Expectation

$Y = g(X)$ is a continuous function.

If X is discrete with $p_k = P\{X = x_k\}$, then $E(Y) = E[g(X)] = \sum_{k=1}^{\infty} g(x_k) p_k$.

If X is continuous with density function $f(x)$, then $E(Y) = E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$.

$Z = g(X, Y)$ is a continuous function.

If X, Y are discrete $p_{ij} = P\{X = x_i, Y = y_j\}$, then

$$E(Z) = E[g(X,Y)] = \sum_1^{\infty} \sum_1^{\infty} g(x_i, y_j) p_{ij} .$$

If X and Y are continuous with density function $f(x, y)$, then

$$E(Z) = E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dxdy$$

Properties:

(1) $E(CX) = CE(X)$.

(2) $E(X+Y) = E(X) + E(Y)$.

(3) If X and Y are independent, then $E(XY) = E(X)E(Y)$.

Variance

$$Var(X) = E\{[X - E(X)]^2\} = E(X^2) - \{E(X)\}^2 .$$

Properties

(1) $Var(CX) = C^2 Var(X)$.

(2) if X and Y are independent, then $Var(X+Y) = Var(X) + Var(Y)$.

Covariance

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

$$\rho_{XY} = Cov(X,Y) \big/ \sqrt{Var(X)} \sqrt{Var(Y)}$$

$$Var(X+Y) = Var(X) + Var(Y) + 2 * Cov(X,Y)$$

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Properties:

(1) $Cov(aX, bY) = abCov(X, Y)$

(2) $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$

Discrete Random Variables

1. Bernoulli(p), $E(X) = p$, $Var(X) = p(1-p)$.

2. $Bin(n, p)$: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $E(X) = np$, $Var(X) = np(1-p)$.

3. $Poisson(\lambda)$. $f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$, $E(X) = \lambda$, $Var(X) = \lambda$.

4. $Geometric(p)$, # trials until first head. $f(x) = p\ (1-p)^{k-1}$.

$E(X) = 1/p$, $Var(X) = (1-p)/p^2$.

5. Negative Binomial: # trials until r-th head.

$f(x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r$. $E(X) = r/p$, $Var(X) = r(1-p)/p^2$

Continuous Random Variables

1. $U(a,b)$, $f(x) = \begin{cases} 1/(b-a), a < x < b \\ 0, \quad o.w. \end{cases}$, $E(X) = (b+a)/2$, $Var(X) = (b-a)^2/12$.

2. $Exp(\lambda)$. $f(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, x \ge 0 \\ 0, \quad o.w. \end{cases}$.

$F(x) = \begin{cases} 1 - e^{-\lambda x}, x \ge 0 \\ 0, \quad o.w. \end{cases}$, $E(X) = 1/\lambda$, $Var(X) = 1/\lambda^2$.

3. $N(\mu, \sigma^2)$. $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

For $Z \sim N(0,1)$, $E(Z^k) = \begin{cases} 0, \quad k \text{ is odd} \\ (k-1)!!, k \text{ is even} \end{cases}$

For $X \sim N(\mu, \sigma^2)$, $X = \mu + \sigma \cdot Z$. $E(X^2) = E(\mu^2 + 2\mu\sigma Z + \sigma^2 Z^2) = \mu^2 + \sigma^2$.

4. $Gamma(\lambda, t)$. $f(x) = \frac{1}{\Gamma(t)} \cdot \lambda^t \cdot x^{t-1} \cdot e^{-\lambda x}$, for $x \geq 0$. $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$.

$E(X) = t/\lambda, Var(X) = t/\lambda^2$.

$X \sim Gamma(\lambda, t) \Leftrightarrow \lambda X \sim Gamma(1, t)$

About $\Gamma(t)$: (1) $\Gamma(1) = 1$;

(2) $\Gamma(z+1) = z\Gamma(z)$;

(3) $\Gamma(k+1) = k!$ for integer k.

(4) $\Gamma(1/2) = \pi$.

Related with other distributions:

(1) $Gamma(\lambda, 1) = Exp(\lambda)$.

(2) $\chi_k^2 = Gamma(1/2, k/2)$

(3) if $t \to \infty$, then $(X - E(X))/\sqrt{\text{var}(X)} \to N(0,1)$

5. Cauchy. $f(x) = \frac{1}{\pi(1+x^2)}, x \in R$. $E(X)$ and $Var(X)$ are not well defined.

6. $Beta(a,b)$. $f(x) = \frac{1}{\beta(a,b)} \cdot x^{a-1} \cdot (1-x)^{b-1}, 0 \leq x \leq 1, a > 0, b > 0$.

$\beta(a,b) = \int_0^1 x^{a-1} \cdot (1-x)^{b-1} dx$. When a=1, b=1, it is U(0,1).

7. $Weibull(\alpha, \beta)$. $F(x) = 1 - e^{-\alpha x^\beta}$ $x \geq 0, \alpha > 0, \beta > 0$. $f(x) = \alpha\beta \cdot x^{\beta-1} e^{-\alpha x^\beta}$.

$\beta = 1 \to Exp(\alpha)$

Distribution of Functions

(1) $\begin{cases} x_1 = X_1(y_1, y_2) \\ x_2 = X_2(y_1, y_2) \end{cases}$

(2) $J(y_1, y_2) = \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_2/\partial y_1 \\ \partial x_1/\partial y_2 & \partial x_2/\partial y_2 \end{vmatrix}$

(3) $f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}\{x_1(y_1,y_2), x_2(y_1,y_2)\} | J |$.

Bivariate Normal Distribution

$$f_{XY}(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}]\}$$

(1) From μ and Σ, to get coefficient matrix $\begin{cases} X_1=\mu_1+\sigma_1 Z_1 \\ X_2=\mu_2+\sigma_2(\rho Z_1+\sqrt{1-\rho^2}Z_2) \end{cases}$

(2) From coefficient matrix, to get μ and Σ. $X \sim N(\mu,\Sigma)$,

$$g(X)=CX+d \sim N(C\mu+d, C\Sigma C')$$

(3) $(X_1,X_2)\sim N(\mu,\Sigma)$,

then $X_2 \mid X_1 \sim N(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1),\sigma_2^2(1-\rho^2))$,

$$X_2 \mid X_1 \sim N(\mu_2+\Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1),\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

(4) $(X_1,X_2)\sim N(\mu,\Sigma)$, then $X_1 \perp X_2 \Leftrightarrow \rho=0$,

(5) Marginal $f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\}$

(6) Multivariate Normal: $f_X(x)=\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\},\Sigma=\mathrm{var}(X)$

Gamma, Beta

$X_1 \perp X_2$, $X_1 \sim Gamma(\lambda,p)$, $X_2 \sim Gamma(\lambda,q)$, $X_1+X_2 \sim Gamma(\lambda,p+q)$,

$\frac{X_1}{X_1+X_2} \sim Beta(p,q)$ (can also extend to Dirichlet)

χ_k^2, T, F

$$\sum_{i=1}^k Z_i^2 \sim \chi_k^2 = Gamma(1/2,k/2)$$

$$t_k = N(0,1)\Big/\sqrt{\chi_k^2/k}$$

$$F_{m,n}=\chi_m^2/m\Big/\chi_n^2/n$$

Sample mean and Sample variance

$X_1, X_2, ..., X_n$ are iid $N(0,\sigma^2)$: (1) $\bar{X} \sim N(0,\sigma^2/n)$ (2) $\bar{X} \perp S^2$ (3) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$X_1, X_2, ..., X_n$ are iid $N(\mu,\sigma^2)$: (1) $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$ (2) $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1)$

Central Limit Theorem

If $X_1, \cdots, X_n$ are iid with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ which is finite, then

$Z_n = \sqrt{n}(\bar{X}-\mu)/\sigma$ converges to $N(0,1)$.

Convergence modes:

(1) In Law/Distribution: CDF converges

(2) In probability: $P(|Z_n - Z| > \varepsilon) = 0$

(3) Almost surely (with probability 1): $P(w: \lim_{n\to\infty} X_n(w) = X(w)) = 1$

(4) In r-th moment: $E(|X_n - X|^r) \to 0$, higher order convergence -> lower order convergence

Ex: $\bar{X}_n \xrightarrow{2} \mu$ because $E(|\bar{X}_n - \mu|^2) = Var(\bar{X}_n) = \sigma^2/n \to 0$

When is (1) and (2) are equivalent?

If $Z_n \xrightarrow{D} C$, a constant, then $Z_n \xrightarrow{P} C$.

Continuous mapping theorem

If $Z_n \xrightarrow{P} Z$, and $g(.)$ is a continuous function, then $g(Z_n) \xrightarrow{P} g(Z)$

Slutsky theorem

If $Z_n \xrightarrow{D} Z, U_n \xrightarrow{P} C$, then (1) $Z_n + U_n \xrightarrow{D} Z + C$ (2) $Z_n U_n \xrightarrow{D} ZC$

Ex: X_i iid with mean μ, variance σ^2. The limiting distribution of

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S} = \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma}\frac{1}{S/\sigma} \to N(0,1)*1.$$

Law of Large Number

Weak LLN $\overline{X}_n \xrightarrow{D,P} \mu$

Strong LLN $\overline{X}_n \xrightarrow{A.S.} \mu$ only if $E(|X|) < \infty$

Decision Theory

How do we use data to answer questions about the world, about the parameter θ.

Elements of Decision Theory:

(1) State space (parameter space) Θ, $\theta \in \Theta$.

(2) Action Space A

- (i) In estimate, we guess the value of θ.
- (ii) In testing, we decide $\theta \in \Theta_0$ or $\theta \in \Theta_1$, $\mathrm{A} = \{0,1\}$
- (iii) In prediction, $\mathrm{A} = \{all\ \ possible\ mapping\ functions\}$.

(3) Loss function $L(\theta, a)$

- (i) In estimate, $L(\theta, a) = \{a - q(\theta)\}^2$ quadratic loss or

  $L(\theta, a) = |a - q(\theta)|$ absolute loss

- (ii) In testing, 0-1 loss. $L(\theta, a) = \begin{cases} 0, if\ correct \\ 1,\ otherwise \end{cases}$

- (iii) In prediction, mean-square error $E\{\{Y - g(x)\}^2\}$

Decision Rule: A decision rule is a function from data X to an action in A, namely $\delta(X) \in A$.

The **risk of a decision rule**: $R(\theta, \delta) = E[l(\theta, \delta(X))]$

$$R(\theta,\delta) = E[l(\theta,\delta(X))] = \begin{cases} E_X[l(\theta,\delta(X))],\ frequentist \\ E_{X|\theta}[l(\theta,\delta(X))],\ Bayesian \end{cases}$$

Maximum risk: $R_M(\delta) = \sup_{\theta} R(\theta, \delta)$

Bayes risk: $R_B(\delta) = \int R(\theta,\delta)\pi(\theta)d\theta = E_{(X,\theta)}[l(\theta,\delta(X))]$

A decision rule minimizing maximum risk is **minimax rule**.

A decision rule minimizing Bayes risk is **Bayes rule**.

Bias-Variance decomposition

$$E[(x-\alpha)^2] = E[x^2] - 2E[X]\alpha + \alpha^2 = E[x]^2 - 2E[X]\alpha + \alpha^2 + E[x^2] - 2E[x]^2 + E[x]^2$$
$$= (E[X]-a)^2 + E[(x-E[X])^2]$$

The first term is the squared bias of the estimate.
The second term is the variance of the estimate.

Hypothesis test

Null space Θ_0, Alternative space Θ_1. Hypothesis test is to test whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$.

Type 1 error, type 2 error. Power.