§ 1.2. Distribution, expectation and inequalities.

Expectation, also called mean, of a random variable is often referred to as the location or center of the random variable or its distribution. To avoid some non-essential trivialities, unless otherwise stated, the random variables will usually be assumed to take finite values and those taking values $-\infty$ and ∞ are considered as *r.v.*s in extended sense.

(i). *Distribution.*

Recall that, given a probability space $(\Omega, \mathcal{F}, P)$, a random variable (*r.v.*) X is defined as a real-valued function of Ω, satisfying certain measurability condition. The *cumulative distribution function* of X is then

$$F(t) = P(X \le t) = P(\{w \in \Omega : X(w) \le t\}) = P(X^{-1}((-\infty, t])), \quad t \in (-\infty, \infty).$$

$F(\cdot)$ is then a right-continuous function defined on the real line $(-\infty, \infty)$.

Remark. The distribution function of a single *r.v.* may be considered as complete profile/description of the r.v.. The distribution function $F(\cdot)$ defines a probability measure on $(-\infty, \infty)$. This is the *induced measure*, induced by the random variable as a map/function from the probability measure P on $(\Omega, \mathcal{F}, P)$ to $((-\infty, \infty), \mathcal{B}, F)$. In this sense, the original probability space is often left unspecified or seemingly irrelevant when dealing with one single random variable.

We often call a *r.v.* discrete *r.v.* if it takes countable number of values, and call a *r.v.* continuous *r.v.* if the chance it takes any particular value is 0. In statistics, continuous *r.v.* is often, by default, given a density function. In general, continuous *r.v.* may not have a density function (with respect to Legbegue measure). An example is the Cantor measure.

For two random variables X and Y, their joint c.d.f. is

$$F_{X,Y}(t, s) = P(X \le t \,\text{and}\, Y \le s) = P(X^{-1}((-\infty, t]) \cap Y^{-1}((-\infty, s])),\ t, s \in (-\infty, \infty).$$

Joint c.d.f can be extended for finite number of variables in a straightforward fashion. If the (joint) c.d.f. is differentiable, the derivative is then called (joint) density.

(ii). *Expectation.*

Definitions. For a nonnegative *r.v.* X with c.d.f F, its expectation is defined as

$$E(X) \equiv \int_0^\infty x dF(x).$$

In general, let $X^+ = X1_{\{X \ge 0\}}$, $X^- = -X1_{\{X \le 0\}}$,

$$E(X) \equiv E(X^+) - E(X^-).$$

If $E(X^+) = \infty = E(X^-)$, $E(X)$ does not exist.

A more original definition of the expectation is through that of Lebesgue integral: for nonnegative X,

$$\begin{aligned} E(X) &\equiv \int X(w) dP(w) \qquad \text{formally} \\ &\equiv \lim_{m \to \infty} \sum_{k=0}^{\infty} \frac{k}{2^m} P\Big(\frac{k}{2^m} < X \le \frac{n+1}{2^m}\Big). \end{aligned}$$

If X takes ∞ with positive probability, $E(X^+) = \infty$. Note that X has finite mean is equivalent to $E|X| < \infty$. And the mean of X does not exist is the same as $E(X^+) = E(X^-) = \infty$.

The expectation defined above is mathematically an integral or summation with respect to certain probability measure induced by the random variable. In layman's words, it is the weighted "average" of the values taken by the *r.v.*, weighted by chances which sum up to 1.

Some basic properties of expectation:

(1). $E(f(X)) = \int f(x)dF(x)$ where F is the c.d.f. of X.

(2). If $P(X \leq Y) = 1$, then $E(X) \leq E(Y)$. If $P(X = Y) = 1$ then $E(X) = E(Y)$.

(3). $E(X)$ is finite if and only if $E(|X|)$ is finite.

(4). (Linearity) $E(aX + bY) = aE(X) + bE(Y)$.

(5). If $a \leq X \leq b$, then $a \leq E(X) \leq b$.

(iii). *Some typical distributions of random variables.*

(1.) Commonly used discrete distributions:

Bernoulli: $X \sim Bin(1, p)$. $P(X = 1) = p = 1 - P(X = 0)$. $E(X) = p$ and $\text{var}(X) = p(1 - p)$.

Binomial: $X \sim Bin(n, p)$. $X = \sum_{i=1}^{n} x_i$ and x_i are iid with $B(1, p)$ (the number of successes of n Bernoulli trials.

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, ..., n.$$

$E(X) = np$. $\text{var}(X) = np(1 - p)$.

Poisson: $X \sim \mathcal{P}(\lambda)$. $E(X) = \text{var}(X) = \lambda$.

$$P(X = k) = \frac{1}{k!}\lambda^k e^{-\lambda}, \quad k = 0, 1, 2, ...$$

Key fact: $B(n, p) \to \mathcal{P}(\lambda)$ if $n \to \infty$, $np \to \lambda$. (Law of rare events.)

Geometric: $X \sim G(p)$: time to the first success in a series of Bernoulli trials.

$$P(X = k) = (1 - p)^{k-1} p, \qquad k = 1, 2, ...$$

$E(X) = 1/p, \text{var}(X) = (1 - p)/p^2$.

Negative binomial: $X \sim NB(p, r)$: time to the first r successes in a series of Bernoulli trials. Therefore $X = \sum_{j=1}^{r} \xi_j$ where ξ_j are iid $\sim G(p)$.

$$P(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r}, \qquad k = r, r + 1, ...$$

$E(X) = r/p$ and $\text{var}(X) = r(1 - p)/p^2$.

Hyper-geometric: $X \sim HG(r, n, m)$: the number of black balls when r balls are taken without replacement from an urn containing n black balls and m white balls.

$$P(X = k) = \binom{n}{k}\binom{m}{r-k} / \binom{n+m}{k}, \qquad k = 0 \vee (r - m), 1, ..., r \wedge n.$$

$E(X) = rn/(m + n)$ and $\text{var}(X) = rnm(n + m - r)/[(n + m)^2(n + m - 1)]$.

(2) Commonly used continuous distributions:

Uniform: $X \sim Unif[a, b]$

$$f(x) = (b - a)1_{\{x \in [a,b]\}}$$

$E(X) = (a+b)/2$ and $\text{var}(X) = (b-a)^2/12$.

Normal: $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$ and $\text{var}(X) = \sigma^2$. Central limit theorem.

$$f(x) = (2\pi\sigma^2)e^{-(x-\mu)^2/2}, \qquad x \in (-\infty, \infty).$$

Exponential: $X \sim \mathcal{E}(\lambda)$. Density:

$$f(x) = e^{-x/\lambda}/\lambda, \qquad x > 0$$

$E(X) = \lambda$ and $\text{var}(X) = \lambda^2$. No memory: $(X-t) \,|\, \{X \geq t\} \sim \mathcal{E}(\lambda)$.

Gamma: $\Gamma(\alpha, \gamma)$. Density:

$$f(x) = \frac{1}{\Gamma(\alpha)\gamma} x^{\alpha-1} e^{-x/\gamma}, \qquad x > 0.$$

$\mathcal{E}(\lambda) = \Gamma(1, \lambda)$, $\chi^2_n = \Gamma(n/2, 2)$. Sum of independent $\Gamma(\alpha_i, \gamma)$ follows $\Gamma(\sum_i \alpha_i, \gamma)$.

Beta: $B(\alpha, \beta)$. Density:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad x \in [0,1]$$

$\xi/(\xi+\eta) \sim B(\alpha, \beta)$ where $\xi \sim \Gamma(\alpha, \gamma)$ and $\eta \sim \Gamma(\beta, \gamma)$ are independent. $X_{(k)} \sim B(k-1, n-k+1)$ as the k-th smallest of $X_1, ..., X_n$ iid $\sim Unif[0,1]$

Cauchy: density $f(x) = 1/[\pi(1+x^2)]$. Symmetric about 0, but expectation and variance not exist.

χ^2_n (with d.f. n): sum of n i.i.d standard normal r.v.s. χ^2_2 is $\mathcal{E}(2)$.

t_n (with d.f n): $\xi/\sqrt{\eta/n}$ where $\xi \sim N(0,1)$, $\eta \sim \chi^2_n$ and ξ and η are independent.

$F_{m,n}$ (with d.f. (m,n)): $(\xi/m)/(\eta/n)$ where $\xi \sim \chi^2_m$, $\eta \sim \chi^2_n$ and ξ and η are independent.

(iv). Some basic inequalities:

Inequalities are extremely useful tools in theoretical development of probability theory. For simplicity of notation, we use $\|X\|_p$, which is also called L_p norm if $p \geq 1$, to denote $[E(|X|^p)]^{1/p}$ for a r.v. X. In what follows, X and Y are two random variables.

(1) *the Jensen inequality*: Suppose $\psi(\cdot)$ is a convex function and X and $\psi(X)$ have finite expectation. Then $\psi(E(X)) \leq E(\psi(X))$.

Proof. Convexity implies for every a, there exists a constant c such that $\psi(x) - \psi(a) \geq c(x-a)$. Let $a = E(X)$ and $x = X$, the right hand side is mean 0. So Jensen's inequality follows. □

(2). *the Markov inequality*: For any $a > 0$, $P(|X| \geq a) \leq 1/aE(|X|)$.

Proof. $aP(|X| \geq a) = E(a1_{\{|X| \geq a\}} \leq E(|X|1_{\{|X| \geq a\}}) \leq E(|X|)$. □

(3). *the Chebyshev (Tchebychev) inequality*: for $a > 0$,

$$P(|X - E(X)| \geq a) \leq \text{var}(X)/a^2$$

Proof. The inequality holds if $\text{var}(X) = \infty$. Assume $\text{var}(X) < \infty$, then $E(X)$ is finite and $Y \equiv (X - E(X))^2$ is well defined. It follows from the Markov inequality that

$$P(|X - E(X)| \geq a) = P(Y \geq a^2) \leq E(Y)/a^2 = \text{var}(X)/a^2.$$

□

(4). *the Hölder inequality*: for $1/p + 1/q = 1$ with $p > 0$ and $q > 0$,

$$E|XY| \le \|X\|_p \|Y\|_q$$

Proof. Observe that for any two nonnegative numbers a and b, $ab \le a^p/p + b^q/q$. (This is a result of the concavity of the log-function. please DIY.) Let $a = |X|/\|X\|_p$ and $b = |Y|/\|Y\|_q$ and take expectation on both sides. The Hölder inequality follows. □

(5). *the Schwarz inequality*:

$$E(|XY|) \le [E(X^2)E(Y^2)]^{1/2}.$$

Proof. A special case of the Hölder inequality. □

(6). *the Minkowski inequality*: for $p \ge 1$,

$$\|X + Y\|_p \le \|X\|_p + \|Y\|_p.$$

Proof. If $p = 1$, the inequality is trivial. Assume $p > 1$. Let $q = p/(p-1)$. Then $1/p + 1/q = 1$. By the Hölder inequality,

$$E[|X||X+Y|^{p-1}] \le \|X\|_p \||X+Y|^{p-1}\|_q = \|X\|_p \{E[|X+Y|^{(p-1)q}]\}^{1/q} = \|X\|_p \{E[|X+Y|^p]\}^{(p-1)/p}.$$

Likewise,

$$E[|Y||X+Y|^{p-1}] \le \|Y\|_p \{E[|X+Y|^p]\}^{(p-1)/p}.$$

Summing up the above two inequalities leas to

$$E(|X+Y|^p) \le (\|X\|_p + \|Y\|_p)\{E[|X+Y|^p]\}^{(p-1)/p},$$

and the Minkowski inequality follows. □

Remark. Jensen's inequality is a powerful tool. For example, straightforward applications include

$$[E(|X|)]^p \le E(|X|^p), \quad \text{for } p \ge 1,$$

which implies

$$\|X\|_p \le \|Y\|_q, \quad \text{for } 0 < p < q.$$

Moreover,

$$E(\log(|X|)) \le \log(E(|X|)).$$

If $E(X)$ exists,

$$E(e^X) \ge e^{E(X)}.$$

These inequalities are all very commonly used. For example, the validity of the maximum likelihood likelihood estimation essentially rests on the fact,

$$E\log\Big(\frac{f_\theta(X)}{f_{\theta_0}(X)}\Big) \le \log E\Big(\frac{f_\theta(X)}{f_{\theta_0}(X)}\Big) = \log\Big(\int \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x)dx\Big) = \log\Big(\int f_\theta(x)dx\Big) = \log(1) = 0,$$

which is a result of Jensen's inequality. Here $f_\theta(\cdot)$ is a parametric family of density of X with θ_0 being the true value of θ.

The Markov inequality, despite its simplicity, shall be frequently used in the order of a sequence of random variables, especially when coupled with the technique of truncation. The Chebyshev inequality is so mighty that, as an example, it directly proves the weak law of large numbers.

The Schwarz inequality shows that covariance is an inner product, and, furthermore, the space of mean 0 *r.v.*s with finite variances forms a Hilbert space. The Minkowsky inequality is the triangle inequality for L_p norm, without which L_p cannot be a norm.

DIY Exercises.

Exercise 1.5. ⋆⋆ Suppose X is a r.v. taking values on all rational numbers on $[0,1]$, Specifically, $P(X = q_i) = p_i > 0$ where $q_1, q_2, ...$ denotes all rational numbers on $[0,1]$. Then, the c.d.f of X is continuous at irrational numbers and discontinuous at rational numbers.

Exercise 1.6. ⋆⋆⋆ Show $\text{var}(X^+) \leq \text{var}(X)$ and $\text{var}(\min(X, c)) \leq \text{var}(X)$ where c is any constant.

Exercise 1.7. ⋆ (*Generalizing Jensen's inequality*). Suppose $g(\cdot)$ is a convex function and X is a random variable with finite mean. Then, for any constant c,

$$Eg(X - E(X) + c) \geq g(c).$$

Exercise 1.8. ⋆⋆⋆ *Lyapunov (Liapounov)* : Show that the function $\log E(|X|^p)$ is a convex function of p on $[0, \infty)$. Or, equivalently, for any $0 < s < m < l$, show

$$E(|X|^m) \leq [E(|X|^s)]^r [E(|X|^l)]^{1-r}$$

where $r = (l - m)/(l - s)$. (Hint: use the H‡older inequality on

$$E(|X|^{\lambda p_1 + (1-\lambda)p_2}) \leq [E(|X|^{p_1})]^{\lambda} [E(|X|^{p_2})]^{1-\lambda}$$

for positive p_1, p_2 and $0 < \lambda < 1$.)