)
    ub = (sum_x + 2 + 2*sqrt(1 + sum_x))/(n)
    if (theta > lb && theta < ub) {
      included = included + 1
    }
  }
  thetas[j] = theta
  coverage[j] = included/nreps
}
```

```r
data = data.frame(thetas, coverage)
ggplot(data, aes(x=thetas, y=coverage)) + geom_line()
```



6 Problem 2d

Use the relationship between the Gamma and Poisson distributions to show the following equality involving their cumulative distribution functions:

`pgamma(t, k, lambda) = pgamma(lambda, k, t ) = 1 - ppois(k-1, lambda*t)`

Let's think about gamma r.v. T as the total waiting time for multiple successes, and poisson r.v. Y as the count of successes in a particular0 interval of time. Let's assume that λ is the rate that an event occurs, t be the real waittime for at least k events.

At least k events happening in time t means $Y \geq k$, which is the same as the wait time for the k-th event is less than t ($T \leq t$). Therefore:

$$P(Y \geq k) = P(T \leq t)$$

Therefore, putting it into code, we have $P(T \leq t) =$`pgamma(t, k, lambda)`, and $P(Y \geq k) =$`1 - ppois(k-1, lambda*t)`, and they are equal.

We may also notice that `pgamma(lambda, k, t ) = 1 - ppois(k-1, lambda*t)` numerically if we strip away the meaning. Therefore, we arrive at the equation above.

6.1 Construct an exact 95% CI for θ.

To construct 95% CI $(\theta_{lo}, \theta_{hi})$, we need to find: $P(Y \geq y|\theta_{lo}) = P(Y \leq y|\theta_{hi}) = 0.025$

$$\begin{aligned}
P(Y \geq y|\theta_{lo}) &= pgamma(\theta_{lo}, y, t) = 0.025 \\
&\Rightarrow \theta_{lo} = qgamma(0.025, y, t) \\
P(Y \leq y|\theta_{hi}) &= 1 - pgamma(\theta_{lo}, y+1, t) = 0.025 \\
&\Rightarrow pgamma(\theta_{lo}, y+1, \theta_u) = 0.975 \\
&\Rightarrow \theta_{hi} = qgamma(0.975, y+1, t)
\end{aligned}$$

Using our example data:

```
t = 200 #n
y = 122
theta_lo= qgamma(0.025, y, t)
theta_hi= qgamma(0.975, y+1, t)
sprintf("0.95 CI: ( %.2f , %.2f )", theta_lo, theta_hi)
```

```
[1] "0.95 CI: ( 0.51 , 0.73 )"
```