1 Probability

1.1 Combinatorics

Repetition, ordered : n^k

No repetiton, Ordered : permutation of length k of n distinct object, $\frac{n!}{(n-k)!} = n(n-1)\dots(n-k+1)$

Repetition, not ordered : $\binom{n-1+k}{k}$, number of vectors of $n_1+\dots+n_k = n$, $n_i >= 0\,\forall i$

No repetition, not ordered : combination of k objects from set of n distinct ones, $C_n^k = \binom{n}{k} = \frac{n!}{(n-k)!k!}$

Permutations : ordered set, given $n = \sum_{i=1}^r n_i$ objects of r different types : $\frac{n!}{n_1!n_2!\dots n_r!}$

Combinations : non ordered selection, ways of distributing n distinct objects in r distinct groups of size $n_1,\dots,n_r$: $\frac{n!}{n_1!n_2!\dots n_r!}$

Properties : $\binom{n}{k} = \binom{n}{n-k}$, $\binom{n+1}{k} = \binom{n}{r-1} + \binom{n}{r}$, $\sum_{j=0}^r \binom{m}{j}\binom{n}{r-j} = \binom{m+n}{r}$, $(a+b)^n = \sum_{r=0}^n \binom{n}{r} a^r b^{n-r}$, $(1-x)^{-n} = \sum_{j=0}^\infty \binom{n+j-1}{j} x^j$ $(|x|<1)$, $\lim_{n\to\infty} n^{-r}\binom{n}{r} = \frac{1}{r!}$ $(r\in\mathbb{N})$

Partition int. : number of vectors of $n_1+\dots+n_r = n$, $n_i > 0\,\forall i$: $\binom{n-1}{r-1}$

Geometric series : $\sum_{i=0}^n a\theta^i = ,\theta \neq 1 :$ $a\frac{1-\theta^{n+1}}{1-\theta}$, $\theta = 1 : a(n+1)$; $|\theta|<1 : \sum_{i=0}^\infty \theta^i = \frac{1}{1-\theta}$

Exponential series : $\exp(x) = \sum_{n=0}^\infty \frac{x^n}{n!}$

1.2 Probability Spaces

Probability space : $(\Omega,\mathcal{F},P)$, Ω sample space (universe) of all possible results ω(non-empty), $\mathcal{F}$ event space (non-empty) : events (subsets of Ω), $P : \mathcal{F}\to[0,1]$ probability distribution

Equiprobable : finite Ω, $\forall\omega\in\Omega\; P(\omega) = \frac{1}{|\Omega|}$, $\forall A\subset\Omega\; P(A) = \frac{|A|}{|\Omega|}$

Event space : $A\in\mathcal{F}\Rightarrow A^c\in\mathcal{F}$, $\{A_i\}_{i=1}^\infty \in \mathbb{F} \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathcal{F}$, $\Omega\in\mathcal{F}$, $\emptyset\in\mathcal{F}$

Probability distribution : $A\in\mathcal{F}\Rightarrow 0\le P(A)\le 1$, $P(\Omega) = 1$, P of pairwise disjoint events $= \sum_{i=1} P(A_i)$, continuous set function

Limits : $A_1\subset A_2\subset\dots\Rightarrow \lim_{n\to\infty} P(A_n) = P(\bigcup_{i=1}^\infty A_i)$; $A_1\supset A_2\supset\dots\Rightarrow\lim_{n\to\infty}P(A_n) = P(\bigcap_{i=1}^\infty P(A_i))$

Properties : $P(\emptyset) = 0$, $P(A^c) = 1-P(A)$, $P(A\cup B) = P(A)+P(B)-P(A\cap B)$, $A\subset B\Rightarrow P(A)\le P(B)\,\&\,P(B\setminus A) = P(B)-P(A)$, $P(\bigcup_{i=1}^\infty A_i)\le\sum_{i=1}^\infty P(A_i)$

Inclusion-exclusion : $P(\bigcup_{i=1}^n A_i) = \sum_{r=1}^n (-1)^{r+1}\sum_{1\le i_1<\dots<i_r\le n} P(A_{i_1}\cap\dots\cap A_{i_r})$

Union bound : $P(\bigcup_{i=1}^{n+1} A_i)\le\sum_{i=1}^{n+1}P(A_i)$

1.3 Conditional Probability

Conditional probability : A given B, $P(A|B)P(B) = P(A\cap B)$, $P(A) = P(A\cap B) + P(A\cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$

Distributions : $B\in\mathcal{F}$ s.t. $P(B)>0$, $Q(A) = P(A|B)$, $(\Omega,\mathcal{F},Q)$ is probability space

Total probability TP : $\{B_i\}_{i=1}^\infty$ $(B_i\cap B_{j\neq i} = \emptyset)$, $A\subset\bigcup_{i=1}^\infty B_i$, $P(A) = \sum_{i=1}^\infty P(A\cap B_i)$

Bayes : TP cond.s, $P(A)>0$; $P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^\infty P(A|B_i)P(B_i)}$, $P(A|B)P(B) = P(B|A)P(A)$

Multiple conditioning : $P(\bigcap_{i=1}^n A_i) = P(A_1)\prod_{i=2}^n P(A_i|\bigcap_{j=1}^{i-1}A_j)$

1.4 Independence

$A\perp\!\!\!\perp B\Leftrightarrow P(A|B) = P(A)\Leftrightarrow P(A\cap B) = P(A)P(B)$

Mutually : $P(\bigcap_F A_i) = \prod_F P(A_i)$

Pairwise : $P(A_{1\le i}\cap A_{i\le j\le n}) = P(A_i)P(A_j)$

Conditional : $P(\bigcap_F A_i|B) = \prod_F P(A_i|B)$

1.4.1 Circuits

Parallel : $P_P(S) = P(\bigcap_{i=1}^n F_i) = \prod_{i=1}^n p_i$

Series : $P_S(S) = P(\bigcup_{i=1}^n F_i) = 1-\prod_{i=1}^n(1-p_i)$

$\exists p_+,p_-, 1>p_+>p_i>p_->0, n\to\infty\Rightarrow P_P(S)\to 0$, $P_S(S)\to 1$

2 Random Variables

Random variable r.v. : $X:\Omega\to\mathbb{R}$

Support : $D_X = \{x\in\mathbb{R} : \exists\omega\in\Omega$ s.t. $X(\omega) = x\}$, D_X countable $\Rightarrow$ X discrete r.v.

Associate prob. : $P(X\in S) = P(\{x\in\Omega : X(\omega)\in S\})$, $A_x = \{\omega\in\Omega : X(\omega) = x\}$, must have $\forall x\; A_x\in\mathcal{F}$ for $P(X=x)$

Bernoulli/Indicator r.v. : binary 0 or 1

Probability Mass Function PMF : X discrete, $f_X(x) = P(X=x) = P(A_x)$, $x\in\mathbb{R}$; $f_X(x)\ge 0$, $x\in D_X$ $f_X(x)\ge 0\Leftrightarrow D_X$ is support of f_X, $\sum_{x\in D_x} f_X(x) = 1$; $f_X\equiv f$, $D_X\equiv D$

Binomial r.v. : PMF $f_X(x) = \binom{n}{x}p^x(1-p)^{n-x}$; $x\in[0,n]\subset\mathbb{N}$, $n\in\mathbb{N}$, $p\in[0,1]\subset\mathbb{R}$; $X\sim B(n,p)$, n denominator, p prob. of success, $n=1$: Bernoulli var. (nb. success with n peeks)

Geometric distr. : PMF $f_X(x) = p(1-p)^{x-1}$; $x\in\mathbb{N}^*$, $p\in[0,1]\subset\mathbb{R}$; $X\sim\text{Geom}(p)$, p : prob. of success, (nb. trials till success)

Memorylessness : $X\sim\text{Geom}(p)\Rightarrow P(X>n+m|X>m) = P(X>n)$

Negative binomial distr. : PMF $f_X(x) = \binom{x-1}{n-1}p^n(1-p)^{x-n}$; $x\in[n:)\subset\mathbb{N}$, $p\in[0,1]\subset\mathbb{R}$; $X\sim\text{NegBin}(n,p)$, $n=1 : X\sim\text{Geom}(p)$, (nb. trials till n successes)

Gamma func. : $\alpha > 0$, $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}\,du$; $\Gamma(1) = 1$, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(n) = (n-1)!$, $\Gamma(1/2) = \sqrt{\pi}$

Hypergeometric distr. : $P(X=x) = \frac{\binom{w}{x}\binom{b}{m-x}}{\binom{w+b}{m}}$, $X\sim\text{HyperGeom}(w,b,m)$; $x\in[\max(0,m-b),\min(w,m)]\subset\mathbb{N}$, (nb. w found with m take in $\{w+b\}$)

Discrete uniform : X discrete uniform, PMF $f_X(x) = \frac{1}{b-a+1}$; $x\in[a,b]\subset\mathbb{N}$, $a<b$, $a,b\in\mathbb{Z}$, $U\sim DU(a,b)$, a : start, b : end, (res from uniform pick in $[a,b]$)

Poisson distr. : PMF $f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}$; $x\in\mathbb{N}$, $\lambda>0$, $X\sim\text{Pois}(\lambda)$, λ mean nb. per interval, (nb. of events during time)

Cumulative distr. func. CDF : $F_X(x) = P(X\le x)$, $x\in\mathbb{R}$; X discrete : $F_X(x) = \sum_{x_i\in D_X, x_i\le x} P(X=x_i)$

CDF properties : $\lim_{x\to-\infty}F_X(x) = 0$, $\lim_{x\to\infty}F_X(x) = 1$, F_X non-decreasing, $\lim_{x\to 0^+}F_X(x+t) = F_X(x)$, $P(X>x) = 1-F_X(x)$, $x<y\Rightarrow P(x<X\le y) = F_X(y)-F_X(x)$

CDF descrete : X discrete, $f(x) = F(X) - \lim_{y\to x^-}F(y)$, $D_X\subset\mathbb{Z}\Rightarrow f(x) = F(x)-F(x-1)$, $x\in\mathbb{Z}$

Transformations discrete : $Y = g(X)$, $f_Y(y) = \sum_{x:g(x)=y} f_X(x) = P(Y=y) = \sum_{x:g(x)=y}P(X=x)$

2.1 Expectation

Expectation : X discrete s.t. $\sum_{x\in D_X}|x|f_X(x) < \infty$, $E[X] = \sum_{x\in D_X} xP(X=x) = \sum_{x\in D_X} xf_X(x)$

Function : g real-valued func., $\sum_{x\in D_X}|g(x)|f_X(x) < \infty$, $E[g(X)] = \sum_{x\in D_x} g(x)f_X(x)$

Properties : if $E[\cdot]$ finite; $E[aX+b] = aE[X]+b$, $E[g(X)+h(X)] = E[g(X)]+E[h(X)]$, $P(X=b) = 1\Rightarrow E[X] = b$, $E[X]^2\le E[X^2]$

Moment : PMF $f(x)$ s.t. $\sum_x |x|^r f(x) < \infty$: r-th moment $E[X^r]$, r-th central moment $E[(X-E[X])^r]$, variance $\text{var}(X) = E[(X-E[X])^2]$, standard deviation $\sqrt{\text{var}(X)}$, r-th factorial moment $E[X(X-1)\dots(x-r+1)]$

Variance properties : $\text{var}(X) = E[X^2]-E[X]^2 = E[X(X-1)]+E[X]-E[X]^2$, $\text{var}(aX+b) = a^2\text{var}(X)$, $\text{var}(X) = 0\Rightarrow X$ constant with prob. 1

Variance properties : X takes $0,1,\dots,r\ge 2$, $E[X]<\infty$; $E[X] = \sum_{x=1}^\infty P(X\ge x)$, $E[X(X-1)\dots(X-r+1)] = r\sum_{x=r}^\infty(x-1)\dots(x-r+1)P(X\ge x)$

2.2 Conditional Probability Distributions

$B\in\mathcal{F}$ s.t. $P(B)>0$, $f_X(x\mid B) = P(X=x\mid B) = P(A_x\cap B)/P(B)$

Properties : $f_X(x\mid B)\ge 0$, $\sum_x f_X(x\mid B) = 1$

Event : B is event like $X\in\mathcal{B}$, $\mathcal{B}\subset\mathbb{R}$; $f_X(x\mid B) = P(X=x, X\in\mathcal{B})/P(X\in\mathcal{B}) = I(x\in\mathcal{B})/P(X\in\mathcal{B})f_X(x)$

Conditional expected value : $\sum_x |g(x)|f_X(x\mid B)<\infty$, $E[g(X)|B] = \sum_x g(x)f_X(x\mid B)$

Expected sum : $P(B^c)>0$, $E[X] = E[X\mid B]P(B)+E[X\mid B^c]P(B^c)$; $\{B_i\}_{i=1}^\infty$ partition Ω, $\forall i\; P(B_i)>0$, $E[X] = \sum_{i=1}^\infty E[X\mid B_i]P(B_i)$

2.3 Notions of Convergence

Convergence of distr.s : $(\{X_n\},X)$, rdm. var.s with CDFs $(\{F_n\},F)$, $\{X_n\}$ converge in distr. (law) to X $(X_n\xrightarrow{D}X)\Leftrightarrow\forall x\in\mathbb{R}$ s.t. F continuous : $F_{n\to\infty}(x)\to F(x)$; $D_X\subset\mathbb{Z}$, $\forall x\; f_{n\to\infty}(x)\to f(x)\Leftrightarrow F_n(x)\to F(x)$

Law small nb. : $X_n\sim B(n,p_n)$, $np_n\to\lambda>0$ $(n\to\infty)$, $X_n\xrightarrow{D}X$, $X\sim\text{Pois}(\lambda)$

3 Continuous Random Variables (CRV)

Ω, D_X is not countable

Probability densitiy func. : X is continuous $\Leftrightarrow\exists f(x)$ (PDF) s.t. $P(X\le x) = F(x) = \int_{-\infty}^x f(u)\,du$, $x\in\mathbb{R}$; $f(x)\ge 0$, $\int_{-\infty}^\infty f(x)\,dx = 1$; $f(x) = \frac{dF(x)}{dx}$, $P(X=x) = 0$

Uniform distr. : $a<b$, $f(u) = \frac{1}{b-a}$ $a\le u\le b$, 0; $U\sim U(a,b)$; finite interval, equal probability

Exponential distr. : $\lambda>0$, $f(x) = \lambda e^{-\lambda x}$ $x>0$, 0; $X\sim\exp(\lambda)$, memorylessness; waiting time, positive quantities

Gamma distr. : $\alpha,\lambda>0$, $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$ $x>0$, 0; $X\sim\text{Gamma}(\alpha,\lambda)$; λ rate, $\alpha = 1$ exp. density; more flexible

Laplace distr. : $x\in\mathbb{R}$, $\eta\in\mathbb{R}$, $\lambda>0$, $f(x) = \frac{\lambda}{2}e^{-\lambda|x-\eta|}$

Pareto dis. : $\alpha,\beta>0$ $F(x) = 1-\left(\frac{\beta}{x}\right)^\alpha$ $x\ge\beta$, 0; lie in (β,∞), financial loss, threshold

Moments : $E[|g(X)|]<\infty$, $E[g(X)] = \int_{-\infty}^\infty g(x)f(x)\,dx$, $E[X] = \int_{-\infty}^\infty xf(x)\,dx$, $\text{var}(X) = \int_{-\infty}^\infty(x-E[X])^2f(x)\,dx$

Conditional densities : $A\subset\mathbb{R}$, $F_X(x\mid X\in A) = \frac{\int_{A_x}f(y)\,dy}{P(X\in A)}$, $A_x = \{y : y\le x, y\in A\}$; $f_X(x\mid X\in A) = \frac{f_X(x)}{P(X\in A)}$ $x\in A$, 0; $E[g(X)\mid X\in A] = \frac{E[g(X)\mathbb{1}(X\in A)]}{P(X\in A)}$

Quantile p : $0<p<1$, $x_p = \inf\{x : F(x)\ge p\}$; most CRVs $x_p = F^{-1}(p)$ unique, $P(X\le x_p) = p$, 0.5 quantile : median of F

Transformations : $g:\mathbb{R}\to\mathbb{R}$, $B\subset\mathbb{R}$, $g^{-1}(B)\subset\mathbb{R}$ set s.t. $g(g^{-1}(B)) = B$

Theorem : $Y = g(X)$, $B_y = (-\infty,y]$, X CRV, $F_Y(y) = P(Y\le y) = \int_{g^{-1}}f_X(x)\,dx$, $g^{-1}(B_y) = \{x\in\mathbb{R} : g(x)\le y\}$; g monotone, g^{-1} differentiable, $y\in\mathbb{R}$, $F_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X(g^{-1}(y))$

3.1 Normal distribution

Normal/Gaussian distribution : $x\in\mathbb{R}$, $\mu\in\mathbb{R}$, $\sigma>0$, $f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$; expectation μ, variance σ^2, std. $\sigma>0$, $X\sim\mathcal{N}(\mu,\sigma^2)$; average many small effects, measure subject to error

Standard normal : $Z\sim\mathcal{N}(0,1)$ standard normal, density $\phi(z) = \sqrt{2\pi}e^{-z^2/2}$, $z\in\mathbb{R}$, $F_Z(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{-z^2/2}\,dz$, $f(x) = \sigma^{-1}\phi((x-\mu)/\sigma)\,\forall x\in\mathbb{R}$

Standard normal density : bell curve $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$, $z\in\mathbb{R}$

Properties : density $\phi(z)$, CDF $\Phi(z)$, quantiles z_p : of $Z\sim\mathcal{N}(0,1)$, $\forall z\in\mathbb{R}$: $\phi(z) = \phi(-z)$, $P(Z\le z) = \Phi(z) = 1-\Phi(-z) = 1-P(Z\ge z)$, $z_p = -z_{1-p}$ $0<p<1$, $z^r\phi(z_{\to\pm\infty})\to 0$ $r>0$ so $E[Z^r]$ exists $\forall r\in\mathbb{R}$, $\phi'(z) = -z\phi(z)$, $\phi''(z) = (z^2-1)\phi(z)$, $E[Z] = 0$, $\text{var}(Z) = 1$, $E[Z^3] = 0$, $X\sim N(\mu,\sigma^2)\Rightarrow Z = (X-\mu)/\sigma\sim N(0,1)$, $X = \mu+\sigma Z$

Moivre-Laplace : $X_n\approx B(n,p)$, $0<p<1$, $\mu_n = E[X_n] = np$, $\sigma_n^2 = \text{var}(X_n) = np(1-p)$, $Z\sim\mathcal{N}(0,1)$, $z\in\mathbb{R}$, $n\to\infty$, $\frac{X_n-\mu_n}{\sigma_n}\xrightarrow{D}Z$; approx. $P(X_n\le r) = \Phi(\frac{r-\mu_n}{\sigma_n})$, $X_n\sim\mathcal{N}(np,np(1-p))$, valid when $\min(np,np(1-p))\ge 5$; instead of normal when outliers

4 Several Random Variables

Joint PMF : (X,Y) discrete r.v., $D = \{(x,y)\in\mathbb{R} : P((X,Y) = (x,y))>0\}$, $(x,y)\in\mathbb{R}^2$, $f_{X,Y}(x,y) = P((X,Y) = (x,y))$

Joint CDF : $F_{X,Y}(x,y) = P(X\le x, Y\le y)$

Continuous : jointly continuous if $\exists f_{X,Y}(x,y)$ joint density of (X,Y) s.t. $P((X,Y)\in A) = \int\int_{(u,v)\in A} f_{X,Y}(u,v)\,du\,dv$, $A\subset\mathbb{R}^2$; $A = \{(u,v) : u\le x, v\le y\}$, joint CDF $F_{X,Y}(x,y) = \int_{-\infty}^x\int_{-\infty}^y f_{X,Y}(u,v)\,du\,dv$, $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y)$

Exponential family distr. : $(X_1,\dots,X_n)$ discrete or continuous r.v., $(x_1,\dots,x_n)\in D\subset\mathbb{R}^n$, $(\theta_1,\dots,\theta_p)\in\Theta\subset\mathbb{R}^p$, $f(x_1,\dots,x_n) = \exp\left(\sum_{i=1}^p s_i(x)\theta_i - \kappa(\theta_1,\dots,\theta_p) + c(x_1,\dots,x_n)\right)$

Marginal PMF : discrete, $x\in\mathbb{R}$, $f_X(x) = \sum_y f_{X,Y}(x,y)$

Marginal density func. : continuous, $x\in\mathbb{R}$, $f_X(x) = \int_{-\infty}^\infty f_{X,Y}(x,y)\,dy$

Conditional PMF : $f_X(x)>0$ (undefined for $f_X(x) = 0$), $y\in\mathbb{R}$, $f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x)$; (X,Y) discrete : $f_X(x) = P(X=x)$, $f_{Y|X}(y|x) = P(Y=y|X=x)$

Multivariate r.v.s : $X_1,\dots,X_n$ r.v.s on same proba. spcae, joint cumulative distr. func.

$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = P(X_1 \le x_1,\dots,X_n \le x_n)$,
Joint PMF (discrete) $f_{X_1,\dots,X_n}(x_1,\dots,x_n) = P(X_1=x_1,\dots,X_n=x_n)$
Joint density func. (continuous) $f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \frac{\partial^n F_{X_1,\dots,X_n}(x_1,\dots,x_n)}{\partial x_1,\dots,\partial x_n}$

Multinomial distr.: m denominator, $(p_1,\dots,p_k)$ probabilities, $x_1,\dots,x_k \in \{0,\dots,m\}, \sum_{j=1}^k x_j = m, m\in\mathbb{N}, \sum_{i=1}^k p_i = 1, f(x_1,\dots,x_k) = \frac{m!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$

Independence: X,Y on same prob. spcae, $A,B\subset\mathbb{R}, P(X\in A, Y\in B) = P(X\in A)P(X\in B)$; $A=(-\infty,x], B=(-\infty,y]$, $f_{X,Y}(x,y)=f_X(x)f_Y(y)\,\forall x,y\in\mathbb{R}$; X,Y indep. $\Rightarrow \forall x$ s.t. $f_X(x)>0$: $f_{Y|X}(y|x)=f_Y(y)$ (symm. with x)

Indep. and identically distrib. (iid.): random sample size n from distr. F, density f,, write $X_1,\dots,X_n \sim^{iid} F$ (or $\sim^{iid} f$), $f_{X_1,\dots,X_n}(x_1,\dots,x_n)=\prod_{j=1}^n f_X(x_j)$

4.1 Dependence

Expectation: $E[|g(X,Y)|]<\infty, E[g(X,Y)] = \sum_{x,y} g(x,y)f_{X,Y}(x,y)$ (discrete), $\iint g(x,y)f_{X,Y}(x,y)\,dx\,dy$ (continuous)

Joint moments: $E[X^rY^s]$

Joint central moments: $E[(X-E(X))^r(Y-E(Y))^s], r,s\in\mathbb{N}$

Covariance: $\text{cov}(X,Y) = E[X-E(X)(Y-E(Y))] = E(XY)-E(X)E(Y)$

Properties: $a,b,c,d\in\mathbb{R}$, X,Y,Z r.v.s $\text{cov}(X,X)=\text{var}(X), \text{cov}(a,X)=0$, $\text{cov}(X,Y)=\text{cov}(Y,X)$, $\text{cov}(a+bX+cY,Z) = b\cdot\text{cov}(X,Z)+c\cdot\text{cov}(Y,Z)$, $\text{cov}(a+bX,c+dY)=bd\,\text{cov}(X,Y)$, $\text{var}(a+bX+cY)=b^2\text{var}(X)+2bc\cdot\text{cov}(X,Y)+c^2\cdot\text{var}(Y)$, $\text{cov}(X,Y)^2\le\text{var}(X)\text{var}(Y)$

Independence: X,Y independent, $\exists$ expectation $g(X),h(Y)$, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$, X,Y indep. $\Rightarrow \text{cov}(X,Y)=0$ (converse false)

Average: $\overline{X} = n^{-1}\sum_{j=1}^n X_j$; mean μ, var. σ^2, $E[\overline{X}]=\mu, \text{var}(\overline{X})=\sigma^2/n$

Correlation: dimensionless dependence, $\text{corr}(X,Y)=\frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$

Properties: $\rho=\text{corr}(X,Y), -1\le\rho\le1$, $\rho=\pm1\Rightarrow\exists a,b,c\in\mathbb{R}$ s.t. $aX+bY+c=0$ $(a,b)\ne(0,0)$ (X,Y linearly dependent), X,Y indep. $\Rightarrow\text{corr}(X,Y)=0$, $\text{corr}(a+bX,c+dY)=\text{sign}(bd)\text{corr}(X,Y)$

Corr. Limitations: measures linear dep. (strong nonlin. dep., corr. 0), corr. can be strong but specious (2 sub-groups), corr. $\ne$ causation

Conditional expectation: $f_X(x)>0, E[(|g(X,Y)|)\mid X=x]<\infty$, $E[g(X,Y)|X=x]=\sum_y g(x,y)f_{Y|X}(y|x)$ (discrete), $\int_{-\infty}^\infty g(x,y)f_{Y|X}(y|x)\,dy$ (continuous), func. of x

Conditioning: required Es exist, $E[g(X,Y)] = E_X[E[g(X,Y)|X=x]]$, $\text{var}(g(X,Y)) = E_X[\text{var}(g(X,Y)|X=x)]+\text{var}_X(E[g(X,Y)|X=x])$

4.2 Generating Functions

Moment-generating MGF: $t\in\mathbb{R}, M_X(t)<\infty$, $M_X(t)=E[e^{tX}]$, called Laplace transform of $f_X(x)$, $M_X(t)=E[\sum_{r=0}^\infty\frac{t^rX^r}{r!}]=\sum_{r=0}^\infty\frac{t^r}{r!}E[X^r]$

Theorems: $M_X(0)=1, M_{a+bX}(t)=e^{at}M_X(bt), E[X^r]=\frac{\partial^r M_X(t)}{\partial t^r}\Big|_{t=0}$, $E[X]=M'_X(0), \text{var}(X)=M''_X(0)-M'_X(0)^2$, $\exists$ injection btw. $F_X(x)$ and $M_X(t)$

Linear cominations: $a,b_1,\dots\in\mathbb{R}$, $X_1,\dots$ indep. r.v.s, $Y=a+b_1X_1+\dots+b_nX_n$, $M_Y(t)=e^{ta}\prod_{j=1}^n M_{X_j}(tb_j)$; $X_1,\dots$ random sample, $S=X_1+\dots+X_n$, $M_S(t)=M_X(t)^n$

Continuity: $\{X_n\}, X$ r.v.s with distr. fun. $\{F_n\}, F$ MGFs $M_n(t), M(t)$ exists for $0\le|t|<b, |t|\le a<b, M_{n\to\infty}(t)\to M(t)\Rightarrow X_n\xrightarrow{D}X$ i.e. $F_n(x)\to F(x)$ at each $x\in\mathbb{R}$ where F continuous

Matrix: $X=(X_1,\dots,X_p)^T$, expectation (mean vector) $E[X]_{p\times1} = E[X[X_1]\dots E(X_p)]^T$, co-variance matrix $\text{var}(X)_{p\times p,(i,j)}=\text{cov}(X_i,X_j)$ (positive semi-definite)

Moment-gen. func. mlti.var. MGF: $X_{p\times1}=(X_1,\dots,X_p)^T, t\in\mathcal{T}=\{t\in\mathbb{R}^p: M_X(t)<\infty\}, M_X(t)=E[e^{t^TX}]=E[e^{\sum_{r=1}^p t_rX_r}]$

MGF Properties: $0\in\mathcal{T}$ so $M_X(0)=1$, $E[X]_{p\times1}=M'_X(0)=\frac{\partial M_X(t)}{\partial t}\Big|_{t=0}$, $\text{var}(X)_{p\times p}=\frac{\partial^2M_X(t)}{\partial t\partial t^T}\Big|_{t=0}-M'_X(0)M'_X(0)^T$, $\mathcal{A}\cup\mathcal{B}=\{1,\dots,p\}$ and $\mathcal{A}\cap\mathcal{B}\ne\emptyset$: $X_\mathcal{A}$ subvector of X containing $\{X_j: j\in\mathcal{A}\}$ then $X_\mathcal{A}$ indep. $X_\mathcal{B}\Leftrightarrow M_X(t)=E[e^{t_\mathcal{A}^TX_\mathcal{A}+t_\mathcal{B}^TX_\mathcal{B}}]=M_{X_\mathcal{A}}(t_\mathcal{A})M_{X_\mathcal{B}}(t_\mathcal{B}), t\in\mathcal{T}$

4.3 Multivariate Normal Distribution

Mult. var. normal distr.: $X=(X_1,\dots,X_p)^T$, $\exists\mu=(\mu_1,\dots,\mu_p)^T\in\mathbb{R}^p, p\times p$ matrix Ω (positive semi-definite), $u^TX\sim\mathcal{N}(u^T\mu,u^T\Omega u), u\in\mathbb{R}^p; X\sim\mathcal{N}_p(\mu,\Omega)$, $\Omega_{i,j}=\omega_{ij}, E[X_j]=\mu_j, \text{var}(X_j)=\omega_{jj}, \text{cov}(X_j,X_k), j\ne k$, mean vector μ, covariance matrix Ω, $M_X(u)=\exp(u^T+\frac12u^T\Omega u)$, $\mathcal{A}\cup\mathcal{B}=\{1,\dots,p\}$ and $\mathcal{A}\cap\mathcal{B}=\emptyset$: $X_\mathcal{A}\perp\!\!\!\perp X_\mathcal{B}\Leftrightarrow\Omega_{\mathcal{A},\mathcal{B}}=0$, $X_1,\dots,X_n\sim_{iid}\mathcal{N}(\mu,\sigma^2)\implies X_{n\times1}=(X_1,\dots,X_n)^T\sim\mathcal{N}_n(\mu 1_n,\sigma^2I_n)$, $a_{r\times1}+B_{r\times p}X\sim\mathcal{N}_r(a+B\mu,B\Omega B^T)$

Density function: $X\sim N_p(\mu,\Omega)$, iff Ω has rank p, $f(x;\mu,\Omega)=\frac{1}{(2\pi)^{\frac p2}|\Omega|^{\frac12}}\exp(-\frac12(x-\mu)^T\Omega^{-1}(x-\mu)), x\in\mathbb{R}^p$

Marginal/conditional distr.s: $X\sim\mathcal{N}_p(\mu_{p\times1},\Omega_{p\times p}), |\Omega|>0, \mathcal{A},\mathcal{B}\subset\{1,\dots,p\}, |\mathcal{A}|=q<p, |\mathcal{B}|=r<p, \mathcal{A}\cap\mathcal{B}=\emptyset, \mu_\mathcal{A},\Omega_\mathcal{A},\Omega_{\mathcal{AB}}$ be $q\times1$ of μ, $q\times q, q\times r$ submatrices of Ω conformable with $\mathcal{A},\mathcal{A}\times\mathcal{A},\mathcal{A}\times\mathcal{B}$, marginal $X_\mathcal{A}\sim\mathcal{N}_q(\mu_\mathcal{A},\Omega_\mathcal{A})$, conditional $X_\mathcal{A}|X_\mathcal{B}=x_\mathcal{B}\sim\mathcal{N}_q(\mu_\mathcal{A}+\Omega_{\mathcal{AB}}\Omega_\mathcal{B}^{-1}(x_\mathcal{B}-\mu_\mathcal{B}),\Omega_\mathcal{A}-\Omega_\mathcal{B}^{-1}\Omega_{\mathcal{BA}})$

4.4 Transformations

Bivariate: $P(Y\in\mathcal{B}), Y\in\mathbb{R}^d, g:\mathbb{R}^2\to\mathbb{R}^d, \mathcal{B}\subset\mathbb{R}^d, g^{-1}(\mathcal{B})\subset\mathbb{R}^2$ set for which $g(g^{-1}(\mathcal{B}))=\mathbb{B}, P(Y\in\mathcal{B})=P(g(X)\in\mathcal{B})=P(X\in g^{-1}(\mathcal{B}))$

Joint continu. densities: $X=(X_1,X_2)\in\mathbb{R}^2$ conti. r.v., $Y=(Y_1,Y_2), Y_1=g_1(X_1,X_2), Y_2=g_2(X_1,X_2)$, system equations $y_1=g_1(x_1,x_2), y_2=g_2(x_1,x_2)$ can be solved $\forall(y_1,y_2)$ giving the sol.s $x_1=h_1(y_1,y_2), x_2=h_2(y_1,y_2)$; g_1,g_2 conti. differentiable with Jacobian $J(x_1,x_2)=|\frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2}-\frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1}|$, $f_{Y_1,Y_2}(y_1,y_2)=f_{X_1,X_2}(x_1,x_2)\times|J(x_1,x_2)|^{-1}\big|_{x_1=h(y_1,y_2),x_2=h_2(y_1,y_2)}$ (positive if $f_{X_1,X_2}(x_1,x_2)>0$)

Sums of independent: X,Y independant r.v.s, $S=X+Y; f_S(s)=f_X*f_Y(s)=$, $\int_{-\infty}^\infty f_X(x)f_Y(s-x)\,dx$ continuous, $\sum_x f_X(x)f_Y(s-x)$ discrete

Convolution: $X_1,\dots,X_n$ indep. r.v.s, $S=\sum_i X_i, f_S(s)=f_{X_1}*\dots*f_{X_n}(s)$

4.5 Order Statistics

Order statistics: ordered values $X_{(1)}\le\dots\le X_{(n)}$, not equal in continuous (<), min. $X_{(1)}$, max. $X_{(n)}$, median $X_{(m+1)}$ $(n=2m+1$ odd) $\frac12(X_{(m)}+X_{(m+1)})$ $(n=2m$ even)

Theorem: $X_1,\dots,X_n\sim_{iid}F$ continuous density f, $P(X_{(n)}\le x)=F(x)^n$, $P(X_{(1)}\le x)=1-(1-F(x))^n$

5 Approximation & Convergence

Inequalities: X r.v., $a>0$ contsant, h non-negative func., g convex func., $P(h(X)\ge a)\le E[h(x)]/a, P(|X|\ge a)\le E[|X|]/a, P(|X|\ge a)\le E[X^2]/a^2$, $E[g(X)]\ge g(E[X]), P(|X-E[X]|\ge a)\le\text{var}(X)/a^2$

Hoeffding inequality: $Z_1,\dots,Z_n$ indep. r.v.s s.t. $E[Z_i]=0$ and $a_i\le Z_i\le b_i$ for const. $a_i\le b_i; \epsilon>0,\forall t>0, P(\sum_{i=1}^n Z_i\ge\epsilon)\le e^{-t\epsilon}\prod_{i=1}^n e^{t^2(b_i-a_i)^2/8}$

5.1 Convergence

Deterministic convergence: $x_1,\dots,x\in\mathbb{R}$, $x_n\to x\Leftrightarrow\forall\epsilon>0\,\exists N_\epsilon$ s.t. $|x_n-x|<\varepsilon\,\forall n>N_\epsilon, X_n\to X$ if either $(n\to\infty)$, $P(X_n\le x)\to P(X\le x)\,x\in\mathbb{R}$, or $E[X_n]\to E[X]$

Modes of convergence of r.v.s: $X,X_1,\dots$ r.v.s with CDF $F,F_1,\dots$, almost surely $X_n\xrightarrow{a.s.}X$ if $P(\lim_{n\to\infty}X_n=X)=1$, in mean square $X_n\xrightarrow{2}X$ if $\lim_{n\to\infty}E[(X_n-X)^2]=0$ $E[X_n^2],E[X^2]<\infty$, in probability $X_n\xrightarrow{P}X$ if $\forall\epsilon>0\lim_{n\to\infty}P(|X_n-X|>\epsilon)=0$, in distribution $X_n\xrightarrow{D}$ if $\lim_{n\to\infty}F_n(x)=F(x)$ ($F(x)$ continuous at each pt. x)

Relations: $(\xrightarrow{a.s.}$ or $\xrightarrow{2})\Rightarrow\xrightarrow{P}\Rightarrow\xrightarrow{D}$

Limits of maxima: $X_1,\dots,X_n\sim^{iid}F, M_n=\max(X_1,\dots,X_n), P(M_n\le x)=P(X_1\le x,\dots,X_n\le x)=F(x)^n$, 0 when $F(x)<1$, 1 when $F(x)=1; Y_n=(M_n-b_n)/a_n$

5.2 Laws of Large Numbers

Weak law: iid, finite expectation μ, $\bar X = n^{-1}(X_1+\dots+X_n), \bar X\xrightarrow{P}\mu, \forall\epsilon>0$ $P(|\bar X-\mu|>\epsilon)\to0, n\to\infty$

5.3 Central limit theorem (CLT)

Standardisation average: $\text{var}(X_j)<\infty$, $E[\bar X]=\mu, \text{var}(\bar X)=\sigma^2/n$, $Z_n=n^{\frac12}(\bar X-\mu)/\sigma$ has expected of 0 and variance of 1

CLT: $X_1,\dots$ iid. expectation μ, var. $0<\sigma^2<\infty, Z_n\xrightarrow{D}Z, n\to\infty$ where $Z\sim N(0,1)$; $P(Z_n\le z)\to\Phi(z)$ for large n

Use: sums of indep. r.v.s, $n>25$, $E[\sum_{j=1}^nX_j]=n\mu, \text{var}(\sum_{j=1}^nX_j)=n\sigma^2$, $P(\sum_{j=1}^nX_j\le x)=\Phi(\frac{x-n\mu}{(n\sigma^2)^{1/2}})$

6 Statistical interference

Induction: observed event A, say something about probability space $(\Omega,\mathcal{F},P): A\Rightarrow^?(\Omega,\mathcal{F},P)$; say something about a process based on the data

Data: y (observed), Y (potential)

Statistical model: proba. distr. $f(y)$ chosen or constructed to learn from data; $f(y)=f(y;\theta)$ parameter θ of finite dimension (parametric model), known model is called simple, otherwise composite

Statistic: $T=t(Y)$ known function of data Y

Sampling distribution: of statistic $T=t(Y)$ is its distrib. when $Y\sim f(y)$

Random sample: set of iid. r.v.s $Y_1,\dots,Y_n$ or their realizations $y_1,\dots,y_n$

6.1 Point estimation

Study set of individuals elements (population), based on a subset (sample).

Statistical Model: unknown distribution F or density f of Y

Parametric statistical model: the distribution of Y is known except for values of parameters θ, $F(y)=F(y;\theta)$, with θ unknown

Sample: must be representative of population, $y_1,\dots,y_n$, supposed to be random sample $(Y_1,\dots,Y_n\sim^{iid}F)$

Statistic: any func. $T=t(Y_1,\dots,Y_n)$ of r.v.s $Y_1,\dots,Y_n$

Estimator: a statistic $\hat\theta$ used to estimate a parameter θ of f

Method of moments: (simple, can be inefficient), $\tilde\theta$ match the theoretical/empirical moments; p unknown param.s, $E(Y^r)=\int y^rf(y;\theta)\,dy=\frac1n\sum_{j=1}^ny_j^r$ $(r=1,\dots,p)$, we need as many moments of underlying model as unknowns, use the first r moments

Likelihood: for θ is $L(\theta)=f(y_1,\dots,y_n;\theta)=f(y_1;\theta)\times\dots\times f(y_n;\theta)$

Maximum likelihood estimation (MLE): (general, optimal in many param. models), $\hat\theta$ value that gives observed data the highest likelihood, $L(\hat\theta)\ge l(\theta)\,\forall\theta$

Calculation of MLE: maximizing $l(\theta)=\log(L(\theta))$, calculate $l(\theta)$ (plot), find value $\hat\theta$ maximizing $l(\theta)$ using derivative = 0, second derivative < 0

M-estimation: (more general, robust, loses efficiency), maxim. $\rho(\theta;Y)=\sum_{j=1}^n\rho(\theta;Y_j)$, $\rho(\theta;y)$ (if possible) concave of θ for all y, $\rho(\theta;y)=\log(f(y;\theta))$ gives maxi. likelihood estimator

Bias: compare estimators, bias of estimator $\hat\theta$ of $\theta: b(\theta)=E[\hat\theta]-\theta; b(\theta)\forall\theta:<0\,\hat\theta$ underestimates θ, >0 overestimates, $=0$ unbiased; $b(\theta)\approx0$ then $\hat\theta$ in the right place on average

Mean square error MSE: $\text{MSE}(\hat\theta)=E[(\hat\theta-\theta)^2]=\text{var}(\hat\theta)+b(\theta)^2$ (average squared distance)

More efficient: $\hat\theta_1,\hat\theta_2$ unbiased of θ, $\text{MSE}(\hat\theta_{1/2})=\text{var}(\hat\theta_{1/2}), \hat\theta_1$ more efficient $\hat\theta_2$ if $\text{var}(\hat\theta_1)\le\text{var}(\hat\theta_2)$

6.2 Interval estimation

Pivot: $Y=(Y_1,\dots,Y_n)$ from distr. F, function $Q=q(Y,\theta)$, distr. of Q known (does not depend on θ)

Confidence intervals (CI): (L,U) for θ, lower L, upper U, rnd. interval contain θ with probability called confidence level; $L=l(Y)$, $U=u(Y)$ statistics from data (do not depend on θ); $P(\theta<L)=\alpha_L, P(U<\theta)=\alpha_U$, level: $P(L\le\theta\le U)=1-\alpha_L-\alpha_U$; $\alpha_L=\alpha_U=\alpha/2$ equi-tailed level $(1-\alpha)$

CI construction: find pivot $Q=q(Y,\theta)$, quantiles $q_{\alpha_U}\,q_{1-\alpha_L}$ of Q, transform $P(q_{\alpha_U}\le q(Y,\theta)\le q_{1-\alpha_L})=(1-\alpha_L)-\alpha_U$ into $P(L\le\Theta\le U)=1-\alpha_L-\alpha_U$, L,U depend on Y & q_α not on θ

One-sided : $(-\infty, U)$ or (L, ∞), take $\alpha_U = 0$ or $\alpha_L = 0$, replace unwanted limit by $\pm\infty$

Standard errors : approximate pivots, $T = t(Y_1, \dots, Y_n)$ estimator of θ, $\tau_n^2 = \text{var}(T)$, $V = v(Y_1, \dots, Y_n)$ estimator of τ_n^2, $V^{1/2}$ or $v^{1/2}$ (realization) a standard error for T

Theorem : $\frac{T-\theta}{\tau_n} \to^D Z$, $\frac{V}{\tau_n^2} \to^P 1, n \to \infty$, $Z \sim \mathcal{N}(0,1)$, $\frac{T-\theta}{V^{1/2}} = \frac{T-\theta}{\tau_n} \times \frac{\theta_n}{V^{1/2}} \to^D Z, n \to \infty$

Normal random sample : $Y_1, \dots, Y_n \sim^{\text{iid}} \mathcal{N}(\mu, \sigma^2)$, $\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ indep. $(n-1)S^2 = \sum_{j=1}^n (Y_j - \bar{Y})^2 \sim \sigma^2 \xi_{n-1}^2$

Comments : in most cases $U - L \propto \sqrt{V} \propto n^{-1/2}$, normal models exact CI available

6.3 Hypothesis Tests

Confidence intervals and tests : value θ^0 of θ, θ^0 lies inside $(1-\alpha)$ CI : cannot reject hypothesis that $\theta = \theta^0$ at significance level α, outside, we reject at level α; cannot prove (only reject)

Null and alternative hypotheses : null hypothesis H_0 model to test, alternative H_1 what happens if H_0 is false, type 1 error (false positive) H_0 true but wrongly reject (choose H_1), type 2 error (false negative) H_1 true but we wrongly accept H_0

Taxonomy of hypotheses : simple hypothesis entirely fixes distr. of data Y, composite does not fix

Receiver operating characteristic (ROC) curve : of test plots $\beta(t)$ against $\alpha(t)$ as cutt-off t varies, $(P_0(T \geq t), P_1(T > t)), t \in \mathbb{R}$

Size and power : μ increases : easier detect H_0 false, densities under H_0 and H_1 separated, H_0 and H_1 same ($\mu = 0$) curve lies on diagonal (cannot distinguish), often μ unknown so fix α and accept resulting $\beta(\alpha)$; false positive probability the size α, true positive probability power β; size $\alpha = P_0(\text{reject } H_0)$, power $\beta = P_1(\text{reject } H_0)$

Power and CI : size is probability α, usually width of (L, U) satisfies $U - L \propto n^{-1/2}$

Pearson statistic (chi-square) : $O_1, \dots, O_k$ nb. observations of random sample size $n = n_1 + \dots + n_k$ falling into categories $1, \dots, k$, expected numbers $E_1, \dots, E_k$, $E_i > 0$, $T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$

Chi-square distr. : with ν degrees of freedom, $Z_1, \dots, Z_\nu \sim^{\text{iid}} \mathcal{N}(0,1)$, $W = Z_1^2 + \dots + Z_\nu^2$ is chi-square, $f_W(w) = \frac{w^{\nu/2-1} e^{-w/2}}{2^{\nu/2} T(\nu/2)}$, $w > 0, \nu \in \mathbb{R}^*, \Gamma(a) = \int_0^\infty u^{a-1} e^{-u} \, du$, $a > 0$

Pearson rationale : $O_i \approx E_i \; \forall i$, T small otherwise tend to be bigger, joint distr. $O_1, \dots, O_k$ multinomial with denominator $n \, p_i = E_i/n$, $O_i \sim B(n, p_i)$, $E(O_i) = np_i = E_i$, $\text{var}(O_i) = E_i(1 - E_i/n) \approx E_i$, $Z_i = (O_i - E_i)/\sqrt{E_i} \sim \mathcal{N}(0,1)$ for large n

Evidence and P-values : observed value of T is t_{obs}, $p_{\text{obs}} = P_0(T \geq t_{\text{obs}})$, p small suggest H_0 is true but something unlikely occurred or H_0 false, $p < \alpha$ test is significant at level α, reject H_0 if $p < \alpha$, provisionally accept H_0 if $P \geq \alpha$

Decision procedure : choose level α, test H_0, reject H_0 if P-value is less than α or do not reject

Measure of evidence : against H_0, small values of p_{obs} suggesting stronger evidence against H_0; H_1 need not be explicit, seek for implicit choice of T

Choice of α : 0.05, 0.01, 0.001

6.4 Comparison of Tests

Parametric tests : based on parametric statistical model (nearly optimal test)

Non-parametric tests : based on general statistical model

ROC curve : good test will have ROC close to upper left corner, useless test have ROC diagonal

Most powerful tests : aim test statistic T to maximise the power of test for given size, partitioning sample space Ω containing data Y into rejection region $\mathcal{Y}$ and its complement $\bar{\mathcal{Y}}$; $Y \in \mathcal{Y} \Rightarrow$ reject H_0, $Y \in \bar{\mathcal{Y}} \Rightarrow$ accept H_0; aim to choose $\mathcal{Y}$ such that $P_1(Y \in \mathcal{Y})$ is largest possible such that $P_0(Y \in \mathcal{Y}) = \alpha$

Neyman-Pearson : $f_0(y), f_1(y)$ densities of Y under simple null and alternative hypotheses, if it exists, the set $\mathcal{Y}_\alpha = \{y \in \Omega : f_1(y)/f_0(y) > t\}$ such that $P_0(Y \in \mathcal{Y}_\alpha) = \alpha$ maximises $P_1(Y \in \mathcal{Y}_\alpha)$ amongst all the $\mathcal{Y}'$ such that $P_0(Y \in \mathcal{Y}') \leq \alpha$, base the decision on $\mathcal{Y}_\alpha$