# Probability and Statistics

## Cookbook

Version 0.2.7

27th July, 2021

http://statistics.zone/

# Contents



---

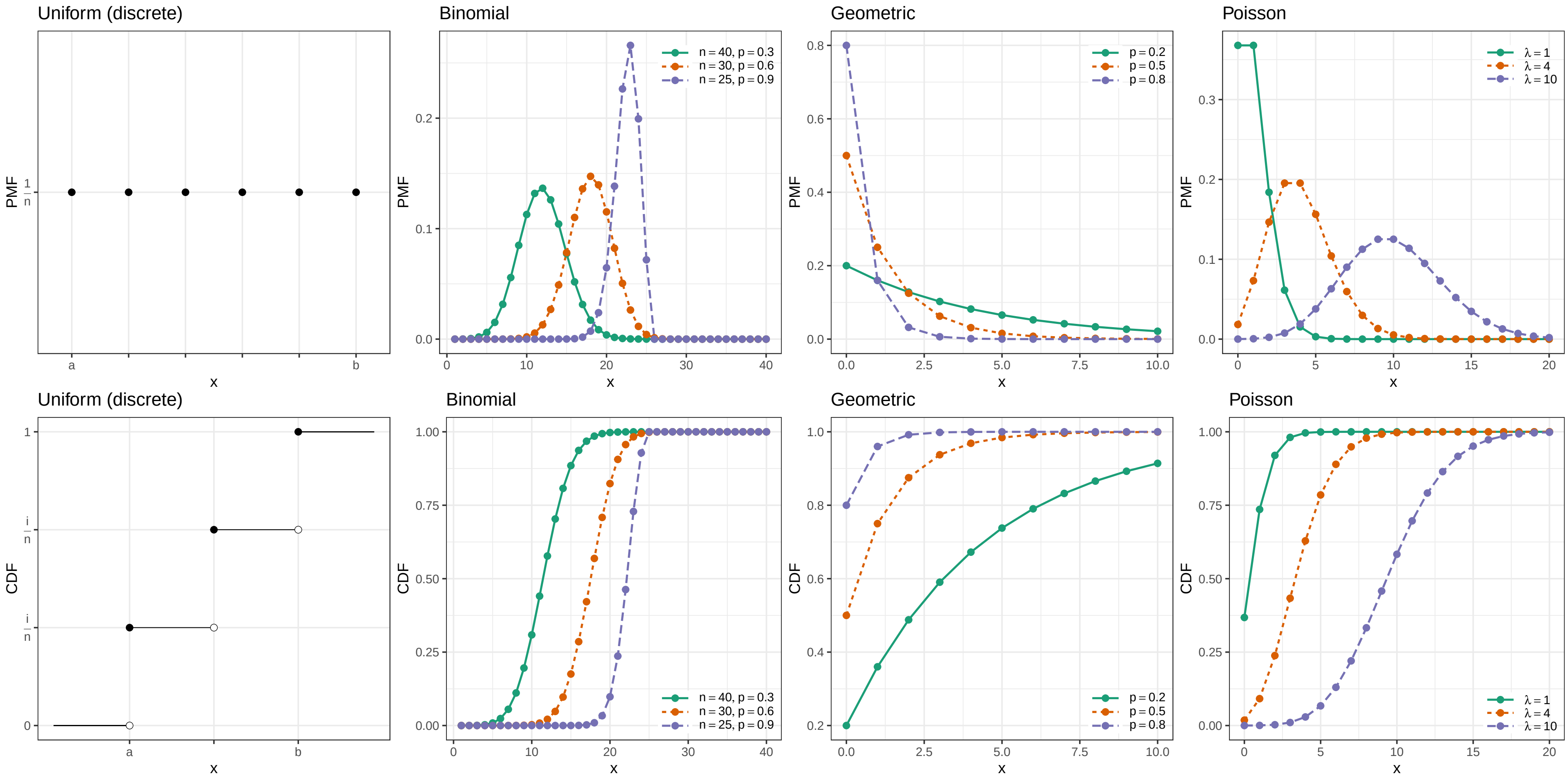
This cookbook integrates various topics in probability theory and statistics, based on literature [1, 6, 3] and in-class material from courses of the statistics department at the University of California in Berkeley but also influenced by others [4, 5]. If you find errors or have suggestions for improvements, 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 | $\text{Unif}\{a,\ldots,b\}$ | $\begin{cases} 0 & x<a \\ \frac{\lfloor x\rfloor-a+1}{b-a} & a\le x\le b \\ 1 & x>b \end{cases}$ | $\dfrac{I(a\le x\le b)}{b-a+1}$ | $\dfrac{a+b}{2}$ | $\dfrac{(b-a+1)^2-1}{12}$ | $\dfrac{e^{as}-e^{-(b+1)s}}{s(b-a)}$ |
| 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 | $\text{Mult}(n,p)$ | | $\dfrac{n!}{x_1!\ldots x_k!}p_1^{x_1}\cdots p_k^{x_k} \quad \sum_{i=1}^k x_i=n$ | $\begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$ | $\begin{pmatrix} np_1(1-p_1) & -np_1p_2 \\ -np_2p_1 & \ddots \end{pmatrix}$ | $\left(\sum_{i=0}^k p_ie^{s_i}\right)^n$ |
| Hypergeometric | $\text{Hyp}(N,m,n)$ | $\approx\Phi\left(\dfrac{x-np}{\sqrt{np(1-p)}}\right)$ | $\dfrac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$ | $\dfrac{nm}{N}$ | $\dfrac{nm(N-n)(N-m)}{N^2(N-1)}$ | |
| Negative Binomial | $\text{NBin}(r,p)$ | $I_p(r,x+1)$ | $\binom{x+r-1}{r-1}p^r(1-p)^x$ | $r\dfrac{1-p}{p}$ | $r\dfrac{1-p}{p^2}$ | $\left(\dfrac{pe^s}{1-(1-p)e^s}\right)^r$ |
| Geometric | $\text{Geo}(p)$ | $1-(1-p)^x \quad x\in\mathbb{N}^+$ | $p(1-p)^{x-1} \quad x\in\mathbb{N}^+$ | $\dfrac{1}{p}$ | $\dfrac{1-p}{p^2}$ | $\dfrac{pe^s}{1-(1-p)e^s}$ |
| Poisson | $\text{Po}(\lambda)$ | $e^{-\lambda}\sum_{i=0}^x\dfrac{\lambda^i}{i!}$ | $\dfrac{\lambda^xe^{-\lambda}}{x!}$ | $\lambda$ | $\lambda$ | $e^{\lambda(e^s-1)}$ |

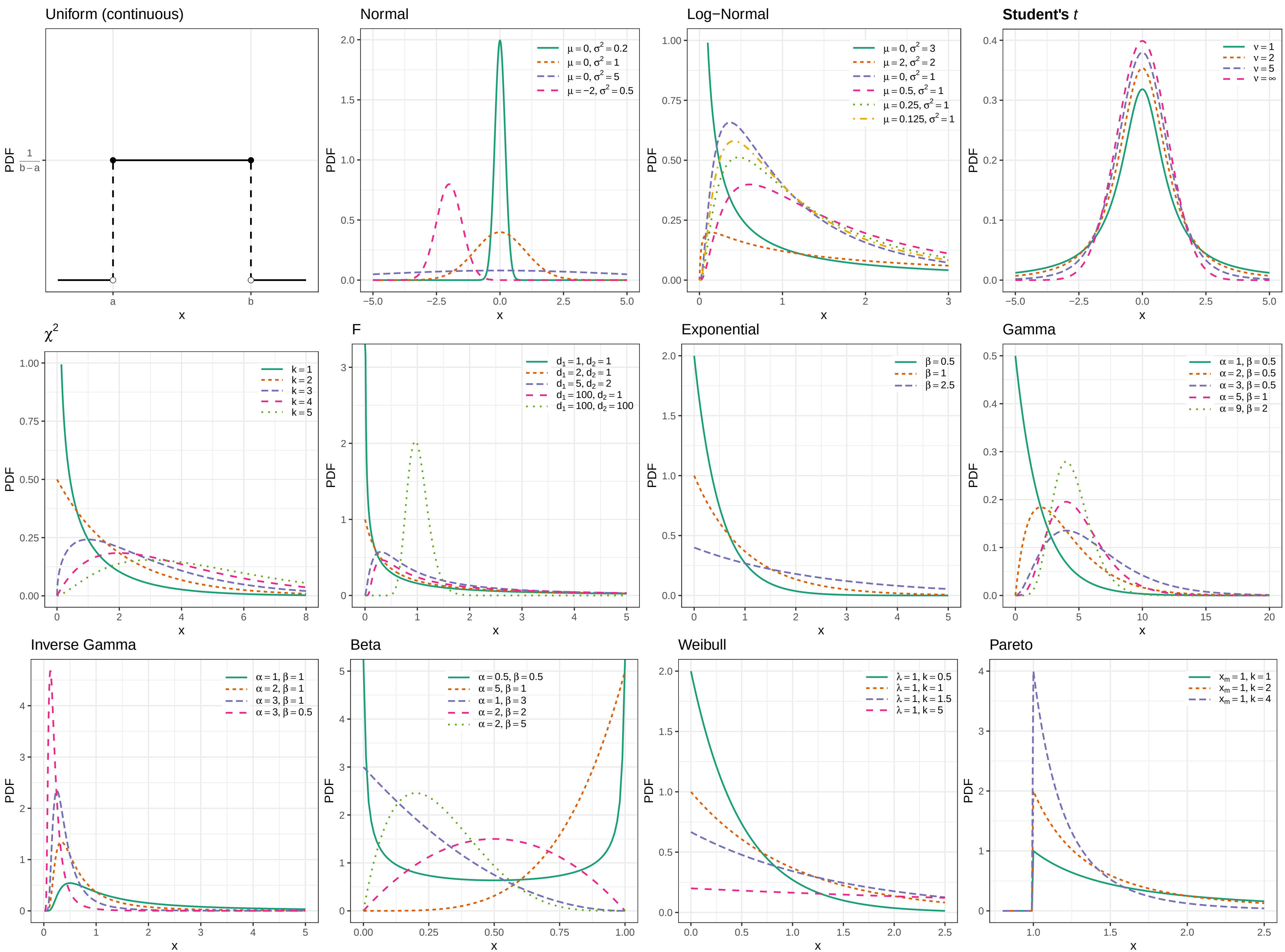
[1] We use the notation $\gamma(s,x)$ and $\Gamma(x)$ to refer to the Gamma functions (see §22.1), and use $\mathrm{B}(x,y)$ and $I_x$ to refer to the Beta functions (see §22.2).


Uniform (discrete)
Binomial
Geometric
Poisson
PMF
CDF
x
n = 40, p = 0.3
n = 30, p = 0.6
n = 25, p = 0.9
p = 0.2
p = 0.5
p = 0.8
λ = 1
λ = 4
λ = 10

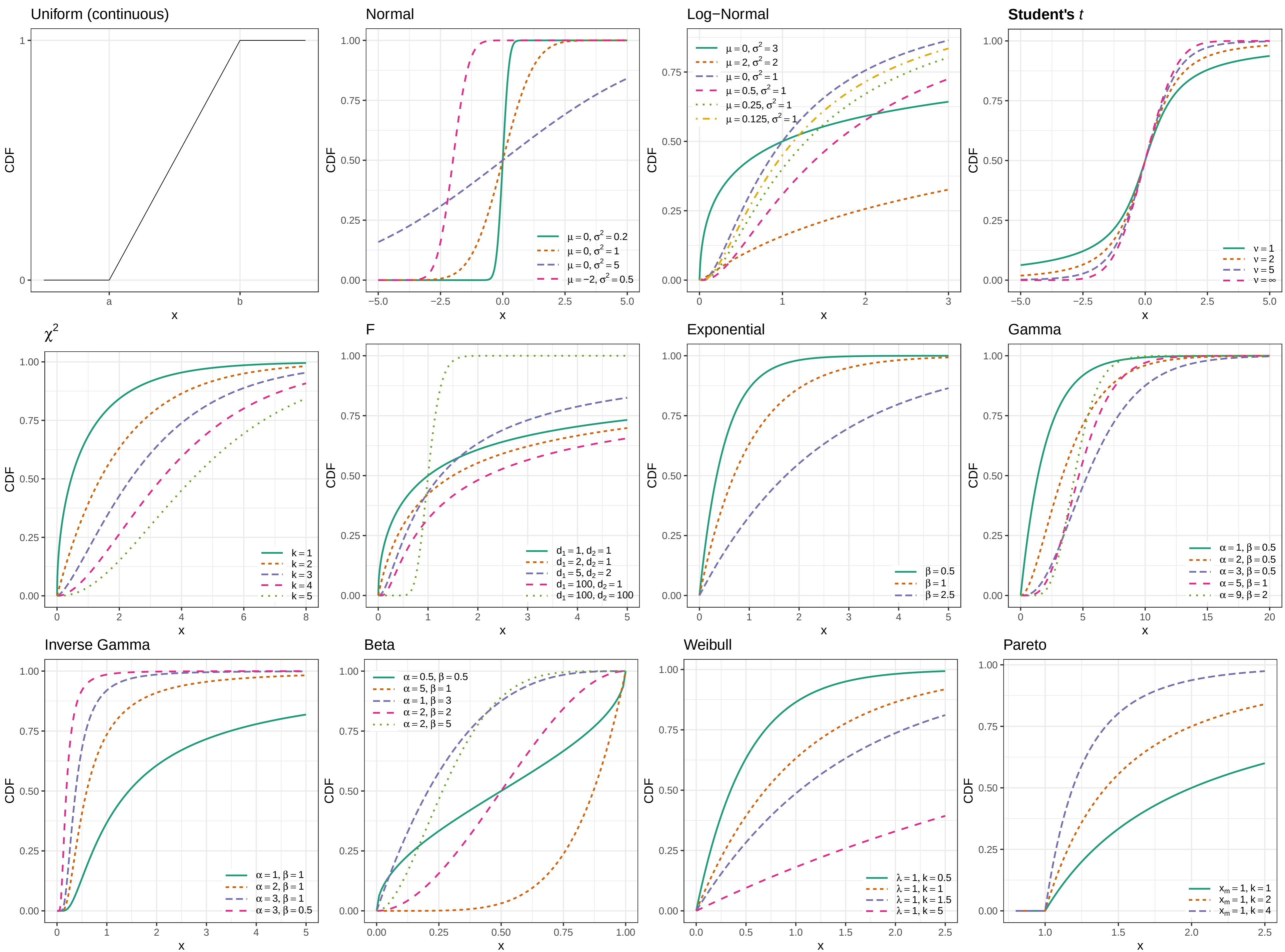
## 1.2 Continuous Distributions

| | Notation | $F_X(x)$ | $f_X(x)$ | $\mathbb{E}[X]$ | $\mathbb{V}[X]$ | $M_X(s)$ |
|---|---|---|---|---|---|---|
| Uniform | $\text{Unif}(a,b)$ | $\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$ | $\dfrac{I(a < x < b)}{b-a}$ | $\dfrac{a+b}{2}$ | $\dfrac{(b-a)^2}{12}$ | $\dfrac{e^{sb}-e^{sa}}{s(b-a)}$ |
| Normal | $\mathcal{N}(\mu,\sigma^2)$ | $\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt$ | $\phi(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$ | $\mu$ | $\sigma^2$ | $\exp\left\{\mu s + \dfrac{\sigma^2 s^2}{2}\right\}$ |
| Log-Normal | $\ln\mathcal{N}(\mu,\sigma^2)$ | $\dfrac{1}{2} + \dfrac{1}{2}\,\text{erf}\left[\dfrac{\ln x - \mu}{\sqrt{2\sigma^2}}\right]$ | $\dfrac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\dfrac{(\ln x-\mu)^2}{2\sigma^2}\right\}$ | $e^{\mu+\sigma^2/2}$ | $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$ | |
| Multivariate Normal | $\text{MVN}(\mu,\Sigma)$ | | $(2\pi)^{-k/2}\lvert\Sigma\rvert^{-1/2}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ | $\mu$ | $\Sigma$ | $\exp\left\{\mu^T s + \dfrac{1}{2}s^T\Sigma s\right\}$ |
| Student's $t$ | $\text{Student}(\nu)$ | $I_x\left(\dfrac{\nu}{2},\dfrac{\nu}{2}\right)$ | $\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-(\nu+1)/2}$ | $0 \quad \nu > 1$ | $\begin{cases} \dfrac{\nu}{\nu-2} & \nu > 2 \\ \infty & 1 < \nu \le 2 \end{cases}$ | |
| Chi-square | $\chi^2_k$ | $\dfrac{1}{\Gamma(k/2)}\gamma\left(\dfrac{k}{2},\dfrac{x}{2}\right)$ | $\dfrac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$ | $k$ | $2k$ | $(1-2s)^{-k/2} \; s < 1/2$ |
| F | $\text{F}(d_1,d_2)$ | $I_{\frac{d_1 x}{d_1 x+d_2}}\left(\dfrac{d_1}{2},\dfrac{d_2}{2}\right)$ | $\dfrac{\sqrt{\frac{(d_1x)^{d_1}d_2^{d_2}}{(d_1x+d_2)^{d_1+d_2}}}}{x\mathrm{B}\left(\frac{d_1}{2},\frac{d_1}{2}\right)}$ | $\dfrac{d_2}{d_2-2}$ | $\dfrac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ | |
| Exponential* | $\text{Exp}(\beta)$ | $1-e^{-x/\beta}$ | $\dfrac{1}{\beta}e^{-x/\beta}$ | $\beta$ | $\beta^2$ | $\dfrac{1}{1-\frac{s}{\beta}}\,(s < \beta)$ |
| Gamma* | $\text{Gamma}(\alpha,\beta)$ | $\dfrac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)}$ | $\dfrac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ | $\dfrac{\alpha}{\beta}$ | $\dfrac{\alpha}{\beta^2}$ | $\left(\dfrac{1}{1-\frac{s}{\beta}}\right)^\alpha (s < \beta)$ |
| Inverse Gamma | $\text{InvGamma}(\alpha,\beta)$ | $\dfrac{\Gamma\left(\alpha,\frac{\beta}{x}\right)}{\Gamma(\alpha)}$ | $\dfrac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$ | $\dfrac{\beta}{\alpha-1}\;\alpha > 1$ | $\dfrac{\beta^2}{(\alpha-1)^2(\alpha-2)}\;\alpha > 2$ | $\dfrac{2(-\beta s)^{\alpha/2}}{\Gamma(\alpha)}K_\alpha\left(\sqrt{-4\beta s}\right)$ |
| Dirichlet | $\text{Dir}(\alpha)$ | | $\dfrac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)}\prod_{i=1}^k x_i^{\alpha_i-1}$ | $\dfrac{\alpha_i}{\sum_{i=1}^k \alpha_i}$ | $\dfrac{\mathbb{E}[X_i](1-\mathbb{E}[X_i])}{\sum_{i=1}^k \alpha_i + 1}$ | |
| Beta | $\text{Beta}(\alpha,\beta)$ | $I_x(\alpha,\beta)$ | $\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ | $\dfrac{\alpha}{\alpha+\beta}$ | $\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ | $1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\dfrac{\alpha+r}{\alpha+\beta+r}\right)\dfrac{s^k}{k!}$ |
| Weibull | $\text{Weibull}(\lambda,k)$ | $1-e^{-(x/\lambda)^k}$ | $\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$ | $\lambda\Gamma\left(1+\dfrac{1}{k}\right)$ | $\lambda^2\Gamma\left(1+\dfrac{2}{k}\right)-\mu^2$ | $\sum_{n=0}^{\infty}\dfrac{s^n\lambda^n}{n!}\Gamma\left(1+\dfrac{n}{k}\right)$ |
| Pareto | $\text{Pareto}(x_m,\alpha)$ | $1-\left(\dfrac{x_m}{x}\right)^\alpha \; x \ge x_m$ | $\alpha\dfrac{x_m^\alpha}{x^{\alpha+1}} \quad x \ge x_m$ | $\dfrac{\alpha x_m}{\alpha-1}\;\alpha > 1$ | $\dfrac{x_m^2\alpha}{(\alpha-1)^2(\alpha-2)}\;\alpha > 2$ | $\alpha(-x_m s)^\alpha\Gamma(-\alpha,-x_m s)\; s < 0$ |

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



Uniform (continuous)
Normal
Log−Normal
Student's t
χ2
F
Exponential
Gamma
Inverse Gamma
Beta
Weibull
Pareto
PDF
x
μ = 0, σ2 = 0.2
μ = 0, σ2 = 1
μ = 0, σ2 = 5
μ = −2, σ2 = 0.5
μ = 0, σ2 = 3
μ = 2, σ2 = 2
μ = 0, σ2 = 1
μ = 0.5, σ2 = 1
μ = 0.25, σ2 = 1
μ = 0.125, σ2 = 1
ν = 1
ν = 2
ν = 5
ν = ∞
k = 1
k = 2
k = 3
k = 4
k = 5
d1 = 1, d2 = 1
d1 = 2, d2 = 1
d1 = 5, d2 = 2
d1 = 100, d2 = 1
d1 = 100, d2 = 100
β = 0.5
β = 1
β = 2.5
α = 1, β = 0.5
α = 2, β = 0.5
α = 3, β = 0.5
α = 5, β = 1
α = 9, β = 2
α = 1, β = 1
α = 2, β = 1
α = 3, β = 1
α = 3, β = 0.5
α = 0.5, β = 0.5
α = 5, β = 1
α = 1, β = 3
α = 2, β = 2
α = 2, β = 5
λ = 1, k = 0.5
λ = 1, k = 1
λ = 1, k = 1.5
λ = 1, k = 5
xm = 1, k = 1
xm = 1, k = 2
xm = 1, k = 4



Uniform (continuous)
CDF
x
a
b
Normal
μ = 0, σ² = 0.2
μ = 0, σ² = 1
μ = 0, σ² = 5
μ = −2, σ² = 0.5
Log−Normal
μ = 0, σ² = 3
μ = 2, σ² = 2
μ = 0, σ² = 1
μ = 0.5, σ² = 1
μ = 0.25, σ² = 1
μ = 0.125, σ² = 1
Student's t
ν = 1
ν = 2
ν = 5
ν = ∞
χ²
k = 1
k = 2
k = 3
k = 4
k = 5
F
d1 = 1, d2 = 1
d1 = 2, d2 = 1
d1 = 5, d2 = 2
d1 = 100, d2 = 1
d1 = 100, d2 = 100
Exponential
β = 0.5
β = 1
β = 2.5
Gamma
α = 1, β = 0.5
α = 2, β = 0.5
α = 3, β = 0.5
α = 5, β = 1
α = 9, β = 2
Inverse Gamma
α = 1, β = 1
α = 2, β = 1
α = 3, β = 1
α = 3, β = 0.5
Beta
α = 0.5, β = 0.5
α = 5, β = 1
α = 1, β = 3
α = 2, β = 2
α = 2, β = 5
Weibull
λ = 1, k = 0.5
λ = 1, k = 1
λ = 1, k = 1.5
λ = 1, k = 5
Pareto
xm = 1, k = 1
xm = 1, k = 2
xm = 1, k = 4

# 2 Probability Theory

Definitions

- Sample space $\Omega$
- Outcome (point or element) $\omega \in \Omega$
- Event $A \subseteq \Omega$
- $\sigma$-algebra $\mathcal{A}$
  1. $\varnothing \in \mathcal{A}$
  2. $A_1, A_2, \ldots, \in \mathcal{A} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$
  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. $\mathbb{P}\left[\bigsqcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$
- Probability space $(\Omega, \mathcal{A}, \mathbb{P})$

Properties

- $\mathbb{P}[\varnothing] = 0$
- $B = \Omega \cap B = (A \cup \neg A) \cap B = (A \cap B) \cup (\neg A \cap B)$
- $\mathbb{P}[\neg A] = 1 - \mathbb{P}[A]$
- $\mathbb{P}[B] = \mathbb{P}[A \cap B] + \mathbb{P}[\neg A \cap B]$
- $\mathbb{P}[\Omega] = 1 \qquad \mathbb{P}[\varnothing] = 0$
- $\neg(\bigcup_n A_n) = \bigcap_n \neg A_n \quad \neg(\bigcap_n A_n) = \bigcup_n \neg A_n$ 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

- $A_1 \subset A_2 \subset \ldots \implies \lim_{n\to\infty} \mathbb{P}[A_n] = \mathbb{P}[A]$ where $A = \bigcup_{i=1}^{\infty} A_i$
- $A_1 \supset A_2 \supset \ldots \implies \lim_{n\to\infty} \mathbb{P}[A_n] = \mathbb{P}[A]$ 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

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \qquad \mathbb{P}[B] > 0$$

Law of Total Probability

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

Bayes' Theorem

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

Inclusion-Exclusion Principle

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

# 3 Random Variables

Random Variable (RV)

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

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)

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

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

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

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

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

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

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 $\varphi$ 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

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

# 4 Expectation

Definition and properties

- $\mathbb{E}[X] = \mu_X = \int x\,dF_X(x) = \begin{cases} \sum_x x f_X(x) & \text{X discrete} \\ \int x f_X(x)\,dx & \text{X continuous} \end{cases}$
- $\mathbb{P}[X = c] = 1 \implies \mathbb{E}[X] = c$
- $\mathbb{E}[cX] = c\,\mathbb{E}[X]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[XY] = \int_{X,Y} xy f_{X,Y}(x,y)\,dF_X(x)\,dF_Y(y)$
- $\mathbb{E}[\varphi(Y)] \neq \varphi(\mathbb{E}[X])$ (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

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

# 5 Variance

Definition and properties

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

Standard deviation

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

Covariance

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

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

Correlation

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

Independence

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

Sample variance

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

Conditional variance

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

# 6 Inequalities

Cauchy-Schwarz

$$\mathbb{E}\left[XY\right]^2 \le \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$$

Markov

$$\mathbb{P}\left[\varphi(X) \ge t\right] \le \frac{\mathbb{E}\left[\varphi(X)\right]}{t}$$

Chebyshev

$$\mathbb{P}\left[\left|X - \mathbb{E}\left[X\right]\right| \ge t\right] \le \frac{\mathbb{V}\left[X\right]}{t^2}$$

Chernoff

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

Hoeffding

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

$$\mathbb{P}\left[\bar{X} - \mathbb{E}\left[\bar{X}\right] \ge t\right] \le e^{-2nt^2} \quad t > 0$$

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

Jensen

$$\mathbb{E}\left[\varphi(X)\right] \ge \varphi(\mathbb{E}\left[X\right]) \quad \varphi \text{ convex}$$

# 7 Distribution Relationships

Binomial

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

Negative Binomial

- $X \sim \mathrm{NBin}\left(1, p\right) = \mathrm{Geo}\left(p\right)$
- $X \sim \mathrm{NBin}\left(r, p\right) = \sum_{i=1}^{r} \mathrm{Geo}\left(p\right)$
- $X_i \sim \mathrm{NBin}\left(r_i, p\right) \implies \sum X_i \sim \mathrm{NBin}\left(\sum r_i, p\right)$
- $X \sim \mathrm{NBin}\left(r, p\right).\ Y \sim \mathrm{Bin}\left(s + r, p\right) \implies \mathbb{P}\left[X \le s\right] = \mathbb{P}\left[Y \ge r\right]$

Poisson

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

Exponential

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

Normal

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

Gamma

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

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

Beta

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

# 8 Probability and Moment Generating Functions

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

# 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

$$(Y \mid X = x) \sim \mathcal{N}\left(\rho x, 1-\rho^2\right) \qquad \text{and} \qquad (X \mid Y = y) \sim \mathcal{N}\left(\rho y, 1-\rho^2\right)$$

Independence

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

## 9.2 Bivariate Normal

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

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

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

Conditional mean and variance

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

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

## 9.3 Multivariate Normal

Covariance matrix $\Sigma$ (Precision matrix $\Sigma^{-1}$)

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

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

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

Properties

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

# 10 Convergence

Let $\{X_1, X_2, \ldots\}$ 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): $X_n \xrightarrow{\text{D}} X$

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

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

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

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

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

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

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

Relationships

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

Slutzky's Theorem

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

### 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)

$$\bar{X}_n \xrightarrow{\text{P}} \mu \qquad n \to \infty$$

Strong (SLLN)

$$\bar{X}_n \xrightarrow{\text{as}} \mu \qquad n \to \infty$$

### 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 $\mathbb{V}[X_1] = \sigma^2$.

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

$$\lim_{n\to\infty} \mathbb{P}[Z_n \le z] = \Phi(z) \qquad z \in \mathbb{R}$$

CLT notations

$$\begin{aligned} Z_n &\approx \mathcal{N}(0,1) \\ \bar{X}_n &\approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \\ \bar{X}_n - \mu &\approx \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \\ \sqrt{n}(\bar{X}_n - \mu) &\approx \mathcal{N}\left(0, \sigma^2\right) \\ \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} &\approx \mathcal{N}(0,1) \end{aligned}$$

Continuity correction

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

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

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 $X_1, \cdots, X_n \overset{iid}{\sim} F$ if not otherwise noted.

### 11.1 Point Estimation

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

- Mean squared error: $\text{MSE} = \mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right] = \mathsf{bias}(\widehat{\theta}_n)^2 + \mathbb{V}\left[\widehat{\theta}_n\right]$
- $\lim_{n\to\infty} \mathsf{bias}(\widehat{\theta}_n) = 0 \wedge \lim_{n\to\infty} \mathsf{se}(\widehat{\theta}_n) = 0 \implies \widehat{\theta}_n$ is consistent
- Asymptotic normality: $\dfrac{\widehat{\theta}_n - \theta}{\mathsf{se}} \xrightarrow{\text{D}} \mathcal{N}(0,1)$
- Slutzky's Theorem often lets us replace $\mathsf{se}(\widehat{\theta}_n)$ by some (weakly) consistent estimator $\widehat{\sigma}_n$.

## 11.2 Normal-Based Confidence Interval

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

$$C_n = \widehat{\theta}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}$$

## 11.3 Empirical distribution

Empirical Distribution Function (ECDF)

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

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

Properties (for any fixed $x$)

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

Dvoretzky-Kiefer-Wolfowitz (DKW) inequality ($X_1, \ldots, X_n \sim F$)

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

Nonparametric $1-\alpha$ confidence band for $F$

$$L(x) = \max\{\widehat{F}_n - \epsilon_n, 0\}$$
$$U(x) = \min\{\widehat{F}_n + \epsilon_n, 1\}$$
$$\epsilon = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

$$\mathbb{P}\left[L(x) \le F(x) \le U(x)\ \forall x\right] \ge 1-\alpha$$

## 11.4 Statistical Functionals

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

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

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

# 12 Parametric Inference

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

## 12.1 Method of Moments

$j^{\text{th}}$ moment

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

$j^{\text{th}}$ sample moment

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

Method of Moments estimator (MoM)

$$\begin{aligned} \alpha_1(\theta) &= \widehat{\alpha}_1 \\ \alpha_2(\theta) &= \widehat{\alpha}_2 \\ \vdots &= \vdots \\ \alpha_k(\theta) &= \widehat{\alpha}_k \end{aligned}$$

Properties of the MoM estimator

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

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

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

## 12.2 Maximum Likelihood

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

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

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

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

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

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: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$
- Equivariance: $\widehat{\theta}_n$ is the MLE $\Longrightarrow \varphi(\widehat{\theta}_n)$ 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. $\mathsf{se} \approx \sqrt{1/I_n(\theta)}$
  $$\frac{(\widehat{\theta}_n - \theta)}{\mathsf{se}} \xrightarrow{\text{D}} \mathcal{N}(0, 1)$$
  2. $\widehat{\mathsf{se}} \approx \sqrt{1/I_n(\widehat{\theta}_n)}$
  $$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\mathsf{se}}} \xrightarrow{\text{D}} \mathcal{N}(0, 1)$$
- Asymptotic optimality

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

- Approximately the Bayes estimator

### 12.2.1 Delta Method

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

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

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

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

## 12.3 Multiparameter Models

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

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

Fisher information matrix

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

Under appropriate regularity conditions

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

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

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

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

### 12.3.1 Multiparameter delta method

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

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

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

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

where

$$\widehat{\mathsf{se}}(\widehat{\tau}) = \sqrt{\left(\widehat{\nabla}\varphi\right)^T \widehat{J}_n \left(\widehat{\nabla}\varphi\right)}$$

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

## 12.4 Parametric Bootstrap

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

# 13 Hypothesis Testing

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

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: $H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$
- One-sided test: $H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$
- 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 | $\surd$ | Type I Error ($\alpha$) |
| $H_1$ true | Type II Error ($\beta$) | $\surd$ (power) |

p-value

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

| 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 $W = \dfrac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}}$
- $\mathbb{P}\left[|W| > z_{\alpha/2}\right] \to \alpha$
- p-value $= \mathbb{P}_{\theta_0}[|W| > |w|] \approx \mathbb{P}[|Z| > |w|] = 2\Phi(-|w|)$

Likelihood ratio test

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

Multinomial LRT

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

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 $H_0$
- $T \xrightarrow{\text{D}} \chi^2_{k-1}$
- p-value $= \mathbb{P}\left[\chi^2_{k-1} > T(x)\right]$
- Faster $\xrightarrow{\text{D}} X^2_{k-1}$ than LRT, hence preferable for small $n$

Independence testing

- $I$ rows, $J$ columns, $\mathbf{X}$ multinomial sample of size $n = I * J$
- MLEs unconstrained: $\widehat{p}_{ij} = \frac{X_{ij}}{n}$
- MLEs under $H_0$: $\widehat{p}_{0ij} = \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}$
- LRT: $\lambda = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}\log\left(\frac{nX_{ij}}{X_{i\cdot}X_{\cdot j}}\right)$
- 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 $\xrightarrow{\text{D}} \chi^2_k\nu$, where $\nu = (I-1)(J-1)$

# 14 Exponential Family

Scalar parameter

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

Vector parameter

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

Natural form

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

# 15 Bayesian Inference

Bayes' Theorem

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

Definitions

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

  In particular, $X^n$ IID $\implies f(x^n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \mathcal{L}_n(\theta)$
- Posterior density $f(\theta \mid x^n)$
- Normalizing constant $c_n = f(x^n) = \int f(x \mid \theta)f(\theta)\,d\theta$
- Kernel: part of a density that depends on $\theta$
- Posterior mean $\bar{\theta}_n = \int \theta f(\theta \mid x^n)\,d\theta = \frac{\int \theta\mathcal{L}_n(\theta)f(\theta)d\theta}{\int \mathcal{L}_n(\theta)f(\theta)\,d\theta}$

## 15.1 Credible Intervals

Posterior interval

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

Equal-tail credible interval

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

Highest posterior density (HPD) region $R_n$

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

$R_n$ is unimodal $\implies R_n$ is an interval

## 15.2 Function of parameters

Let $\tau = \varphi(\theta)$ and $A = \{\theta : \varphi(\theta) \le \tau\}$.
Posterior CDF for $\tau$

$$H(r \mid x^n) = \mathbb{P}\left[\varphi(\theta) \le \tau \mid x^n\right] = \int_A f(\theta \mid x^n)\, d\theta$$

Posterior density

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

Bayesian delta method

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

## 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: $f(\theta) \propto constant$
- Proper: $\int_{-\infty}^{\infty} f(\theta)\, d\theta = 1$
- Improper: $\int_{-\infty}^{\infty} f(\theta)\, d\theta = \infty$
- Jeffrey's Prior (transformation-invariant):

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

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

### 15.3.1 Conjugate Priors

| Continuous likelihood (subscript $c$ denotes constant) | | |
|---|---|---|
| Likelihood | Conjugate prior | Posterior hyperparameters |
| $\text{Unif}\,(0,\theta)$ | $\text{Pareto}(x_m, k)$ | $\max\left\{x_{(n)}, x_m\right\}, k+n$ |
| $\text{Exp}\,(\lambda)$ | $\text{Gamma}\,(\alpha,\beta)$ | $\alpha+n, \beta+\sum_{i=1}^n x_i$ |
| $\mathcal{N}\left(\mu, \sigma_c^2\right)$ | $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$ | $\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma_c^2}\right) / \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma_c^2}\right)$, $\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma_c^2}\right)^{-1}$ |
| $\mathcal{N}\left(\mu_c, \sigma^2\right)$ | Scaled Inverse Chi-square$(\nu, \sigma_0^2)$ | $\nu+n, \frac{\nu\sigma_0^2+\sum_{i=1}^n (x_i-\mu)^2}{\nu+n}$ |
| $\mathcal{N}\left(\mu, \sigma^2\right)$ | Normal-scaled Inverse Gamma$(\lambda,\nu,\alpha,\beta)$ | $\frac{\nu\lambda+n\bar{x}}{\nu+n}, \quad \nu+n, \quad \alpha+\frac{n}{2}$, $\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2+\frac{\gamma(\bar{x}-\lambda)^2}{2(n+\gamma)}$ |
| $\text{MVN}(\mu, \Sigma_c)$ | $\text{MVN}(\mu_0, \Sigma_0)$ | $\left(\Sigma_0^{-1}+n\Sigma_c^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0+n\Sigma^{-1}\bar{x}\right)$, $\left(\Sigma_0^{-1}+n\Sigma_c^{-1}\right)^{-1}$ |
| $\text{MVN}(\mu_c, \Sigma)$ | Inverse-Wishart$(\kappa, \Psi)$ | $n+\kappa, \Psi+\sum_{i=1}^n (x_i-\mu_c)(x_i-\mu_c)^T$ |
| $\text{Pareto}(x_{m_c}, k)$ | $\text{Gamma}\,(\alpha,\beta)$ | $\alpha+n, \beta+\sum_{i=1}^n \log\frac{x_i}{x_{m_c}}$ |
| $\text{Pareto}(x_m, k_c)$ | $\text{Pareto}(x_0, k_0)$ | $x_0, k_0 - kn$ where $k_0 > kn$ |
| $\text{Gamma}\,(\alpha_c, \beta)$ | $\text{Gamma}\,(\alpha_0, \beta_0)$ | $\alpha_0+n\alpha_c, \beta_0+\sum_{i=1}^n x_i$ |

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

## 15.4 Bayesian Testing

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

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

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

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

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

Marginal likelihood

$$f(x^n \mid H_i) = \int_\Theta f(x^n \mid \theta, H_i) f(\theta \mid H_i)\,d\theta$$

Posterior odds (of $H_i$ relative to $H_j$)

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

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 |

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

# 16 Sampling Methods

## 16.1 Inverse Transform Sampling

Setup

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

Algorithm

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

## 16.2 The Bootstrap

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

1. Estimate $\mathbb{V}_F[T_n]$ with $\mathbb{V}_{\widehat{F}_n}[T_n]$.
2. Approximate $\mathbb{V}_{\widehat{F}_n}[T_n]$ using simulation:
   (a) Repeat the following $B$ times to get $T^*_{n,1}, \ldots, T^*_{n,B}$, an IID sample from the sampling distribution implied by $\widehat{F}_n$
      i. Sample uniformly $X^*_1, \ldots, X^*_n \sim \widehat{F}_n$.
      ii. Compute $T^*_n = g(X^*_1, \ldots, X^*_n)$.
   (b) Then

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

### 16.2.1 Bootstrap Confidence Intervals

Normal-based interval

$$T_n \pm z_{\alpha/2}\widehat{\mathsf{se}}_{boot}$$

Pivotal interval

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

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

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R^*_{n,b} \le r)$$

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

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

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: $h(\theta) = \dfrac{k(\theta)}{\int k(\theta)\, d\theta}$
- Envelope condition: we can find $M > 0$ such that $k(\theta) \le Mg(\theta) \quad \forall\theta$

Algorithm

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

Example

- We can easily sample from the prior $g(\theta) = f(\theta)$
- Target is the posterior $h(\theta) \propto k(\theta) = f(x^n \mid \theta) f(\theta)$
- Envelope condition: $f(x^n \mid \theta) \le f(x^n \mid \widehat{\theta}_n) = \mathcal{L}_n(\widehat{\theta}_n) \equiv M$
- Algorithm
  1. Draw $\theta^{cand} \sim f(\theta)$
  2. Generate $u \sim \text{Unif}(0,1)$
  3. Accept $\theta^{cand}$ if $u \le \dfrac{\mathcal{L}_n(\theta^{cand})}{\mathcal{L}_n(\widehat{\theta}_n)}$

## 16.4 Importance Sampling

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

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

# 17 Decision Theory

Definitions

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

Loss functions

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

## 17.1 Risk

Posterior risk

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

(Frequentist) risk

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

Bayes risk

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

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

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

## 17.2 Admissibility

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

$$\forall\theta : R(\theta,\widehat{\theta}') \le R(\theta,\widehat{\theta})$$

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

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

## 17.3 Bayes Rule

Bayes rule (or Bayes estimator)

- $r(f,\widehat{\theta}) = \inf_{\widetilde{\theta}} r(f,\widetilde{\theta})$
- $\widehat{\theta}(x) = \inf r(\widehat{\theta}\,|\,x)\ \forall x \implies r(f,\widehat{\theta}) = \int r(\widehat{\theta}\,|\,x) f(x)\,dx$

Theorems

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

## 17.4 Minimax Rules

Maximum risk

$$\bar{R}(\widehat{\theta}) = \sup_{\theta} R(\theta,\widehat{\theta}) \qquad \bar{R}(a) = \sup_{\theta} R(\theta,a)$$

Minimax rule

$$\sup_{\theta} R(\theta,\widehat{\theta}) = \inf_{\widetilde{\theta}} \bar{R}(\widetilde{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta,\widetilde{\theta})$$

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

Least favorable prior

$$\widehat{\theta}^f = \text{Bayes rule} \;\wedge\; R(\theta,\widehat{\theta}^f) \le r(f,\widehat{\theta}^f)\ \forall\theta$$

# 18 Linear Regression

Definitions

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

## 18.1 Simple Linear Regression

Model

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

Fitted line

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

Predicted (fitted) values

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

Residuals

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

Residual sums of squares (RSS)

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

Least square estimates

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

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

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

$$\mathbb{E}\left[\widehat{\beta}\,|\,X^n\right] = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$$

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

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

$$\widehat{\mathsf{se}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}$$

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

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

- Asymptotic normality:

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

- Approximate $1-\alpha$ confidence intervals for $\beta_0$ and $\beta_1$:

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

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

R$^2$

$$R^2 = \frac{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2}{\sum_{i=1}^n(Y_i-\overline{Y})^2} = 1-\frac{\sum_{i=1}^n\hat{\epsilon}_i^2}{\sum_{i=1}^n(Y_i-\overline{Y})^2} = 1-\frac{\textsc{rss}}{\textsc{tss}}$$

Likelihood

$$\mathcal{L} = \prod_{i=1}^n f(X_i,Y_i) = \prod_{i=1}^n f_X(X_i) \times \prod_{i=1}^n f_{Y|X}(Y_i \mid X_i) = \mathcal{L}_1 \times \mathcal{L}_2$$
$$\mathcal{L}_1 = \prod_{i=1}^n f_X(X_i)$$
$$\mathcal{L}_2 = \prod_{i=1}^n f_{Y|X}(Y_i \mid X_i) \propto \sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_i\Big(Y_i-(\beta_0-\beta_1X_i)\Big)^2\right\}$$

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

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

## 18.2 Prediction

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

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

Prediction interval

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

## 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

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

$$\textsc{rss} = (y-X\beta)^T(y-X\beta) = \|Y-X\beta\|^2 = \sum_{i=1}^N(Y_i-x_i^T\beta)^2$$

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

$$\widehat{\beta} = (X^TX)^{-1}X^TY$$
$$\mathbb{V}\left[\widehat{\beta} \mid X^n\right] = \sigma^2(X^TX)^{-1}$$
$$\widehat{\beta} \approx \mathcal{N}\left(\beta,\sigma^2(X^TX)^{-1}\right)$$

Estimate regression function

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

Unbiased estimate for $\sigma^2$

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

MLE

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

$1-\alpha$ Confidence interval

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

## 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)

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

Prediction risk

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

Training error

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

$R^2$

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

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

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

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

Adjusted $R^2$

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

Mallow's $C_p$ statistic

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

Akaike Information Criterion (AIC)

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

Bayesian Information Criterion (BIC)

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

Validation and training

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

Leave-one-out cross-validation

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

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

# 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)

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

Frequentist risk

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

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

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

### 19.1.1 Histograms

Definitions

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

Histogram estimator

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

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

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

### 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

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

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 $\mathbb{E}\left[J(h)\right]$

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

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

## 19.2 Non-parametric Regression

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

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

$k$-nearest Neighbor Estimator

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

Nadaraya-Watson Kernel Estimator

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

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

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

## 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

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

Least squares estimator

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

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

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

# 20 Stochastic Processes

Stochastic Process

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

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

## 20.1 Markov Chains

Markov chain

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

Transition probabilities

$$p_{ij} \equiv \mathbb{P}\left[X_{n+1} = j \mid X_n = i\right]$$
$$p_{ij}(n) \equiv \mathbb{P}\left[X_{m+n} = j \mid X_m = i\right] \quad \text{n-step}$$

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

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

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

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

Marginal probability

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

## 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:

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

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

Homogeneous Poisson process

$$\lambda(t) \equiv \lambda \implies X_t \sim \mathrm{Po}\left(\lambda t\right) \qquad \lambda > 0$$

Waiting times

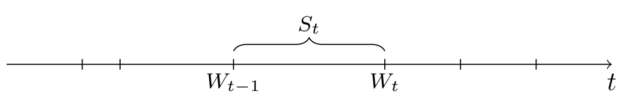
$$W_t := \text{time at which } X_t \text{ occurs}$$

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

Interarrival times

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

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




# 21 Time Series

Mean function

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

Autocovariance function

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

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

Autocorrelation function (ACF)

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

Cross-covariance function (CCV)

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

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

Random walk

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

Symmetric moving average

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

## 21.1 Stationary Time Series

Strictly stationary

$$\mathbb{P}\left[x_{t_1} \le c_1, \ldots, x_{t_k} \le c_k\right] = \mathbb{P}\left[x_{t_1+h} \le c_1, \ldots, x_{t_k+h} \le c_k\right]$$

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

Weakly stationary

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

Autocovariance function

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

Autocorrelation function (ACF)

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

Jointly stationary time series

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

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

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

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

Sample autocovariance function

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

Sample autocorrelation function

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

Sample cross-variance function

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

Sample cross-correlation function

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

Properties

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

## 21.3 Non-Stationary Time Series

Classical decomposition model

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

- $\mu_t$ = 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 $\widehat{\mu}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2$
3. Residuals $\triangleq$ noise $w_t$

Moving average

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

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

- 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

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z_p \qquad z \in \mathbb{C} \wedge \phi_p \neq 0$$

Autoregressive operator

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

Autoregressive model order $p$, $\mathsf{AR}(p)$

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

$\mathsf{AR}(1)$

- $x_t = \phi^k (x_{t-k}) + \sum_{j=0}^{k-1} \phi^j (w_{t-j}) \overset{k\to\infty, |\phi|<1}{=} \underbrace{\sum_{j=0}^{\infty} \phi^j (w_{t-j})}_{\text{linear process}}$
- $\mathbb{E}[x_t] = \sum_{j=0}^{\infty} \phi^j (\mathbb{E}[w_{t-j}]) = 0$
- $\gamma(h) = \text{Cov}[x_{t+h}, x_t] = \frac{\sigma_w^2 \phi^h}{1-\phi^2}$
- $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h$
- $\rho(h) = \phi \rho(h-1) \quad h = 1, 2, \ldots$

Moving average polynomial

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z_q \qquad z \in \mathbb{C} \wedge \theta_q \neq 0$$

Moving average operator

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

$\mathsf{MA}(q)$ (moving average model order $q$)

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

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

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

$\mathsf{MA}(1)$

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

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

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

$\mathsf{ARMA}(p, q)$

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

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

Partial autocorrelation function (PACF)

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

$\mathsf{ARIMA}(p, d, q)$

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

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

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

### 21.4.1 Causality and Invertibility

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

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

$\mathsf{ARMA}\,(p,q)$ 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

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

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

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

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

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

| | $\mathsf{AR}\,(p)$ | $\mathsf{MA}\,(q)$ | $\mathsf{ARMA}\,(p,q)$ |
|---|---|---|---|
| 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 $\omega$ (cycles per unit time), period $1/\omega$
- Amplitude $A$
- Phase $\phi$
- $U_1 = A\cos\phi$ and $U_2 = A\sin\phi$ often normally distributed RV's

Periodic mixture

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

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

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

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

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

Spectral density

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

- Needs $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty \implies \gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h}f(\omega)\,d\omega \quad h = 0, \pm1, \dots$
- $f(\omega) \ge 0$
- $f(\omega) = f(-\omega)$
- $f(\omega) = f(1-\omega)$
- $\gamma(0) = \mathbb{V}\left[x_t\right] = \int_{-1/2}^{1/2} f(\omega)\,d\omega$
- 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)

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

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

## 22.2 Beta Function

- Ordinary: $\mathrm{B}(x,y) = \mathrm{B}(y,x) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$
- Incomplete: $\mathrm{B}(x;\,a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,dt$
- Regularized incomplete:
  $$I_x(a,b) = \frac{\mathrm{B}(x;\,a,b)}{\mathrm{B}(a,b)} \overset{a,b\in\mathbb{N}}{=} \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!} x^j (1-x)^{a+b-1-j}$$
- $I_0(a,b) = 0 \qquad I_1(a,b) = 1$
- $I_x(a,b) = 1 - I_{1-x}(b,a)$

## 22.3 Series

Finite

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

Binomial

- $\displaystyle\sum_{k=0}^{n} \binom{n}{k} = 2^n$
- $\displaystyle\sum_{k=0}^{n} \binom{r+k}{k} = \binom{r+n+1}{n}$
- $\displaystyle\sum_{k=0}^{n} \binom{k}{m} = \binom{n+1}{m+1}$
- Vandermonde's Identity:
  $\displaystyle\sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k} = \binom{m+n}{r}$
- Binomial Theorem:
  $\displaystyle\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n$

Infinite

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

## 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, $2^{nd}$ kind

$$\left\{ {n \atop k} \right\} = k \left\{ {n-1 \atop k} \right\} + \left\{ {n-1 \atop k-1} \right\} \qquad 1 \le k \le n \qquad \left\{ {n \atop 0} \right\} = \begin{cases} 1 & n = 0 \\ 0 & \text{else} \end{cases}$$

Partitions

$$P_{n+k,k} = \sum_{i=1}^{n} P_{n,i} \qquad k > n : \; P_{n,k} = 0 \qquad n \ge 1 : \; P_{n,0} = 0, \; P_{0,0} = 1$$

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

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

# References


[1] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Probability Theory*. Brooks Cole, 1972.

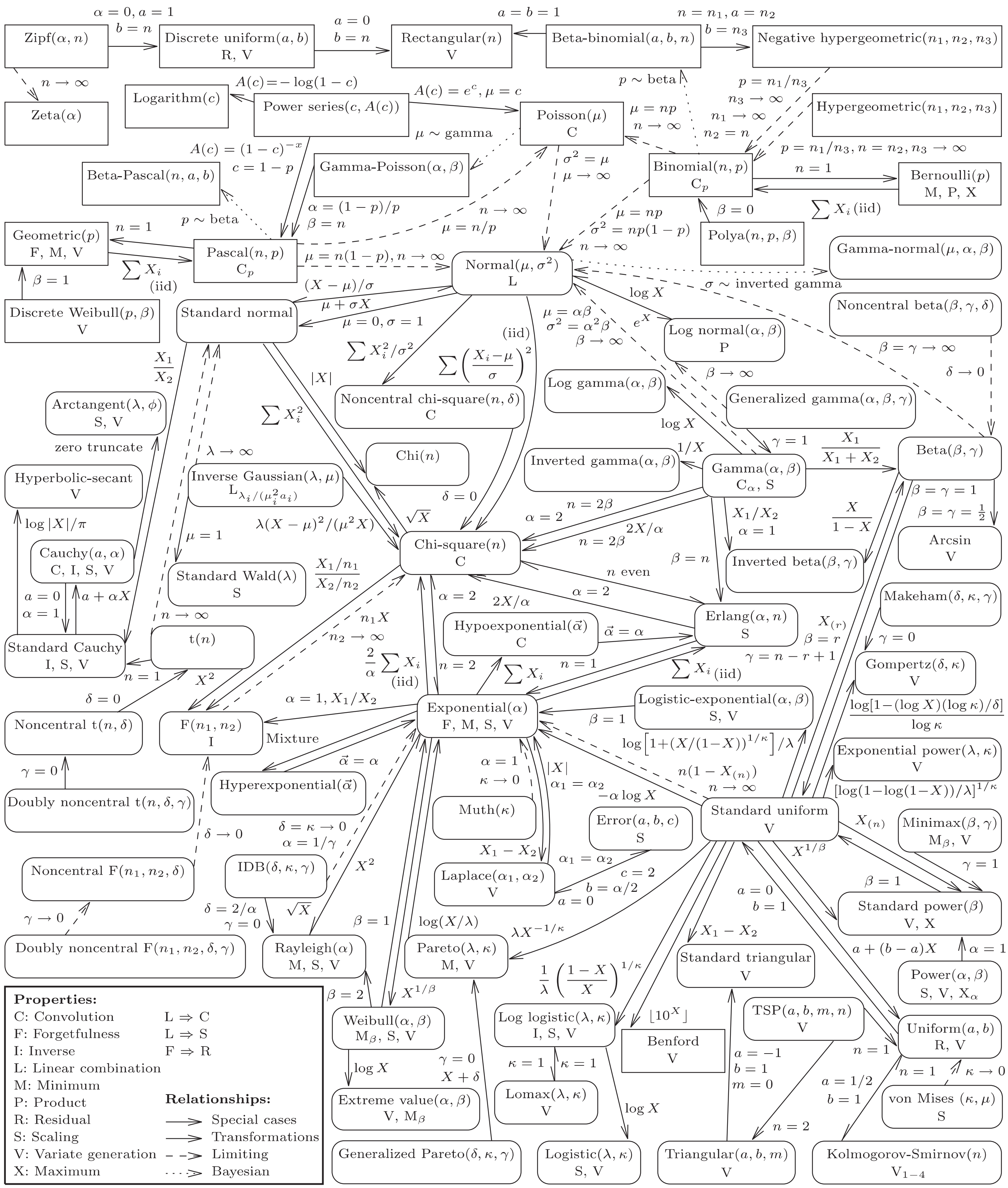
[2] L. M. Leemis and J. T. McQueston. Univariate Distribution Relationships. *The American Statistician*, 62(1):45–53, 2008.

[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].