Statistical Inference I	**Prof. Chin-Tsang Chiang**
Lecture Notes 9	
November 12, 2015	**Scribe: Wei-Chang Lee, Chi-Ning Chou**

0.1 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 1 (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 1 *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* $a \leq \mathbb{E}[g(X)] \leq b$.

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,

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

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:

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

0.2 Moment

Definition 2 (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 3 (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 2 *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: *(*$\Leftarrow$*)*

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

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

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

It implies that

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

■

Property 3 *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. ■

0.3 Moment generating and characteristic function

Definition 4 (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 5 (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 1 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 $(1 + \sin(2\pi \log x))\frac{e^{-(\log x)^2/2}}{\sqrt{2\pi}x}$