ESC Spring 2021 Week2

One-parameter Model and Review of Mathematical Statistics

이재현, 이청파

연세대학교 통계데이터사이언스 석사과정
leejaehyun@yonsei.ac.kr
leechungpa@naver.com

March 10, 2021

Overview

Review

Binomial Model Review

$$Y_i|\theta \sim Ber(\theta)$$

$$\sum_{i=1}^{n} Y_i|\theta \sim Binom(n, \theta)$$

$$\theta \sim Beta(a, b)$$

$$\theta|data \ \sim Beta(a + \sum_{i=1}^{n} y_i, b + n - \sum_{i=1}^{n} y_i)$$

$$\tilde{Y}|\theta \sim Ber(\theta)$$

$$\tilde{Y}|data \sim Ber(\frac{a + \sum_{i=1}^{n} y_i}{a + b + n})$$

Poisson Distribution

Poisson Distribution Overview

Probability mass function

$$X|\theta \sim Poi(\theta)$$

$$p(x|\theta) = \frac{e^{-\theta}\theta^{x}}{x!} \qquad x = 0, 1, 2, ... \ for \ \ \theta > 0$$

$$\sum_{x=0}^{\inf} p(x|\theta) = 1$$

Mean and Variance

$$\mu = \theta, \quad \sigma^2 = \theta$$

Proof. Plug in theta

$$e^{\square} = 1 + \frac{1}{1!}\square + \frac{1}{2!}\square^2 + \frac{1}{3!}\square^3 + \cdots = \sum_{k=0}^{\frac{\square^k}{k!}}$$

$$\because when\ f(x) = e^x,\ \ f^{(k)}(0) = 1\ for\ k = 0, 1, 2, ...$$

Poisson Process

Let g(x,w) denote the the probability of x changes in each (time) interval of length w. Let the symbol o(h) represent any function such that

$$\lim_{h \to 0} \frac{o(h)}{h} = 0$$

For example,

$$h^2 = o(h)$$

$$o(h) + o(h) = o(h)$$

The Poisson postulates are the following(next slide).

1. For a positive constant lambda,

$$g(1,h) = \lambda h + o(h) \qquad h > 0$$

2. $$\sum_{x=2} g(x,h) = g(2,h) + g(3,h) + ... = o(h)$$

3. The numbers of changes in non-overlapping intervals are independent. Then, we can show by mathematical induction that

$$g(x,w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}$$

Hence the number of changes in X in an interval of length w has a Poisson distribution with parameter lambda*w

Bayesian Poisson Model

Bayesian Poisson Model

In Bayesian Poisson model,

$$Y_i|\theta \sim Poi(\theta)$$

$$p(y_i|\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!} \qquad y_i = 0, 1, 2, ... \quad for\ \theta > 0$$

$$\sum_{i=1}^{n} Y_i|\theta \sim Poi(n\theta) \qquad \because conditionally\ i.i.d.$$

What do Y and theta mean?
Suppose a large thick book that has N pages.
It could be intuitively interpreted that,

$$Y_i = \#typo\ on\ page\ i$$

$$\sum_{i=1}^{n} Y_i = \#typo\ on\ pages$$

$$\theta = expected\ \#typo\ per\ page = \frac{\sum_{i=1}^{N} Y_i}{N}$$

$$Y_i|\theta = \#typo\ on\ page\ i\ when\ \theta\ is\ known$$

$$\sum_{i=1}^{n} Y_i|\theta = \#typo\ on\ pages\ when\ \theta\ is\ known$$

Likelihood

$$Pr(Y_1 = y_n, ..., Y_n = y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n}(\frac{1}{y_i!}\theta^{y_i}e^{-\theta}) = \frac{1}{\prod_{i=1}^{n} y_i!}\theta^{\sum_{i=1}^{n} y_i}e^{-n\theta}$$

Compute the following

$$\frac{p(\theta_a|data)}{p(\theta_b|data)} =$$

Also from

$$\sum_{i=1}^{n} Y_i|\theta \sim Poi(n\theta) \qquad \because conditionally\, i.i.d.$$

Compare the results

$$Pr(\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} y_i|\theta) = \frac{1}{(\sum_{i=1}^{n} y_i)!}(n\theta)^{\sum_{i=1}^{n} y_i}e^{-n\theta}$$

Conjugate Prior

Prior Distribution

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} \quad for\ \theta, a, b > 0$$

$$\theta \sim Gamma(a, b)$$

Then by Bayes Theorem,

$$\theta | data \sim Gamma(a + \sum_{i=1}^{n} y_i, b + n)$$

Proof.

Interpretation

Continuing form slide 'Bayesian Poisson Model',
Prior Distribution

$$\theta \sim Gamma(a, b)$$

$$a = \#typo\ I\ encountered\ in\ the\ past$$

$$b = \#pages\ I\ read\ in\ the\ past$$

How would I expect typo per page?

Mean and Variance of Gamma(a,b)

$$\mu = a/b, \quad \sigma^2 = a/b^2$$

Interpretation

Now a research has been conducted.

$$n = \#pages\ researchers\ have\ read$$

$$y_i = \#typo\ on\ page\ i$$

$$\sum_{i=1}^{n} y_i = \#typo\ on\ pages$$

Overall,

$$b + n = \#pages\ I\ know$$

$$a + \sum_{i=1}^{n} y_i = \#typo\ I\ know$$

How would I expect *typo* per page?

$$\theta|data \sim Gamma(a + \sum_{i=1}^{n} y_i, b + n)$$

$$\mu = \frac{a + \sum_{i=1}^{n} y_i}{b + n} = \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum_{i=1}^{n} y_i}{n}$$

$$\sigma^2 = \frac{a + \sum_{i=1}^{n} y_i}{(b+n)^2}$$

Posterior Prediction

$$\tilde{Y}|\theta \sim Poi(\theta)$$

$$\tilde{Y}|data \sim ??$$

Deriving the distribution

$$p(\tilde{y}|data) = \int_0 p(\tilde{y}, \theta|data)\, d(\theta) = \int_0 p(\tilde{y}|\theta, data)p(\theta|data)\, d(\theta)$$

$$= \int_0 p(\tilde{y}|\theta)p(\theta|data)\, d(\theta) = \ \dots \ = \frac{\Gamma(a + \sum_{i=1}^n y_i + \tilde{y})}{\Gamma(\tilde{y}+1)\Gamma(a + \sum_{i=1}^n y_i)}(1-p)^{a+\sum_{i=1}^n y_i} p^{\tilde{y}}$$

$$p = \frac{1}{b+n+1}$$

$$\therefore \tilde{Y}|data \sim NB(a + \sum_{i=1}^n y_i,\ p), \quad p = \frac{1}{b+n+1}$$

Interpretation

Suppose the following.

$$X \sim NB(r, p)$$

$$p(X) = \binom{r-1+x}{x}(1-p)^r p^x$$

what does X mean?

$$X = \#Successes\, before\, r\, th\, Failure$$

$$p = Success\, rate\, of\, a\, trial$$

Similarly,

$$\tilde{Y}|data \sim NB(a + \sum_{i=1}^{n} y_i,\ p), \quad p = \frac{1}{b+n+1}$$

$$\tilde{Y}|data = \#Successes\ \ before\ (a + \sum_{i=1}^{n} y_i)th\ Failure$$

$$Success = a\, typo\, belongs\, to\, the\, new\, page$$

$$Failure = a\, typo\, belongs\, to\, page\, 1\, to\, b+n$$

Mean and Variance

$$\mu = \frac{rp}{1-p} = \frac{a + \sum_{i=1}^{n} y_i}{b+n} = E[\theta|data]$$

$$\sigma^2 = \frac{rp}{(1-p)^2} = \frac{a + \sum_{i=1}^{n} y_i}{b+n} \frac{b+n+1}{b+n} = E[\theta|data](1 + \frac{1}{b+n+1})$$

Overdispersion in Poisson Model

Overdispersion : Possibility of variation beyond that of the assumed sampling distribution.

"The data are too dispersed than expected!"

Under the Poisson model,

$$Y_i|\theta \sim Poi(\theta)$$

$$Var(Y_i|\theta) = \theta$$

However, an overdispersion problem exists when data of y's imply greater variance than theta.
To handle overdispersion, the following alternative methods are suggested.

1. Use Negative Binomial Distribution instead, a two-parameter model that can fit mean and variance separately.
2. Use Hierarchical Normal Model instead.

Bayesian Normal Model

Normal Model with known variance

Likelihood of sampling distribution is

$$p(y|\theta;\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(y-\theta)^2}{2\sigma^2})$$

And prior for a normal mean is

$$p(\theta) \sim N(\mu_0, \tau_0^2) \qquad \Longrightarrow \qquad p(\theta) \propto \exp(-\frac{(\theta-\mu_0)^2}{2\tau_0^2})$$

Thus, posterior distribution is

$$\begin{aligned} p(\theta|y) &\propto \exp(-\frac{(y-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\tau_0^2}) \\ &\propto \exp(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}) \end{aligned}$$

$$p(\theta|y) \propto \exp(-\frac{(\theta-\mu_1)^2}{2\tau_1^2})$$

By reparameterize

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \qquad \text{and} \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Interpretation of posterior variance

The posterior variance is expressed as

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- a sum of the prior precision and the precision of the obsered y

Interpretation of posterior mean

The posterior mean is expressed as

$$\begin{aligned}
\mu_1 &= \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \\
&= \mu_0 + (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2} \\
&= y - (y - \mu_0)\frac{\sigma^2}{\sigma^2 + \tau_0^2}
\end{aligned}$$

- a weighted average of the prior mean and the observe value with weights proportional to the precisions
- the prior mean adjusted toward the observed y
- the data 'shrunk' toward the prior mean

Each formulation represents the posterior mean as a compromise between the prior mean and the observed value.

Normal model with multiple parameter

Likelihood of sampling distribution and prior are

$$\prod_{i=1}^{n} p(y_i|\theta;\sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n}\prod_{i=1}^{n}\exp(-\frac{(y_i-\theta)^2}{2\sigma^2})$$

$$p(\theta) \sim N(\mu_0,\tau_0^2) \qquad \Longrightarrow \quad p(\theta) \propto \exp(-\frac{(\theta-\mu_0)^2}{2\tau_0^2})$$

Thus, posterior distribution is

$$p(\theta|y) \propto \exp(-\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\tau_0^2}) \qquad \propto \exp(-\frac{(\theta-\mu_n)^2}{2\tau_n^2})$$

By reparameterize

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \qquad \text{and} \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

Exponential Families

Exponential Families

$$f(x;\theta) = \begin{cases} \exp[p(\theta)K(x) + s(x) + q(\theta)] & x \in S \\ 0 \quad o.w \end{cases}$$

1. S does not depend on θ
2. $p(\theta)$ is a nontrivial continuous function of $\theta \in \Omega$
3. If X is continuous, $K'(x) \neq 0$ and $s(x)$ is continuous function.
   If X is discrete, $K(x)$ is nontrivial function.

Or, same as,

$$f(y|\phi) = \begin{cases} h(y)c(\phi)\exp[\phi K(y)] & x \in S \\ 0 \quad o.w \end{cases}$$

1. S does not depend on ϕ
2. ϕ is a nontrivial continuous function of $\theta \in \Omega$
3. If Y is continuous, $K'(y) \neq 0$ and $h(y)$ is continuous function.
   If Y is discrete, $K(y)$ is nontrivial function.

Sufficient statistic for θ

$X_1, X_2, \cdots, X_n \sim iid\ f(x;\theta)$ is a regular exponential class
Then,
$Y_1 = \sum_{i=1}^{n} K(X_i)$ is complete sufficient statistic for θ

What is sufficient?
$Y_1 = u_1(X_1, \cdots, X_n)$ is sufficient statistic for θ
if and only if
$p(X_1, X_2, \cdots, X_n | Y_1 = y_1)$ does not depend on θ

Example of Exponential Families

Binomial distribution $B(n, \theta)$ is

$$
\begin{aligned}
p(y_1, y_2, \cdots, y_n|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\
&= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\
&= \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{\sum_{i=1}^{n}(1-y_i)} \\
&= \theta^{y}(1-\theta)^{n-y} \qquad \text{where } y = \sum_{i=1}^{n} y_i \\
&= \left(\frac{\theta}{1-\theta}\right)^{y}(1-\theta)^{n} \\
&= e^{\phi y}(1+e^{\phi})^{-n} \qquad \text{where } \phi = \log\frac{\theta}{1-\theta}
\end{aligned}
$$

$$p(y_1, y_2, \cdots, y_n|\theta) = e^{\phi y}(1+e^{\phi})^{-n}$$

Compare the above result with exponential family

$$f(y|\phi) = e^{\phi K(y)}h(y)c(\phi) \qquad \text{for } x \in S$$

Then,

$$\phi = \log \frac{\theta}{1-\theta}$$
$$K(y) = \sum_{i=1}^{n} y_i$$
$$h(y) = 1$$
$$c(\phi) = (1+e^{\phi})^{-n}$$

which is

1. $S = \{1, 2, \cdots, n\}$ does not depend on ϕ
2. ϕ is a nontrivial continuous function of $\phi \in \Omega$
3. If Y is discrete, $K(y)$ is nontrivial function.

Conjugacy

Exponential Families (revisit)

$$f(y|\phi) = \begin{cases} h(y)c(\phi)e^{\phi K(y)} & x \in S \\ 0 & o.w \end{cases}$$

If

1. S does not depend on ϕ
2. ϕ is a nontrivial continuous function of $\theta \in \Omega$
3. If Y is continuous, $K'(y) \neq 0$ and $h(y)$ is continuous function.
   If Y is discrete, $K(y)$ is nontrivial function.

$$f(y|\phi) = h(y)c(\phi)e^{\phi K(y)} \qquad x \in S$$

What is conjugate?
If the posterior distributions $p(\theta|x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions. (by wikipedia)

Conjugacy with Exponential Families

- Likelihood :

$$
\begin{aligned}
f(y_1, y_2, \cdots, y_n|\phi) &= \prod_{i=1}^{n} h(y_i)c(\phi)e^{\phi K(y_i)} \\
&\propto c(\phi)^n e^{\phi \sum_{i=1}^{n} K(y_i)}
\end{aligned}
$$

- Prior :

$$
\begin{aligned}
p(\phi) &= k(n_0, t_0)c(\phi)^{n_0} e^{n_0 t_0 \phi} \\
&\propto c(\phi)^{n_0} e^{n_0 t_0 \phi}
\end{aligned}
$$

- Posterior :

$$
\begin{aligned}
p(\phi|y) &\propto p(\phi)f(y|\phi) \\
&\propto c(\phi)^{n_0} e^{n_0 t_0 \phi} \quad c(\phi)^n e^{\phi \sum_{i=1}^{n} K(y_i)}
\end{aligned}
$$

$$
\begin{aligned}
p(\phi|y) &\propto p(\phi) f(y|\phi) \\
&\propto c(\phi)^{n_0} e^{n_0 t_0 \phi} \quad c(\phi)^n e^{\phi \sum_{i=1}^n K(y_i)} \\
&\propto c(\phi)^{n_0+n} \exp\left[n_0 t_0 \phi + \phi \sum_{i=1}^n K(y_i) \right] \\
&\propto c(\phi)^{n_0+n} \exp\left[\phi \left(n_0 t_0 + n \frac{\sum_{i=1}^n K(y_i)}{n} \right) \right]
\end{aligned}
$$

Which has the same class with prior

$$
p(\phi) \propto c(\phi)^{n_0} e^{n_0 t_0 \phi}
$$

- n_0 can be interpreted as a **prior sample size**
- t_0 as a **prior guess** of $K(Y)$.

Homework

Homework

1 FCB Exercises 3.3
Tumor counts: A cancer laboratory is estimating the rate of tumorigenesis in two strains of mice, A and B. They have tumor count data for 10 mice in strain A and 13 mice in strain B. Type A mice have been well studied, and information from other laboratories suggests that type A mice have tumor counts that are approximately Poisson-distributed with a mean of 12. Tumor count rates for type B mice are unknown, but type B mice are related to type A mice. (이하 생략)

cf. week2_lab.ipynb 참고 : Birth rates for Poisson Model (FCB P.48 $\sim$ 50)

2 Data가 binomial distribution일때, Likelihood를 Exponential Families 형태로 변환해 보기. 또한 왜 Beta distribution이 Conjugacy인지 생각해 보기.

cf. Appendix 참고

3 Relationship between Poisson distribution and Negative Binomial Distribution

$$X \sim NB(r,p) \qquad \text{where } p(X=x) = \binom{r-1+x}{x}(1-p)^r p^x$$

$$\text{Let mean } \frac{pr}{1-p} = \lambda \qquad \rightarrow \qquad p = \frac{\lambda}{r+\lambda}$$

3.1 Prove the following.

$$Poi(\lambda) = \lim_{r\rightarrow} NB(r, \frac{\lambda}{r+\lambda})$$

3.2 Compare the variance of each distribution. Show that the Negative Binomial distribution is always overdispersed.

3.3 Likewise, prove the following.

$$Y \sim Binom(n,p) \qquad \text{where } p(y) = \binom{n}{x} p^y p^{n-y}$$

$$\text{Let mean } np = \lambda \qquad \rightarrow \qquad p = \frac{\lambda}{n}$$

$$Poi(\lambda) = \lim_{n\rightarrow\infty} Binom(n, \frac{\lambda}{n})$$

Appendix. Binomial Conjugate

- Likelihood :

$$p(y_1, y_2, \cdots, y_n|\theta) = e^{\phi y}(1+e^{\phi})^{-n} \qquad \text{where } \phi = \log\frac{\theta}{1-\theta} \quad y = \sum_{i=1}^{n} y_i$$

- Prior : $p(\theta) \sim \text{Beta}(n_0 t_0,\ n_0(1-t_0))$

$$\begin{aligned} p(\theta) &\propto \theta^{n_0 t_0 - 1}(1-\theta)^{n_0(1-t_0)-1} \\ &\propto e^{\phi(n_0 t_0 - 1)}(1+e^{\phi})^{2-n_0} \end{aligned}$$

- Posterior : $p(\theta|y) \sim \text{Beta}(n_0 t_0 + y,\ n_0(1-t_0) + (n-y))$

$$\begin{aligned} p(\phi|y) &\propto e^{\phi(n_0 t_0 - 1)}(1+e^{\phi})^{2-n_0}\ e^{\phi y}(1+e^{\phi})^{-n} \\ &\propto e^{\phi(n_0 t_0 - 1 + y)}(1+e^{\phi})^{2-n_0-n} \\ &\propto \theta^{(n_0 t_0 + y)-1}(1-\theta)^{n_0(1-t_0)+(n-y)-1} \end{aligned}$$