Categorical Data Analysis
WS 2014/15
-
Homework 2

by A good Student
January 14, 2015

1.)

Let $Y_i \overset{iid}{\sim} Poi(\mu)$, so $\mathbb{P}(Y = y) = \frac{\mu^y \cdot e^{-\mu}}{y!}$, and let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ denote the observed data vector.

(a)

The MLE of μ then follows from the Likelihood function

$$L(\mu; \mathbf{y}) = \prod_{i=1}^{n} \frac{\mu^{y_i} \cdot e^{-\mu}}{y_i!}$$

resp. the log-Likelihood function

$$l(\mu; \mathbf{y}) = \log L(\mu; \mathbf{y}) = \sum_{i=1}^{n} (y_i \log(\mu) - \mu - \log(y_i!))$$

by setting the first derivation to zero

$$\begin{aligned} \frac{d\, l(\mu; \mathbf{y})}{d\mu} = \sum_{i=1}^{n} \left(y_i \cdot \frac{1}{\mu} \right) - n &\overset{!}{=} 0 \\ \sum_{i=1}^{n} y_i - n\hat{\mu} &= 0 \\ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i &= \bar{y}. \end{aligned}$$

This is a maximum, because the second derivative is strict negativ:

$$\frac{d^2 l(\mu; \mathbf{y})}{d\mu^2} = -\sum_{i=1}^{n} y_i \cdot \frac{1}{\mu^2} < 0.$$

The Information matrix is given by

$$I(\mu) = \mathbb{E}\left[-\left(-\sum_{i=1}^{n} \frac{y_i}{\mu^2} \right) \right] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{y_i}{\mu^2} \right] = \frac{1}{\mu^2} \sum_{i=1}^{n} \underbrace{\mathbb{E}[y_i]}_{\mu} = \frac{n \cdot \mu}{\mu^2} = \frac{n}{\mu}.$$

So the standard error from $\hat{\mu}$ resp. its variance is

$$\begin{aligned} \mathrm{Var}(\hat{\mu}) &= I^{-1}(\mu) = \left(\frac{n}{\mu} \right)^{-1} = \frac{\mu}{n} \\ \widehat{\mathrm{Var}}(\hat{\mu}) &= I^{-1}(\hat{\mu}) = \frac{\hat{\mu}}{n} = \frac{\bar{y}}{n}. \end{aligned}$$

(b)

The Wald test statistic for testing $H_0 : \mu = \mu_0$ is

$$z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} = \frac{\hat{\mu} - \mu_0}{se(\hat{\mu})} = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\mu}/n}} = \frac{\bar{y} - \mu_0}{\sqrt{\bar{y}/n}} \overset{H_0}{\sim} N(0,1)$$

and the null hypothesis is rejected if $|z| > z_{1-\alpha/2}$.

(c)

$$\begin{aligned}
\text{Wald CI for } \mu &\Leftrightarrow \text{set of all } \mu_0\text{'s for which I fail to rejecet } H_0 : \mu = \mu_0 \\
&\Leftrightarrow \{\mu_0 : \text{p-value for testing } H_0 : \mu = \mu_0 \text{ is } > \alpha\} \\
&\Leftrightarrow \left\{\mu_0 : \text{ test statistic } |z| \leq z_{1-\alpha/2}\right\} \\
&\Leftrightarrow \left\{\mu_0 : \left|\frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\mu}/n}}\right| \leq z_{1-\alpha/2}\right\} \\
&\Leftrightarrow \left\{\mu_0 : -z_{1-\alpha/2} \leq \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\mu}/n}} \leq z_{1-\alpha/2}\right\} \\
&\Leftrightarrow \left\{\mu_0 : \hat{\mu} - \sqrt{\hat{\mu}/n} \cdot z_{1-\alpha/2} \leq \mu_0 \leq \hat{\mu} + \sqrt{\hat{\mu}/n} \cdot z_{1-\alpha/2}\right\} \\
&\Leftrightarrow \left\{\mu_0 : \bar{y} - \sqrt{\bar{y}/n} \cdot z_{1-\alpha/2} \leq \mu_0 \leq \bar{y} + \sqrt{\bar{y}/n} \cdot z_{1-\alpha/2}\right\}
\end{aligned}$$

This is the interval

$$\hat{\mu} \pm \sqrt{\hat{\mu}/n} \cdot z_{1-\alpha/2} = \bar{y} \pm \sqrt{\bar{y}/n} \cdot z_{1-\alpha/2}.$$

(d)

The score for $Y_i \overset{iid}{\sim} Poi(\mu)$ is

$$s(\theta) = s(\mu) = \frac{d\, l(\mu)}{d\mu} = \sum_{i=1}^{n} \frac{y_i}{\mu} - n = \frac{n}{\mu} \cdot (\bar{y} - \mu) = \frac{n}{\mu} \cdot (\hat{\mu} - \mu)$$

and so the score test statistic for testing $H_0 : \mu = \mu_0$ is

$$z = \frac{s(\mu_0)}{\sqrt{I(\mu_0)}} = \frac{\frac{n}{\mu_0} \cdot (\hat{\mu} - \mu_0)}{\sqrt{n/\mu_0}} = \frac{\hat{\mu} - \mu_0}{\sqrt{\mu_0/n}} = \frac{\bar{y} - \mu_0}{\sqrt{\mu_0/n}} \overset{H_0}{\sim} N(0,1).$$

The null hypothesis is rejected if $|z| > z_{1-\alpha/2}$.

(e)

$$\begin{aligned}
\text{Score CI for } &\Leftrightarrow \{\mu_0 : \text{p-value for testing } H_0 : \mu = \mu_0 \text{ is } > \alpha\} \\
&\Leftrightarrow \left\{\mu_0 : \text{ test statistic } |z| \leq z_{1-\alpha/2}\right\} \\
&\Leftrightarrow \left\{\mu_0 : \left|\frac{\hat{\mu} - \mu_0}{\sqrt{\mu_0/n}}\right| \leq z_{1-\alpha/2}\right\} \\
&\Leftrightarrow \left\{\mu_0 : -z_{1-\alpha/2} \leq \frac{\hat{\mu} - \mu_0}{\sqrt{\mu_0/n}} \leq z_{1-\alpha/2}\right\}
\end{aligned}$$

To get a closed formula it is necessary to solve the inequality for the upper and lower bound. By squaring the two equations this can be solved simultaneously for the upper and lower bound:

$$\frac{\hat{\mu} - \mu_0}{\sqrt{\mu_0/n}} = \pm z_{1-\alpha/2} =: K \quad \Leftrightarrow \quad \mu_0^2 - \mu_0(-2\hat{\mu} - K^2/n) - \hat{\mu}^2 = 0$$

So the two solutions for μ_0 are

$$\begin{aligned}
\mu_0 &= \hat{\mu} + \frac{K^2}{2n} \pm \sqrt{\frac{4\hat{\mu}^2 + 4\hat{\mu}K^2/n + K^4/n^2}{4} - \hat{\mu}^2} \\
&= \hat{\mu} + \frac{K^2}{2n} \pm \sqrt{\frac{\hat{\mu}K^2}{n} + \frac{K^4}{n^2}} \\
&= \hat{\mu} + \frac{K^2}{2n} \cdot \left(1 \pm \sqrt{1 + \frac{\hat{\mu} \cdot 4n}{K^2}}\right) \\
&= \hat{\mu} + \frac{z_{1-\alpha/2}^2}{2n} \cdot \left(1 \pm \sqrt{1 + \frac{\hat{\mu} \cdot 4n}{z_{1-\alpha/2}^2}}\right),
\end{aligned}$$

where obviously the "+"-solution leads to the upper bound and the "-"-solution to the lower bound. The score confidence interval is therefore

$$\left\{\mu_0 : \bar{y} + \frac{z_{1-\alpha/2}^2}{2n} \cdot \left(1 - \sqrt{1 + \frac{\bar{y} \cdot 4n}{z_{1-\alpha/2}^2}}\right) \leq \mu_0 \leq \bar{y} + \frac{z_{1-\alpha/2}^2}{2n} \cdot \left(1 + \sqrt{1 + \frac{\bar{y} \cdot 4n}{z_{1-\alpha/2}^2}}\right)\right\}.$$

(f)

The likelihood ratio test for the null hypothesis $H_0 : \mu = \mu_0$ is based on the likelihood ratio (LR), which for the Poisson case is

$$\lambda = \frac{\max\limits_{\mu \in H_0} L(\mu; \mathbf{y})}{\max\limits_{\mu} L(\mu; \mathbf{y})}$$

and the LR-statistic is therefore given by

$$\begin{aligned}
-2\log\lambda &= -2\left(\max_{H_0} l(\mu; \mathbf{y}) - \max_{H_0 \cup H_A} l(\mu; \mathbf{y})\right) \\
&= -2\left(\sum_{i=1}^{n} (y_i \log\mu_0 - \mu_0 - \log(y_i!)) - \sum_{i=1}^{n} (y_i \log\hat{\mu} - \hat{\mu} - \log(y_i!))\right) \\
&= -2\left(\sum_{i=1}^{n} y_i \cdot \log\left(\frac{\mu_0}{\hat{\mu}}\right) - n(\mu_0 - \hat{\mu})\right) \\
&= 2n \cdot \left(\bar{y}\log\left(\frac{\bar{y}}{\mu_0}\right) + (\mu_0 - \bar{y})\right) = G^2 \overset{H_0}{\sim} \chi_1^2.
\end{aligned}$$

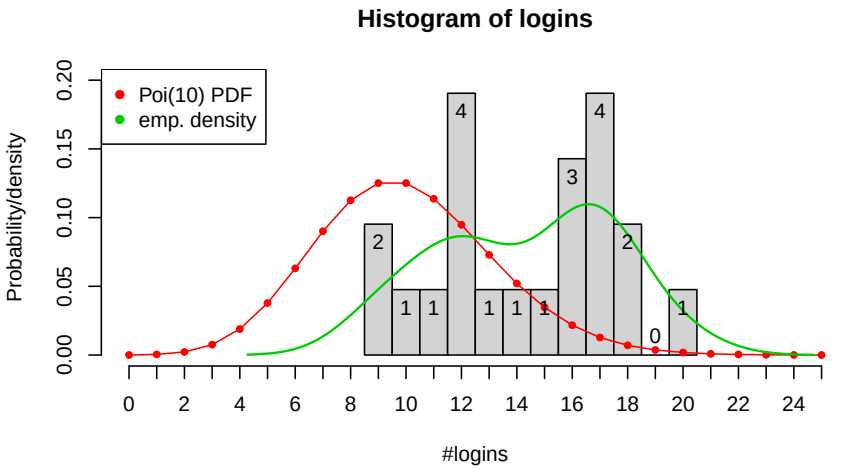
The null hypothesis is rejected if $G^2 > \chi_{1,1-\alpha}^2$.

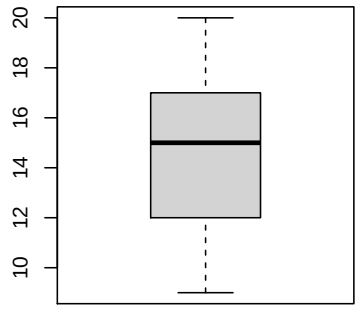
(g)

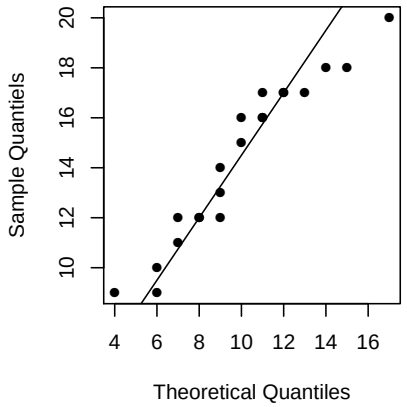
A short look at the data shows that the number of logins doen't seem to be $Poi(\lambda = 10)$ distributed.

```
> logins <- c(9,16,17,14,17,11,12,15,13,9,12,17,18,18,17,10,12,12,16,20,16)
> summary(logins)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   12.00   15.00   14.33   17.00   20.00
```








```
> layout(matrix(c(1,1,2,3),2,2,byrow=TRUE))
> hist(logins,freq=FALSE,xlab="#logins",ylab="Probability/density",col="lightgray",ylim=c(0,0.2),
+      breaks=(min(logins)-0.5):(max(logins)+0.5),xlim=xax <- c(0,25),xaxt="n")
> axis(1,xax[1]:xax[2],xax[1]:xax[2])
> lines(0:25,dpois(0:25,10),col=2,type="o",pch=20)
> lines(density(logins),col=3,lwd=1.5)
> hx <- hist(logins,plot=FALSE,breaks=(min(logins)-0.5):(max(logins)+0.5)) # histogram lables
> text(hx$mids, hx$density, label=c(hx$counts),pos=ifelse(hx$counts >= 1,1,3))
> legend("topleft",c("Poi(10) PDF","emp. density"),col=c(2,3),pch=16)
> #
> boxplot(logins,col="lightgray",main="Boxplot of logins")
> #
> plot(qpois((((1:(n <- length(logins)))-0.5)/n),10),sort(logins),pch=16, xlab="Theoretical Quantiles",
+      ylab="Sample Quantiels", main="Poi(10)-QQ-Plot of logins")
> abline(lm(quantile(logins,c(0.25,0.75))~c(qpois(0.25,10),qpois(0.75,10))))
```

Now, to confirm this impression, the hypothesis $H_0 : \mu = 10$ is tested against $H_A : \mu \neq 10$.

```
> mu.hat <- mean(logins); mu0 <- 10; n <- length(logins); alpha <- 0.05
> #
> wald <- c(z.wald<-(mu.hat-mu0)/sqrt(mu.hat/n),
+           qnorm(1-alpha/2),2*(1-pnorm(z.wald)))
> score <- c(z.score <- (mu.hat - mu0)/sqrt(mu0/n),
+            qnorm(1-alpha/2),2*(1-pnorm(z.score)))
> lr.stat <- c(G2<-2*n*(mu.hat*log(mu.hat/mu0)+(mu0-mu.hat)),
+              qchisq(1-alpha,1),1-pchisq(G2,1))
```

The following table shows the results of the three tests and all three clearly reject the null hypothesis:

	test statistic	critical value	p-value
Wald	5.25	1.96	1.56E-07
Score	6.28	1.96	3.39E-10
LR-stat	34.72	3.84	3.80E-09

The 95% confidence intervals obviously do not contain $\mu = 10$:

```
> # Wald CI
> alpha <- 0.05
> (ci.wald <- mu.hat + c(-1,1)*sqrt(mu.hat/n)*qnorm(1-alpha/2))

[1] 12.71409 15.95258

> c(mean(ci.wald), (ci.wald[2]-ci.wald[1])/2)

[1] 14.333333  1.619243

> #-------------------------------------------------------------------------------
> # score CI
> (ci.score <- mu.hat + qnorm(1-alpha/2)^2/(2*n)*
+                         (1+c(-1,1)*sqrt(1+(mu.hat*4*n)/qnorm(1-alpha/2)^2)))

[1] 12.80297 16.04662

> c(mean(ci.score), (ci.score[2]-ci.score[1])/2)

[1] 14.424797  1.621824

> #-------------------------------------------------------------------------------
> # LR stat function shifted by chi2
> G2.shift<-function(mu){2*n*(mu.hat*log(mu.hat/mu)+(mu-mu.hat))- qchisq(1-alpha,1)}
> # LR CI
> (ci.lr <- c(uniroot(G2.shift,c(0,mu.hat))$root,
+             uniroot(G2.shift,c(mu.hat,.Machine$integer.max))$root))

[1] 12.77451 16.01409

> c(mean(ci.lr), (ci.lr[2]-ci.lr[1])/2)

[1] 14.394301  1.619792
```

2.)

The likelihood ratio confidence interval for the binomial proportion π is calculated with the following function:

```
> LRint <- function(y,n,alpha){
+  G2.shift <- function(pi0){2*(y*ifelse(y==0,0,log(y/(n*pi0)))+
+                                 (n-y)*ifelse(y==n,0,log((n-y)/(n*(1-pi0))))) -
+                              qchisq(1-alpha,1)}
+  LB <- ifelse(y==0,0,uniroot(G2.shift,c(0,y/n))$root)
+  UB <- ifelse(y==n,1,uniroot(G2.shift,c(y/n,1))$root)
+  return(c(LB,UB))
+ }
```

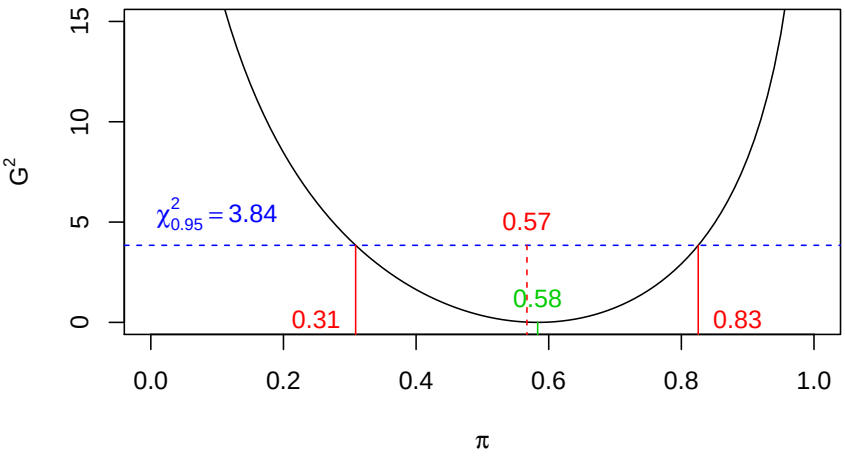
Because there is no analytic solution, the roots of the shifted G^2-function are calculated. The first root is to the left of the minimum/MLE in $\hat{\pi} = y/n$ and the second root is to the right. For $y = 7$ successes out of $n = 12$ trials the likelihood ratio confidence interval therefore is:

```
> y <- 7; n <- 12; alpha <- 0.05
> LRint(y,n,alpha)
```

```
[1] 0.3088470 0.8254632
```

```
> c(y/n,mean(LRint(y,n,alpha)))
```

```
[1] 0.5833333 0.5671551
```




```
> y <- 7; n <- 12; alpha <- 0.05
> G2 <- function(pi0){2*(y*ifelse(rep(y,length(pi0))==0,0,log(y/(n*pi0)))+
+                          (n-y)*ifelse(rep(y,length(pi0))==n,0,log((n-y)/(n*(1-pi0)))))}
> chi2 <- qchisq(1-alpha,1)
> lr.int <- LRint(y,n,alpha)
> mu.hat <- y/n
> plot(x <- seq(0,1,0.01),G2(x),type="l",ylim=c(0,15),xlab=expression(pi),ylab=expression(G^2))
> abline(h=chi2,col=4,lty=2)
> text(0.1,chi2,label=bquote(chi[0.95]^2==.(round(chi2,2))),pos=3,col=4)
> segments(x0=c(lr.int[1:2],mu.hat,mean(lr.int)),y0=-10,
+          y1=c(G2(c(lr.int[1:2],mu.hat)),chi2),col=c(2,2,3,2),lty=c(1,1,1,2))
> text(c(lr.int[1:2],mu.hat,mean(lr.int)),c(0,0,0,chi2),
+      label=round(c(lr.int[1:2],mu.hat,mean(lr.int)),2),pos=c(2,4,3,3),col=c(2,2,3,2))
```