Distributions

Wei-Chang Lee, Chi-Ning Chou

November 1, 2015

Contents

Chapter 1

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

Lecture Notes 11

November 1, 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 1, 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* ot *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,

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

2.3 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 1, 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 $Y \overset{d}{=} Poisson(\lambda)$

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