Transformation

Wei-Chang Lee, Chi-Ning Chou

October 30, 2015

Contents

Chapter 1

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

Lecture Notes 7

October 30, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

In the second chapter, we're going to talk about **transformation** and **expectation**. In short, transformation induces new random variable with a known random variable and expectation is the functional a random variable and somehow provides a weighted average sense.

1.1 Transformation

Recall that random variable maps a specific probability space to a universal probability space: $(\mathcal{R}, \mathcal{B}, \mathcal{L})$ while preserving the measurablility. Now, we want to extend the idea and map $(\mathcal{R}, \mathcal{B})$ to itself!

1.1.1 Measurable function

Definition 1 (Borel measurable function) *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$, then g is a Borel measurable function from $(\mathcal{X}, \mathcal{B})$ to $(\mathcal{Y}, \mathcal{B})$ if $\forall B \in \mathcal{B}$ $g^{-1}(B) \in \mathcal{B}$.*

Property 1 *If $X : \Omega \to \mathcal{X}$ is a random variable and g is a Borel measurable function from $(\mathcal{X}, \mathcal{B})$ to $(\mathcal{Y}, \mathcal{B})$. Then, $g(X)$ is a random variable from Ω to $(\mathcal{Y}, \mathcal{B})$.*

Proof: Simply check the sufficient condition of random variable. $\forall B \in \mathcal{B}$ consider

$$\begin{aligned} \{w : g(X(w)) \in B\} &= \{w : X(w) \in g^{-1}(B)\} \\ &= \{w : w \in X^{-1}[g^{-1}(B)]\} \\ &\in \mathcal{B} \end{aligned}$$

Since g is Borel measurable, $g^{-1}(B) \in \mathcal{B}$. Moreover, X is a random variable, thus $X^{-1}[g^{-1}(B)] \in \mathcal{B}$. ■

If we go back to the construction of random variable X, we can see that what X really does is inducing a probability space from $(\Omega, \mathcal{A}, P)$ to $(\mathcal{X}, \mathcal{B}, \mu_X)$, where μ_X is the induced measure defined as $\mu_X(B) = P(X \in B),\ \forall B \in \mathcal{B}$.

Now, let's turn to a transformation of random variable, $Y = g(X)$, it's clearly that Y also induces a new probability space $(\mathcal{Y}, \mathcal{B}, \mu_Y)$. Moreover, here we can have two ways to define the measure μ_Y:

$$\begin{aligned} \mu_Y(B) &= P(g(X) \in B) && \text{(from P)} \\ &= \mu_X(X \in g^{-1}(B)) && (\text{from } \mu_X) \end{aligned}$$

Intuition (transformation)

$$(\Omega, \mathcal{A}, P) \xrightarrow{X} (\mathcal{X}, \mathcal{B}, \mu_X) \xrightarrow{g} (\mathcal{Y}, \mathcal{B}, \mu_Y)$$

1.1.2 p.m.f., p.d.f., and cumulative distribution

In the very beginning, let's define some notations here. For convenience, from now on X denotes a random variable from $(\Omega, \mathcal{A}, P)$ to $(\mathcal{X}, \mathcal{B}, \mu_X)$. And g is a Borel measurable function from $(\mathcal{X}, \mathcal{B})$ to $(\mathcal{Y}, \mathcal{B})$ that induces a random variable $Y = g(X)$ with measure μ_Y.

As long as we have the definition of induced probability measure, we can further define the density function and cumulative distribution.

Property 2 *The cumulative distribution of $Y = g(X)$ is*

$$\begin{aligned} F_Y(y) &= P(g(X) \le y),\ \forall y \in \mathbb{R} \\ &= P(\{x \in \mathcal{X} : g(x) \le y\}) \\ &= \int_{\{x \in \mathcal{X} : g(x) \le y\}} f_X(x) dx \end{aligned}$$

Remark: Note that the integration form might not be directly for a discrete random variable. Here, we use the expectation of $g(X)$ to present a way to represent discrete random variable in an integration form:

$$E[g(X)] = \int_{x \in \mathcal{X}} g(x) dF_X(x)$$

Note that here we define $dF_X(x) = F_X(x^- + dx) - F_X(x^-)$. For continuous r.v., this is the density at point x. As to discrete r.v., when there's some mass on point x, then the value will have a jump (step) at that point. One can see that this representation is equivalent to sum over the point that have mass.

Now, we might want to have a more convenient way to calculate the cumulative distribution and density function. However, we have to find the preimage of each region and compute the probability of that region over the measure of X. And for most of the cases this is not quite easy. Consider the case that $g(x) = x^2$, $\forall y \ge 0$, $\{x \in \mathcal{X} : g(x) \le y\} = [-y, y]$. We cannot find a similar form ($\le y$) in the context of X.

But this is not the end of the world, soon we can see that as g is monotone, the relation can be easily build up. Intuitively, the monotonicity preserves the direction and result in the convenience of calculation.

Property 3 *Suppose $g(y)$ is* **strictly monotone** *on $\mathcal{X}$, then*

$$\begin{aligned} F_Y(y) &= F_X(g^{-1}(y)) && , \ g \text{ is increasing} \\ F_Y(y) &= 1 - F_X(g^{-1}(y)) && , \ g \text{ is decreasing} \end{aligned}$$

Proof: Since we have the monotonicity, $\forall y \in \mathcal{Y}$

$$\begin{aligned} \{x \in \mathcal{X} : g(x) \leq y\} &= \{x \in \mathcal{X} : x \leq g^{-1}(y)\} && , \ g \text{ is increasing} \\ &= \{x \in \mathcal{X} : x \geq g^{-1}(y)\} && , \ g \text{ is decreasing} \end{aligned}$$

Intuitively, monotonicity preserves the relative position of points in $\mathcal{X}$ and $\mathcal{Y}$. ■

Property 4 *Suppose $g(y)$ is monotone, then*

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|\mathbf{1}_Y(y)$$

Intuitively, monotonicity also gives us a sense of **scaling** from X to Y. We can regard the term $|\frac{df_X(g^{-1}(y))}{dy}|$ simply as a scaling factor which is the relative exchange in the domain of $\mathcal{X}$ w.r.t. the domain of $\mathcal{Y}$.

Chapter 2

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 8	
October 30, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

Today we are going to talk about some examples of transformation in random variable and probability integral transformation.

2.1 Examples of transformation

2.1.1 Binomial distribution

Let X be a random variable with binomial distribution, then $f_X(x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x(1-p)^{n-x}\mathbf{1}_{0,1,\ldots,n}(x)$.
Suppose $Y = n - X$, then Y is also a random variable.

2.1.2 Exponential

Let X be a random variable with exponential distribution, then $f_X(x) = \lambda e^{-\lambda x}\mathbf{1}_{(-\infty,\infty)}(x)$. Suppose $Y = X^\gamma$, $\gamma > 0$, then Y is also a random variable. Furthermore, the we can derive the distribution of Y as follow:

$$F_Y(y) = \int_{\{x:\ x^\gamma \le y\}} = \int_0^{y^{1/\gamma}} f_X(x)dx f_Y(y) = \frac{\lambda}{\gamma} y^{\frac{1}{\gamma}-1} e^{-\lambda y^{\frac{1}{\gamma}}} \mathbf{1}_{(0,\infty)}(y)$$

Remark 1 Power transformation v.s. natural logarithm transformation

Box and Cox introduced the power transformation:

$$\frac{X^\gamma - 1}{\gamma}$$

Moreover, we can see that as $\gamma \to 0$, the power transformation becomes natural logarithm transformation:

$$\ln X$$

These two transformations are similar but with different properties. In practice, we sometimes need to perform test to fit one of them to our model.

Remark 2 Degree of freedom

When using **linear model** plus **normal-family distribution**, degree of freedom refers to the rank of the power of our description variable. Most of the time, the degree of freedom is $n-k$, where n is the number of samples and k is the number of statistics we use.

Remark 3 Mean or median?

Mean and median are two different statistics to describe the central location of a data. We want to use them to estimate the population center. However, we might afraid of the existence of outliers so that the result will be biased. Thus, we can propose other methods to inference the central location as long as it's meaningful.

Remark 4 Box-Muller transformation: Generating normal distribution

Since the inverse of normal distribution cannot be written down in a close form, we cannot simply plugging uniformly distributed random variable to generate normal random variable. Instead, there's a method based on the intuition to calculate $\int_{-\infty}^{\infty} e^{-x^2/2}dx$ called Box-Muller transformation. Intuitively, we consider the polar coordinate (R,θ), and let

- $R^2 = X^2 + Y^2$, where X and Y are independent normal distribution.
- $\theta = 2\pi U_1$, where U_1 is a uniform distribution on [0,1].

As R^2 is actually a chi-square distribution with degree of freedom 2, we can write it as $R^2 = -2\ln U_2$. As a result, we can generate a normal distributed random variable as

$$Z = R\cos\theta = \sqrt{-2\ln U_2}\cos(2\pi U_1)$$

Remark 5 Poisson approximation v.s. normal approximation

The close form of binomial distribution involves lots of binomial terms. As a result, when n is large, it's computationally inefficient to directly compute the cumulative distribution or pmf from the close form. We need some computationally efficient approximation. There are two kinds of approximations for binomial random variable: Poisson and normal. What's the difference? Here, we state two approximations and compare their intuitions.

Poisson approximation

$$\begin{aligned}\binom{n}{k}p^k(1-p)^{n-k} &= \frac{n\cdot(n-1)\cdots(n-k+1)}{k!}p^k(1-p)^{n-k}\\ (\text{let } \lambda = np) &= \frac{1}{k!}\frac{n\cdot(n-1)\cdots(n-k+1)}{n\cdot n\cdots n}\lambda^k(1-\frac{\lambda}{n})^{n-k}\\ (n \text{ large}) &\approx \frac{\lambda^k e^{-\lambda}}{k!}\end{aligned}$$

As p is small and n is large, Poisson approximation can give a good results.

Normal approximation As n grows large, by central limit theorem, we have

$$\binom{n}{k}p^k(1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi}}e^{-\frac{(k/n-p)^2}{2}}$$

Chapter 3

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 9	
October 30, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

3.1 Expectation

Expectation is simply a functional of the distribution. It maps a distribution to a certain real value to represent the behavior, shape, or other properties. Formally, we define the expectation of a random variable X as follow:

Definition 2 (expectation) *Let X be a r.v. and g be a measurable function. Then, the expectation of $g(X)$, which is also a r.v., is denoted as $\mathbb{E}[g(X)]$,i.e.,*

$$\mathbb{E}[g(X)] = \int_x g(x) dF_X(x)$$

Note that the expectation of $\mathbb{E}[g(X)]$ exists provided that $\mathbb{E}[|g(X)|] < \infty$.

Remark: If the distribution is not a mixture of both discrete and continuous distribution, then we can represent it as

- If X is discrete, $\mathbb{E}[g(X)] = \sum_x g(x) f_X(x)$.
- If X is continuous, $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) dF_X(x)$.

However, not all the distribution has expectation! Cauchy distribution is a beautiful example:
Example: (**Cauchy distribution has no mean**)
The pdf of Cauchy distribution is

$$f(x) = \frac{1}{\pi(1+x^2)} dx$$

With simple integration, we can check that $\mathbb{E}(|X|) = 2\int_0^{\infty} \frac{x}{\pi(1+x^2)} dx = \infty$. Thus, the expectation of Cauchy distribution does not exist. As a remark, Cauchy is a bell-shaped distribution with median 0. And actually, the cumulative distribution of Cauchy is the arc tangent function!

Property 5 *Let X be a r.v. and a,b,c be constants. Moreover, $g_1(X)$, $g_2(X)$ be any r.v. with expectation. Then,*

1. *(Preserve linear combination)* $\mathbb{E}[]ag_1(X) + bg_2(X) + c] = a\mathbb{E}[g_1(X)] + b\mathbb{E}[g_2(X)] + c$.
2. *(Preserve non-negativity) If* $f(x) \geq 0,\ \forall x$, *then* $\mathbb{E}[g(X)]$.
3. *(Preserve dominance) If* $g_1(x) \geq g_2(x),\ \forall x$, *then* $\mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)]$.
4. *(Existence of bounded r.v.) If* $a \leq g(x) \leq b,\ \forall x$, *then* $a \leq \mathbb{E}[g(X)] \leq \mathbb{E}[g(X)]b$.

Now, we turn to an useful and interesting application of expectation.

Example: (**The expectation of indicator function is probability**) Consider I_A to be an indicator function of a set $A \subseteq \mathbb{R}$, then

$$\mathbb{E}[I_A(X)] = P(A)$$

Moreover, we can regard the above equation as a **binary response**. That is, the indicator separate the space $\mathbb{R}$ into two parts: $\{x : x \in A\}$ and $\{x : x \notin A\}$ and the expectation is a functional to see the response of such partition.

For example, consider the following indicator function $I(X \leq x)$. We can see that $\mathbb{E}[I(X \leq x)] = F_X(x)$. And this representation gives us a broad way to describe the data. Suppose now we are concerning the probability $Pr[X = x|Z_1, Z_2, ..., Z_p]$, the most simply way is to use a general model to describe it, say

$$Pr[X = x|Z_1, Z_2, ..., Z_p] = G(x, \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p)$$

As we choose to use the expectation representation: $\mathbb{E}[I(X \leq x)|Z_1, Z_2, ..., Z_p]$, the impact of Z_is can somehow depends on the value of x and become even more general. In other words, the linear parameter β_is can be depended on x. For example,

$$x_1 : \{x : X \leq x_1\} \leftrightarrow \beta_{11} Z_1 + \beta_{12} Z_2 + ... + \beta_{1p} Z_p$$
$$x_2 : \{x : X \leq x_2\} \leftrightarrow \beta_{21} Z_1 + \beta_{22} Z_2 + ... + \beta_{2p} Z_p$$

With the above concept, we can simply show the inclusion-exclusion theorem with the help of indicator function and its expectation. First consider two facts:

- $\mathbf{1}_{A \cap B} = \mathbf{1}_A \mathbf{1}_B$ and $\mathbf{1}_{A \cup B} = 1 - \mathbf{1}_{A^C \cap B^C}$
- $\mathbf{1}_{\cup_i A_i} = 1 - \prod_i (1 - \mathbf{1}_{A_i})$

Now, we can derive the inclusion-exclusion theorem:

$$\begin{aligned} P(\cup A_i) &= 1 - \mathbb{E}[\prod_i (1 - \mathbf{1}_{A_i})] \\ &= 1 - \mathbb{E}[1 - \sum_i \mathbf{1}_{A_i} + \sum_{i,j} \mathbf{1}_{A_i} \mathbf{1}_{A_j} - ... + (-1)^k \sum_{i_1,...,i_k} \mathbf{1}_{A_{i_1}} \cdots \mathbf{1}_{A_{i_k}} \pm ... \pm \mathbf{1}_{\cap A_i}] \\ &= \sum_i P(A_i) - \sum_{i,j} P(A_i \cap A_j) + ... + (-1)^{k-1} \sum_{i_1,...,i_k} P(A_{i_1} \cap \cdots \cap A_{i_k}) \pm ... \pm P(\cup_i A_i) \end{aligned}$$

3.2 Moment

Definition 3 (Moment and Central moment) *For each integer n, the n^{th} moment of X is $\mu'_n = E[X^n]$, and the n^{th} central moment of X is $\mu_n = E[(X-\mu)^n]$ where $\mu = \mu'_1$.*

Definition 4 (Variance and Standard deviation) *The variance of a r.v. denoted by Var(x), is μ_2. The positive square root of Var(x), denoted by σ_x, is called the standard deviation/error of X.*

Moment carries some information of the distribution, some useful moments are as below.

1. mean: $E[X] = \mu$
2. variance: $E[(X-\mu)^2] = \sigma^2$
3. skewness: $E[(\frac{X-\mu}{\sigma})^3] = \frac{E[X^3]-3\mu\sigma^2-\mu^3}{\sigma^3} = \gamma_1$
4. kurtosis: $E[(\frac{X-\mu}{\sigma})^4] = \gamma_2$ (Kurt[N(0,1)]=3)

Property 6 *Minimum variance*

1. *(a) $argmin(E[(X-a)^2]) = \mu$ and minimum $E[(X-a)^2] = Var(x)$.*
2. *(b) $Var(X) = 0 \Leftrightarrow P(|X - E[X]| < \epsilon) = 1 \ \forall \epsilon > 0$.*

Proof: *($\Leftarrow$)*

$$\begin{aligned} Var(X) &= \int_{x\in\mathcal{R}} (x-E[X])^2 P(X=x) dF_X(x) \\ &= \int_{x-E[X]<\epsilon} (x-E[X])^2 P(X=x) dF(x) + \int_{x-E[X]\geq\epsilon} (x-E[X])^2 P(X=x) dF_X(x) \\ &\leq \epsilon^2 + 0 \ (pick \ \epsilon \downarrow 0) \\ &= 0 \end{aligned}$$

($\Rightarrow$) $\forall \epsilon > 0$

$$\begin{aligned} 0 = Var(X) &= \int_{x\in\mathcal{R}} (x-E[X])^2 P(X=x) dF_X(x) \\ &\geq \int_{x-E[X]\geq\epsilon} (x-E[X])^2 P(X=x) dF_X(x) \\ &\geq \epsilon^2 \times P(|x-E[X]| \geq \epsilon) \end{aligned}$$

It implies that

$$P(|x-E[X]| \geq \epsilon) = 0 \rightarrow P(|x-E[X]| < \epsilon) = 1$$

■

Property 7 *If X has finite variance, $Var(aX \pm b) = a^2 Var(x) \ \forall a, b \in \mathcal{R}$.*

Proof: Simply expands it and use linearity of expectation to rearrange it. ■

3.3 Moment generating and characteristic function

Definition 5 (Moment generating function) *The moment generating function of X is defined to be $M_X(t) = E[e^{tX}]$ provided that the expectation exists for t in some $\mathcal{B}_r(0)$.*

Definition 6 (Characteristic function) *The characteristic function of X is defined to be $\phi_X(t) = E[e^{itX}] = E[cos(tx)] + iE[sin(tx)]$.*

Remark:

1. $\int |cos(tx)| dF_X(x) \leq \int dF_X(x) = 1$ and $\int |sin(tx)| dF_X(x) \leq \int dF_X(x) = 1$
2. The characteristic function does much more than the moment generating function does. The characteristic function always exists and completely determines the distribution.

Example: Consider the lognormal distribution:

$$f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{\frac{-(\ln x)^2}{2}} \mathbb{1}_{(0,\infty)}(x)$$

$$f_2(x) = f_1(x)(1 + sin(2\pi \ln x))\mathbb{1}_{(0,\infty)}(x)$$

let $u = \ln x$ and $v = u - r$, one derives that

$$\begin{aligned}
E[X_1^r] &= \int_0^\infty \frac{x^{r-1}}{\sqrt{2\pi}} e^{\frac{-(\ln x)^2}{2}} dx = \int_{-\infty}^\infty \frac{x^r}{\sqrt{2\pi}} e^{\frac{-(\ln x)^2}{2}} du \\
&= \int_{-\infty}^\infty \frac{e^{ru}}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = \int_{-\infty}^\infty \frac{e^{r(v+r)}}{\sqrt{2\pi}} e^{\frac{-(v+r)^2}{2}} dv \\
&= e^{\frac{r^2}{2}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} dv = e^{\frac{r^2}{2}} \\
E[X_2^r] &= \int_0^\infty \frac{x^{r-1}}{\sqrt{2\pi}} e^{\frac{-(\ln x)^2}{2}} (1 + sin(2\pi \ln x)) dx \\
&= e^{\frac{r^2}{2}} + e^{\frac{r^2}{2}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} sin(2\pi(v + r)) dv \\
&= e^{\frac{r^2}{2}} + e^{\frac{r^2}{2}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} (sin(2\pi v)cos(2\pi r) + sin(2\pi r)cos(2\pi v)) dv \\
&= e^{\frac{r^2}{2}} \text{ (sin is a odd function and taks on zero value when } r \in \mathcal{Z})
\end{aligned}$$

Remark 6 Also noticed that the moment generating function of log normal distribution does not exist since $E[e^{tX}] = \int_0^\infty \frac{e^{tx}}{\sqrt{2\pi}x} e^{\frac{-(\ln x)^2}{2}} dx$ diverges. $e^{tx} e^{\frac{-(\ln x)^2}{2}}$ diverges as $n \to \infty$.

Conclusion

Determine distribution Basically, mgf is a **stronger** of a r.v. The following lists the positive results and negative results of mgf:
Positive:

1. If the support is **bounded** and two r.v.s share every moment, then they will have the same distribution.
2. As the two mgfs are the same in a neighborhood of 0, then they will have the same distribution.
3. Convergence in mgf implies the convergence of distribution.
4. Characteristic function always exists and completely determines the distribution.

Negative:

1. Even all moments exists does not imply m.g.f exists. *e.g., log-normal* distribution.
2. Two distributions might have same moments but have different distribution. *e.g., log-normal* distribution and $(1+\sin(2\pi\log x))\frac{e^{-(\log x)^2/2}}{\sqrt{2\pi}x}$

Chapter 4

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 10	
October 30, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

Lemma 1 (inversion theorem) *Let X be a random variable with characteristic function $\phi_X(t)$ and $a, b \in \mathbb{R}$ with $a < b$, then*

1. *For any random variable X,*

$$P(a < X < b) + \frac{1}{2}P(X = a) + \frac{1}{2}P(X = b) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t)dt$$

2. *If X is a continuous random variable, then*

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t)dt \quad a.e.$$

4.1 Characteristic function

Intuition (inversion theorem)

Inversion theorem provides a *isomorphism* between **distribution** and **characteristic function**.

Proof: The proof is divided into four steps. The first step introduce a integration tool for latter usage. The second step provides a clean form of the original term. The third step uses dominant theorem to show the convergence. The last step gives the density function.

1. **claim**: $\int_0^\infty \frac{\sin \alpha x}{x} dx = \frac{\pi}{2}\text{sign}(\alpha)$

$$\begin{aligned}
\int_0^\infty \frac{\sin \alpha x}{x} dx &= \int_0^\infty \int_0^\infty \sin \alpha x \ e^{-ux} du dx \\
(\text{by Fubini's thm}) &= \int_0^\infty \int_0^\infty \sin \alpha x \ e^{-ux} dx du \\
&= \dots \text{ some change of integrals } \dots \\
&= \frac{\pi}{2} \text{sign}(\alpha)
\end{aligned}$$

2. Consider

$$\begin{aligned}
\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t) dt &= \frac{1}{2\pi} \int_{-T}^{T} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} e^{itx} dF_X(x) dt \\
&= \frac{1}{2\pi} \int_{-T}^{T} \int_{-\infty}^{\infty} \frac{\cos t(x-a) + i \sin t(x-a)}{it} dF_X(x) dt \\
&= \frac{1}{2\pi} \int_{-T}^{T} \int_{-\infty}^{\infty} \frac{\cos t(x-b) + i \sin t(x-b)}{it} dF_X(x) dt \\
(\because \text{symmetry}) &= \frac{1}{\pi} \int_{0}^{T} \int_{-\infty}^{\infty} \frac{\sin t(x-a) - \sin t(x-b)}{t} dF_X(x) dt \\
(\text{by Fubini's thm}) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \int_{0}^{T} \frac{\sin t(x-a) - \sin t(x-b)}{t} dt dF_X(x)
\end{aligned}$$

3. By dominant convergence theorem,

$$\begin{aligned}
\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t) dt &\xrightarrow{T \to \infty} \frac{1}{\pi} \int_{-\infty}^{\infty} \int_{0}^{T} \frac{\sin t(x-a) - \sin t(x-b)}{t} dt dF_X(x) \\
&= \frac{1}{\pi} \int_{x \in (-\infty, a)} \int_{0}^{T} \frac{\sin t(x-a) - \sin t(x-b)}{t} dt dF_X(x) \\
&+ \frac{1}{\pi} \int_{x=a} \int_{0}^{T} \frac{\sin t(x-a) - \sin t(x-b)}{t} dt dF_X(x) \\
&+ \frac{1}{\pi} \int_{x \in (a,b)} \int_{0}^{T} \frac{\sin t(x-a) - \sin t(x-b)}{t} dt dF_X(x) \\
&+ \frac{1}{\pi} \int_{x=b} \int_{0}^{T} \frac{\sin t(x-a) - \sin t(x-b)}{t} dt dF_X(x) \\
&+ \frac{1}{\pi} \int_{x \in (b, \infty)} \int_{0}^{T} \frac{\sin t(x-a) - \sin t(x-b)}{t} dt dF_X(x) \\
(\text{by the tool in 1.}) &= 0 + \frac{1}{2} P(X = a) + P(a < X < b) + \frac{1}{2} P(X = b) + 0
\end{aligned}$$

4. Suppose X is continuous and $\int_{\mathcal{R}} |\phi_X(t)dt| < \infty$,

$$\begin{aligned}
\int_a^b f(x)dx &= \frac{1}{2\pi}\int_a^b \int_{\mathcal{R}} e^{-itx}\phi_X(t)dtdx \\
&= \frac{1}{2\pi}\int_{\mathcal{R}} (\int_a^b e^{-itx}dx)\phi_X(t)dt \\
&= \frac{1}{2\pi}\int_{\mathcal{R}} \frac{e^{-ita}-e^{-itb}}{it}\phi_X(t)dt \\
&= \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita}-e^{-itb}}{it}\phi_X(t)dt \\
&= \frac{1}{2}P(X=a) + P(a<X<b) + \frac{1}{2}P(X=b) \\
&= P(a<X<b)(\text{ further set b=x, a}\downarrow -\infty)
\end{aligned}$$

■

Corollary 1 *For probability measures μ_X and μ_Y on $\mathcal{B}(\mathcal{R})$, the equality $\phi_{\mu_X} = \phi_{\mu_Y}$ implies that $\mu_X = \mu_Y$.*

Proof: From inversion theorem, we have $\mu_X((a,b)) = \mu_Y((a,b))\ \forall a, b \in C$, where C is the set of all $z \in \mathcal{R}$ such that $\mu_X(\{z\}) = \mu_Y(\{z\}) = 0$. Since C^c is at most countable. The family of $\{(a,b) : a.b \in C\}$ of intervals is a π-system generating $\mathcal{B}(\mathcal{R})$. μ_X and μ_Y agrees on a π-system also agrees on the σ-algebra generated by it. ■

4.2 Convergence

Theorem 1 *Let $\{X_n\}$ be a sequence of random variables with characteristic functions $\phi_{X_n}(t)$. Suppose that*

- $\lim_{n\to\infty}\phi_{X_n}(t) = \phi_X(t)$ *for t in a neighborhood of 0. (pairwise convergence)*
- $\phi_X(t)$ *is a characteristic function of some random variable X.*

Then,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),\ \forall x \text{ such that } F_X(x) \text{ is continuous (weakly convergence)}$$

Proof: Let a and b be continuous points of $F_X(x)$ and $F_{X_n}(x)$ for $n \geq N_0$ for some $n \in \mathcal{N}$

$$\begin{aligned}
F_X(b) - F_X(a) &= \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi_X(t)dt \\
&= \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\lim_{n\to\infty}\phi_{X_n}(t)dt \\
&= \lim_{n\to\infty}\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi_{X_n}(t)dt \text{ (since } |\phi_{X_n}(t)| \leq 1, \text{ dominated)} \\
&= \lim_{n\to\infty}(F_{X_n}(b) - F_{X_n}(a))
\end{aligned}$$

By setting b=x, a$\downarrow -\infty$ one obtains $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$. ■

Remark: We don't have to worry about the discrete points since they must converge to the right value.

Intuition (convergence of characteristic function)

Theorem 1 tells us that if r.v.s converge in characteristic functions, then r.v.s also converge in distribution.

Remark: Tips for calculating MGF: consider the MGF of binomial distributed random variable X such that $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \mathbf{1}_{0,1,\ldots,n}$. We have

$$\begin{aligned}
M_{X_n}(t) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} e^{tx} \\
&= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} \frac{(pe^t)^x (1-p)^{n-x}}{[(pe^t)+(1-p)]^n} [(pe^t)+(1-p)]^n \\
(\text{set } p' = \frac{pe^t}{pe^t+(1-p)}) &= \sum_{x=0}^{n} \binom{n}{x} p'^x (1-p')^{n-x} [(pe^t)+(1-p)]^n \\
&= [(pe^t)+(1-p)]^n \sum_{x=0}^{n} \binom{n}{x} p'^x (1-p')^{n-x} \\
&= [(pe^t)+(1-p)]^n
\end{aligned}$$

Chapter 5

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

Lecture Notes 11

October 30, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

5.1 Convergence of m.g.f

Example: Consider $f_{X_n}(x) = \binom{n}{x} P_n^x (1-P_n)^{n-x} \mathbb{1}_{\{0,1,2\ldots n\}}(x)$, the corresponding m.g.f $M_{X_n}(t) = (P_n e^t + (1-P_n))^n$. As $n \to \infty$, $nP_n \to \lambda$, we have $P_n = \frac{\lambda}{n}(1+O(1))$ and

$$M_{X_n}(t) \to M_X(t) = e^{\lambda(e^t-1)}$$

which is the m.g.f of Poisson distribution.

Proof: Let $y = (P_n e^t + (1-P_n))^n$ we have $\ln y = \frac{\ln(P_n e^t + (1-P_n))}{\frac{1}{n}} = \frac{\ln(\frac{\lambda}{n} e^t (1+O(1)) + (1-\frac{\lambda}{n}(1+O(1)))}{\frac{1}{n}}$ applying $L'hôpital's$ rule we get

$$\lim_{n\to\infty} \ln y = \lambda(e^t - 1)$$

so

$$M_{X_n}(t) \to M_X(t) = e^{\lambda(e^t-1)} \; as \; n \to \infty \; nP_n \to \lambda$$

■

Intuition (From the basic view)

The example construct a relationship between Poisson distribution and Binomial distribution. And it is quite reasonable from the definition of Poisson distribution.
Recall the definition of Poisson:

- It's a **counting process**. That is, $N(t)$ that counts the number of appearances before time t.
- (**Boundary condition**) $N(0) = 0$
- (**Stationary**) $\forall t_1 < t_2$, $N(t_2) - N(t_1) \sim N(t_2 - t_1)$
- (**Independence**) $\forall t_1 < t_2 < t_3 < t_4$, $N(t_4) - N(t_3) \sim N(t_2) - N(t_1)$
- (**Fixed frequency**) $\lim_{\Delta \to 0^+} \frac{Pr[N(\Delta)-N(0)=1]}{\Delta} = \lambda$, and $\lim_{\Delta \to 0^+} \frac{Pr[N(\Delta)-N(0)>1]}{\Delta} = 0$
- (**Density function**) $f_\lambda(t, k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \mathbf{1}_{\{k=0,1,2,\ldots\}}$

Since $\lim_{\Delta \to 0^+} \frac{Pr[N(\Delta)=1}{\Delta} = \lambda$ simply let $\Delta\lambda$ be the success probability of P_n.

5.2 Basic property of characteristic function

Property 8 (relation to moment) *Let X be a random variable. If $E[|X^n|] < \infty$, then $\frac{d^n}{(dt)^n}\phi_X(t)$ exists for all t and*

$$\frac{d^n}{(dt)^n}\phi_X(t) = E[e^{itX}(iX)^n]$$

so the lower moments are

$$E[X^n] = (-i)^n \frac{d^n}{(dt)^n}\phi_X(0)$$

Property 9 (Basic) *Let X and Y be random variables.*

1. *$\phi_X(0) = 1$ and $|\phi_X(t)| \leq 1$ $\forall t$.*
2. *$\phi_{-X}(t) = \overline{\phi_X(t)}$ where bar denotes complex conjugation.*
3. *$\phi_{aX+b}(t) = e^{itb}\phi_X(at)$.*
4. *If X and Y are independent, $\phi_{X+Y}(t) = \phi_X(t) \times \phi_Y(t)$.*

Appendix A

Statistical Inference I	**Prof. Chin-Tsang Chiang**
Tex book studies	
October 30, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

In calculating expectation in generating function or characteristic function we sometimes face problems of exchanging summation/integration/differentiation/limits. We give some theory as the guide for us to exchange the operator and give some example applying the technique of interchanging operator for Casella 2.4. We neglect the proof since it is beyond our scope and may relies on measure theory from real analysis.

A.1 Interchanging Integration and Differentiation

Theorem 2 (Leibnitz's Rule) *If $f(x,\theta), a(\theta),$ and $b(\theta)$ are differentiable with respect to θ, then*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - f(a(\theta),\theta)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta}f(x,\theta)dx$$

where $a(\theta)$ and $b(\theta)$ are constant,

$$\frac{d}{d\theta}\int_{a}^{b} f(x,\theta)dx = \int_{a}^{b} \frac{\partial}{\partial\theta}f(x,\theta)dx$$

We get the desired interchanging order of integration and differentiation on a finite range. If the range is unbounded, problem can arise. The question of whether interchanging the order of differentiation and integration relies on whether limits and integration can be interchanged since derivation comes from limits:

$$\frac{\partial}{\partial\theta}f(x,\theta) = \lim_{\delta\to 0}\frac{f(x,\theta+\delta) - f(x,\theta)}{\delta}$$

We have

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \lim_{\delta\to 0}\int_{-\infty}^{\infty}\frac{f(x,\theta+\delta) - f(x,\theta)}{\delta}dx$$

and

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)dx = \int_{-\infty}^{\infty} \lim_{\delta\to 0} \frac{f(x,\theta+\delta)-f(x,\theta)}{\delta} dx$$

A.1.1 Interchanging Limits

Theorem 3 (Dominated Convergence) *Suppose the function h(x,y) is continuous at y_0 for each x, and there exists a function g(x) satisfying*

1. $|h(x,y)| \leq g(x)$ *for all x and y,*
2. $\int_{-\infty}^{\infty} g(x)dx < \infty$.

Then

$$\lim_{y\to y_0} \int_{-\infty}^{\infty} h(x,y)dx = \int_{-\infty}^{\infty} \lim_{y\to y_0} h(x,y)dx$$

.

Existence of a dominating function g(x) with a finite integral ensures that the integrals cannot be too badly behaved. We can use dominated convergence theorem to state the interchange of differentiation and integration.

Theorem 4 *Suppose f(x,θ) is differentiable at $\theta = \theta_0$, that is,*

$$\lim_{\delta\to 0} \frac{f(x,\theta+\delta)-f(x,\theta)}{\delta} = \frac{\partial}{\partial\theta} f(x,\theta)|_{\theta=\theta_0}$$

exists for every x and there exists a function $g(x,\theta_0)$ and a constant $\delta_0 > 0$ such that

1. $|\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}| \leq g(x,\theta_0)$, *for all x and* $|\delta| \leq \delta_0$,
2. $\int_{-\infty}^{\infty} g(x,\theta_0)dx < \infty$.

Then

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x,\theta)dx|_{\theta=\theta_0} = \int_{-\infty}^{\infty} [\frac{\partial}{\partial\theta} f(x,\theta)|_{\theta=\theta_0}]dx$$

Typically $f(x,\theta)$ is differentiable at all θ. The condition 1 can be replaced by easier verifying version by mean value theorem:

$$|\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\theta}| = |\frac{\partial}{\partial\theta} f(x,\theta)|_{\theta=\theta'} \leq g(x,\theta) \ \forall \ \theta' \text{ such that } |\theta'-\theta| \leq \delta_0$$

Corollary 2 *Suppose $f(x,\theta)$ is differentiable in θ and there exists a dominated function g(x,θ) integrable. Then*

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x,\theta)dx|_{\theta=\theta_0} = \int_{-\infty}^{\infty} [\frac{\partial}{\partial\theta} f(x,\theta)|_{\theta=\theta_0}]dx$$

Example The derivation of m.g.f at 0 is the moment is justified.

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}E[e^{tX}] = E[\frac{\partial}{\partial t}e^{tX}] = E[Xe^{tX}]$$

Example The moment of exponential distribution E(λ) and N(μ,1) can be quickly constructed from lower moment since:

$$E[X^{n+1}] = \lambda E[X^n] + \lambda^2 \frac{d}{d\lambda}E[X^n]$$

and

$$E[X^{n+1}] = \mu E[X^n] - \frac{d}{d\mu}E[X^n]$$

Remark 7 One have to find the dominated function and verify it is integrable. We omit the details.

A.2 Interchanging Summation and Differentiation/Integration

Theorem 5 *Suppose that the series* $sum_{x=0}^{\infty} h(\theta, x)$ *converges for all* θ *in a real interval (a,b) and*

1. $\frac{\partial}{\partial\theta}h(\theta, x)$ *is continuous in* θ *for each x,*
2. $\sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(\theta, x)$ *converges uniformly on every closed bounded subinterval of (a,b).*

Then

$$\frac{d}{d\theta}\sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(\theta, x).$$

And the convergence will be uniform on [a,b], if given $\epsilon > 0$*, we can find an N such that*

$$n > N \Rightarrow |\sum_{x=0}^{n}\frac{\partial}{\partial\theta}h(\theta, x) - \sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(\theta, x)| < \epsilon \ \forall\theta \in [a, b].$$

Theorem 6 *Suppose the series* $\sum_{x=0}^{\infty} h(\theta, x)$ *converges uniformly on [a,b] and that, for each x,* $h(\theta, x)$ *is a continuous function of* θ*. Then*

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x)d\theta = \sum_{x=0}^{\infty}\int_a^b h(\theta, x)d\theta$$

A stronger result can be given by Tonelli Theorem.

Theorem 7 (Tonelli Theorem) *If* $f_n(x) \geq 0$ *for all n,x then,*

$$\sum\int f_n(x)dx = \int\sum f_n(x)dx$$

.