Distributions

Wei-Chang Lee, Chi-Ning Chou

November 22, 2015

Contents

Chapter 1

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 11	
November 22, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

1.1 Discrete Distributions

Definition 1 (Discrete Uniform Distribution) *Suppose X follows discrete uniform distribution then it has density*

$$f_X(|N) = \frac{1}{N}\mathbb{1}_{\{1,2,3\ldots N\}}(x)$$

where N is an integer, with notation X∼Discrete Uniform(N).

Property 1 *Given N, X follows discrete uniform distribution then,*

1. $E[X|N] = \sum_{i=1}^{N} P(X=i)i = \frac{N+1}{2}$
2. $Var(X|N) = E[X^2|N] - E[X|N]^2 = \frac{N^2-1}{12}$

Intuition (Usage in statistics)

How can we test the two given data group X,Y follows the same distribution?
$X_1, X_2...X_n \overset{iid}{\sim} F_1(x)$ and $Y_1, Y_2...Y_m \overset{iid}{\sim} F_1(x)$ want to test $H_0 : F_1(x) = F_2(x)\ \forall x$
Kolmogrov statistics: Using empirical distribution of $F_1(x), F_2(x)$

$$\hat{F_1}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$$

$$\hat{F_2}(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(Y_i \leq x)$$

We have Kolmogrov statistics:

$$\sup_x |\hat{F_1}(x) - \hat{F_2}(x)|$$

which can not be too big if X and Y following same distributions.
Rank statistics(Wilcoxon test): Instead of using true value as the predictor. We use the order i.e. rank of the data in the group. We combine X and Y and sort them to give rank

$$W = \frac{1}{n}\sum_{i=1}^{n} Rank(X_i)$$

To prevent the issue that extreme values dominated the statistics. And $X \sim Y$ if W is not too big or too small.

Definition 2 (Bernoulli Distribution) *X follows Bernoulli distribution then it has density*

$$f_X(x|p) = p^x(1-p)^{1-x}\mathbb{1}_{\{0,1\}}(x)$$

where $0 \leq p \leq 1$ *denoting as X∼Bernoulli(p).*

Property 2 *Given p, X follows binomial distribution then,*

1. $E[X^m|p] = E[X|p] = p$
2. $Var(X|p) = p - p^2 = p(1-p)$
3. $F_X(x) = P(X \leq x) = E[I(X \leq x)] = E[N(x)]$

Definition 3 (Binomial Distribution) $X_1, X_2...X_n$ *i.i.d follows Bernoulli(p), let* $X = \sum_{i=1}^{n} X_n$, *X follows binomial distribution having density*

$$f_X(x) = \binom{n}{x} p^x(1-p)^x \mathbb{1}_{\{0,1,2,3...n\}}(x)$$

Intuition (Independent)

$X_1, X_2...X_n$ are independent iff

$$P(X_1 = x_1, X_2 = x_2...X_n = x_n) = \prod_{i=1}^{n} f(x_i|p)$$

The mutually independent property automatically satisfied since we can think of $\Omega = \Omega_1 \times \Omega_2... \times \Omega_n$ where $\Omega_i = \{0, 1\}$ for the ith Bernoulli trial. And $\bigcup(A_i \in \Omega_1)(\Omega_2... \times \Omega_n)$ augmented Ω.

Remark 1 In reality, $X_1, X_2...X_n$ are not i.i.d.. Since we sometimes sample population with common factors. They may affect each other, within positive or negative relation.

1. Over-dispersion binomial distribution: There are positive correlation among populations. That is, if the event happens on one member, then other members will have higher tendency to success.

2. Under-dispersion binomial distribution: There are negative correlation among populations.

Formally, if the variance of a random variable look like: $var[X] = \phi p(1-p)$. If $\phi > 1$, we say X is a over-dispersion binomial and on the contrary if $\phi < 1$, then we say X is an under-dispersion binomial. Note that, although we call them "binomial", they are definitely not a binomial random variable!

Chapter 2

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 12	
November 22, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

Today we talk about the discrete distributions related to Bernoulli distribution.

2.1 Big picture

Bernoulli distribution is a single event with two possible outcome: yes/no. The probability is p for the yes result and $(1-p)$ for the no. Intuitively, we can view a Bernoulli distribution as an indicator to identify whether an event has happened.

What if we want to consider more than one event?

Imagine the following scenario, there is a large population containing N elements and M of them are label as *type-I* and the rest $N-M$ are labeled as *type-II*. Now, as a statistician, we want to draw some inference about the population, but we have only limited access to the population, say k samples. What can we know from the experiment?

Basically, we can categorize the above scenario with two different properties:

- Draw *with replacement* or *without replacement*.
- Draw *fix number of samples*, or keep drawing *until a certain event happens*?

With these two factors, we can extend Bernoulli distribution into the following three discrete distribution:

	Replacement	Draw	Goal
Hypergeometric	Without	k times	Number of yes
Binomial	With	k times	Number of yes
Negative Binomial	With	Wait until r yes	Number of no

2.2 Hypergeometric distribution

2.2.1 Definition

Hypergeometric distribution describes the probability of the number of *yes* result under k samples **without replacement**. The density function consists of three parameters: (N, M, k) and the pdf is

$$f(x|N,M,k) = \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} \mathbf{1}_{(\max(0,k-(N-M)),min(M,k))}(x)$$

Here, we discuss the meaning of each term:

- $\binom{N}{k}$ in the denominator is the number of possible k samples outcome.
- $\binom{M}{x}$ in the numerator is the number of possible combinations of k yes instances.
- $\binom{N-M}{M-x}$ in the numerator is the number of possible combinations of $x-k$ no instances.

2.2.2 Basic properties

Here, we list the mean and variance of hypergeometric distribution and discuss the idea of reparametrize techniques.

- $\mathbb{E}[X|N,M,k] = k\frac{M}{N}$
- $var[X|N,M,k] = k\frac{M}{N}\frac{N-M}{N}\frac{N-k}{N-1}$

In the following, we are going to prove the above results via reparametrize techniques and factorial moment. **Proof**:

- The mean of $X \sim$ *Hypergeometric*(N,M,k)

$$\begin{aligned}
\mathbb{E}[X|N,M,k] &= \sum_{x=\max(0,k-(N-M))}^{\min(M,k)} x\frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} = \sum_{x=\max(0,k-(N-M))}^{\min(M,k)} M\frac{\binom{M-1}{x-1}\binom{N-M}{k-x}}{\binom{N}{k}} \\
&= \sum_{x=\max(0,k-(N-M))}^{\min(M,k)} M\frac{\binom{M-1}{x-1}\binom{N-(M-1)}{(k-1)-(x-1)}}{\binom{N-1}{k-1}\times\frac{N}{k}} \\
&= k\frac{M}{N}\sum_{x}\frac{\binom{M-1}{x-1}\binom{N-(M-1)}{(k-1)-(x-1)}}{\binom{N-1}{k-1}} = k\frac{M}{N}
\end{aligned}$$

- The variance of $X \sim$ *Hypergeometric*(N,M,k)

$$var[X|N,M,k] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2$$

As we know $\mathbb{E}[X]$, it suffices to find $\mathbb{E}[X(X-1)]$. The trick that computing the expectation of $X(X-1)$ instead of that of X^2 is called *factorial moment*, which is computation-friendly when having lots of binomial terms. As a result,

$$\mathbb{E}[X(X-1)]$$

2.2.3 Approximation

Hypergeometric distribution can be approximated by binomial distribution as $N, M \to \infty$ and $\frac{M}{N} \to p$. Intuitively, the population size grows to infinite and thus each draw has negligible influence to the population, which makes the whole process to be identical. As the drawing process is uniform, the process becomes independent. Thus, we can view it as a binomial. The derivation is simple:

$$\begin{aligned} P(x|N,M,k) &= \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} \approx \frac{\frac{M^x}{x!}\frac{(N-M)^{k-x}}{(k-x)!}}{\frac{N^k}{k!}} \\ &= \frac{k!}{x!(k-x)!}\frac{M^x(N-M)^{k-x}}{N^k} \\ (\frac{M}{N} = p) &= \binom{k}{x}p^x(1-p)^{k-x} \end{aligned}$$

2.3 Binomial distribution

2.3.1 Definition

The binomial distribution describe the probability of the number of *yes* results with a fixed number of i.i.d. drawing **with replacement**. The density function consists of two parameters: (N, p) and the pdf is

$$f(x|N,p) = \binom{N}{x}p^x(1-p)^{N-x}\mathbf{1}_{0,1,\ldots,N}(x)$$

Note that, the difference between the definitions of hypergeometric distribution and binomial distribution is not only with/without replacement, the underlying mechanism of binomial distribution is not a fix finite sample space as hypergeometric. For example, the number of drawing can be unbounded, or the *yes* probability should not be necessarily a rational number.

2.3.2 Basic properties

Suppose $X \sim Binomial(N, p)$, the following is the mean and variance of X:

- $\mathbb{E}[X|N,p] = Np$
- $var[X|N,p] = Np(1-p)$

2.4 Negative binomial distribution

2.4.1 Definition

The negative binomial distribution describes the probability of the number of *no* instances before certain number of *yes* results in a sequence of i.i.d. drawing. Formally speaking, for a negative binomial distribution with parameters: (p, r) where p is the probability of *yes* and r is the number of *yes* instances we are waiting for, the pdf is

$$f(x|p,r) = \binom{x+r-1}{r-1}p^r(1-p)^x\mathbf{1}_{0,1,\ldots}(x)$$

2.4.2 Basic properties

Suppose $X \sim$ *Negative Binomial*(p, r)

- $\mathbb{E}[X|p, r] = \frac{pr}{1-p}$
- $var[X|p, r] = \frac{pr}{(1-p)^2}$
- When $r = 1$, it is called *geometric* distribution.
- The drawing process is memoryless. For example, the distribution of number of *no* will remain the same as we conditioned on the number of *no* instances before.
- As we let $p \to 1$, the *yes* result will tend to happen and some how the distribution will converge to Poisson distribution similarly to binomial distribution. (Detail discussion next time)

Chapter 3

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 13	
November 22, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

3.1 Poisson distribution

Poisson random variable is defined with a parameter λ denoting the rate or intensity of a counting process. As Poisson distribution is **memoryless**, these two notions don't conflict. We define the probability density function of Poisson(λ) as follow:

$$f_X(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}\mathbf{1}_{\{0,1,\ldots\}}(x)$$

The following is the basic properties of Poisson distribution:

- $\mathbb{E}[x|\lambda] = \lambda$
- $var[x|\lambda] = \lambda$
- $M_X(t) = e^{-\lambda(1-e^t)}$

Now, let's consider a theorem that connects the intuition of Poisson process with Poisson distribution.

Theorem 1 (Poisson process) *Let N_t be a nondecreasing integer-valued random variable satisfying*

1. $N_0 = 0$
2. $\forall 0 < t_1 < t_2 < t_3 < t_4$, $N_{t_2} - Nt_1 \sim N_{t_2 - t_1}$ (**identical**). $N_{t_2} - N_{t_1}$ *is independent to* $N_{t_4} - N_{t_3}$
3. $\lim_{n\to\infty} \frac{Pr[N_0=1]}{h} = \lambda$ *and* $\lim_{n\to\infty} \frac{Pr[N_0\geq 2]}{h} = 0$

Then, $Pr[N_t = k] = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$

Proof: First, we consider the case where $k = 0$. Then we use induction to prove the result for all k. In the following proof, denote $P_n(t) = Pr[N_t = n]$

1. Suppose $n = 0$, we have $\forall t > 0$

$$\begin{aligned} P_0(t+h) &= Pr[N_t = 0 \text{ and } N_{t+h} - N_t = 0] \\ (\because \text{independent and stationary}) &= P_0(t)P_0(h) \\ &= P_0(t)(1 - \lambda h + o(h)) \end{aligned}$$

Subtract $P_(t)$ on both side and divide by h, let $h \to 0$ we have

$$\begin{aligned} P_0'(t) &= \lim_{h\to 0} \frac{P_0(t+h) - P_0(t)}{h} \\ &= \lim_{h\to 0} -\lambda P_0(h) + \frac{o(h)}{h} \\ &= -\lambda P_0(t) \end{aligned}$$

This is equivalent as solving $\frac{d}{dt} \ln P_0(t) = -\lambda$. With the boundary condition $P_0(0) = 0$, we have

$$P_0(t) = e^{-\lambda t}$$

2. Now, consider $n \geq 1$. We have

$$\begin{aligned} P_n(t+h) &= Pr[N_t = n-1 \text{ and } N_{t+h} - N_t = 1] + Pr[N_t = n \text{ and } N_{t+h} - N_t = 0] \\ &\quad + Pr[N_{t+h} - N_t \geq 2] \\ &= P_{n-1}(t)(\lambda h + o(h)) + P_n(t)(1 - \lambda h + o(h)) + o(h) \end{aligned}$$

Subtract $P_n(t)$ on both side and divide by h, let $h \to 0$ we have,

$$\begin{aligned} P_n'(t) &= \lim_{h\to 0} \frac{P_n(t+h) - P_n(t)}{h} \\ &= \lim_{h\to 0} \lambda P_{n-1}(t) - \lambda P_n(t) + \frac{o(h)}{h} \\ &= \lambda P_{n-1}(t) - \lambda P_n(t) \end{aligned}$$

Consider $n = 1$, we have $P_1'(t) = \lambda e^{-\lambda t} - \lambda P_1(t)$, which is equivalent as solving $\frac{d}{dt}(e^{\lambda t} P_1(t)) = \lambda$. With boundary condition $P_1(0) = 0$, we have

$$P_1(t) = \lambda t e^{-\lambda t}$$

With induction hypothesis $P_{n-1}(t) = \frac{(\lambda t)^{n-1} e^{-\lambda t}}{(n-1)!}$, the problem is equivalent as solving $\frac{d}{dt} e^{\lambda t} P_n(t) = \lambda \frac{(\lambda t)^{n-1}}{(n-1)!}$. With boundary condition $P_n(0) = 0$, we have

$$P_n(t) = \frac{(\lambda t)^n e^{\lambda t}}{n!}$$

■

3.1.1 Counting process and Stopping time

In fact, counting process and stopping time are the two side of a coin. The following shows how to interchange from one to another.

Stopping time T → Counting process $\{N(t), t \geq 0\}$
For a given stopping T, we can define a corresponding zero-one counting process: $N_T(t) := \mathbf{1}_{\{T<t\}}$

Counting process $\{N(t), t \geq 0\}$ → Stopping time T
For a counting process $\{N(t), t \geq 0\}$, we can define a stopping time T as $Pr[T > t] = Pr[N(t) = 0]$ so for a Poisson counting process:

$$\begin{aligned} 1 - F_T(t) &= e^{-\lambda t} \\ f_T(t) &= \lambda e^{-\lambda t} \mathbf{1}_{\{0,1,2,\ldots\}}(t) \end{aligned}$$

for Gamma distribution:

$$\begin{aligned} T^* &= \sum_{j=1}^{m} T_j^* \\ f_{T^*}(t|m,\lambda) &= \frac{t^{m-1}\lambda^m e^{-\lambda t}}{\tau(m)} \mathbb{1}_{\{0,\infty\}}(t) \end{aligned}$$

3.2 Relationship between distribution

Example 1 Let X∼Poisson(λ) and Y∼Binomial(n,p), then $f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ and we can expand

$$f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} = \frac{n-y+1}{y} \frac{p}{1-p} \binom{n}{y-1} p^{y-1}(1-p)^{n-y+1} = \frac{np - yp + p}{y - yp} f_Y(y-1|n,p)$$

so when $p \to 0, n \to \infty, np \to \lambda$, we have $Y \overset{d}{=} X$

$$\begin{aligned} f_Y &= \frac{\lambda}{y} f_Y(y-1|n,p) \\ &= \prod_{i=1}^{y} \frac{\lambda}{i} f_Y(0|n,p) \\ &= \frac{\lambda^y}{y!}(1 - \frac{np}{n})^n \\ &= \frac{\lambda^y e^{-\lambda}}{y!} \end{aligned}$$

Example 2 Y∼Negative Binomial(r,p), then $f_Y(y) = \binom{y+r-1}{r-1} p^r (1-p)^y$

when $r \to \infty, p \to 1, r(1-p) \to \lambda$, we have $Y \stackrel{d}{=} Poisson(\lambda)$

$$\begin{aligned} M_Y(t) = E[e^{tY}] &= \sum_{y=0}^{\infty} \binom{y+r-1}{r-1} p^r (1-p)^y e^{ty} \\ &= \sum_{y=0}^{\infty} \binom{y+r-1}{r-1} p^r ((1-p)e^t)^y \\ &= (\frac{p}{1-(1-p)e^t})^r \\ &= (1 + \frac{1}{r} \frac{r(1-p)(e^t-1)}{1-(1-p)e^t})^r \\ &= e^{\lambda(e^t-1)} \end{aligned}$$

Chapter 4

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 14	
November 22, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

4.1 Continuous distribution

4.1.1 Uniform distribution

In continuous regime, we define a uniform random variable on a close interval $[a, b]$, where $a < b$, and denote it as Uni(a, b). If $X \sim$Uni(a, b), then

- $f_X(x|a, b) = \frac{1}{b-a}$
- $\mathbb{E}[X|a, b] = \frac{a+b}{2}$
- $Var[X|a, b] = \frac{(b-a)^2}{12}$

4.1.2 Exponential family

Here, we define three highly related continuous random variables: exponential, Weibull, and gamma. We first write down their distribution respectively, then introduce their relationship and properties.

Exponential: Exponential random variable captures a single interleaving time of a Poisson process with frequency $1/\beta$. If $X \sim$exponential(β)

$$f_X(x|\beta) = \frac{1}{\beta} e^{-x/\beta} \mathbf{1}_{(0,\infty)}(x)$$

Weibull: If $Y = X^{1/\gamma}$, where $X \sim$exponential(β) and $\gamma > 0$, we say Y has a Weibull(β, γ) distribution.

$$f_Y(y|\beta, \gamma) = \frac{\gamma}{\beta} y^{\gamma-1} e^{-y^\gamma/\beta} \mathbf{1}_{(0,\infty)}(y)$$

In other words, Weibull random variable is a power transformed version of exponential random variable. And as $\gamma = 1$, the Weibull degenerates to exponential.
Gamma: Intuitively, gamma distribution captures the total interleaving time up to more than $\alpha + 1$ appearances. We use two parameters α, β to define a gamma random variable and denote it as Gamma(α, β). If $X \sim$Gamma(α, β), then

$$f_X(x|\alpha, \beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}\mathbf{1}_{(0,\infty)}$$

As we mentioned earlier, these three distributions are highly related to Poisson distribution in the sense that they describe the waiting time of a counting process given the number of desired observations while Poisson distribution captures the number of appearances given the amount of observing time. The two aspects are just two side of a coin, and we can use the following equation to relate them all together: Let $\{N(t)\}$ be a counting process and T be the corresponding waiting time for a single event to happen. We have

$$\{T > t\} = \{N(t) = 0\}$$

If we write down the probability of each side and do some computation, we can derive a relationship between exponential distribution and Poisson distribution.

4.1.3 Exponential family to Poisson

Exponential: Let $F_X(x)$ be the distribution function of a random variable X$\sim$exponential(β) and t_1 be the interleaving time of next arrival. Exponential distribution captures the probability of interleaving time less than or equal to t. We can interpreted it as at least one arrival happened up to time t. So it follows:

$$F_X(x) = P(t_1 \leq x) = 1 - P(t_1 > x) = 1 - P(N(t) = 0) = \frac{(\lambda x)^0 e^{-\lambda x}}{0!}$$

We have

$$f_X(x|\beta) = \frac{1}{\beta}e^{-x/\beta}\mathbf{1}_{(0,\infty)}(x)$$

Gamma:Let $F_X(x)$ be the distribution function of a random variable X$\sim$gamma(α, β). The distribution captures the probability that there are at least α arrival up to time t. We can interpreted it as the probability a poisson process with intensity $\frac{1}{\beta}$ up to time t with at least α arrivals.
If we have $Y \sim Poisson(x/\beta)$ then,

$$P(X \leq x|\alpha, \beta) = P(Y \geq \alpha|\frac{x}{\beta})$$

Property 3 *let X$\sim$Gamma(α, β)*

1. *E[X]$|\alpha, \beta]$=$\alpha\beta$*
2. *Var[X]$|\alpha, \beta]$=$\alpha\beta^2$*
3. *$M_X(t) = (\frac{1}{1-\beta t})^\alpha$, $t < \frac{1}{\beta}$*

4. $f_X(x|\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}\mathbf{1}_{(0,\infty)}$

Inverse Gamma: Let X∼Gamma(α,β) then $Y = \frac{1}{X}$ follows Inverse Gamma distribution(α,β). Its moments can be expressed as:

$$E[Y^n] = \frac{T(\alpha-n)\beta^{\alpha-n}}{T(\alpha)\beta^\alpha}$$

Chi-square: Let X be a random variable follows Chi-square distribution with k degrees of freedom.

$$X \overset{d}{=} Gamma(\frac{k}{2}, 2)$$

4.1.4 Normal distribution

Normal: If X∼Normal(μ,σ^2) then X has density function:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathbb{1}_{x\in\mathcal{R}}(x)$$

Log-normal: If $\ln X \sim N(\mu,\sigma^2)$ then $X \sim Lognormal(\mu,\sigma)\mathbb{1}_{x\in\mathcal{R}^+}(x)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}x\sigma}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

Cauchy: Cauchy is a symmetric distribution with more heavy tails than normal distribution, if X∼Cauchy(0,1) then:

$$f_X(x) = \frac{1}{\pi(1+x^2)}\mathbb{1}_{x\in\mathcal{R}}(x)$$

Logistic: Logistic has tails between Cauchy and Normal, if X∼Logistic(0,1) then:

$$F_X(x) = \frac{e^x}{1+e^x}\mathbb{1}_{x\in\mathcal{R}}(x)$$

Property 4 *If X∼N(μ,σ^2) and $Z = \frac{x-\mu}{\sigma} \sim (0,1)$ then the m.g.f of X is:*

$$\begin{aligned}
M_X(t) &= E[e^{t(\mu+\sigma Z)}] = e^{\mu t}M_Z(\sigma t)\\
&= e^{\mu t}\int_{-\infty}^{\infty} e^{\sigma z t}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz\\
&= e^{\mu t}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{(z-\sigma t)^2}{2}}e^{\frac{\sigma^2t^2}{2}}dz\\
&= e^{\mu t+\frac{\sigma^2t^2}{2}}
\end{aligned}$$

Theorem 2 *Let X∼ $N(\mu,\sigma^2)$ and g(x) be differentiable function satisfying $E[|g'(x)|] < \infty$ then we have*

$$E[g(x)(x-\mu)] = \sigma^2 E[g'(x)]$$

Proof:

$$\begin{aligned}
\sigma^2 E[g'(x)] &= \sigma^2 \int_{-\infty}^{\infty} g'(x) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \frac{\sigma^2}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} g'(x) dx \int_{-\infty}^{x} -\frac{z-\mu}{\sigma^2} e^{\frac{-(z-\mu)^2}{2\sigma^2}} dz \\
&= \int_{-\infty}^{\infty} g'(x) dx \int_{-\infty}^{x} -(z-\mu) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(z-\mu)^2}{2\sigma^2}} dz \\
&= \int_{-\infty}^{\infty} -(z-\mu) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(z-\mu)^2}{2\sigma^2}} dz \int_{z}^{\infty} g'(x) dx \\
&= \int_{-\infty}^{\infty} (z-\mu) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(z-\mu)^2}{2\sigma^2}} \lim_{b\to\infty} (g(z) - g(b)) dz \\
&= E[g(x)(x-\mu)]
\end{aligned}$$

■

Chapter 5

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 15	
November 22, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

5.1 Logistic distribution

The cumulative distribution function of logistic distribution is

$$F_X(x) = \frac{e^x}{1+e^x}\mathbf{1}_{(-\infty,\infty)}(x)$$

The importance of logistic distribution is that it lies between Cauchy distribution and Gaussian distribution. That is, the decaying rate of logistic distribution is somewhere between $O(2^{-n})$ and $O(2^{-n^2})$.

Another important application of logistic distribution is *logistic regression*. Here, we sketch the formulation of logistic regression:

- Binary response: $Y \in \{0, 1\}$
- Explanatory variables: $Z_1, ..., Z_p$
- Odds ratio: $\frac{P(Y=1|Z_1,...,Z_p)}{1-P(Y=1|Z_1,...,Z_p)}$
- Logistic transformation:

$$\ln \frac{P(Y=1|Z_1,...,Z_p)}{1-P(Y=1|Z_1,...,Z_p)} = \beta_0 + \beta_1 Z_1 + ... + \beta_p Z_p$$

- Positive probability:

$$P(Y=1|Z_1,...,Z_p) = \frac{e^{\beta_0+\beta_1 Z_1+...+\beta_p Z_p}}{1+e^{\beta_0+\beta_1 Z_1+...+\beta_p Z_p}}$$

To sum up, logistic regression is a special case of generalized linear model and aims to predict the probability of certain outcome. **Generalized linear model**:

$$P(Y|Z_1,..,Z_p) = F_0(\beta_0 + \beta_1 Z_1 + ... + \beta_p Z_p)$$

, where F_0 is a cumulative distribution function.

5.2 Beta distribution

Beta distribution is a distribution that describes the battle between 0 and 1. We can create various of distribution in the interval [0,1] with beta distribution. With this abundance property, beta distribution can be used as a prior function in Bayesian analysis. Formally speaking, the density function of beta distribution is

$$f_X(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\text{Beta}(\alpha,\beta)}\mathbf{1}_{[0,1]}(x)$$

, where $\alpha, \beta > 0$ and $\text{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
Now, let's see some basic properties of beta distribution:

- $\mathbb{E}[X|\alpha,\beta] = \frac{\alpha}{\alpha+\beta}$
- $Var[X|\alpha,\beta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha\beta+1)}$
- Shape:
  - $\alpha > 1, \beta = 1$, increasing.
  - $\alpha = 1, \beta > 1$, decreasing.
  - $\alpha < 1, \beta < 1$, U shape.
  - $\alpha > 1, \beta > 1$, unimodal.
  - $\alpha = \beta$, symmetric.
  - $\alpha = \beta = 1$, uniform.
- The relation between beta and binomial: $X \sim$Beta(α,β), $Y \sim$Binomial(n,p)

$$P(X \geq p|x+1, n-x) = P(Y \leq x|n,p)$$

5.3 Double exponential distribution (Laplace distribution)

The density function of Laplace distribution is

$$f_X(x|\mu,\sigma^2) = \frac{1}{2\sigma}e^{-|x-\mu|/\sigma}\mathbf{1}_{(-\infty,\infty)}$$

The following is some basic properties:

- $\mathbb{E}[X|\mu,\sigma^2] = \mu$
- $Var[X|\mu,\sigma^2] = 2\sigma^2$

Note that here σ^2 is not variance.

5.4 Log-normal distribution

Let X be a log-normal distribution with parameter (μ, σ^2), then

$$Y = \ln X \sim N(\mu, \sigma^2)$$

The following is some basic properties:

- $\mathbb{E}[X|\mu, \sigma^2] \neq \mu$
- median $= \mu$
- pdf: $f_X(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \frac{1}{x} e^{-(\ln x - \mu)^2/2\sigma^2} \mathbf{1}_{(0,\infty)}(x)$
- Have moments but no mgf.
- $\mathbb{E}[X|\mu, \sigma^2] = e^{\mu + \sigma^2/2}$
- $Var[X|\mu, \sigma^2] = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$

5.5 Cauchy distribution

The density function of Cauchy distribution is

$$f_X(x|\mu, \sigma) = \frac{1}{\pi\sigma(1 + (\frac{x-\mu}{\sigma})^2)} \mathbf{1}_{(-\infty,\infty)}(x)$$

The following is some basic properties

- $\mathbb{E}[|X| \mid \mu, \sigma^2] = \infty$
- median $= \mu$
- If X, Y are two independent $N(0, 1)$, then $\frac{X}{Y} \sim$Cauchy(0,1).

Chapter 6

Statistical Inference I **Prof. Chin-Tsang Chiang**

Lecture Notes 16

November 22, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

6.1 Exponential families

A family of distributions is a collection of pdf (pmf) sharing some common properties. The exponential families is a family of distributions in the following form:

$$f_X(x|\theta) = h(x)c(\theta)\exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$$

, where $h, c \geq 0$ and w_i, t_i are real-valued function.
Here, we can think of these functions as:

- $t_i(x)$: functionals of x, or empirical moment.
- $w_i(\theta)$: linear weight.
- $h(x), c(\theta)$: normalization term.

Remark:

1. When $dim(\theta) < k$, we call it a *curved* exponential family. Otherwise, we call it a *full* exponential family. Intuitively, parameters in a curved exponential family have some correlation, which can be geometrically considered as a curve.
2. In fact, we can **re-parametrize** an exponential family from parameter space θ to a *natural* parameter space η in the following sense:

$$f_X(x|\eta) = h(x)c^*(\eta)\exp(\sum_{i=1}^{k} \eta_i t_i(x))$$

, where $\mathcal{H} = \{\eta : \int_{-\infty}^{\infty} h(x)\exp(\sum_{i=1}^{k}\eta_i t_i(x))dx < \infty\}$ is the natural parameter space. Intuitively, parameters in η are independent to each others.

Now, you might wonder, what's the benefit we can get from exponential family? Why we want to discuss it? For now, we can benefit from the following theorem.

Theorem 3 (properties of exponential families)
Let X has the pdf (pmf) $f_X(x|\theta) = h(x)c(\theta)\exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$, then

- $\mathbb{E}[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)] = \frac{-\partial}{\partial\theta_j}\ln c(\theta)$
- $Var[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)] = \frac{-\partial^2}{\partial\theta_j^2}\ln c(\theta) - \mathbb{E}[\sum_{i=1}^{k}\frac{\partial^2 w_i(\theta)}{\partial\theta_j^2}t_i(X)]$

Proof: Start with the observation that

$$1 = \int_{-\infty}^{\infty} f_X(x|\theta)dx$$
$$0 = \frac{\partial}{\partial\theta_j}\int_{-\infty}^{\infty} f_X(x|\theta)dx$$

We have

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta_j}\int_{-\infty}^{\infty} f_X(x|\theta)dx \\
(\because \text{Fubini}) &= \int_{-\infty}^{\infty}\frac{\partial f_X(x|\theta)}{\partial\theta_j}dx = \int_{-\infty}^{\infty}\frac{\partial \ln f_X(x|\theta)}{\partial\theta_j}f_X(x|\theta)dx \\
&= \int_{-\infty}^{\infty}\frac{1}{f_X(x|\theta)}[h(x)\frac{\partial c(\theta)}{\partial\theta_j}\exp(\sum_{i=1}^{k}w_i(\theta)t_i(x)) + f_X(x|\theta)\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(x)]f_X(x|\theta)dx \\
&= \int_{-\infty}^{\infty} h(x)c(\theta)\exp(\sum_{i=1}^{k}w_i(\theta)t_i(x))\frac{\partial \ln c(\theta)}{\partial\theta_j}dx + \mathbb{E}[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(x)] \\
&= \mathbb{E}[\frac{\partial \ln c(\theta)}{\partial\theta_j}] + \mathbb{E}[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(x)] = \frac{\partial \ln c(\theta)}{\partial\theta_j} + \mathbb{E}[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(x)]
\end{aligned}$$

As a result,

$$\mathbb{E}[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(x)] = -\frac{\partial \ln c(\theta)}{\partial\theta_j}$$

Similarly, we can show

$$Var[\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)] = \frac{-\partial^2}{\partial\theta_j^2}\ln c(\theta) - \mathbb{E}[\sum_{i=1}^{k}\frac{\partial^2 w_i(\theta)}{\partial\theta_j^2}t_i(X)]$$

■

Remark: $\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(x)$ is the variation of the descriptive term w.r.t parameter θ_j.

6.1.1 Generalized linear model

Generalized linear model is composed of two parts: *random component* and *systematic component*. Basically, we have a response random variable Y and a set of explanatory random variables $\{Z_1, ..., Z_p\}$.

- Random component: it is from exponential family

$$f_Y(y|\theta, \phi) = \exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi))$$

  , where $\theta = \theta(Z_1, ..., Z_p)$ which is a functional of explanatory random variables and ϕ is the dispersion parameter.

- Systematic component: it describes the mean of the response random variable with a *link function* of a linear combination of explanatory random variables.

$$\mathbb{E}[Y|\theta(Z_1, ..., Z_p)] = h(\beta^T \mathbf{Z})$$

  , where h is the link function.

From Theorem 3, we have

- $\mathbb{E}[Y|\theta, \phi] = \frac{d}{d\theta} b(\theta)$
- $Var[Y|\theta, \phi] = a(\phi)\frac{d^2}{d\theta^2} b(\theta) = a(\phi)\frac{d}{d\theta}\mathbb{E}[Y|\theta, \phi]$

Proof: By Theorem 3, we have

$$\mathbb{E}[Y|\theta, \phi] = \mathbb{E}[\frac{\partial}{\partial\theta}\frac{y\theta}{a(\phi)}]a(\phi) = -a(\phi)\frac{\partial}{\partial\theta}\ln\exp(\frac{-b(\theta)}{a(\phi)}) = \frac{d}{d\theta}b(\theta)$$

$$Var[Y|\theta, \phi] = Var[\frac{\partial}{\partial\theta}\frac{y\theta}{a(\phi)}]a^2(\phi) = -a^2(\phi)[\frac{\partial^2}{\partial\theta^2}\ln\exp(\frac{-b(\theta)}{a(\phi)}) - 0] = a(\phi)\frac{\partial^2}{\partial\theta^2}b(\theta)$$

■

Chapter 7

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 17	
November 22, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

7.1 Rose: Poisson process and Renewal process

Nonhomogeneous Poisson process deals with the situation that the rate is not fixes. Instead, it is a function of time, that is, $\lambda(t)$. Professor Chiang showed three examples about nonhomogeneous Poisson process as follow:

7.1.1 How to generate nonhomogeneous Poisson process

Suppose we want to generate $N^*(t) \sim$Poisson$(\lambda(t))$, what can we do? By definition, we have

$$P[N^*(t+h) - N^*(t) = 1] = \lambda(t)h + o(h)$$

Let $m(t) = \int_0^t \lambda(u)du$, then we can show that

$$P[N^*(t+s) - N^*(t) = n] = \frac{(m(t+s) - m(t))^n e^{-(m(t+s)-m(t))}}{n!}$$

Now, we consider a homogeneous Poisson process $N(t) \sim$Poisson(λ). Similarly, we have

$$P[N(t+s) - N(t)] = \frac{\lambda^n e^{-\lambda}}{n!}$$

By definition, the rate of nonhomogeneous Poisson process varies by times. Intuitively, we can think of this phenomenon in another way: the time unit is scaling time by time. For example, in the small time interval $[t, t+h]$ where the rate almost being constant. We have

$$P[N(t+h) - N(t) = 1] = h\lambda + o(h)$$

and

$$\begin{aligned} P[N^*(t+h) - N^*(t) = 1] &= h\lambda(t) + o(h) \\ &= (h\frac{\lambda(t)}{\lambda})\lambda + o(h\frac{\lambda(t)}{\lambda}) \\ &= P[N(t + h\frac{\lambda(t)}{\lambda}) - N(t) = 1] \end{aligned}$$

To sum up, we will have

$$P[N^*(t+s) - N^*(t) = n] = P[N(\frac{m(t+s)}{\lambda}) - N(\frac{m(t)}{\lambda}) = n]$$

Namely, we can generate a nonhomogeneous Poisson process $N^*(t) \sim$Poisson$(\lambda(t))$ with a standard Poisson process $N(t) \sim$Poisson(λ) in the following way

$$N^*(t) = N(\frac{m(t)}{\lambda})$$

7.1.2 Latent effects will be ruled out as fixing observation time

7.2 Location-scale family

Definition 4 (location-scale family) *We say a family of distribution with standard pdf f is*

$$\mathcal{F} = \{g : \mu \in \mathbb{R}, \sigma > 0, \forall x \in \mathbb{R}, \sigma g(\frac{x-\mu}{\sigma}) = f(x)\}$$

We can also characterize a location-scale family in another way presented in the following theorem.

Theorem 4 *Let f be a pdf and $\mu \in \mathbb{R}, \sigma > 0$, then*

$$X \sim \frac{1}{\sigma} f(\frac{x-\mu}{\sigma}) \Leftrightarrow \exists Z \in f(z) \text{ s.t. } X = \sigma Z + \mu$$

Proof:
($\Leftarrow$) Define $g(z) = \sigma z + \mu$. As $g(\cdot)$ is monotone and measurable, we can conclude that $X = g(Z)$ is a random variable with pdf $f_X(x = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma}))$
($\Rightarrow$) Define $g(x) = \frac{x-\mu}{\sigma}$. As $g(\cdot)$ is monotone and measurable, we can conclude that $Z = g(X)$ is a random variable with pdf $f_Z(z) = f(z)$. ■

With Theorem 4, we can see that when we know the behavior of Z, we can use some linear transformation to infer the behavior of random variables that follow the distribution in the same location-scale family. For example, we have

- $\mathbb{E}[X] = \sigma\mathbb{E}[Z] + \mu$
- $Var[X] = \sigma^2 Var[Z]$

Moreover, location-scale family is *stochastically increasing*. Recall that

Definition 5 (stochastically increasing) *A family of distribution* $\{f(x|\theta) : \theta \in \Theta\}$ *is stochastically increasing in* θ *is* $\forall \theta_1 > \theta_2$*,*

$$F(x|\theta_1) < F(x|\theta_2), \ \forall x$$

In other words,

$$\textit{Tail probability of } \theta_1 > \textit{Tail probability of } \theta_2$$

Remark: Geometrically, the cdf of θ_2 is strictly above that of θ_1.

7.3 Probability inequalities

7.3.1 Markov inequality

When a random variable X is **nonnegative**, Markov inequality provides an upper bound for the tail probability of X as follow

$$P[X \geq r] \leq \frac{\mathbb{E}[X]}{r}$$

Proof: The proof is one line:

$$\mathbb{E}[X] = \int_0^\infty x dF_X(x) \geq \int_r^\infty x dF_X(x) \geq r P[X \geq r]$$

■

Intuition (Markov inequality)

The idea of Markov inequality and other related inequalities is quite elegant. We separate the integral in two parts, then upper bound one of them by the nonnegativity of probability and upper the other by the monotonicity of the support.

7.3.2 Chebyshev inequality

We can generalize Markov inequality by transforming random variables into nonnegative random variables. And this is exactly the central idea of Chebyshev inequality.

Theorem 5 (Chebyshev inequality) *Let* X *be a random variable and* $g(x)$ *be a nonnegative measurable function. Then,* $\forall r > 0$*, we have*

$$P[g(X) \geq r] \leq \frac{\mathbb{E}[g(X)]}{r}$$

However, we can see that the probability inequalities in this fashion are very loose, just like Boole's inequality. As a result, they will definitely overestimate. Thus, the most important applications of these inequalities are not bounding error. Instead, they have excellent performance on bounding *error rate*.