

PROBABILITY AND STATISTICS

COOKBOOK

Version 0.2.6

19th December, 2017

<http://statistics.zone/>

Copyright © Matthias Vallentin

Contents

1	Distribution Overview	3	14	Exponential Family	16	21.5 Spectral Analysis	28	
1.1	Discrete Distributions	3	15	Bayesian Inference	16	22	Math	29
1.2	Continuous Distributions	5	15.1	Credible Intervals	16	22.1	Gamma Function	29
2	Probability Theory	8	15.2	Function of parameters	17	22.2	Beta Function	29
3	Random Variables	8	15.3	Priors	17	22.3	Series	29
3.1	Transformations	9	15.3.1	Conjugate Priors	17	22.4	Combinatorics	30
4	Expectation	9	15.4	Bayesian Testing	18			
5	Variance	9	16	Sampling Methods	18			
6	Inequalities	10	16.1	Inverse Transform Sampling	18			
7	Distribution Relationships	10	16.2	The Bootstrap	18			
8	Probability and Moment Generating Functions	11	16.2.1	Bootstrap Confidence Intervals	18			
9	Multivariate Distributions	11	16.3	Rejection Sampling	19			
9.1	Standard Bivariate Normal	11	16.4	Importance Sampling	19			
9.2	Bivariate Normal	11	17	Decision Theory	19			
9.3	Multivariate Normal	11	17.1	Risk	19			
10	Convergence	11	17.2	Admissibility	20			
10.1	Law of Large Numbers (LLN)	12	17.3	Bayes Rule	20			
10.2	Central Limit Theorem (CLT)	12	17.4	Minimax Rules	20			
11	Statistical Inference	12	18	Linear Regression	20			
11.1	Point Estimation	12	18.1	Simple Linear Regression	20			
11.2	Normal-Based Confidence Interval	13	18.2	Prediction	21			
11.3	Empirical distribution	13	18.3	Multiple Regression	21			
11.4	Statistical Functionals	13	18.4	Model Selection	22			
12	Parametric Inference	13	19	Non-parametric Function Estimation	22			
12.1	Method of Moments	13	19.1	Density Estimation	22			
12.2	Maximum Likelihood	14	19.1.1	Histograms	23			
12.2.1	Delta Method	14	19.1.2	Kernel Density Estimator (KDE)	23			
12.3	Multiparameter Models	14	19.2	Non-parametric Regression	23			
12.3.1	Multiparameter delta method	15	19.3	Smoothing Using Orthogonal Functions	24			
12.4	Parametric Bootstrap	15	20	Stochastic Processes	24			
13	Hypothesis Testing	15	20.1	Markov Chains	24			
			20.2	Poisson Processes	25			
			21	Time Series	25			
			21.1	Stationary Time Series	26			
			21.2	Estimation of Correlation	26			
			21.3	Non-Stationary Time Series	26			
			21.3.1	Detrending	27			
			21.4	ARIMA models	27			
			21.4.1	Causality and Invertibility	28			

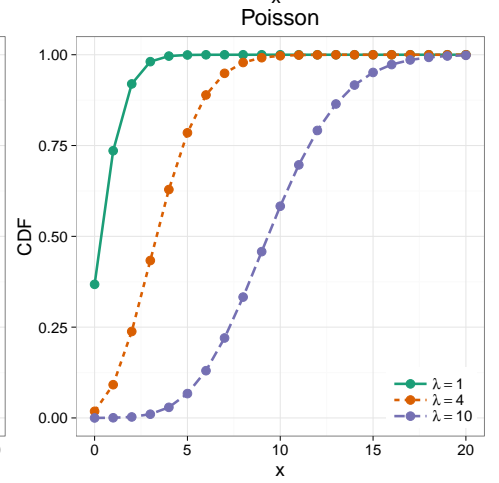
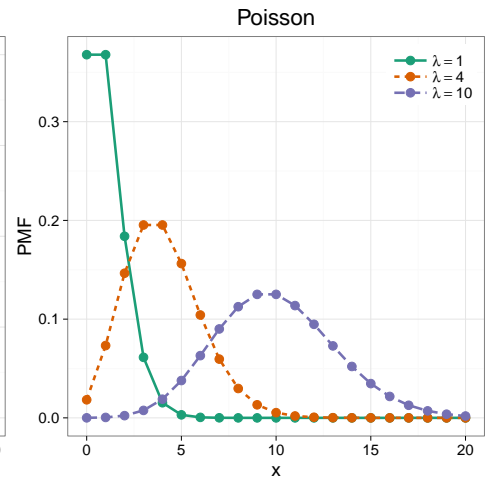
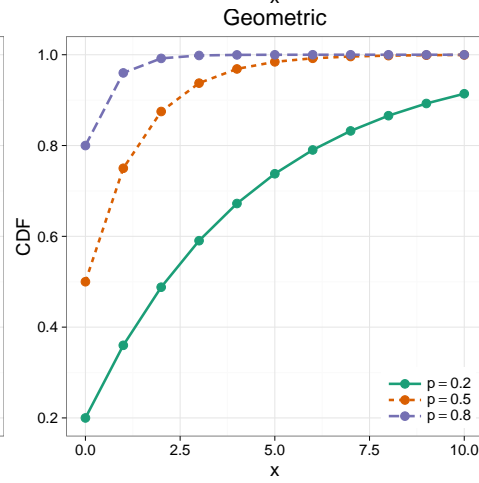
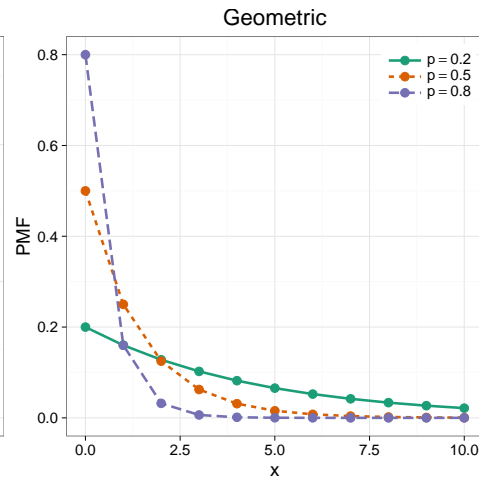
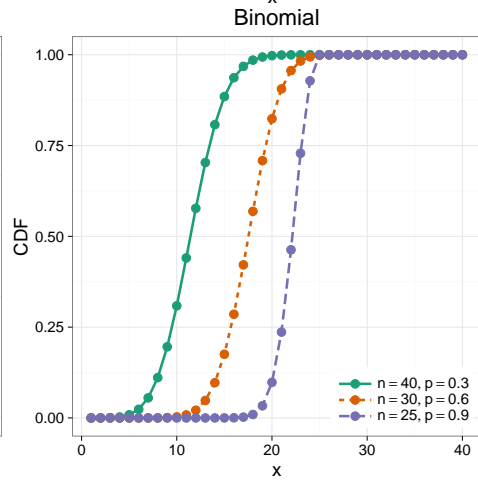
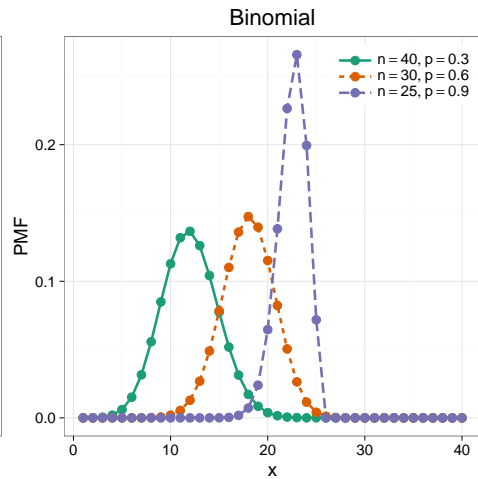
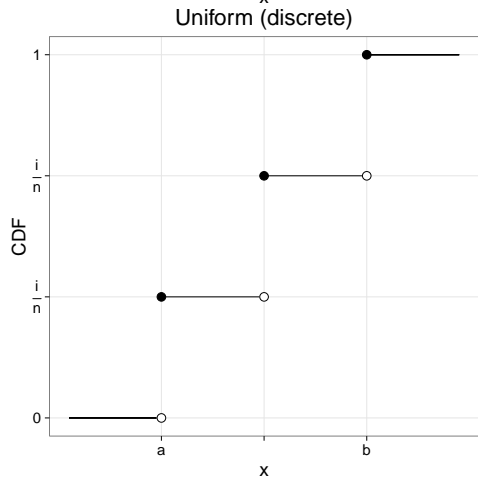
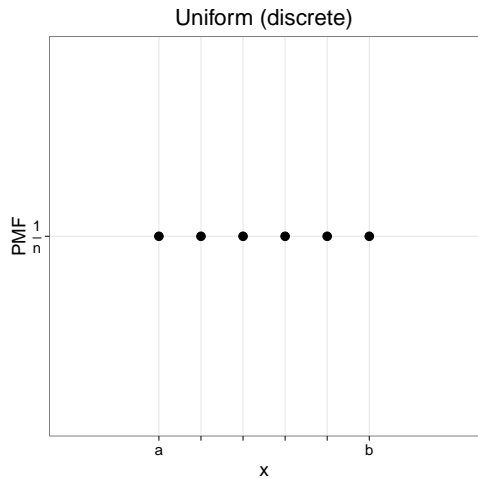
This cookbook integrates various topics in probability theory and statistics, based on literature [1, 6, 3] and in-class material from courses of the statistics department at the University of California in Berkeley but also influenced by others [4, 5]. If you find errors or have suggestions for improvements, please get in touch at <http://statistics.zone/>.

1 Distribution Overview

1.1 Discrete Distributions

	Notation ¹	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform	$\text{Unif}\{a, \dots, b\}$	$\begin{cases} 0 & x < a \\ \frac{ x - a + 1}{b - a} & a \leq x \leq b \\ 1 & x > b \end{cases}$	$\frac{I(a \leq x \leq b)}{b - a + 1}$	$\frac{a + b}{2}$	$\frac{(b - a + 1)^2 - 1}{12}$	$\frac{e^{as} - e^{-(b+1)s}}{s(b - a)}$
Bernoulli	$\text{Bern}(p)$	$(1 - p)^{1-x}$	$p^x (1 - p)^{1-x}$	p	$p(1 - p)$	$1 - p + pe^s$
Binomial	$\text{Bin}(n, p)$	$I_{1-p}(n - x, x + 1)$	$\binom{n}{x} p^x (1 - p)^{n-x}$	np	$np(1 - p)$	$(1 - p + pe^s)^n$
Multinomial	$\text{Mult}(n, p)$		$\frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \quad \sum_{i=1}^k x_i = n$	$\begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$	$\begin{pmatrix} np_1(1 - p_1) & -np_1p_2 \\ -np_2p_1 & \ddots \end{pmatrix}$	$\left(\sum_{i=0}^k p_i e^{s_i} \right)^n$
Hypergeometric	$\text{Hyp}(N, m, n)$	$\approx \Phi \left(\frac{x - np}{\sqrt{np(1 - p)}} \right)$	$\frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}}$	$\frac{nm}{N}$	$\frac{nm(N - n)(N - m)}{N^2(N - 1)}$	
Negative Binomial	$\text{NBin}(r, p)$	$I_p(r, x + 1)$	$\binom{x + r - 1}{r - 1} p^r (1 - p)^x$	$r \frac{1 - p}{p}$	$r \frac{1 - p}{p^2}$	$\left(\frac{p}{1 - (1 - p)e^s} \right)^r$
Geometric	$\text{Geo}(p)$	$1 - (1 - p)^x \quad x \in \mathbb{N}^+$	$p(1 - p)^{x-1} \quad x \in \mathbb{N}^+$	$\frac{1}{p}$	$\frac{1 - p}{p^2}$	$\frac{pe^s}{1 - (1 - p)e^s}$
Poisson	$\text{Po}(\lambda)$	$e^{-\lambda} \sum_{i=0}^x \frac{\lambda^i}{i!}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	$e^{\lambda(e^s - 1)}$

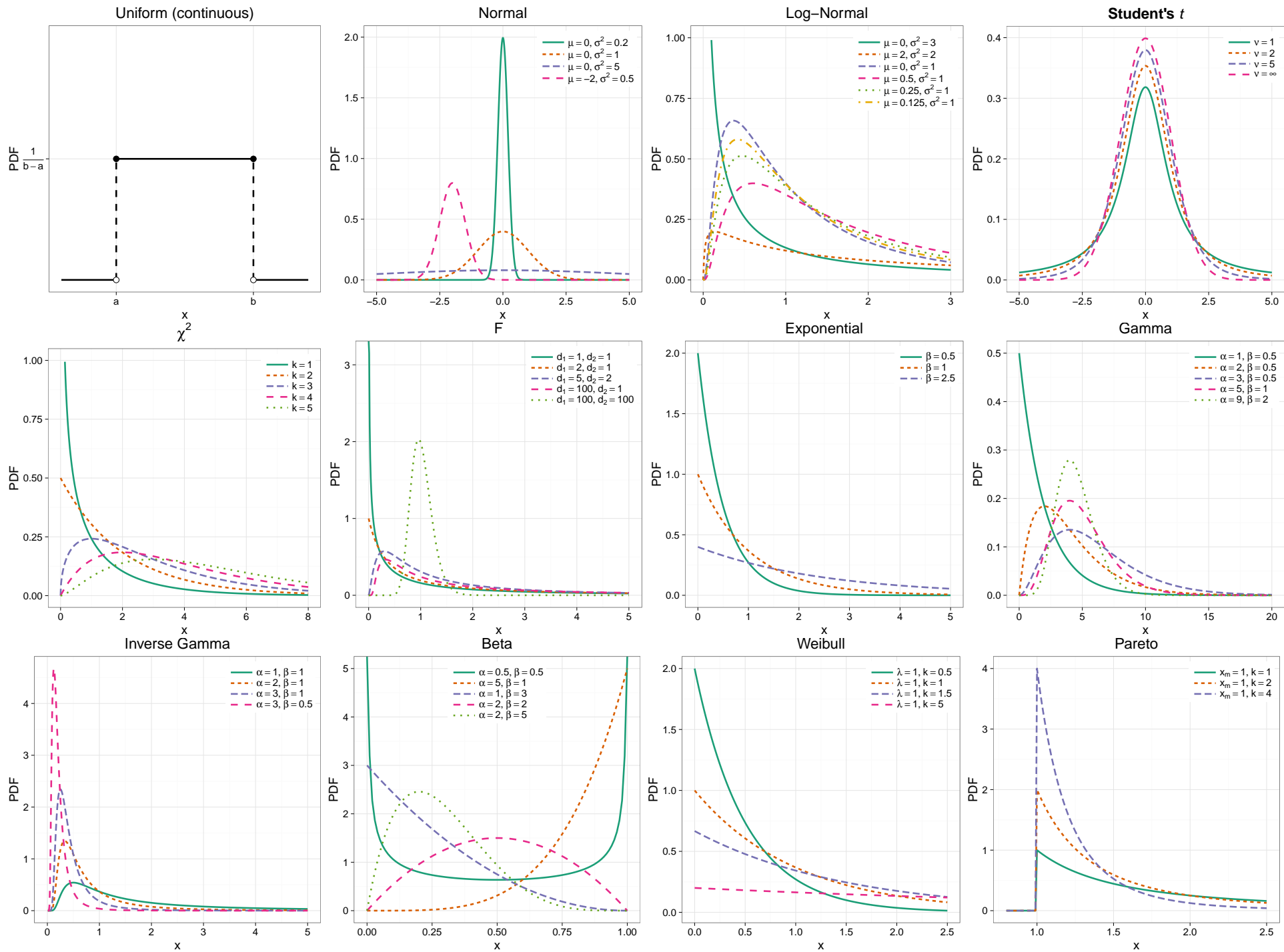
¹We use the notation $\gamma(s, x)$ and $\Gamma(x)$ to refer to the Gamma functions (see §22.1), and use $B(x, y)$ and I_x to refer to the Beta functions (see §22.2).

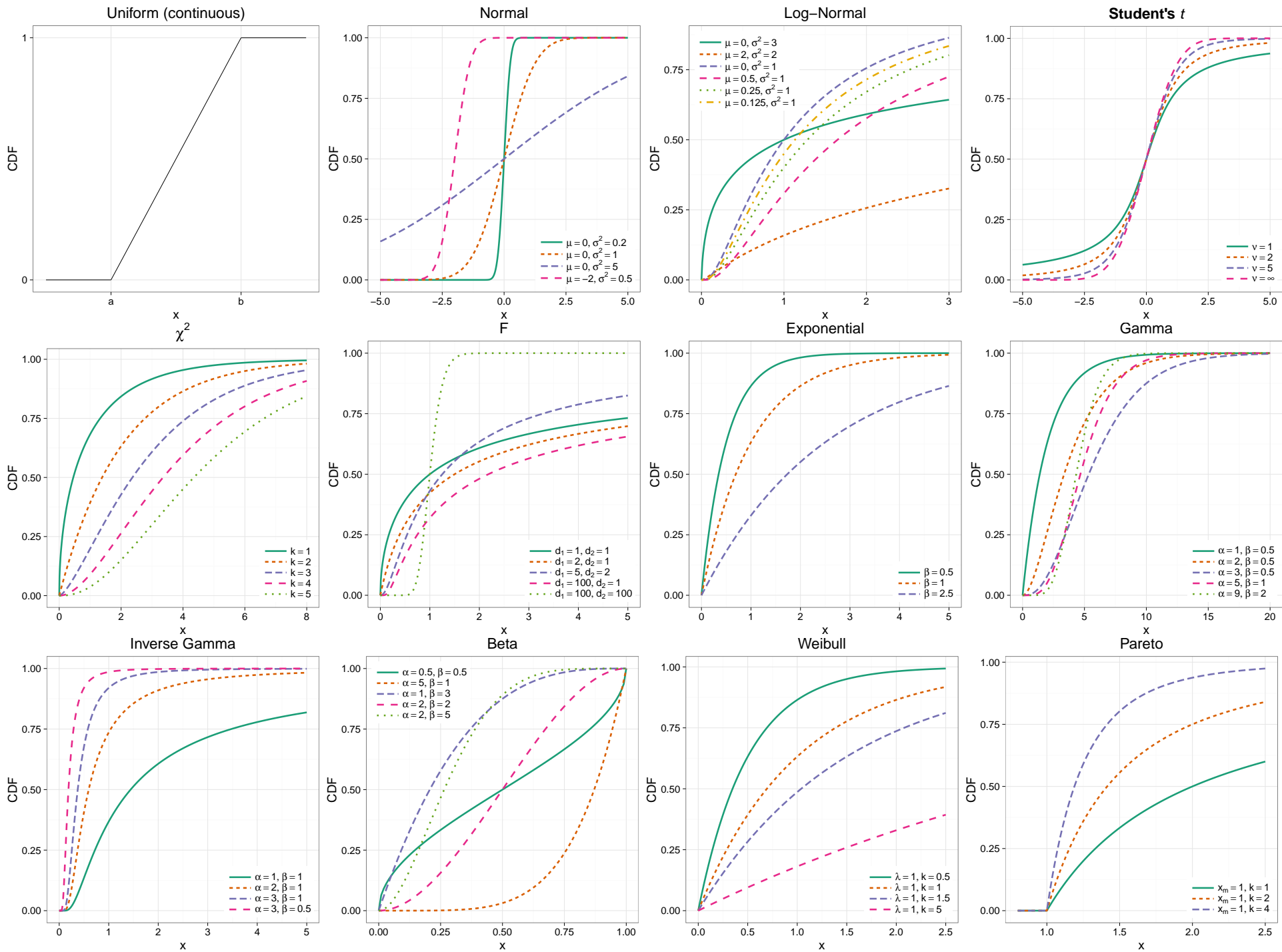


1.2 Continuous Distributions

	Notation	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	$M_X(s)$
Uniform	$\text{Unif}(a, b)$	$\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$	$\frac{I(a < x < b)}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{sb} - e^{sa}}{s(b-a)}$
Normal	$\mathcal{N}(\mu, \sigma^2)$	$\Phi(x) = \int_{-\infty}^x \phi(t) dt$	$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$\exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}$
Log-Normal	$\ln \mathcal{N}(\mu, \sigma^2)$	$\frac{1}{2} + \frac{1}{2} \operatorname{erf}\left[\frac{\ln x - \mu}{\sqrt{2\sigma^2}}\right]$	$\frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$	
Multivariate Normal	$\text{MVN}(\mu, \Sigma)$		$(2\pi)^{-k/2} \Sigma ^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$	μ	Σ	$\exp\left\{\mu^T s + \frac{1}{2} s^T \Sigma s\right\}$
Student's t	$\text{Student}(\nu)$	$I_x\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	$0 \quad \nu > 1$	$\begin{cases} \frac{\nu}{\nu-2} & \nu > 2 \\ \infty & 1 < \nu \leq 2 \end{cases}$	
Chi-square	χ_k^2	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$	k	$2k$	$(1-2s)^{-k/2} s < 1/2$
F	$F(d_1, d_2)$	$I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$	$\frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$	$\frac{d_2}{d_2 - 2}$	$\frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)}$	
Exponential*	$\text{Exp}(\beta)$	$1 - e^{-x/\beta}$	$\frac{1}{\beta} e^{-x/\beta}$	β	β^2	$\frac{1}{1 - \frac{s}{\beta}} (s < \beta)$
Gamma*	$\text{Gamma}(\alpha, \beta)$	$\frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{1}{1 - \frac{s}{\beta}}\right)^\alpha (s < \beta)$
Inverse Gamma	$\text{InvGamma}(\alpha, \beta)$	$\frac{\Gamma(\alpha, \frac{\beta}{x})}{\Gamma(\alpha)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\frac{\beta}{\alpha-1} \quad \alpha > 1$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \quad \alpha > 2$	$\frac{2(-\beta s)^{\alpha/2}}{\Gamma(\alpha)} K_\alpha(\sqrt{-4\beta s})$
Dirichlet	$\text{Dir}(\alpha)$		$\frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1}$	$\frac{\alpha_i}{\sum_{i=1}^k \alpha_i}$	$\frac{\mathbb{E}[X_i](1 - \mathbb{E}[X_i])}{\sum_{i=1}^k \alpha_i + 1}$	
Beta	$\text{Beta}(\alpha, \beta)$	$I_x(\alpha, \beta)$	$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$\frac{\alpha}{\alpha + \beta}$	$\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$	$1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r}\right) \frac{s^k}{k!}$
Weibull	$\text{Weibull}(\lambda, k)$	$1 - e^{-(x/\lambda)^k}$	$\frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$	$\lambda \Gamma\left(1 + \frac{1}{k}\right)$	$\lambda^2 \Gamma\left(1 + \frac{2}{k}\right) - \mu^2$	$\sum_{n=0}^{\infty} \frac{s^n \lambda^n}{n!} \Gamma\left(1 + \frac{n}{k}\right)$
Pareto	$\text{Pareto}(x_m, \alpha)$	$1 - \left(\frac{x_m}{x}\right)^\alpha \quad x \geq x_m$	$\alpha \frac{x_m^\alpha}{x^{\alpha+1}} \quad x \geq x_m$	$\frac{\alpha x_m}{\alpha - 1} \quad \alpha > 1$	$\frac{x_m^2 \alpha}{(\alpha - 1)^2(\alpha - 2)} \quad \alpha > 2$	$\alpha(-x_m s)^\alpha \Gamma(-\alpha, -x_m s) \quad s < 0$

* We use the *rate* parameterization where $\beta = \frac{1}{\lambda}$. Some textbooks use β as *scale* parameter instead [6].





2 Probability Theory

Definitions

- Sample space Ω
- Outcome (point or element) $\omega \in \Omega$
- Event $A \subseteq \Omega$
- σ -algebra \mathcal{A}
 1. $\emptyset \in \mathcal{A}$
 2. $A_1, A_2, \dots, \in \mathcal{A} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$
 3. $A \in \mathcal{A} \implies \neg A \in \mathcal{A}$
- Probability Distribution \mathbb{P}
 1. $\mathbb{P}[A] \geq 0 \quad \forall A$
 2. $\mathbb{P}[\Omega] = 1$
 3. $\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$
- Probability space $(\Omega, \mathcal{A}, \mathbb{P})$

Properties

- $\mathbb{P}[\emptyset] = 0$
- $B = \Omega \cap B = (A \cup \neg A) \cap B = (A \cap B) \cup (\neg A \cap B)$
- $\mathbb{P}[\neg A] = 1 - \mathbb{P}[A]$
- $\mathbb{P}[B] = \mathbb{P}[A \cap B] + \mathbb{P}[\neg A \cap B]$
- $\mathbb{P}[\Omega] = 1 \quad \mathbb{P}[\emptyset] = 0$
- $\neg(\bigcup_n A_n) = \bigcap_n \neg A_n \quad \neg(\bigcap_n A_n) = \bigcup_n \neg A_n \quad \text{DEMORGAN}$
- $\mathbb{P}[\bigcup_n A_n] = 1 - \mathbb{P}[\bigcap_n \neg A_n]$
- $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$
 $\implies \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$
- $\mathbb{P}[A \cup B] = \mathbb{P}[A \cap \neg B] + \mathbb{P}[\neg A \cap B] + \mathbb{P}[A \cap B]$
- $\mathbb{P}[A \cap \neg B] = \mathbb{P}[A] - \mathbb{P}[A \cap B]$

Continuity of Probabilities

- $A_1 \subset A_2 \subset \dots \implies \lim_{n \rightarrow \infty} \mathbb{P}[A_n] = \mathbb{P}[A] \quad \text{where } A = \bigcup_{i=1}^{\infty} A_i$
- $A_1 \supset A_2 \supset \dots \implies \lim_{n \rightarrow \infty} \mathbb{P}[A_n] = \mathbb{P}[A] \quad \text{where } A = \bigcap_{i=1}^{\infty} A_i$

Independence $\perp\!\!\!\perp$

$$A \perp\!\!\!\perp B \iff \mathbb{P}[A \cap B] = \mathbb{P}[A] \mathbb{P}[B]$$

Conditional Probability

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \quad \mathbb{P}[B] > 0$$

Law of Total Probability

$$\mathbb{P}[B] = \sum_{i=1}^n \mathbb{P}[B|A_i] \mathbb{P}[A_i] \quad \Omega = \bigsqcup_{i=1}^n A_i$$

BAYES' THEOREM

$$\mathbb{P}[A_i|B] = \frac{\mathbb{P}[B|A_i] \mathbb{P}[A_i]}{\sum_{j=1}^n \mathbb{P}[B|A_j] \mathbb{P}[A_j]} \quad \Omega = \bigsqcup_{i=1}^n A_i$$

Inclusion-Exclusion Principle

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{r=1}^n (-1)^{r-1} \sum_{i_1 < \dots < i_r \leq n} \left| \bigcap_{j=1}^r A_{i_j} \right|$$

3 Random Variables

Random Variable (RV)

$$X : \Omega \rightarrow \mathbb{R}$$

Probability Mass Function (PMF)

$$f_X(x) = \mathbb{P}[X = x] = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}]$$

Probability Density Function (PDF)

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x) dx$$

Cumulative Distribution Function (CDF)

$$F_X : \mathbb{R} \rightarrow [0, 1] \quad F_X(x) = \mathbb{P}[X \leq x]$$

1. Nondecreasing: $x_1 < x_2 \implies F(x_1) \leq F(x_2)$
2. Normalized: $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$
3. Right-Continuous: $\lim_{y \downarrow x} F(y) = F(x)$

$$\mathbb{P}[a \leq Y \leq b | X = x] = \int_a^b f_{Y|X}(y|x) dy \quad a \leq b$$

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Independence

1. $\mathbb{P}[X \leq x, Y \leq y] = \mathbb{P}[X \leq x] \mathbb{P}[Y \leq y]$
2. $f_{X,Y}(x, y) = f_X(x) f_Y(y)$

3.1 Transformations

Transformation function

$$Z = \varphi(X)$$

Discrete

$$f_Z(z) = \mathbb{P}[\varphi(X) = z] = \mathbb{P}[\{x : \varphi(x) = z\}] = \mathbb{P}[X \in \varphi^{-1}(z)] = \sum_{x \in \varphi^{-1}(z)} f_X(x)$$

Continuous

$$F_Z(z) = \mathbb{P}[\varphi(X) \leq z] = \int_{A_z} f(x) dx \quad \text{with } A_z = \{x : \varphi(x) \leq z\}$$

Special case if φ strictly monotone

$$f_Z(z) = f_X(\varphi^{-1}(z)) \left| \frac{d}{dz} \varphi^{-1}(z) \right| = f_X(x) \left| \frac{dx}{dz} \right| = f_X(x) \frac{1}{|J|}$$

The Rule of the Lazy Statistician

$$\mathbb{E}[Z] = \int \varphi(x) dF_X(x)$$

$$\mathbb{E}[I_A(x)] = \int I_A(x) dF_X(x) = \int_A dF_X(x) = \mathbb{P}[X \in A]$$

Convolution

- $Z := X + Y \quad f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) dx \stackrel{X,Y \geq 0}{=} \int_0^z f_{X,Y}(x, z-x) dx$
- $Z := |X - Y| \quad f_Z(z) = 2 \int_0^{\infty} f_{X,Y}(x, z+x) dx$
- $Z := \frac{X}{Y} \quad f_Z(z) = \int_{-\infty}^{\infty} |y| f_{X,Y}(yz, y) dy \stackrel{\perp}{=} \int_{-\infty}^{\infty} |y| f_X(yz) f_Y(y) dy$

4 Expectation

Definition and properties

$$\bullet \mathbb{E}[X] = \mu_X = \begin{cases} \sum_x x f_X(x) & \text{X discrete} \\ \int x f_X(x) dx & \text{X continuous} \end{cases}$$

- $\mathbb{P}[X = c] = 1 \implies \mathbb{E}[X] = c$
- $\mathbb{E}[cX] = c \mathbb{E}[X]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

- $\mathbb{E}[XY] = \int_{X,Y} xy f_{X,Y}(x, y) dF_X(x) dF_Y(y)$
- $\mathbb{E}[\varphi(Y)] \neq \varphi(\mathbb{E}[X])$ (cf. JENSEN inequality)
- $\mathbb{P}[X \geq Y] = 1 \implies \mathbb{E}[X] \geq \mathbb{E}[Y]$
- $\mathbb{P}[X = Y] = 1 \implies \mathbb{E}[X] = \mathbb{E}[Y]$
- $\mathbb{E}[X] = \sum_{x=1}^{\infty} \mathbb{P}[X \geq x]$ X discrete

Sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Conditional expectation

- $\mathbb{E}[Y | X = x] = \int y f(y | x) dy$
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]]$
- $\mathbb{E}_{\varphi(X,Y) | X=x} [=] \int_{-\infty}^{\infty} \varphi(x, y) f_{Y|X}(y | x) dx$
- $\mathbb{E}[\varphi(Y, Z) | X = x] = \int_{-\infty}^{\infty} \varphi(y, z) f_{(Y,Z)|X}(y, z | x) dy dz$
- $\mathbb{E}[Y + Z | X] = \mathbb{E}[Y | X] + \mathbb{E}[Z | X]$
- $\mathbb{E}[\varphi(X)Y | X] = \varphi(X) \mathbb{E}[Y | X]$
- $\mathbb{E}[Y | X] = c \implies \text{Cov}[X, Y] = 0$

5 Variance

Definition and properties

- $\mathbb{V}[X] = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $\mathbb{V}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{V}[X_i] + \sum_{i \neq j} \text{Cov}[X_i, X_j]$
- $\mathbb{V}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{V}[X_i]$ if $X_i \perp\!\!\!\perp X_j$

Standard deviation

$$\text{sd}[X] = \sqrt{\mathbb{V}[X]} = \sigma_X$$

Covariance

- $\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]$
- $\text{Cov}[X, a] = 0$
- $\text{Cov}[X, X] = \mathbb{V}[X]$
- $\text{Cov}[X, Y] = \text{Cov}[Y, X]$

- $\text{Cov}[aX, bY] = ab\text{Cov}[X, Y]$
- $\text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]$
- $\text{Cov}\left[\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right] = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}[X_i, Y_j]$

Correlation

$$\rho[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$$

Independence

$$X \perp\!\!\!\perp Y \implies \rho[X, Y] = 0 \iff \text{Cov}[X, Y] = 0 \iff \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Conditional variance

- $\mathbb{V}[Y|X] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X] = \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2$
- $\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$

6 Inequalities

CAUCHY-SCHWARZ

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

MARKOV

$$\mathbb{P}[\varphi(X) \geq t] \leq \frac{\mathbb{E}[\varphi(X)]}{t}$$

CHEBYSHEV

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathbb{V}[X]}{t^2}$$

CHERNOFF

$$\mathbb{P}[X \geq (1 + \delta)\mu] \leq \left(\frac{e^\delta}{(1 + \delta)^{1 + \delta}}\right) \quad \delta > -1$$

HOEFFDING

$$X_1, \dots, X_n \text{ independent} \wedge \mathbb{P}[X_i \in [a_i, b_i]] = 1 \wedge 1 \leq i \leq n$$

$$\mathbb{P}[\bar{X} - \mathbb{E}[\bar{X}] \geq t] \leq e^{-2nt^2} \quad t > 0$$

$$\mathbb{P}[|\bar{X} - \mathbb{E}[\bar{X}]| \geq t] \leq 2 \exp\left\{-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\} \quad t > 0$$

JENSEN

$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]) \quad \varphi \text{ convex}$$

7 Distribution Relationships

Binomial

- $X_i \sim \text{Bern}(p) \implies \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$
- $X \sim \text{Bin}(n, p), Y \sim \text{Bin}(m, p) \implies X + Y \sim \text{Bin}(n + m, p)$
- $\lim_{n \rightarrow \infty} \text{Bin}(n, p) = \text{Po}(np) \quad (n \text{ large, } p \text{ small})$
- $\lim_{n \rightarrow \infty} \text{Bin}(n, p) = \mathcal{N}(np, np(1-p)) \quad (n \text{ large, } p \text{ far from 0 and 1})$

Negative Binomial

- $X \sim \text{NBin}(1, p) = \text{Geo}(p)$
- $X \sim \text{NBin}(r, p) = \sum_{i=1}^r \text{Geo}(p)$
- $X_i \sim \text{NBin}(r_i, p) \implies \sum X_i \sim \text{NBin}(\sum r_i, p)$
- $X \sim \text{NBin}(r, p) \cdot Y \sim \text{Bin}(s + r, p) \implies \mathbb{P}[X \leq s] = \mathbb{P}[Y \geq r]$

Poisson

- $X_i \sim \text{Po}(\lambda_i) \wedge X_i \perp\!\!\!\perp X_j \implies \sum_{i=1}^n X_i \sim \text{Po}\left(\sum_{i=1}^n \lambda_i\right)$
- $X_i \sim \text{Po}(\lambda_i) \wedge X_i \perp\!\!\!\perp X_j \implies X_i \left| \sum_{j=1}^n X_j \sim \text{Bin}\left(\sum_{j=1}^n X_j, \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}\right)\right.$

Exponential

- $X_i \sim \text{Exp}(\beta) \wedge X_i \perp\!\!\!\perp X_j \implies \sum_{i=1}^n X_i \sim \text{Gamma}(n, \beta)$
- Memoryless property: $\mathbb{P}[X > x + y | X > y] = \mathbb{P}[X > x]$

Normal

- $X \sim \mathcal{N}(\mu, \sigma^2) \implies \left(\frac{X - \mu}{\sigma}\right) \sim \mathcal{N}(0, 1)$
- $X \sim \mathcal{N}(\mu, \sigma^2) \wedge Z = aX + b \implies Z \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \wedge X_i \perp\!\!\!\perp X_j \implies \sum_i X_i \sim \mathcal{N}(\sum_i \mu_i, \sum_i \sigma_i^2)$
- $\mathbb{P}[a < X \leq b] = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$
- $\Phi(-x) = 1 - \Phi(x) \quad \phi'(x) = -x\phi(x) \quad \phi''(x) = (x^2 - 1)\phi(x)$
- Upper quantile of $\mathcal{N}(0, 1)$: $z_\alpha = \Phi^{-1}(1 - \alpha)$

Gamma

- $X \sim \text{Gamma}(\alpha, \beta) \iff X/\beta \sim \text{Gamma}(\alpha, 1)$
- $\text{Gamma}(\alpha, \beta) \sim \sum_{i=1}^\alpha \text{Exp}(\beta)$
- $X_i \sim \text{Gamma}(\alpha_i, \beta) \wedge X_i \perp\!\!\!\perp X_j \implies \sum_i X_i \sim \text{Gamma}(\sum_i \alpha_i, \beta)$

- $\frac{\Gamma(\alpha)}{\lambda^\alpha} = \int_0^\infty x^{\alpha-1} e^{-\lambda x} dx$

Beta

- $\frac{1}{\mathbb{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$
- $\mathbb{E}[X^k] = \frac{\mathbb{B}(\alpha + k, \beta)}{\mathbb{B}(\alpha, \beta)} = \frac{\alpha + k - 1}{\alpha + \beta + k - 1} \mathbb{E}[X^{k-1}]$
- Beta(1, 1) \sim Unif(0, 1)

8 Probability and Moment Generating Functions

- $G_X(t) = \mathbb{E}[t^X] \quad |t| < 1$
- $M_X(t) = G_X(e^t) = \mathbb{E}[e^{Xt}] = \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{(Xt)^i}{i!}\right] = \sum_{i=0}^{\infty} \frac{\mathbb{E}[X^i]}{i!} \cdot t^i$
- $\mathbb{P}[X = 0] = G_X(0)$
- $\mathbb{P}[X = 1] = G'_X(0)$
- $\mathbb{P}[X = i] = \frac{G_X^{(i)}(0)}{i!}$
- $\mathbb{E}[X] = G'_X(1^-)$
- $\mathbb{E}[X^k] = M_X^{(k)}(0)$
- $\mathbb{E}\left[\frac{X!}{(X-k)!}\right] = G_X^{(k)}(1^-)$
- $\mathbb{V}[X] = G''_X(1^-) + G'_X(1^-) - (G'_X(1^-))^2$
- $G_X(t) = G_Y(t) \implies X \stackrel{d}{=} Y$

9 Multivariate Distributions

9.1 Standard Bivariate Normal

Let $X, Y \sim \mathcal{N}(0, 1) \wedge X \perp\!\!\!\perp Z$ where $Y = \rho X + \sqrt{1 - \rho^2} Z$

Joint density

$$f(x, y) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left\{-\frac{x^2 + y^2 - 2\rho xy}{2(1 - \rho^2)}\right\}$$

Conditionals

$$(Y | X = x) \sim \mathcal{N}(\rho x, 1 - \rho^2) \quad \text{and} \quad (X | Y = y) \sim \mathcal{N}(\rho y, 1 - \rho^2)$$

Independence

$$X \perp\!\!\!\perp Y \iff \rho = 0$$

9.2 Bivariate Normal

Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$.

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} \exp\left\{-\frac{z}{2(1 - \rho^2)}\right\}$$

$$z = \left[\left(\frac{x - \mu_x}{\sigma_x}\right)^2 + \left(\frac{y - \mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x - \mu_x}{\sigma_x}\right)\left(\frac{y - \mu_y}{\sigma_y}\right)\right]$$

Conditional mean and variance

$$\mathbb{E}[X | Y] = \mathbb{E}[X] + \rho \frac{\sigma_X}{\sigma_Y} (Y - \mathbb{E}[Y])$$

$$\mathbb{V}[X | Y] = \sigma_X \sqrt{1 - \rho^2}$$

9.3 Multivariate Normal

Covariance matrix Σ (Precision matrix Σ^{-1})

$$\Sigma = \begin{pmatrix} \mathbb{V}[X_1] & \cdots & \text{Cov}[X_1, X_k] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_k, X_1] & \cdots & \mathbb{V}[X_k] \end{pmatrix}$$

If $X \sim \mathcal{N}(\mu, \Sigma)$,

$$f_X(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\}$$

Properties

- $Z \sim \mathcal{N}(0, 1) \wedge X = \mu + \Sigma^{1/2} Z \implies X \sim \mathcal{N}(\mu, \Sigma)$
- $X \sim \mathcal{N}(\mu, \Sigma) \implies \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, 1)$
- $X \sim \mathcal{N}(\mu, \Sigma) \implies AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$
- $X \sim \mathcal{N}(\mu, \Sigma) \wedge \|a\| = k \implies a^T X \sim \mathcal{N}(a^T \mu, a^T \Sigma a)$

10 Convergence

Let $\{X_1, X_2, \dots\}$ be a sequence of RV's and let X be another RV. Let F_n denote the CDF of X_n and let F denote the CDF of X .

Types of Convergence

1. In distribution (weakly, in law): $X_n \xrightarrow{D} X$

$$\lim_{n \rightarrow \infty} F_n(t) = F(t) \quad \forall t \text{ where } F \text{ continuous}$$

2. In probability: $X_n \xrightarrow{P} X$

$$(\forall \varepsilon > 0) \lim_{n \rightarrow \infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0$$

3. Almost surely (strongly): $X_n \xrightarrow{as} X$

$$\mathbb{P}\left[\lim_{n \rightarrow \infty} X_n = X\right] = \mathbb{P}\left[\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right] = 1$$

4. In quadratic mean (L_2): $X_n \xrightarrow{qm} X$

$$\lim_{n \rightarrow \infty} \mathbb{E}[(X_n - X)^2] = 0$$

Relationships

- $X_n \xrightarrow{qm} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$
- $X_n \xrightarrow{as} X \implies X_n \xrightarrow{P} X$
- $X_n \xrightarrow{D} X \wedge (\exists c \in \mathbb{R}) \mathbb{P}[X = c] = 1 \implies X_n \xrightarrow{P} X$
- $X_n \xrightarrow{P} X \wedge Y_n \xrightarrow{P} Y \implies X_n + Y_n \xrightarrow{P} X + Y$
- $X_n \xrightarrow{qm} X \wedge Y_n \xrightarrow{qm} Y \implies X_n + Y_n \xrightarrow{qm} X + Y$
- $X_n \xrightarrow{P} X \wedge Y_n \xrightarrow{P} Y \implies X_n Y_n \xrightarrow{P} XY$
- $X_n \xrightarrow{P} X \implies \varphi(X_n) \xrightarrow{P} \varphi(X)$
- $X_n \xrightarrow{D} X \implies \varphi(X_n) \xrightarrow{D} \varphi(X)$
- $X_n \xrightarrow{qm} b \iff \lim_{n \rightarrow \infty} \mathbb{E}[X_n] = b \wedge \lim_{n \rightarrow \infty} \mathbb{V}[X_n] = 0$
- $X_1, \dots, X_n \text{ IID} \wedge \mathbb{E}[X] = \mu \wedge \mathbb{V}[X] < \infty \iff \bar{X}_n \xrightarrow{qm} \mu$

SLUTZKY'S THEOREM

- $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c \implies X_n + Y_n \xrightarrow{D} X + c$
- $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c \implies X_n Y_n \xrightarrow{D} cX$
- In general: $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} Y \not\implies X_n + Y_n \xrightarrow{D} X + Y$

10.1 Law of Large Numbers (LLN)

Let $\{X_1, \dots, X_n\}$ be a sequence of IID RV's, $\mathbb{E}[X_1] = \mu$.

Weak (WLLN)

$$\bar{X}_n \xrightarrow{P} \mu \quad n \rightarrow \infty$$

Strong (SLLN)

$$\bar{X}_n \xrightarrow{as} \mu \quad n \rightarrow \infty$$

10.2 Central Limit Theorem (CLT)

Let $\{X_1, \dots, X_n\}$ be a sequence of IID RV's, $\mathbb{E}[X_1] = \mu$, and $\mathbb{V}[X_1] = \sigma^2$.

$$Z_n := \frac{\bar{X}_n - \mu}{\sqrt{\mathbb{V}[\bar{X}_n]}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} Z \quad \text{where } Z \sim \mathcal{N}(0, 1)$$

$$\lim_{n \rightarrow \infty} \mathbb{P}[Z_n \leq z] = \Phi(z) \quad z \in \mathbb{R}$$

CLT notations

$$Z_n \approx \mathcal{N}(0, 1)$$

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X}_n - \mu \approx \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\bar{X}_n - \mu) \approx \mathcal{N}(0, \sigma^2)$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx \mathcal{N}(0, 1)$$

Continuity correction

$$\mathbb{P}[\bar{X}_n \leq x] \approx \Phi\left(\frac{x + \frac{1}{2} - \mu}{\sigma/\sqrt{n}}\right)$$

$$\mathbb{P}[\bar{X}_n \geq x] \approx 1 - \Phi\left(\frac{x - \frac{1}{2} - \mu}{\sigma/\sqrt{n}}\right)$$

Delta method

$$Y_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \varphi(Y_n) \approx \mathcal{N}\left(\varphi(\mu), (\varphi'(\mu))^2 \frac{\sigma^2}{n}\right)$$

11 Statistical Inference

Let $X_1, \dots, X_n \stackrel{iid}{\sim} F$ if not otherwise noted.

11.1 Point Estimation

- Point estimator $\hat{\theta}_n$ of θ is a RV: $\hat{\theta}_n = g(X_1, \dots, X_n)$
- bias($\hat{\theta}_n$) = $\mathbb{E}[\hat{\theta}_n] - \theta$
- Consistency: $\hat{\theta}_n \xrightarrow{P} \theta$
- Sampling distribution: $F(\hat{\theta}_n)$
- Standard error: $\text{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}[\hat{\theta}_n]}$

- Mean squared error: $\text{MSE} = \mathbb{E} [(\hat{\theta}_n - \theta)^2] = \text{bias}(\hat{\theta}_n)^2 + \mathbb{V} [\hat{\theta}_n]$
- $\lim_{n \rightarrow \infty} \text{bias}(\hat{\theta}_n) = 0 \wedge \lim_{n \rightarrow \infty} \text{se}(\hat{\theta}_n) = 0 \implies \hat{\theta}_n$ is consistent
- Asymptotic normality: $\frac{\hat{\theta}_n - \theta}{\text{se}} \xrightarrow{D} \mathcal{N}(0, 1)$
- SLUTZKY'S THEOREM often lets us replace $\text{se}(\hat{\theta}_n)$ by some (weakly) consistent estimator $\hat{\sigma}_n$.

11.2 Normal-Based Confidence Interval

Suppose $\hat{\theta}_n \approx \mathcal{N}(\theta, \hat{\text{se}}^2)$. Let $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$, i.e., $\mathbb{P}[Z > z_{\alpha/2}] = \alpha/2$ and $\mathbb{P}[-z_{\alpha/2} < Z < z_{\alpha/2}] = 1 - \alpha$ where $Z \sim \mathcal{N}(0, 1)$. Then

$$C_n = \hat{\theta}_n \pm z_{\alpha/2} \hat{\text{se}}$$

11.3 Empirical distribution

Empirical Distribution Function (ECDF)

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n}$$

$$I(X_i \leq x) = \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases}$$

Properties (for any fixed x)

- $\mathbb{E} [\hat{F}_n] = F(x)$
- $\mathbb{V} [\hat{F}_n] = \frac{F(x)(1 - F(x))}{n}$
- $\text{MSE} = \frac{F(x)(1 - F(x))}{n} \xrightarrow{D} 0$
- $\hat{F}_n \xrightarrow{P} F(x)$

DVORETZKY-KIEFER-WOLFOWITZ (DKW) inequality ($X_1, \dots, X_n \sim F$)

$$\mathbb{P} \left[\sup_x |F(x) - \hat{F}_n(x)| > \varepsilon \right] = 2e^{-2n\varepsilon^2}$$

Nonparametric $1 - \alpha$ confidence band for F

$$L(x) = \max\{\hat{F}_n - \epsilon_n, 0\}$$

$$U(x) = \min\{\hat{F}_n + \epsilon_n, 1\}$$

$$\epsilon = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$$

$$\mathbb{P}[L(x) \leq F(x) \leq U(x) \forall x] \geq 1 - \alpha$$

11.4 Statistical Functionals

- Statistical functional: $T(F)$
- Plug-in estimator of $\theta = (F)$: $\hat{\theta}_n = T(\hat{F}_n)$
- Linear functional: $T(F) = \int \varphi(x) dF_X(x)$
- Plug-in estimator for linear functional:

$$T(\hat{F}_n) = \int \varphi(x) d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \varphi(X_i)$$

- Often: $T(\hat{F}_n) \approx \mathcal{N}(T(F), \hat{\text{se}}^2) \implies T(\hat{F}_n) \pm z_{\alpha/2} \hat{\text{se}}$
- p^{th} quantile: $F^{-1}(p) = \inf\{x : F(x) \geq p\}$
- $\hat{\mu} = \bar{X}_n$
- $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
- $\hat{\kappa} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^3}{\hat{\sigma}^3}$
- $\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$

12 Parametric Inference

Let $\mathfrak{F} = \{f(x; \theta) : \theta \in \Theta\}$ be a parametric model with parameter space $\Theta \subset \mathbb{R}^k$ and parameter $\theta = (\theta_1, \dots, \theta_k)$.

12.1 Method of Moments

j^{th} moment

$$\alpha_j(\theta) = \mathbb{E}[X^j] = \int x^j dF_X(x)$$

j^{th} sample moment

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^n X_i^j$$

Method of Moments estimator (MoM)

$$\alpha_1(\theta) = \hat{\alpha}_1$$

$$\alpha_2(\theta) = \hat{\alpha}_2$$

$$\vdots = \vdots$$

$$\alpha_k(\theta) = \hat{\alpha}_k$$

Properties of the MoM estimator

- $\widehat{\theta}_n$ exists with probability tending to 1
- Consistency: $\widehat{\theta}_n \xrightarrow{P} \theta$
- Asymptotic normality:

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{D} \mathcal{N}(0, \Sigma)$$

where $\Sigma = g\mathbb{E}[YY^T]g^T$, $Y = (X, X^2, \dots, X^k)^T$,
 $g = (g_1, \dots, g_k)$ and $g_j = \frac{\partial}{\partial \theta} \alpha_j^{-1}(\theta)$

12.2 Maximum Likelihood

Likelihood: $\mathcal{L}_n : \Theta \rightarrow [0, \infty)$

$$\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i; \theta)$$

Log-likelihood

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^n \log f(X_i; \theta)$$

Maximum likelihood estimator (MLE)

$$\mathcal{L}_n(\widehat{\theta}_n) = \sup_{\theta} \mathcal{L}_n(\theta)$$

Score function

$$s(X; \theta) = \frac{\partial}{\partial \theta} \log f(X; \theta)$$

Fisher information

$$I(\theta) = \mathbb{V}_{\theta} [s(X; \theta)]$$

$$I_n(\theta) = nI(\theta)$$

Fisher information (exponential family)

$$I(\theta) = \mathbb{E}_{\theta} \left[-\frac{\partial}{\partial \theta} s(X; \theta) \right]$$

Observed Fisher information

$$I_n^{obs}(\theta) = -\frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \log f(X_i; \theta)$$

Properties of the MLE

- Consistency: $\widehat{\theta}_n \xrightarrow{P} \theta$

- Equivariance: $\widehat{\theta}_n$ is the MLE $\implies \varphi(\widehat{\theta}_n)$ is the MLE of $\varphi(\theta)$
- Asymptotic optimality (or efficiency), i.e., smallest variance for large samples. If $\widetilde{\theta}_n$ is any other estimator, the asymptotic relative efficiency is:

$$1. \text{ se} \approx \sqrt{1/I_n(\theta)}$$

$$\frac{(\widehat{\theta}_n - \theta)}{\text{se}} \xrightarrow{D} \mathcal{N}(0, 1)$$

$$2. \widehat{\text{se}} \approx \sqrt{1/I_n(\widehat{\theta}_n)}$$

$$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\text{se}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

- Asymptotic optimality

$$\text{ARE}(\widehat{\theta}_n, \widetilde{\theta}_n) = \frac{\mathbb{V} \left[\widehat{\theta}_n \right]}{\mathbb{V} \left[\widetilde{\theta}_n \right]} \leq 1$$

- Approximately the Bayes estimator

12.2.1 Delta Method

If $\tau = \varphi(\widehat{\theta})$ where φ is differentiable and $\varphi'(\theta) \neq 0$:

$$\frac{(\widehat{\tau}_n - \tau)}{\widehat{\text{se}}(\widehat{\tau})} \xrightarrow{D} \mathcal{N}(0, 1)$$

where $\widehat{\tau} = \varphi(\widehat{\theta})$ is the MLE of τ and

$$\widehat{\text{se}} = \left| \varphi'(\widehat{\theta}) \right| \widehat{\text{se}}(\widehat{\theta}_n)$$

12.3 Multiparameter Models

Let $\theta = (\theta_1, \dots, \theta_k)$ and $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_k)$ be the MLE.

$$H_{jj} = \frac{\partial^2 \ell_n}{\partial \theta_j^2} \quad H_{jk} = \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta_k}$$

Fisher information matrix

$$I_n(\theta) = - \begin{bmatrix} \mathbb{E}_{\theta} [H_{11}] & \cdots & \mathbb{E}_{\theta} [H_{1k}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_{\theta} [H_{k1}] & \cdots & \mathbb{E}_{\theta} [H_{kk}] \end{bmatrix}$$

Under appropriate regularity conditions

$$(\widehat{\theta} - \theta) \approx \mathcal{N}(0, J_n)$$

with $J_n(\theta) = I_n^{-1}$. Further, if $\hat{\theta}_j$ is the j^{th} component of θ , then

$$\frac{(\hat{\theta}_j - \theta_j)}{\widehat{\text{se}}_j} \xrightarrow{D} \mathcal{N}(0, 1)$$

where $\widehat{\text{se}}_j^2 = J_n(j, j)$ and $\text{Cov}[\hat{\theta}_j, \hat{\theta}_k] = J_n(j, k)$

12.3.1 Multiparameter delta method

Let $\tau = \varphi(\theta_1, \dots, \theta_k)$ and let the gradient of φ be

$$\nabla\varphi = \begin{pmatrix} \frac{\partial\varphi}{\partial\theta_1} \\ \vdots \\ \frac{\partial\varphi}{\partial\theta_k} \end{pmatrix}$$

Suppose $\nabla\varphi|_{\theta=\hat{\theta}} \neq 0$ and $\hat{\tau} = \varphi(\hat{\theta})$. Then,

$$\frac{(\hat{\tau} - \tau)}{\widehat{\text{se}}(\hat{\tau})} \xrightarrow{D} \mathcal{N}(0, 1)$$

where

$$\widehat{\text{se}}(\hat{\tau}) = \sqrt{(\widehat{\nabla}\varphi)^T \widehat{J}_n(\widehat{\nabla}\varphi)}$$

and $\widehat{J}_n = J_n(\hat{\theta})$ and $\widehat{\nabla}\varphi = \nabla\varphi|_{\theta=\hat{\theta}}$.

12.4 Parametric Bootstrap

Sample from $f(x; \hat{\theta}_n)$ instead of from \widehat{F}_n , where $\hat{\theta}_n$ could be the MLE or method of moments estimator.

13 Hypothesis Testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

Definitions

- Null hypothesis H_0
- Alternative hypothesis H_1
- Simple hypothesis $\theta = \theta_0$
- Composite hypothesis $\theta > \theta_0$ or $\theta < \theta_0$
- Two-sided test: $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
- One-sided test: $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$

- Critical value c
- Test statistic T
- Rejection region $R = \{x : T(x) > c\}$
- Power function $\beta(\theta) = \mathbb{P}[X \in R]$
- Power of a test: $1 - \mathbb{P}[\text{Type II error}] = 1 - \beta = \inf_{\theta \in \Theta_1} \beta(\theta)$
- Test size: $\alpha = \mathbb{P}[\text{Type I error}] = \sup_{\theta \in \Theta_0} \beta(\theta)$

	Retain H_0	Reject H_0
H_0 true	✓	Type I Error (α)
H_1 true	Type II Error (β)	✓ (power)

p-value

- p-value = $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta [T(X) \geq T(x)] = \inf\{\alpha : T(x) \in R_\alpha\}$
- p-value = $\sup_{\theta \in \Theta_0} \underbrace{\mathbb{P}_\theta [T(X^*) \geq T(X)]}_{1 - F_\theta(T(X))} = \inf\{\alpha : T(X) \in R_\alpha\}$ since $T(X^*) \sim F_\theta$

p-value	evidence
< 0.01	very strong evidence against H_0
0.01 – 0.05	strong evidence against H_0
0.05 – 0.1	weak evidence against H_0
> 0.1	little or no evidence against H_0

Wald test

- Two-sided test
- Reject H_0 when $|W| > z_{\alpha/2}$ where $W = \frac{\hat{\theta} - \theta_0}{\widehat{\text{se}}}$
- $\mathbb{P}[|W| > z_{\alpha/2}] \rightarrow \alpha$
- p-value = $\mathbb{P}_{\theta_0}[|W| > |w|] \approx \mathbb{P}[|Z| > |w|] = 2\Phi(-|w|)$

Likelihood ratio test

- $T(X) = \frac{\sup_{\theta \in \Theta} \mathcal{L}_n(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}_n(\theta)} = \frac{\mathcal{L}_n(\hat{\theta}_n)}{\mathcal{L}_n(\hat{\theta}_{n,0})}$
- $\lambda(X) = 2 \log T(X) \xrightarrow{D} \chi_{r-q}^2$ where $\sum_{i=1}^k Z_i^2 \sim \chi_k^2$ and $Z_1, \dots, Z_k \stackrel{iid}{\sim} \mathcal{N}(0, 1)$
- p-value = $\mathbb{P}_{\theta_0}[\lambda(X) > \lambda(x)] \approx \mathbb{P}[\chi_{r-q}^2 > \lambda(x)]$

Multinomial LRT

- MLE: $\hat{p}_n = \left(\frac{X_1}{n}, \dots, \frac{X_k}{n} \right)$
- $T(X) = \frac{\mathcal{L}_n(\hat{p}_n)}{\mathcal{L}_n(p_0)} = \prod_{j=1}^k \left(\frac{\hat{p}_j}{p_{0j}} \right)^{X_j}$
- $\lambda(X) = 2 \sum_{j=1}^k X_j \log \left(\frac{\hat{p}_j}{p_{0j}} \right) \xrightarrow{D} \chi_{k-1}^2$
- The approximate size α LRT rejects H_0 when $\lambda(X) \geq \chi_{k-1, \alpha}^2$

Pearson Chi-square Test

- $T = \sum_{j=1}^k \frac{(X_j - \mathbb{E}[X_j])^2}{\mathbb{E}[X_j]}$ where $\mathbb{E}[X_j] = np_{0j}$ under H_0
- $T \xrightarrow{D} \chi_{k-1}^2$
- p-value = $\mathbb{P}[\chi_{k-1}^2 > T(x)]$
- Faster $\xrightarrow{D} \chi_{k-1}^2$ than LRT, hence preferable for small n

Independence testing

- I rows, J columns, \mathbf{X} multinomial sample of size $n = I * J$
- MLEs unconstrained: $\hat{p}_{ij} = \frac{X_{ij}}{n}$
- MLEs under H_0 : $\hat{p}_{0ij} = \hat{p}_i \hat{p}_{\cdot j} = \frac{X_i}{n} \frac{X_{\cdot j}}{n}$
- LRT: $\lambda = 2 \sum_{i=1}^I \sum_{j=1}^J X_{ij} \log \left(\frac{n X_{ij}}{X_i X_{\cdot j}} \right)$
- PearsonChiSq: $T = \sum_{i=1}^I \sum_{j=1}^J \frac{(X_{ij} - \mathbb{E}[X_{ij}])^2}{\mathbb{E}[X_{ij}]}$
- LRT and Pearson $\xrightarrow{D} \chi_k^2 \nu$, where $\nu = (I - 1)(J - 1)$

14 Exponential Family

Scalar parameter

$$\begin{aligned} f_X(x | \theta) &= h(x) \exp \{ \eta(\theta) T(x) - A(\theta) \} \\ &= h(x) g(\theta) \exp \{ \eta(\theta) T(x) \} \end{aligned}$$

Vector parameter

$$\begin{aligned} f_X(x | \theta) &= h(x) \exp \left\{ \sum_{i=1}^s \eta_i(\theta) T_i(x) - A(\theta) \right\} \\ &= h(x) \exp \{ \eta(\theta) \cdot T(x) - A(\theta) \} \\ &= h(x) g(\theta) \exp \{ \eta(\theta) \cdot T(x) \} \end{aligned}$$

Natural form

$$\begin{aligned} f_X(x | \eta) &= h(x) \exp \{ \eta \cdot \mathbf{T}(x) - A(\eta) \} \\ &= h(x) g(\eta) \exp \{ \eta \cdot \mathbf{T}(x) \} \\ &= h(x) g(\eta) \exp \{ \eta^T \mathbf{T}(x) \} \end{aligned}$$

15 Bayesian Inference

BAYES' THEOREM

$$f(\theta | x) = \frac{f(x | \theta) f(\theta)}{f(x^n)} = \frac{f(x | \theta) f(\theta)}{\int f(x | \theta) f(\theta) d\theta} \propto \mathcal{L}_n(\theta) f(\theta)$$

Definitions

- $X^n = (X_1, \dots, X_n)$
- $x^n = (x_1, \dots, x_n)$
- Prior density $f(\theta)$
- Likelihood $f(x^n | \theta)$: joint density of the data

$$\text{In particular, } X^n \text{ IID} \implies f(x^n | \theta) = \prod_{i=1}^n f(x_i | \theta) = \mathcal{L}_n(\theta)$$

- Posterior density $f(\theta | x^n)$
- Normalizing constant $c_n = f(x^n) = \int f(x | \theta) f(\theta) d\theta$
- Kernel: part of a density that depends on θ
- Posterior mean $\bar{\theta}_n = \int \theta f(\theta | x^n) d\theta = \frac{\int \theta \mathcal{L}_n(\theta) f(\theta) d\theta}{\int \mathcal{L}_n(\theta) f(\theta) d\theta}$

15.1 Credible Intervals

Posterior interval

$$\mathbb{P}[\theta \in (a, b) | x^n] = \int_a^b f(\theta | x^n) d\theta = 1 - \alpha$$

Equal-tail credible interval

$$\int_{-\infty}^a f(\theta | x^n) d\theta = \int_b^{\infty} f(\theta | x^n) d\theta = \alpha/2$$

Highest posterior density (HPD) region R_n

1. $\mathbb{P}[\theta \in R_n] = 1 - \alpha$
2. $R_n = \{ \theta : f(\theta | x^n) > k \}$ for some k

R_n is unimodal $\implies R_n$ is an interval

15.2 Function of parameters

Let $\tau = \varphi(\theta)$ and $A = \{\theta : \varphi(\theta) \leq \tau\}$.

Posterior CDF for τ

$$H(r | x^n) = \mathbb{P}[\varphi(\theta) \leq \tau | x^n] = \int_A f(\theta | x^n) d\theta$$

Posterior density

$$h(\tau | x^n) = H'(\tau | x^n)$$

Bayesian delta method

$$\tau | X^n \approx \mathcal{N}\left(\varphi(\hat{\theta}), \widehat{\text{se}}\left|\varphi'(\hat{\theta})\right|\right)$$

15.3 Priors

Choice

- Subjective Bayesianism: prior should incorporate as much detail as possible the research's a priori knowledge—via *prior elicitation*
- Objective Bayesianism: prior should incorporate as little detail as possible (*non-informative* prior)
- Robust Bayesianism: consider various priors and determine *sensitivity* of our inferences to changes in the prior

Types

- Flat: $f(\theta) \propto \text{constant}$
- Proper: $\int_{-\infty}^{\infty} f(\theta) d\theta = 1$
- Improper: $\int_{-\infty}^{\infty} f(\theta) d\theta = \infty$
- JEFFREY'S Prior (transformation-invariant):

$$f(\theta) \propto \sqrt{I(\theta)} \quad f(\theta) \propto \sqrt{\det(I(\theta))}$$

- Conjugate: $f(\theta)$ and $f(\theta | x^n)$ belong to the same parametric family

15.3.1 Conjugate Priors

Continuous likelihood (subscript c denotes constant)		
Likelihood	Conjugate prior	Posterior hyperparameters
Unif(0, θ)	Pareto(x_m, k)	$\max\{x_{(n)}, x_m\}, k + n$
Exp(λ)	Gamma(α, β)	$\alpha + n, \beta + \sum_{i=1}^n x_i$
$\mathcal{N}(\mu, \sigma_c^2)$	$\mathcal{N}(\mu_0, \sigma_0^2)$	$\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma_c^2}\right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_c^2}\right),$ $\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_c^2}\right)^{-1}$
$\mathcal{N}(\mu_c, \sigma^2)$	Scaled Inverse Chi-square(ν, σ_0^2)	$\nu + n, \frac{\nu\sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{\nu + n}$
$\mathcal{N}(\mu, \sigma^2)$	Normal-scaled Inverse Gamma($\lambda, \nu, \alpha, \beta$)	$\frac{\nu\lambda + n\bar{x}}{\nu + n}, \quad \nu + n, \quad \alpha + \frac{n}{2},$ $\beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{\gamma(\bar{x} - \lambda)^2}{2(n + \gamma)}$
MVN(μ, Σ_c)	MVN(μ_0, Σ_0)	$(\Sigma_0^{-1} + n\Sigma_c^{-1})^{-1} (\Sigma_0^{-1}\mu_0 + n\Sigma_c^{-1}\bar{x}),$ $(\Sigma_0^{-1} + n\Sigma_c^{-1})^{-1}$
MVN(μ_c, Σ)	Inverse-Wishart(κ, Ψ)	$n + \kappa, \Psi + \sum_{i=1}^n (x_i - \mu_c)(x_i - \mu_c)^T$
Pareto(x_{m_c}, k)	Gamma(α, β)	$\alpha + n, \beta + \sum_{i=1}^n \log \frac{x_i}{x_{m_c}}$
Pareto(x_m, k_c)	Pareto(x_0, k_0)	$x_0, k_0 - kn$ where $k_0 > kn$
Gamma(α_c, β)	Gamma(α_0, β_0)	$\alpha_0 + n\alpha_c, \beta_0 + \sum_{i=1}^n x_i$

Discrete likelihood		
Likelihood	Conjugate prior	Posterior hyperparameters
Bern (p)	Beta (α, β)	$\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i$
Bin (p)	Beta (α, β)	$\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$
NBin (p)	Beta (α, β)	$\alpha + rn, \beta + \sum_{i=1}^n x_i$
Po (λ)	Gamma (α, β)	$\alpha + \sum_{i=1}^n x_i, \beta + n$
Multinomial(p)	Dir (α)	$\alpha + \sum_{i=1}^n x^{(i)}$
Geo (p)	Beta (α, β)	$\alpha + n, \beta + \sum_{i=1}^n x_i$

15.4 Bayesian Testing

If $H_0 : \theta \in \Theta_0$:

$$\text{Prior probability } \mathbb{P}[H_0] = \int_{\Theta_0} f(\theta) d\theta$$

$$\text{Posterior probability } \mathbb{P}[H_0 | x^n] = \int_{\Theta_0} f(\theta | x^n) d\theta$$

Let $H_0 \dots H_{k-1}$ be k hypotheses. Suppose $\theta \sim f(\theta | H_k)$,

$$\mathbb{P}[H_k | x^n] = \frac{f(x^n | H_k) \mathbb{P}[H_k]}{\sum_{k=1}^K f(x^n | H_k) \mathbb{P}[H_k]},$$

Marginal likelihood

$$f(x^n | H_i) = \int_{\Theta} f(x^n | \theta, H_i) f(\theta | H_i) d\theta$$

Posterior odds (of H_i relative to H_j)

$$\frac{\mathbb{P}[H_i | x^n]}{\mathbb{P}[H_j | x^n]} = \underbrace{\frac{f(x^n | H_i)}{f(x^n | H_j)}}_{\text{Bayes Factor } BF_{ij}} \times \underbrace{\frac{\mathbb{P}[H_i]}{\mathbb{P}[H_j]}}_{\text{prior odds}}$$

Bayes factor

$\log_{10} BF_{10}$	BF_{10}	evidence
0 – 0.5	1 – 1.5	Weak
0.5 – 1	1.5 – 10	Moderate
1 – 2	10 – 100	Strong
> 2	> 100	Decisive

$$p^* = \frac{\frac{p}{1-p} BF_{10}}{1 + \frac{p}{1-p} BF_{10}} \text{ where } p = \mathbb{P}[H_1] \text{ and } p^* = \mathbb{P}[H_1 | x^n]$$

16 Sampling Methods

16.1 Inverse Transform Sampling

Setup

- $U \sim \text{Unif}(0, 1)$
- $X \sim F$
- $F^{-1}(u) = \inf\{x | F(x) \geq u\}$

Algorithm

1. Generate $u \sim \text{Unif}(0, 1)$
2. Compute $x = F^{-1}(u)$

16.2 The Bootstrap

Let $T_n = g(X_1, \dots, X_n)$ be a statistic.

1. Estimate $\mathbb{V}_F[T_n]$ with $\mathbb{V}_{\hat{F}_n}[T_n]$.
2. Approximate $\mathbb{V}_{\hat{F}_n}[T_n]$ using simulation:
 - (a) Repeat the following B times to get $T_{n,1}^*, \dots, T_{n,B}^*$, an IID sample from the sampling distribution implied by \hat{F}_n
 - i. Sample uniformly $X_1^*, \dots, X_n^* \sim \hat{F}_n$.
 - ii. Compute $T_n^* = g(X_1^*, \dots, X_n^*)$.
 - (b) Then

$$v_{boot} = \hat{\mathbb{V}}_{\hat{F}_n} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2$$

16.2.1 Bootstrap Confidence Intervals

Normal-based interval

$$T_n \pm z_{\alpha/2} \hat{\mathbf{s}}e_{boot}$$

Pivotal interval

1. Location parameter $\theta = T(F)$

2. Pivot $R_n = \hat{\theta}_n - \theta$
3. Let $H(r) = \mathbb{P}[R_n \leq r]$ be the CDF of R_n
4. Let $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$. Approximate H using bootstrap:

$$\hat{H}(r) = \frac{1}{B} \sum_{b=1}^B I(R_{n,b}^* \leq r)$$

5. $\theta_\beta^* = \beta$ sample quantile of $(\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,B}^*)$
6. $r_\beta^* = \beta$ sample quantile of $(R_{n,1}^*, \dots, R_{n,B}^*)$, i.e., $r_\beta^* = \theta_\beta^* - \hat{\theta}_n$
7. Approximate $1 - \alpha$ confidence interval $C_n = (\hat{a}, \hat{b})$ where

$$\begin{aligned} \hat{a} &= \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r_{1-\alpha/2}^* = 2\hat{\theta}_n - \theta_{1-\alpha/2}^* \\ \hat{b} &= \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r_{\alpha/2}^* = 2\hat{\theta}_n - \theta_{\alpha/2}^* \end{aligned}$$

Percentile interval

$$C_n = \left(\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*\right)$$

16.3 Rejection Sampling

Setup

- We can easily sample from $g(\theta)$
- We want to sample from $h(\theta)$, but it is difficult
- We know $h(\theta)$ up to a proportional constant: $h(\theta) = \frac{k(\theta)}{\int k(\theta) d\theta}$
- Envelope condition: we can find $M > 0$ such that $k(\theta) \leq Mg(\theta) \quad \forall \theta$

Algorithm

1. Draw $\theta^{cand} \sim g(\theta)$
2. Generate $u \sim \text{Unif}(0, 1)$
3. Accept θ^{cand} if $u \leq \frac{k(\theta^{cand})}{Mg(\theta^{cand})}$
4. Repeat until B values of θ^{cand} have been accepted

Example

- We can easily sample from the prior $g(\theta) = f(\theta)$
- Target is the posterior $h(\theta) \propto k(\theta) = f(x^n | \theta)f(\theta)$
- Envelope condition: $f(x^n | \theta) \leq f(x^n | \hat{\theta}_n) = \mathcal{L}_n(\hat{\theta}_n) \equiv M$
- Algorithm
 1. Draw $\theta^{cand} \sim f(\theta)$

2. Generate $u \sim \text{Unif}(0, 1)$
3. Accept θ^{cand} if $u \leq \frac{\mathcal{L}_n(\theta^{cand})}{\mathcal{L}_n(\hat{\theta}_n)}$

16.4 Importance Sampling

Sample from an importance function g rather than target density h . Algorithm to obtain an approximation to $\mathbb{E}[q(\theta) | x^n]$:

1. Sample from the prior $\theta_1, \dots, \theta_n \stackrel{iid}{\sim} f(\theta)$
2. $w_i = \frac{\mathcal{L}_n(\theta_i)}{\sum_{i=1}^B \mathcal{L}_n(\theta_i)} \quad \forall i = 1, \dots, B$
3. $\mathbb{E}[q(\theta) | x^n] \approx \sum_{i=1}^B q(\theta_i)w_i$

17 Decision Theory

Definitions

- Unknown quantity affecting our decision: $\theta \in \Theta$
- Decision rule: synonymous for an estimator $\hat{\theta}$
- Action $a \in \mathcal{A}$: possible value of the decision rule. In the estimation context, the action is just an estimate of θ , $\hat{\theta}(x)$.
- Loss function L : consequences of taking action a when true state is θ or discrepancy between θ and $\hat{\theta}$, $L : \Theta \times \mathcal{A} \rightarrow [-k, \infty)$.

Loss functions

- Squared error loss: $L(\theta, a) = (\theta - a)^2$
- Linear loss: $L(\theta, a) = \begin{cases} K_1(\theta - a) & a - \theta < 0 \\ K_2(a - \theta) & a - \theta \geq 0 \end{cases}$
- Absolute error loss: $L(\theta, a) = |\theta - a|$ (linear loss with $K_1 = K_2$)
- L_p loss: $L(\theta, a) = |\theta - a|^p$
- Zero-one loss: $L(\theta, a) = \begin{cases} 0 & a = \theta \\ 1 & a \neq \theta \end{cases}$

17.1 Risk

Posterior risk

$$r(\hat{\theta} | x) = \int L(\theta, \hat{\theta}(x))f(\theta | x) d\theta = \mathbb{E}_{\theta|X} [L(\theta, \hat{\theta}(x))]$$

(Frequentist) risk

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x))f(x | \theta) dx = \mathbb{E}_{X|\theta} [L(\theta, \hat{\theta}(X))]$$

Bayes risk

$$r(f, \hat{\theta}) = \iint L(\theta, \hat{\theta}(x)) f(x, \theta) dx d\theta = \mathbb{E}_{\theta, X} [L(\theta, \hat{\theta}(X))]$$

$$r(f, \hat{\theta}) = \mathbb{E}_{\theta} \left[\mathbb{E}_{X|\theta} [L(\theta, \hat{\theta}(X))] \right] = \mathbb{E}_{\theta} [R(\theta, \hat{\theta})]$$

$$r(f, \hat{\theta}) = \mathbb{E}_X \left[\mathbb{E}_{\theta|X} [L(\theta, \hat{\theta}(X))] \right] = \mathbb{E}_X [r(\hat{\theta} | X)]$$

17.2 Admissibility

- $\hat{\theta}'$ dominates $\hat{\theta}$ if

$$\forall \theta : R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta})$$

$$\exists \theta : R(\theta, \hat{\theta}') < R(\theta, \hat{\theta})$$

- $\hat{\theta}$ is inadmissible if there is at least one other estimator $\hat{\theta}'$ that dominates it. Otherwise it is called admissible.

17.3 Bayes Rule

Bayes rule (or Bayes estimator)

- $r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta})$
- $\hat{\theta}(x) = \arg \min_{\theta} r(\theta | x) \forall x \implies r(f, \hat{\theta}) = \int r(\hat{\theta} | x) f(x) dx$

Theorems

- Squared error loss: posterior mean
- Absolute error loss: posterior median
- Zero-one loss: posterior mode

17.4 Minimax Rules

Maximum risk

$$\bar{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta}) \quad \bar{R}(a) = \sup_{\theta} R(\theta, a)$$

Minimax rule

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \bar{R}(\tilde{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

$$\hat{\theta} = \text{Bayes rule} \wedge \exists c : R(\theta, \hat{\theta}) = c$$

Least favorable prior

$$\hat{\theta}^f = \text{Bayes rule} \wedge R(\theta, \hat{\theta}^f) \leq r(f, \hat{\theta}^f) \forall \theta$$

18 Linear Regression

Definitions

- Response variable Y
- Covariate X (aka predictor variable or feature)

18.1 Simple Linear Regression

Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \mathbb{E}[\epsilon_i | X_i] = 0, \quad \mathbb{V}[\epsilon_i | X_i] = \sigma^2$$

Fitted line

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Predicted (fitted) values

$$\hat{Y}_i = \hat{r}(X_i)$$

Residuals

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

Residual sums of squares (RSS)

$$\text{RSS}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n \hat{\epsilon}_i^2$$

Least square estimates

$$\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1)^T : \min_{\hat{\beta}_0, \hat{\beta}_1} \text{RSS}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

$$\mathbb{E}[\hat{\beta} | X^n] = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\mathbb{V}[\hat{\beta} | X^n] = \frac{\sigma^2}{n s_X^2} \begin{pmatrix} n^{-1} \sum_{i=1}^n X_i^2 & -\bar{X}_n \\ -\bar{X}_n & 1 \end{pmatrix}$$

$$\hat{\text{se}}(\hat{\beta}_0) = \frac{\hat{\sigma}}{s_X \sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}$$

$$\hat{\text{se}}(\hat{\beta}_1) = \frac{\hat{\sigma}}{s_X \sqrt{n}}$$

where $s_X^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2$ (unbiased estimate).
Further properties:

- Consistency: $\hat{\beta}_0 \xrightarrow{P} \beta_0$ and $\hat{\beta}_1 \xrightarrow{P} \beta_1$

- Asymptotic normality:

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\text{se}}(\widehat{\beta}_0)} \xrightarrow{D} \mathcal{N}(0, 1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\text{se}}(\widehat{\beta}_1)} \xrightarrow{D} \mathcal{N}(0, 1)$$

- Approximate $1 - \alpha$ confidence intervals for β_0 and β_1 :

$$\widehat{\beta}_0 \pm z_{\alpha/2} \widehat{\text{se}}(\widehat{\beta}_0) \quad \text{and} \quad \widehat{\beta}_1 \pm z_{\alpha/2} \widehat{\text{se}}(\widehat{\beta}_1)$$

- Wald test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$: reject H_0 if $|W| > z_{\alpha/2}$ where $W = \widehat{\beta}_1 / \widehat{\text{se}}(\widehat{\beta}_1)$.

R^2

$$R^2 = \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n \widehat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Likelihood

$$\mathcal{L} = \prod_{i=1}^n f(X_i, Y_i) = \prod_{i=1}^n f_X(X_i) \times \prod_{i=1}^n f_{Y|X}(Y_i | X_i) = \mathcal{L}_1 \times \mathcal{L}_2$$

$$\mathcal{L}_1 = \prod_{i=1}^n f_X(X_i)$$

$$\mathcal{L}_2 = \prod_{i=1}^n f_{Y|X}(Y_i | X_i) \propto \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2 \right\}$$

Under the assumption of Normality, the least squares estimator is also the MLE but the least squares variance estimator is not the MLE.

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \widehat{\epsilon}_i^2$$

18.2 Prediction

Observe $X = x_*$ of the covariate and want to predict their outcome Y_* .

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*$$

$$\mathbb{V}[\widehat{Y}_*] = \mathbb{V}[\widehat{\beta}_0] + x_*^2 \mathbb{V}[\widehat{\beta}_1] + 2x_* \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_1]$$

Prediction interval

$$\widehat{\xi}_n^2 = \widehat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_i (X_i - \bar{X})^2} + 1 \right)$$

$$\widehat{Y}_* \pm z_{\alpha/2} \widehat{\xi}_n$$

18.3 Multiple Regression

$$Y = X\beta + \epsilon$$

where

$$X = \begin{pmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Likelihood

$$\mathcal{L}(\mu, \Sigma) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \text{RSS} \right\}$$

$$\text{RSS} = (y - X\beta)^T (y - X\beta) = \|Y - X\beta\|^2 = \sum_{i=1}^n (Y_i - x_i^T \beta)^2$$

If the $(k \times k)$ matrix $X^T X$ is invertible,

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

$$\mathbb{V}[\widehat{\beta} | X^n] = \sigma^2 (X^T X)^{-1}$$

$$\widehat{\beta} \approx \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

Estimate regression function

$$\widehat{r}(x) = \sum_{j=1}^k \widehat{\beta}_j x_j$$

Unbiased estimate for σ^2

$$\widehat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n \widehat{\epsilon}_i^2 \quad \widehat{\epsilon} = X\widehat{\beta} - Y$$

MLE

$$\widehat{\mu} = \bar{X} \quad \widehat{\sigma}^2 = \frac{n-k}{n} \sigma^2$$

$1 - \alpha$ Confidence interval

$$\widehat{\beta}_j \pm z_{\alpha/2} \widehat{\text{se}}(\widehat{\beta}_j)$$

18.4 Model Selection

Consider predicting a new observation Y^* for covariates X^* and let $S \subset J$ denote a subset of the covariates in the model, where $|S| = k$ and $|J| = n$.

Issues

- Underfitting: too few covariates yields high bias
- Overfitting: too many covariates yields high variance

Procedure

1. Assign a score to each model
2. Search through all models to find the one with the highest score

Hypothesis testing

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0 \quad \forall j \in J$$

Mean squared prediction error (MSPE)

$$\text{MSPE} = \mathbb{E} \left[(\hat{Y}(S) - Y^*)^2 \right]$$

Prediction risk

$$R(S) = \sum_{i=1}^n \text{MSPE}_i = \sum_{i=1}^n \mathbb{E} \left[(\hat{Y}_i(S) - Y_i^*)^2 \right]$$

Training error

$$\hat{R}_{tr}(S) = \sum_{i=1}^n (\hat{Y}_i(S) - Y_i)^2$$

R^2

$$R^2(S) = 1 - \frac{\text{RSS}(S)}{\text{TSS}} = 1 - \frac{\hat{R}_{tr}(S)}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i(S) - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

The training error is a downward-biased estimate of the prediction risk.

$$\mathbb{E} \left[\hat{R}_{tr}(S) \right] < R(S)$$

$$\text{bias}(\hat{R}_{tr}(S)) = \mathbb{E} \left[\hat{R}_{tr}(S) \right] - R(S) = -2 \sum_{i=1}^n \text{Cov} \left[\hat{Y}_i, Y_i \right]$$

Adjusted R^2

$$R^2(S) = 1 - \frac{n-1}{n-k} \frac{\text{RSS}}{\text{TSS}}$$

MALLOW'S C_p statistic

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2k\hat{\sigma}^2 = \text{lack of fit} + \text{complexity penalty}$$

AKAIKE Information Criterion (AIC)

$$\text{AIC}(S) = \ell_n(\hat{\beta}_S, \hat{\sigma}_S^2) - k$$

Bayesian Information Criterion (BIC)

$$\text{BIC}(S) = \ell_n(\hat{\beta}_S, \hat{\sigma}_S^2) - \frac{k}{2} \log n$$

Validation and training

$$\hat{R}_V(S) = \sum_{i=1}^m (\hat{Y}_i^*(S) - Y_i^*)^2 \quad m = |\{\text{validation data}\}|, \text{ often } \frac{n}{4} \text{ or } \frac{n}{2}$$

Leave-one-out cross-validation

$$\hat{R}_{CV}(S) = \sum_{i=1}^n (Y_i - \hat{Y}_{(i)})^2 = \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i(S)}{1 - U_{ii}(S)} \right)^2$$

$$U(S) = X_S (X_S^T X_S)^{-1} X_S \text{ ("hat matrix")}$$

19 Non-parametric Function Estimation

19.1 Density Estimation

Estimate $f(x)$, where $f(x) = \mathbb{P}[X \in A] = \int_A f(x) dx$.

Integrated square error (ISE)

$$L(f, \hat{f}_n) = \int \left(f(x) - \hat{f}_n(x) \right)^2 dx = J(h) + \int f^2(x) dx$$

Frequentist risk

$$R(f, \hat{f}_n) = \mathbb{E} \left[L(f, \hat{f}_n) \right] = \int b^2(x) dx + \int v(x) dx$$

$$b(x) = \mathbb{E} \left[\hat{f}_n(x) \right] - f(x)$$

$$v(x) = \mathbb{V} \left[\hat{f}_n(x) \right]$$

19.1.1 Histograms

Definitions

- Number of bins m
- Binwidth $h = \frac{1}{m}$
- Bin B_j has ν_j observations
- Define $\hat{p}_j = \nu_j/n$ and $p_j = \int_{B_j} f(u) du$

Histogram estimator

$$\begin{aligned}\hat{f}_n(x) &= \sum_{j=1}^m \frac{\hat{p}_j}{h} I(x \in B_j) \\ \mathbb{E}[\hat{f}_n(x)] &= \frac{p_j}{h} \\ \mathbb{V}[\hat{f}_n(x)] &= \frac{p_j(1-p_j)}{nh^2} \\ R(\hat{f}_n, f) &\approx \frac{h^2}{12} \int (f'(u))^2 du + \frac{1}{nh} \\ h^* &= \frac{1}{n^{1/3}} \left(\frac{6}{\int (f'(u))^2 du} \right)^{1/3} \\ R^*(\hat{f}_n, f) &\approx \frac{C}{n^{2/3}} \quad C = \left(\frac{3}{4} \right)^{2/3} \left(\int (f'(u))^2 du \right)^{1/3}\end{aligned}$$

Cross-validation estimate of $\mathbb{E}[J(h)]$

$$\hat{J}_{CV}(h) = \int \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{(-i)}(X_i) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h} \sum_{j=1}^m \hat{p}_j^2$$

19.1.2 Kernel Density Estimator (KDE)

Kernel K

- $K(x) \geq 0$
- $\int K(x) dx = 1$
- $\int xK(x) dx = 0$
- $\int x^2K(x) dx \equiv \sigma_K^2 > 0$

KDE

$$\begin{aligned}\hat{f}_n(x) &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x-X_i}{h}\right) \\ R(f, \hat{f}_n) &\approx \frac{1}{4}(h\sigma_K)^4 \int (f''(x))^2 dx + \frac{1}{nh} \int K^2(x) dx \\ h^* &= \frac{c_1^{-2/5} c_2^{-1/5} c_3^{-1/5}}{n^{1/5}} \quad c_1 = \sigma_K^2, \quad c_2 = \int K^2(x) dx, \quad c_3 = \int (f''(x))^2 dx \\ R^*(f, \hat{f}_n) &= \frac{c_4}{n^{4/5}} \quad c_4 = \underbrace{\frac{5}{4}(\sigma_K^2)^{2/5} \left(\int K^2(x) dx \right)^{4/5}}_{C(K)} \left(\int (f'')^2 dx \right)^{1/5}\end{aligned}$$

EPANECHNIKOV Kernel

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}(1-x^2/5)} & |x| < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

Cross-validation estimate of $\mathbb{E}[J(h)]$

$$\hat{J}_{CV}(h) = \int \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{(-i)}(X_i) \approx \frac{1}{hn^2} \sum_{i=1}^n \sum_{j=1}^n K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh} K(0)$$

$$K^*(x) = K^{(2)}(x) - 2K(x) \quad K^{(2)}(x) = \int K(x-y)K(y) dy$$

19.2 Non-parametric Regression

Estimate $f(x)$ where $f(x) = \mathbb{E}[Y | X = x]$. Consider pairs of points $(x_1, Y_1), \dots, (x_n, Y_n)$ related by

$$\begin{aligned}Y_i &= r(x_i) + \epsilon_i \\ \mathbb{E}[\epsilon_i] &= 0 \\ \mathbb{V}[\epsilon_i] &= \sigma^2\end{aligned}$$

k -nearest Neighbor Estimator

$$\hat{r}(x) = \frac{1}{k} \sum_{i: x_i \in N_k(x)} Y_i \quad \text{where } N_k(x) = \{k \text{ values of } x_1, \dots, x_n \text{ closest to } x\}$$

NADARAYA-WATSON Kernel Estimator

$$\begin{aligned}\widehat{r}(x) &= \sum_{i=1}^n w_i(x) Y_i \\ w_i(x) &= \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)} \in [0, 1] \\ R(\widehat{r}_n, r) &\approx \frac{h^4}{4} \left(\int x^2 K^2(x) dx \right)^4 \int \left(r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right)^2 dx \\ &\quad + \int \frac{\sigma^2 \int K^2(x) dx}{nhf(x)} dx \\ h^* &\approx \frac{c_1}{n^{1/5}} \\ R^*(\widehat{r}_n, r) &\approx \frac{c_2}{n^{4/5}}\end{aligned}$$

Cross-validation estimate of $\mathbb{E}[J(h)]$

$$\widehat{J}_{CV}(h) = \sum_{i=1}^n (Y_i - \widehat{r}_{(-i)}(x_i))^2 = \sum_{i=1}^n \frac{(Y_i - \widehat{r}(x_i))^2}{\left(1 - \frac{K(0)}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}\right)^2}$$

19.3 Smoothing Using Orthogonal Functions

Approximation

$$r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x) \approx \sum_{j=1}^J \beta_j \phi_j(x)$$

Multivariate regression

$$Y = \Phi\beta + \eta$$

$$\text{where } \eta_i = \epsilon_i \quad \text{and} \quad \Phi = \begin{pmatrix} \phi_0(x_1) & \cdots & \phi_J(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_n) & \cdots & \phi_J(x_n) \end{pmatrix}$$

Least squares estimator

$$\begin{aligned}\widehat{\beta} &= (\Phi^T \Phi)^{-1} \Phi^T Y \\ &\approx \frac{1}{n} \Phi^T Y \quad (\text{for equally spaced observations only})\end{aligned}$$

Cross-validation estimate of $\mathbb{E}[J(h)]$

$$\widehat{R}_{CV}(J) = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^J \phi_j(x_i) \widehat{\beta}_{j,(-i)} \right)^2$$

20 Stochastic Processes

Stochastic Process

$$\{X_t : t \in T\} \quad T = \begin{cases} \{0, \pm 1, \dots\} = \mathbb{Z} & \text{discrete} \\ [0, \infty) & \text{continuous} \end{cases}$$

- Notations $X_t, X(t)$
- State space \mathcal{X}
- Index set T

20.1 Markov Chains

Markov chain

$$\mathbb{P}[X_n = x | X_0, \dots, X_{n-1}] = \mathbb{P}[X_n = x | X_{n-1}] \quad \forall n \in T, x \in \mathcal{X}$$

Transition probabilities

$$\begin{aligned}p_{ij} &\equiv \mathbb{P}[X_{n+1} = j | X_n = i] \\ p_{ij}(n) &\equiv \mathbb{P}[X_{m+n} = j | X_m = i] \quad \text{n-step}\end{aligned}$$

Transition matrix \mathbf{P} (n-step: \mathbf{P}_n)

- (i, j) element is p_{ij}
- $p_{ij} > 0$
- $\sum_i p_{ij} = 1$

CHAPMAN-KOLMOGOROV

$$p_{ij}(m+n) = \sum_k p_{ik}(m) p_{kj}(n)$$

$$\mathbf{P}_{m+n} = \mathbf{P}_m \mathbf{P}_n$$

$$\mathbf{P}_n = \mathbf{P} \times \dots \times \mathbf{P} = \mathbf{P}^n$$

Marginal probability

$$\begin{aligned}\mu_n &= (\mu_n(1), \dots, \mu_n(N)) \quad \text{where } \mu_i(i) = \mathbb{P}[X_n = i] \\ \mu_0 &\triangleq \text{initial distribution} \\ \mu_n &= \mu_0 \mathbf{P}^n\end{aligned}$$

20.2 Poisson Processes

Poisson process

- $\{X_t : t \in [0, \infty)\}$ = number of events up to and including time t
- $X_0 = 0$
- Independent increments:

$$\forall t_0 < \dots < t_n : X_{t_1} - X_{t_0} \perp\!\!\!\perp \dots \perp\!\!\!\perp X_{t_n} - X_{t_{n-1}}$$

- Intensity function $\lambda(t)$
 - $\mathbb{P}[X_{t+h} - X_t = 1] = \lambda(t)h + o(h)$
 - $\mathbb{P}[X_{t+h} - X_t = 2] = o(h)$
- $X_{s+t} - X_s \sim \text{Po}(m(s+t) - m(s))$ where $m(t) = \int_0^t \lambda(s) ds$

Homogeneous Poisson process

$$\lambda(t) \equiv \lambda \implies X_t \sim \text{Po}(\lambda t) \quad \lambda > 0$$

Waiting times

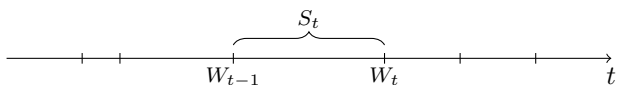
$W_t :=$ time at which X_t occurs

$$W_t \sim \text{Gamma}\left(t, \frac{1}{\lambda}\right)$$

Interarrival times

$$S_t = W_{t+1} - W_t$$

$$S_t \sim \text{Exp}\left(\frac{1}{\lambda}\right)$$



21 Time Series

Mean function

$$\mu_{x_t} = \mathbb{E}[x_t] = \int_{-\infty}^{\infty} x f_t(x) dx$$

Autocovariance function

$$\gamma_x(s, t) = \mathbb{E}[(x_s - \mu_s)(x_t - \mu_t)] = \mathbb{E}[x_s x_t] - \mu_s \mu_t$$

$$\gamma_x(t, t) = \mathbb{E}[(x_t - \mu_t)^2] = \mathbb{V}[x_t]$$

Autocorrelation function (ACF)

$$\rho(s, t) = \frac{\text{Cov}[x_s, x_t]}{\sqrt{\mathbb{V}[x_s] \mathbb{V}[x_t]}} = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s) \gamma(t, t)}}$$

Cross-covariance function (CCV)

$$\gamma_{xy}(s, t) = \mathbb{E}[(x_s - \mu_{x_s})(y_t - \mu_{y_t})]$$

Cross-correlation function (CCF)

$$\rho_{xy}(s, t) = \frac{\gamma_{xy}(s, t)}{\sqrt{\gamma_x(s, s) \gamma_y(t, t)}}$$

Backshift operator

$$B^k(x_t) = x_{t-k}$$

Difference operator

$$\nabla^d = (1 - B)^d$$

White noise

- $w_t \sim wn(0, \sigma_w^2)$
- Gaussian: $w_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$
- $\mathbb{E}[w_t] = 0 \quad t \in T$
- $\mathbb{V}[w_t] = \sigma^2 \quad t \in T$
- $\gamma_w(s, t) = 0 \quad s \neq t \wedge s, t \in T$

Random walk

- Drift δ
- $x_t = \delta t + \sum_{j=1}^t w_j$
- $\mathbb{E}[x_t] = \delta t$

Symmetric moving average

$$m_t = \sum_{j=-k}^k a_j x_{t-j} \quad \text{where } a_j = a_{-j} \geq 0 \text{ and } \sum_{j=-k}^k a_j = 1$$

21.1 Stationary Time Series

Strictly stationary

$$\mathbb{P}[x_{t_1} \leq c_1, \dots, x_{t_k} \leq c_k] = \mathbb{P}[x_{t_1+h} \leq c_1, \dots, x_{t_k+h} \leq c_k]$$

$$\forall k \in \mathbb{N}, t_k, c_k, h \in \mathbb{Z}$$

Weakly stationary

- $\mathbb{E}[x_t^2] < \infty \quad \forall t \in \mathbb{Z}$
- $\mathbb{E}[x_t^2] = m \quad \forall t \in \mathbb{Z}$
- $\gamma_x(s, t) = \gamma_x(s+r, t+r) \quad \forall r, s, t \in \mathbb{Z}$

Autocovariance function

- $\gamma(h) = \mathbb{E}[(x_{t+h} - \mu)(x_t - \mu)] \quad \forall h \in \mathbb{Z}$
- $\gamma(0) = \mathbb{E}[(x_t - \mu)^2]$
- $\gamma(0) \geq 0$
- $\gamma(0) \geq |\gamma(h)|$
- $\gamma(h) = \gamma(-h)$

Autocorrelation function (ACF)

$$\rho_x(h) = \frac{\text{Cov}[x_{t+h}, x_t]}{\sqrt{\mathbb{V}[x_{t+h}] \mathbb{V}[x_t]}} = \frac{\gamma(t+h, t)}{\sqrt{\gamma(t+h, t+h)\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}$$

Jointly stationary time series

$$\gamma_{xy}(h) = \mathbb{E}[(x_{t+h} - \mu_x)(y_t - \mu_y)]$$

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}$$

Linear process

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j} \quad \text{where} \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

$$\gamma(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h} \psi_j$$

21.2 Estimation of Correlation

Sample mean

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x_t$$

Sample variance

$$\mathbb{V}[\bar{x}] = \frac{1}{n} \sum_{h=-n}^n \left(1 - \frac{|h|}{n}\right) \gamma_x(h)$$

Sample autocovariance function

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})$$

Sample autocorrelation function

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

Sample cross-variance function

$$\hat{\gamma}_{xy}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y})$$

Sample cross-correlation function

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}$$

Properties

- $\sigma_{\hat{\rho}_x(h)} = \frac{1}{\sqrt{n}}$ if x_t is white noise
- $\sigma_{\hat{\rho}_{xy}(h)} = \frac{1}{\sqrt{n}}$ if x_t or y_t is white noise

21.3 Non-Stationary Time Series

Classical decomposition model

$$x_t = \mu_t + s_t + w_t$$

- μ_t = trend
- s_t = seasonal component
- w_t = random noise term

21.3.1 Detrending

Least squares

1. Choose trend model, e.g., $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$
2. Minimize RSS to obtain trend estimate $\hat{\mu}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2$
3. Residuals \triangleq noise w_t

Moving average

- The *low-pass* filter v_t is a symmetric moving average m_t with $a_j = \frac{1}{2k+1}$:

$$v_t = \frac{1}{2k+1} \sum_{i=-k}^k x_{t-i}$$

- If $\frac{1}{2k+1} \sum_{i=-k}^k w_{t-j} \approx 0$, a linear trend function $\mu_t = \beta_0 + \beta_1 t$ passes without distortion

Differencing

- $\mu_t = \beta_0 + \beta_1 t \implies \nabla x_t = \beta_1$

21.4 ARIMA models

Autoregressive polynomial

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \quad z \in \mathbb{C} \wedge \phi_p \neq 0$$

Autoregressive operator

$$\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$$

Autoregressive model order p , AR(p)

$$x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t \iff \phi(B)x_t = w_t$$

AR(1)

$$\bullet x_t = \phi^k(x_{t-k}) + \sum_{j=0}^{k-1} \phi^j(w_{t-j}) \stackrel{k \rightarrow \infty, |\phi| < 1}{=} \underbrace{\sum_{j=0}^{\infty} \phi^j(w_{t-j})}_{\text{linear process}}$$

- $\mathbb{E}[x_t] = \sum_{j=0}^{\infty} \phi^j(\mathbb{E}[w_{t-j}]) = 0$
- $\gamma(h) = \text{Cov}[x_{t+h}, x_t] = \frac{\sigma_w^2 \phi^h}{1 - \phi^2}$
- $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h$
- $\rho(h) = \phi \rho(h-1) \quad h = 1, 2, \dots$

Moving average polynomial

$$\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q \quad z \in \mathbb{C} \wedge \theta_q \neq 0$$

Moving average operator

$$\theta(B) = 1 + \theta_1 B + \dots + \theta_p B^p$$

MA(q) (moving average model order q)

$$x_t = w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q} \iff x_t = \theta(B)w_t$$

$$\mathbb{E}[x_t] = \sum_{j=0}^q \theta_j \mathbb{E}[w_{t-j}] = 0$$

$$\gamma(h) = \text{Cov}[x_{t+h}, x_t] = \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & 0 \leq h \leq q \\ 0 & h > q \end{cases}$$

MA(1)

$$x_t = w_t + \theta w_{t-1}$$

$$\gamma(h) = \begin{cases} (1 + \theta^2)\sigma_w^2 & h = 0 \\ \theta\sigma_w^2 & h = 1 \\ 0 & h > 1 \end{cases}$$

$$\rho(h) = \begin{cases} \frac{\theta}{(1+\theta^2)} & h = 1 \\ 0 & h > 1 \end{cases}$$

ARMA(p, q)

$$x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}$$

$$\phi(B)x_t = \theta(B)w_t$$

Partial autocorrelation function (PACF)

- $x_i^{h-1} \triangleq$ regression of x_i on $\{x_{h-1}, x_{h-2}, \dots, x_1\}$
- $\phi_{hh} = \text{corr}(x_h - x_h^{h-1}, x_0 - x_0^{h-1}) \quad h \geq 2$
- E.g., $\phi_{11} = \text{corr}(x_1, x_0) = \rho(1)$

ARIMA(p, d, q)

$$\nabla^d x_t = (1 - B)^d x_t \text{ is ARMA}(p, q)$$

$$\phi(B)(1 - B)^d x_t = \theta(B)w_t$$

Exponentially Weighted Moving Average (EWMA)

$$x_t = x_{t-1} + w_t - \lambda w_{t-1}$$

$$x_t = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1}x_{t-j} + w_t \quad \text{when } |\lambda| < 1$$

$$\tilde{x}_{n+1} = (1-\lambda)x_n + \lambda\tilde{x}_n$$

Seasonal ARIMA

- Denoted by ARIMA $(p, d, q) \times (P, D, Q)_s$
- $\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t$

21.4.1 Causality and Invertibility

ARMA (p, q) is causal (future-independent) $\iff \exists\{\psi_j\} : \sum_{j=0}^{\infty} \psi_j < \infty$ such that

$$x_t = \sum_{j=0}^{\infty} w_{t-j} = \psi(B)w_t$$

ARMA (p, q) is invertible $\iff \exists\{\pi_j\} : \sum_{j=0}^{\infty} \pi_j < \infty$ such that

$$\pi(B)x_t = \sum_{j=0}^{\infty} X_{t-j} = w_t$$

Properties

- ARMA (p, q) causal \iff roots of $\phi(z)$ lie outside the unit circle

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)} \quad |z| \leq 1$$

- ARMA (p, q) invertible \iff roots of $\theta(z)$ lie outside the unit circle

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)} \quad |z| \leq 1$$

Behavior of the ACF and PACF for causal and invertible ARMA models

	AR (p)	MA (q)	ARMA (p, q)
ACF	tails off	cuts off after lag q	tails off
PACF	cuts off after lag p	tails off q	tails off

21.5 Spectral Analysis

Periodic process

$$\begin{aligned} x_t &= A \cos(2\pi\omega t + \phi) \\ &= U_1 \cos(2\pi\omega t) + U_2 \sin(2\pi\omega t) \end{aligned}$$

- Frequency index ω (cycles per unit time), period $1/\omega$
- Amplitude A
- Phase ϕ
- $U_1 = A \cos \phi$ and $U_2 = A \sin \phi$ often normally distributed RV's

Periodic mixture

$$x_t = \sum_{k=1}^q (U_{k1} \cos(2\pi\omega_k t) + U_{k2} \sin(2\pi\omega_k t))$$

- U_{k1}, U_{k2} , for $k = 1, \dots, q$, are independent zero-mean RV's with variances σ_k^2
- $\gamma(h) = \sum_{k=1}^q \sigma_k^2 \cos(2\pi\omega_k h)$
- $\gamma(0) = \mathbb{E}[x_t^2] = \sum_{k=1}^q \sigma_k^2$

Spectral representation of a periodic process

$$\begin{aligned} \gamma(h) &= \sigma^2 \cos(2\pi\omega_0 h) \\ &= \frac{\sigma^2}{2} e^{-2\pi i\omega_0 h} + \frac{\sigma^2}{2} e^{2\pi i\omega_0 h} \\ &= \int_{-1/2}^{1/2} e^{2\pi i\omega h} dF(\omega) \end{aligned}$$

Spectral distribution function

$$F(\omega) = \begin{cases} 0 & \omega < -\omega_0 \\ \sigma^2/2 & -\omega_0 \leq \omega < \omega_0 \\ \sigma^2 & \omega \geq \omega_0 \end{cases}$$

- $F(-\infty) = F(-1/2) = 0$
- $F(\infty) = F(1/2) = \gamma(0)$

Spectral density

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i\omega h} \quad -\frac{1}{2} \leq \omega \leq \frac{1}{2}$$

- Needs $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty \implies \gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h} f(\omega) d\omega \quad h = 0, \pm 1, \dots$
- $f(\omega) \geq 0$
- $f(\omega) = f(-\omega)$
- $f(\omega) = f(1-\omega)$
- $\gamma(0) = \mathbb{V}[x_t] = \int_{-1/2}^{1/2} f(\omega) d\omega$
- White noise: $f_w(\omega) = \sigma_w^2$

- ARMA (p, q) , $\phi(B)x_t = \theta(B)w_t$:

$$f_x(\omega) = \sigma_w^2 \frac{|\theta(e^{-2\pi i\omega})|^2}{|\phi(e^{-2\pi i\omega})|^2}$$

where $\phi(z) = 1 - \sum_{k=1}^p \phi_k z^k$ and $\theta(z) = 1 + \sum_{k=1}^q \theta_k z^k$

Discrete Fourier Transform (DFT)

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^n x_t e^{-2\pi i\omega_j t}$$

Fourier/Fundamental frequencies

$$\omega_j = j/n$$

Inverse DFT

$$x_t = n^{-1/2} \sum_{j=0}^{n-1} d(\omega_j) e^{2\pi i\omega_j t}$$

Periodogram

$$I(j/n) = |d(j/n)|^2$$

Scaled Periodogram

$$\begin{aligned} P(j/n) &= \frac{4}{n} I(j/n) \\ &= \left(\frac{2}{n} \sum_{t=1}^n x_t \cos(2\pi t j/n) \right)^2 + \left(\frac{2}{n} \sum_{t=1}^n x_t \sin(2\pi t j/n) \right)^2 \end{aligned}$$

22 Math

22.1 Gamma Function

- Ordinary: $\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt$
- Upper incomplete: $\Gamma(s, x) = \int_x^\infty t^{s-1} e^{-t} dt$
- Lower incomplete: $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$
- $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$ $\alpha > 1$
- $\Gamma(n) = (n-1)!$ $n \in \mathbb{N}$
- $\Gamma(0) = \Gamma(-1) = \infty$
- $\Gamma(1/2) = \sqrt{\pi}$
- $\Gamma(-1/2) = -2\Gamma(1/2)$

22.2 Beta Function

- Ordinary: $B(x, y) = B(y, x) = \int_0^1 t^{x-1} (1-t)^{y-1} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$
- Incomplete: $B(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt$
- Regularized incomplete:

$$I_x(a, b) = \frac{B(x; a, b)}{B(a, b)} \stackrel{a, b \in \mathbb{N}}{=} \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!} x^j (1-x)^{a+b-1-j}$$
- $I_0(a, b) = 0$ $I_1(a, b) = 1$
- $I_x(a, b) = 1 - I_{1-x}(b, a)$

22.3 Series

Finite

- $\sum_{k=1}^n k = \frac{n(n+1)}{2}$
- $\sum_{k=1}^n (2k-1) = n^2$
- $\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$
- $\sum_{k=1}^n k^3 = \left(\frac{n(n+1)}{2} \right)^2$
- $\sum_{k=0}^n c^k = \frac{c^{n+1} - 1}{c - 1}$ $c \neq 1$

Binomial

- $\sum_{k=0}^n \binom{n}{k} = 2^n$
- $\sum_{k=0}^n \binom{r+k}{k} = \binom{r+n+1}{n}$
- $\sum_{k=0}^n \binom{k}{m} = \binom{n+1}{m+1}$
- VANDERMONDE's Identity:

$$\sum_{k=0}^r \binom{m}{k} \binom{n}{r-k} = \binom{m+n}{r}$$
- Binomial Theorem:

$$\sum_{k=0}^n \binom{n}{k} a^{n-k} b^k = (a+b)^n$$

Infinite

- $\sum_{k=0}^{\infty} p^k = \frac{1}{1-p}$, $\sum_{k=1}^{\infty} p^k = \frac{p}{1-p}$ $|p| < 1$
- $\sum_{k=0}^{\infty} k p^{k-1} = \frac{d}{dp} \left(\sum_{k=0}^{\infty} p^k \right) = \frac{d}{dp} \left(\frac{1}{1-p} \right) = \frac{1}{(1-p)^2}$ $|p| < 1$
- $\sum_{k=0}^{\infty} \binom{r+k-1}{k} x^k = (1-x)^{-r}$ $r \in \mathbb{N}^+$
- $\sum_{k=0}^{\infty} \binom{\alpha}{k} p^k = (1+p)^\alpha$ $|p| < 1, \alpha \in \mathbb{C}$

22.4 Combinatorics

Sampling

k out of n	w/o replacement	w/ replacement
ordered	$n^{\underline{k}} = \prod_{i=0}^{k-1} (n-i) = \frac{n!}{(n-k)!}$	n^k
unordered	$\binom{n}{k} = \frac{n^{\underline{k}}}{k!} = \frac{n!}{k!(n-k)!}$	$\binom{n-1+r}{r} = \binom{n-1+r}{n-1}$

Stirling numbers, 2nd kind

$$\left\{ \begin{matrix} n \\ k \end{matrix} \right\} = k \left\{ \begin{matrix} n-1 \\ k \end{matrix} \right\} + \left\{ \begin{matrix} n-1 \\ k-1 \end{matrix} \right\} \quad 1 \leq k \leq n \quad \left\{ \begin{matrix} n \\ 0 \end{matrix} \right\} = \begin{cases} 1 & n=0 \\ 0 & \text{else} \end{cases}$$

Partitions

$$P_{n+k,k} = \sum_{i=1}^n P_{n,i} \quad k > n : P_{n,k} = 0 \quad n \geq 1 : P_{n,0} = 0, P_{0,0} = 1$$

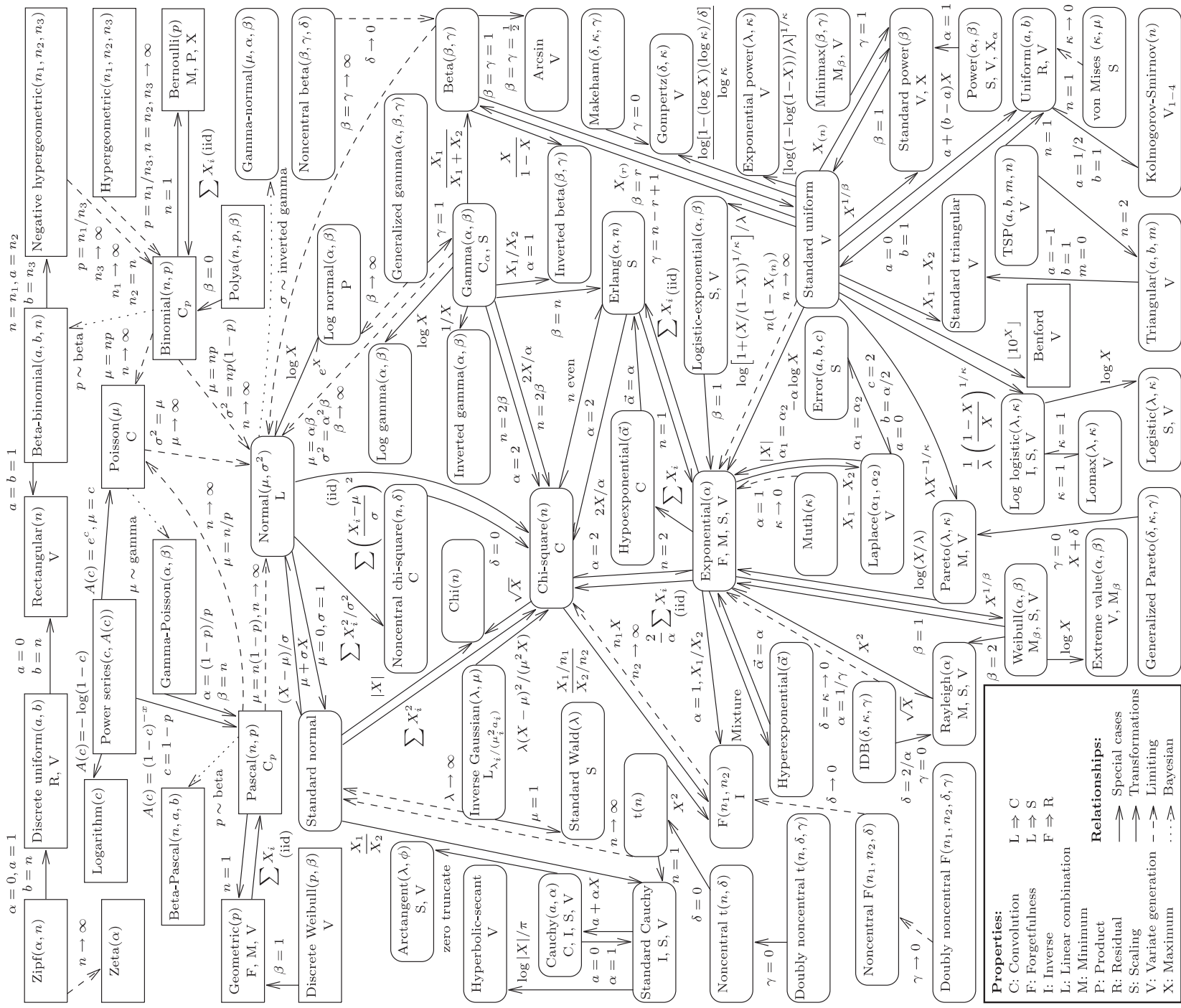
Balls and Urns $f : B \rightarrow U$ $D =$ distinguishable, $\neg D =$ indistinguishable.

$ B = n, U = m$	f arbitrary	f injective	f surjective	f bijective
$B : D, U : D$	m^n	$\begin{cases} m^{\underline{n}} & m \geq n \\ 0 & \text{else} \end{cases}$	$m! \binom{n}{m}$	$\begin{cases} n! & m = n \\ 0 & \text{else} \end{cases}$
$B : \neg D, U : D$	$\binom{m+n-1}{n}$	$\binom{m}{n}$	$\binom{n-1}{m-1}$	$\begin{cases} 1 & m = n \\ 0 & \text{else} \end{cases}$
$B : D, U : \neg D$	$\sum_{k=1}^m \left\{ \begin{matrix} n \\ k \end{matrix} \right\}$	$\begin{cases} 1 & m \geq n \\ 0 & \text{else} \end{cases}$	$\left\{ \begin{matrix} n \\ m \end{matrix} \right\}$	$\begin{cases} 1 & m = n \\ 0 & \text{else} \end{cases}$
$B : \neg D, U : \neg D$	$\sum_{k=1}^m P_{n,k}$	$\begin{cases} 1 & m \geq n \\ 0 & \text{else} \end{cases}$	$P_{n,m}$	$\begin{cases} 1 & m = n \\ 0 & \text{else} \end{cases}$

References

- [1] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Probability Theory*. Brooks Cole, 1972.
- [2] L. M. Leemis and J. T. McQueston. Univariate Distribution Relationships. *The American Statistician*, 62(1):45–53, 2008.

- [3] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications With R Examples*. Springer, 2006.
- [4] A. Steger. *Diskrete Strukturen – Band 1: Kombinatorik, Graphentheorie, Algebra*. Springer, 2001.
- [5] A. Steger. *Diskrete Strukturen – Band 2: Wahrscheinlichkeitstheorie und Statistik*. Springer, 2002.
- [6] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.



Univariate distribution relationships, courtesy Leemis and McQueston [2].