Contents

1 Probability and Statistics Foundations

1.1 Sample Space and Probability Measure

Definition 1 (Probability miscellany)**.** The **sample space** Ω is the collection of all possible outcomes of a random experiment. Elements of the sample space are outcomes. Subsets of the sample space are **events**. $A_1 \dots$ are **pairwise disjoint** if $A_i \cap A_j = \emptyset$ for all $i \neq j$. A **partition** of the sample space is a collection of pairwise disjoint sets such that $\cup_{i=1}^{\infty} A_i = \Omega$.

Definition 2 (σ-algebra)**.** A **σ-algebra**, $\mathcal{F}$ is s collection of subsets satisfying:

(a) $\Omega \in \mathcal{F}, \emptyset \in \mathcal{F}$

(b) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

(c) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$

The sets in $\mathcal{F}$ are said to be measurable and $(\Omega, \mathcal{F})$ are a measurable space.

Definition 3 (Probability measure)**.** A **measure** is a function that takes elements of the σ-algebra and outputs a real number. The **probability measure**, $\mathbb{P}(\cdot) : \mathcal{F} \to [0,1]$, where the number describes the likelihood of the event.

(a) $\mathbb{P}(\Omega) = 1$

(b) $\mathbb{P}(A) \geq 0 \quad \forall A \in \mathcal{F}$

(c) For all mutually exclusive events, $\mathbb{P}\big(\cup_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

1.2 Random Variables

Definition 4 (Random variable)**.** A **random variable** is a function from the sample space, Ω into the real numbers (C&B 1.4.1 pg 27).

Definition 5 (CDF)**.** Every random variable, X, has a **cumulative distribution function, or cdf**, denoted by $F_X(x)$ satisfying:

$$F_X(x) = P_X(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}), \text{ for all } x$$

Note: every cdf is right-continuous (continuous when approached from the right), nondecreasing, $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$, and completely determines the distribution of X.

Definition 6 (PMF/PDF)**.** The probability mass function (PMF) describes the distribution of a discrete RV:

$$p(x) = P(X = x) = F(x) - F(x^-)$$

The probability density function (PDF) describes the distribution of a continuous RV:

$$p(x) = \frac{d}{dx}F(x)$$

When X is continuous, the CDF is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x^{'})dx^{'}$$

Theorem 1 (Theorem 1.6.5 (C&B))**.** A function $f_X(x)$ is a pdf (or pmf) of X iff:

(a) $f_X(x) \geq 0$ for all x

(b) $\sum_x f_X(x) = 1$ for pmf and $\int_{-\infty}^{\infty} f_X(x)dx = 1$ for pdf

1.3 Common distributions

1.3.1 Discrete

Definition 7 (Bernoulli)**.** If $X \sim \text{Ber}(p)$, then $X \in \{0, 1\}$ and $0 \leq p \leq 1$ s.t.:

$$P(X = 1) = p, \quad P(X = 0) = (1 - p)$$

The following are properties of a Bernoulli RV:

Mean:	$E[X] = p$
Variance:	$\text{Var}[X] = p(1 - p)$
MGF:	$M_X(t) = (1 - p) + pe^t$

Definition 8 (Binomial)**.** If $X \sim \text{Bin}(n, p)$, then $X = 0, 1, 2, \ldots$:

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

A $\text{Bin}(n, p)$ variable is the sum of n Bernoulli trials with probability p. The following are properties of a Binomial RV:

Mean:	$E[X] = np$
Variance:	$\text{Var}[X] = np(1 - p)$
MGF:	$M_X(t) = [(1 - p) + pe^t]^n$

Theorem 2 (Binomial Theorem)**.** For any real numbers, x, y and integer $n \geq 0$

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

Definition 9 (Geometric)**.** If $X \sim \text{Geo}(p)$:

$$P(X = n) = (1-p)^{n-1} p$$

for $n = 1, 2, 3, \ldots$. A geometric random variable can be considered the ”number of trials to obtain a success”. The following are properties of a Geometric RV:

Mean:	$E[X] = \frac{1}{p}$
Variance:	$\text{Var}[X] = \frac{1-p}{p^2}$
MGF:	$M_X(t) = \frac{pe^t}{1-(1-p)e^t}$

Definition 10 (Poisson)**.** If $X \sim \text{Poi}(\lambda)$, $X = 0, 1, 2, \ldots$ and:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

A Poisson RV is often used to model a counting process where we are waiting for an occurrence (i.e., we assume that probability of arrival for small intervals is proportional to the length of the time interval). The following are properties of a Poisson RV:

Mean:	$E[X] = \lambda$
Variance:	$\text{Var}[X] = \lambda$
MGF:	$M_X(t) = e^{\lambda(e^t - 1)}$

Property 1 (Sum of Poisson RVs)**.** Given $X_1, \ldots, X_n \sim \text{Poisson}(\lambda_i)$, and $X_i \perp X_j$ for all $i \neq j$, then

$$\sum_{i=1}^{n} X_i \sim \text{Poisson}(\sum_{i=1}^{n} \lambda_i)$$

In other words, the sum of independent Poisson RVs are Poisson with rate equal to the sum of the individual rates.

Definition 11 (Negative binomial)**.** To model the number of failures before the r-th success (which is equivalent to the number of trials to achieve a fixed number of success):

$$P(Y=y)=\binom{r+y-1}{y}p^r(1-p)^y, \quad y=0,1,\dots$$

where

$$\binom{r+y-1}{y}=(-1)^y\binom{-r}{y}$$

The following are properties of a NB RV:

$$\textbf{Mean:} \qquad E[Y]=r\frac{1-p}{p}$$

$$\textbf{Variance:} \qquad \mathrm{Var}[Y]=r\frac{1-p}{p^2}$$

$$\textbf{MGF:} \qquad M_Y(t)=\left(\frac{p}{1-(1-p)e^t}\right)^r$$

1.3.2 Continuous

Definition 12 (Uniform)**.** If $X\sim \mathrm{Unif}[a,b]$ is a continuous RV over $[a,b]$ then:

$$p(x)=\frac{1}{b-a}\mathbb{I}(a\le x\le b)$$

The following are properties of a Uniform RV:

$$\textbf{Mean:} \qquad E[X]=\frac{b+a}{2}$$

$$\textbf{Variance:} \qquad \mathrm{Var}[X]=\frac{(b-a)^2}{12}$$

$$\textbf{MGF:} \qquad M_X(t)=\frac{e^{bt}-e^{at}}{(b-a)t}$$

Property 2 (Minimum of many uniforms)**.** Consider $X_1,\dots,X_n\sim\mathrm{Unif}[0,1]$ and $U=n\cdot\min\{X_1,\dots,X_n\}$. Then:

$$1-F_U(u)=P(\min\{X_1,\dots,X_n\}>\frac{u}{n})=\prod_{i=1}^n P\big(X_i>\frac{u}{n}\big)=\left(1-\frac{u}{n}\right)^n\to e^{-u}$$
$$F_U(u)=1-e^{-u} \quad \& \quad f_U(u)\to e^{-u}=\mathrm{Exp}(1)$$

Property 3 (Uniform in n-dimensions)**.** A distribution that is uniform across an n-dimensional box has marginals that are independent and uniform. When X and Y are associated, the marginals are uniform but not independent.

Consider the case where (X,Y) is uniform over $D = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Then $p_{Y|X} =$ Unif$[0,1]$. Now consider when (X,Y) is uniform over $D = \{(x,y) : x \geq 0, y \geq 0, x+y \leq 1\}$. $p_{Y|X}$ will be Unif$[0, 1-X] \implies p_{Y|X}(y|X) = \frac{1}{1-x} I(0 \leq y \leq 1-x)$

Definition 13 (Normal)**.** If $X \sim \mathrm{N}(\mu, \sigma^2)$:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The following are properties of a Normal RV:

$$\begin{aligned} \textbf{Mean:} &\quad E[X] = \mu \\ \textbf{Variance:} &\quad \mathrm{Var}[X] = \sigma^2 \\ \textbf{MGF:} &\quad M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} \end{aligned}$$

Property 4 (Properties of Normal RV)**.** Here are some properties of Normal random variables:

1. For $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ independent, and $a_1, a_2 \in \mathbb{R}$, $a_1 X + a_2 Y \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$
2. For $X_1, \ldots, X_n$ IID normal from $N(\mu, \Sigma^2)$, then $\bar{X_n} \sim N(\mu, \sigma^2/n)$
3. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(0,1)$. Then $Z_1 = X_1^2$ follows a χ^2 distribution with df=1. And $Z_n = \sum_{i=1}^n X_i^2$ follows a χ^2 distribution with df=n.
4. **Stein's Lemma**: is useful for calculating higher order moments of normal distributions. For $X \sim N(\theta, \sigma^2)$:

$$\begin{aligned} \mathbb{E}(g(X)(X-\theta)) &= \sigma^2 \mathbb{E}[g'(X)] \\ \text{In action: : } \mathbb{E}[X^3] &= \mathbb{E}(X^2(X-\theta+\theta) \\ &= \mathbb{E}[X^2(X-\theta)] + \theta\mathbb{E}[X^2] \\ &= 2\sigma^2\mathbb{E}[X] + \theta(\sigma^2+\theta^2) \\ &= 3\sigma^2\theta + \theta^3 \end{aligned}$$

Definition 14 (Exponential)**.** If $X \sim \mathrm{Exp}(\lambda)$, then $X \in [0, \infty)$:

$$p(x) = \lambda e^{-\lambda x} \mathbb{I}(x \geq 0)$$

The following are properties of an Exponential RV:

Mean:	$E[X] = \frac{1}{\lambda}$
Variance:	$\text{Var}[X] = \left(\frac{1}{\lambda}\right)^2$
MGF:	$M_X(t) = \frac{1}{1-\frac{t}{\lambda}}, \quad t < \lambda$

Property 5 (Memoryless property)**.** Given $X \sim \text{Exp}(\lambda)$:

$$\begin{aligned} P(X > x + y | X > x) &= \frac{P(\{w : X(w) > x + y, X(w) > x)\}}{P(\{x : X(w) > x\})} \\ &= \frac{P(X > x + y)}{P(X > x)} \\ &= \frac{1 - P(X < x + y)}{1 - P(X < x)} \qquad \text{(CDF: } P(X < x) = F(x) = 1 - e^{\lambda x}) \\ &= \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} \\ &= e^{-\lambda y} = P(X > y) \end{aligned}$$

This is the *memoryless property*, as the probability of $X > x + y$ depends only on the increment y, not on x.

Property 6 (Abs. Difference, Minimum, and Ratio)**.** Consider exponential RVs:

(a) Consider $X, Y \sim \text{Exp}(1)$. For $U = |X - Y|$, $U \sim \text{Exp}(1)$ (another memoryless property)

(b) For $X_1, \dots, X_n \sim \text{Exp}(\lambda)$, $U = \min\{X_1, \dots, X_n\} \sim \text{Exp}(n\lambda)$

(c) Consider $X, Y \sim \text{Exp}(1)$. For $U = \frac{X}{X+Y}$, then $U \sim \text{Unif}[0, 1]$

Definition 15 (Cauchy)**.** If $X \sim \text{Cauchy}(\mu, \sigma^2)$:

$$p(x) = \frac{1}{\pi\sigma} \frac{1}{1 + (x - \mu)^2/\sigma^2}$$

The following are properties of a Cauchy RV:

Mean:	Does Not Exist
Variance:	Does Not Exist
MGF:	Does Not Exist

Definition 16 (Gamma)**.** If $X \sim \text{Gamma}(\alpha, \lambda)$ s.t. $X \geq 0$ and $\alpha, \lambda, \beta > 0$:

$$p(x) = \boxed{\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbb{I}(x \geq 0)} \quad = \quad \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

Where $\Gamma(\alpha) = (\alpha - 1)!$ The following are properties of an Gamma RV:

$$\begin{aligned} &\textbf{Mean:} && E[X] = \frac{\alpha}{\lambda} = \alpha\beta \\ &\textbf{Variance:} && \text{Var}[X] = \frac{\alpha}{\lambda^2} = \alpha\beta^2 \\ &\textbf{MGF:} && M_X(t) = \left(\frac{1}{1 - \frac{t}{\lambda}}\right)^\alpha, \quad t < \lambda \quad M_X(t) = \left(\frac{1}{1 - \beta t}\right)^\alpha \quad t < \frac{1}{\beta} \end{aligned}$$

Property 7 (Sum of independent Gammas)**.** For $X_1, \ldots, X_n$ independent and $X_i \sim \text{Gamma}(\alpha_i, \beta)$, and $T = \sum_{i=1}^n X_i$, the distribution of T is:

$$T \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$$

Definition 17 (Beta)**.** If $X \sim \text{Beta}(\alpha, \beta)$ s.t. $X \in [0, 1]$:

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{I}(0 \leq x \leq 1)$$

The following are properties of a Beta RV:

$$\begin{aligned} &\textbf{Mean:} && E[X] = \frac{\alpha}{\alpha + \beta} \\ &\textbf{Variance:} && \text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &\textbf{MGF:} && M_X(t) = 1 + \sum_{k=1}^\infty \left(\prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r}\right) \frac{t^k}{k!} \end{aligned}$$

1.4 Joint/Conditional Probability/CDF/PDF

Definition 18 (Conditional probability)**.** The conditional probability of A given B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Definition 19 (Simpson's Paradox)**.** Consider you have 3 events, A, B, C. If you know the following:

$$P(A|B,C) \geq P(A|B^C,C)$$
$$P(A|B,C^C) \geq P(A|B^C,C^C)$$

Consider A is the event your accepted, B is the event that you're a female, and C denote your program. So even though the probability of acceptance given female and program is higher than the probability of acceptance given male and program, this does not imply that probability of acceptance given you're female is higher than the probability of acceptance given you're a male: i.e., we **cannot conclude** $P(A|B) \geq P(A|B^C)$

Definition 20 (Joint and Conditional CDF)**.** Given two random variables, X, Y, their joint CDF:

$$P_{XY}(x,y) = F(x,y) = P(X \leq x, Y \leq y)$$

The joint pdf:

$$p_{XY}(x,y) = \frac{\partial^2 F}{\partial x \partial y}$$

The conditional PDF of Y given $X = x$:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

Strategy 1 (Solving for conditional pdf)**.** Note that:

$$\begin{aligned} & p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} \\ \implies & p_{Y|X}(y|x) \propto p_{XY}(x,y) \quad \text{B/c } X \text{ can be treated as a constant} \end{aligned}$$

1.5 Independence

Definition 21 (Independence)**.** Two events are independent if $P(A \cap B) = P(A)P(B)$ or equivalently $P(A|B) = P(A)$.

Theorem 3 (Re: Independence)**.** Two events are independent if their joint CDF can be factorized directly into a product of marginal CDFs:

$$F(x,y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

Definition 22 (Mutual Independence)**.** Many random variables are mutually independent if their joint CDF/PDF can be factorized into a product of marginals.

1.6 Total Probability and Bayes Theorem

Theorem 4 (Law of Total Probability)**.** The Law of Total Probability says that if $B_1, \dots, B_k$ form a partition of Ω:

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Theorem 5 (Bayes Rule)**.** The Law of Total Probability says that if $A_1, \dots, A_k$ form a partition of Ω:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(B_j)} = \frac{P(B|A_i)P(A_i)}{P(B)}$$

Theorem 6 (Bayes Theorem)**.** Bayes Theorem generalizes the result of Bayes Rule to RVs.

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{XY}(x,y)}{p_Y(y)} \\ &= \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \\ &= \begin{cases} \frac{p_{Y|X}(y|x)p_X(x)}{\int p_{Y|X}(y|x)p_X(x)} & \text{if continuous} \\ \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')} & \text{if discrete} \end{cases} \end{aligned}$$

1.7 Conditional Independence

Theorem 7 (Conditional independence)**.** For p_{XYZ} a joint pdf/pmf, the following are equivalent:

$$\left.\begin{array}{ll} \text{(a)} & X \perp Y|Z \\ \text{(b)} & p_{XY|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z) \\ \text{(c)} & p_{X|YZ}(x|y,z) = p_{X|Z}(x|z) \end{array}\right\} \text{The most important ones}$$

(d) $p_{XYZ}(x,y,z) = \frac{p_{XZ}(x,z)p_{YZ}(y,z)}{p_Z(z)}$

(e) $p_{XYZ}(x,y,z) = g(x,z)h(y,z)$ for g, h functions.

(f) $p_{X|YZ}(x|y,z) = w(x,z)$ for some function w.

Property 8 (Conditional Independence $\not\Leftrightarrow$ Independence)**.** Consider the case where $X, Y \sim \text{Unif}[0,1]^2 \implies X \perp Y$. But let $Z = 1$ if $x^2 + y^2 \leq 1$, so $XY|Z$.

Next, consider the case that $X, Y|Z = 1 \sim \text{Unif}[0,1]^2$, $X, Y|Z = 0 \sim \text{Unif}[2,3]^2$. $X \perp Y|Z$ but XY

2 Transformations of RVs

2.1 Functions of one RV

Theorem 8 (Theorem 2.1.5 C&B)**.** Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is monotone. Suppose $f_X(x)$ is continuous on $\mathcal{X}$ and $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then pdf of Y is:

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 \quad \text{else} \end{cases}$$

Theorem 9 (Theorem 2.1.8 C&B)**.** Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is monotone. Suppose there exists a partition, $A_0, \dots, A_k$ of the sample space, $\mathcal{X}$ s.t. $P(X \in A_0) = 0$ and f_X is continuous on each A_i. Suppose there exist functions $g_1(x), \dots, g_k(x)$ defined on $A_1, \dots, A_k$ respectively s.t.

(a) $g(x) = g_i(x) \forall x \in A_i$

(b) $g_i(x)$ is monotone on A_i

(c) $\mathcal{Y} = \{y : y = g_i(x) \text{ for some } x \in A_i\}$ is the same for each $i = 1, \dots, k$.

(d) $g_i^{-1}(y)$ has a continous derivative on $\mathcal{Y}$ for each $i = 1, \dots, k$.

Then:

$$f_Y(y) = \begin{cases} \sum_{i=1}^k f_X(g_i^{-1}(y))\left|\frac{d}{dy}g_i^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 \quad \text{else} \end{cases}$$

Theorem 10 (Probability Integral Transform (Thm 2.1.10 C & B))**.**
Let X have continuous cdf $F_X(x)$. Define RV $Y = F_X(x)$. Then $Y \sim \text{Unif}[0, 1]$.

Strategy 2 (When in doubt, work out the CDF)**.** Say $X \sim N(0, 1)$ and $Y = X^2$:

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(X^2 \le y) \\ &= P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$$

Then you can differentiate wrt y to get the pdf.

2.2 Functions of two or more RVs

Property 9 (Max of many independent RVs)**.** Let $U = \max\{X_1, \dots, X_n\}$. Then:

$$F_U(u) = P(U \leq u) = P(\max\{X_1, \dots, X_n\} \leq u) = P(X_1 \leq u, \dots, X_n \leq u) = P(X_1 \leq u) \cdot \dots P(X_n \leq u)$$

Property 10 (Min of many independent RVs)**.** Let $U = \min\{X_1, \dots, X_n\}$. Then:

$$\begin{aligned} 1 - F_U(u) = P(\min\{X_1, \dots, X_n\} > u) = P(X_1 > u, \dots, X_n > u) = P(X_1 > u) \dots P(X_n > u) \\ F_U(u) = 1 - P(X_1 > u) \dots P(X_n > u) \end{aligned}$$

3 Expectation and moments

3.1 Expectation

Definition 23 (Expected value)**.** The *expected value* or *mean* of a random variable, $g(X)$, denoted by $E(g(X))$ is:

$$E(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \text{ continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) P(X = x) & \text{if } X \text{ discrete} \end{cases}$$

Property 11 (Properties of Expectation)**.** The following are some useful properties of expectation:

(a) Decomposable under addition: $E\Big(\sum_{j=1}^{k} c_j g_j(X)\Big) = \sum_{j=1}^{k} c_j E(g_j(X))$

(b) Decomposable under multiplication with independence: if $X_1, \ldots, X_n$ are independent, $E(\prod_{i=1}^{n} g_i(X_i)) = \prod_{i=1}^{n} E(g_i(X_i))$

Definition 24 (Variance, Covariance, and Pearson's Correlation)**.** The *variance* is the second-centered moment of X. It is also the second uncentered moment minus the square of the first uncentered moment.

$$\text{Var}(X) = \underbrace{E((X - E(X))^2)}_{\text{second centered moment}} = \underbrace{E(X^2) - (E(X))^2}_{\text{second moment- first moment squared}}$$

Given two RV X, Y, their *covariance* is:

$$\text{Cov}(X, Y) = E((X - E(X)(Y - E(Y)) = E(XY) - E(X)E(Y)$$

Given two RV, X, Y, their *Pearson's correlation* is:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Note that when $X \perp Y \implies \text{Cov}(X, Y) = \rho(X, Y) = 0$. However, $\text{Cov}(X, Y) = \rho(X, Y) = 0 \not\Longrightarrow X \perp Y$.

Property 12 (Properties of Variance)**.** The following are properties of the variance:

(a) $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X, Y)$

(b) If $X \perp Y$, $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$

(c) $\text{Var}\Big(\sum_{i=1}^{n} a_i X_i\Big) = \sum_{i=1}^{k} a_i^2 \text{Var}(X_i)$ when X_i are independent

(d) $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$ when X, Y are dependent.

Property 13 (Lower bound on covariance using Cauchy-Schwartz)**.**

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))\\ &\overset{\text{C-S}}{\leq} \sqrt{\mathbb{E}[(X-\mathbb{E}(X))^2]\mathbb{E}[(Y-\mathbb{E}(Y))^2]}\\ &\leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}\end{aligned}$$

3.2 Moments and Moment Generating Functions

Definition 25 (Moments and Centered Moments)**.** The r-th momemt of X is defined as:

$$E(X^r) = \begin{cases}\int x^r p_X(x)dx & \text{if } X \text{ cont.}\\ \sum_{x\in\mathcal{X}} x^r p_X(x) & \text{if } X \text{ disc.}\end{cases}$$

$E((X-E(X))^r)$ is called the r-th centered moment. Note that $\mathrm{Var}(X)$, $\mathrm{Skew}(X)$, and $\mathrm{Kurtosis}(X)$ are the second, third, and fourth centered moments.

Definition 26 (Moment Generating Function)**.** The *Moment generating function* is a powerful function that describes the underlying features of a RV:

$$M_X(t) = E(e^{tX}) = 1 + tE(X) + \frac{t^2E(X^2)}{2!} + \frac{t^3E(X^3)}{3!} + \dots$$

The j-th moment of X is then:

$$E(X^j) = M^{(j)}(0) = \frac{d^jM_X(t)}{dt^j}\Big|_{t=0}$$

Property 14 (Properties of MGFs)**.** MGFs have the following properties:

(a) Location-scale: $M_{aX+b} = e^{bt}M_X(at)$

(b) Multiplicity: $M_{X+Y}(t) = E(e^{Xt}e^{YT})$. So if $X \perp Y \implies M_{X+Y}(t) = M_X(t)M_Y(t)$

Property 15 (MGFs uniquely determine distribution)**.** For RVs X and Y, if they have the same MGF, then their distributions (CDFs) are the same.

Theorem 11 (Convergence of MGFs)**.** Suppose $\{X_i, i = 1, 2, \dots\}$ is a sequence of RVs, each with MGF

$M_{X_i}(t)$. Suppose that:

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t) \qquad \text{For all } t \text{ in neighborhood of } 0$$

Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and for all x where $F_X(x)$ is continous, we have:

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x)$$

4 Convergence Theory

Definition 27 (Convergence in distribution)**.** Consider a sequence of RVs, $X_1, X_2, \dots$ with corresponding CDFs $F_1, F_2, \dots$. For a random variable X with CDF F, we say X_n **converges in distribution** to X, i.e., $X_n \xrightarrow{D} X$ if for every x:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

This can be interpreted as the CDFs of a sequence of random variables converging to the CDF of a fixed RV.

Note: convergence in distribution is often used to construct a confidence interval or perform a hypothesis test.

Definition 28 (Convergence in probability)**.** Consider a sequence of RVs, $X_1, X_2, \dots$. We say X_n **converges in probability**, i.e., $X_n \xrightarrow{P} X$, to another random variable X if for any $\epsilon > 0$:

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

$$\text{Alternatively: } \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$$

Note that convergence in probability implies convergence in distribution.

Note: an estimator is consistent if it converges in probability towards its target population quantity.

4.1 Inequalities, Weak Law of Large Numbers, and Convergence Theorems

4.1.1 Key Inequalities

Theorem 12 (Markov's Inequality)**.** Let X be a non-negative RV. Then for any $\epsilon > 0$:

$$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

This inequality implies that *convergence in expectation implies convergence in probability.*

Theorem 13 (Chebychev's Inequality)**.** Let X be a RV and let g be a nonnegative function. For any $\epsilon > 0$:

$$P(g(X) \geq \epsilon) \leq \frac{E[g(X)]}{\epsilon}$$

The specific case of $g(X) = |X - E(X)|$ yields:

$$P(|X - E(X)| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}$$

Demonstrating that any sequence of random variables with vanishing variance converges in probability to their mean.

Theorem 14 (Jensen's Inequality)**.** Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then:

(a) If X is a RV, then:

$$\phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X))$$

(b) If g is a function s.t. $\mathbb{E}(g(X)) < \infty$

$$\phi(\mathbb{E}(g(X))) \leq E(\phi(g(X)))$$

(c) Suppose $f : [a, b] \to \mathbb{R}$ is integrable on $[a, b]$.

$$\phi(\mathbb{E}(f(x))) = \phi\left(\frac{1}{b-a}\int_a^b f(x)dx\right) \leq \frac{1}{b-a}\int_a^b \phi(f(x))dx$$

Theorem 15 (Cauchy-Schartz Inequality)**.**

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

Theorem 16 (Triangle/Reverse Triangle inequality)**.** Triangle inequality:

$$\mathbb{E}(|X+Y|) \leq E(|X|) + E(|Y|)$$

Triangle inequality with norms: quantity in parentheses taken to squared power:

$$||x+y|| \leq ||x|| + ||y||$$

Revers triangle inequality:

$$|||x|| - ||y||| \leq ||x-y||$$

4.1.2 Weak Law of Large Numbers

Theorem 17 (Weak Law of Large Numbers)**.** v Let $X_1, \ldots, X_n \sim F$ and $\mu = E[X_1]$, If $\mu, (X_1) = \sigma^2 < \infty$, the sample average:

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

converges in probability to μ:

$$\lim_{n\to\infty} P(|\bar{X_n} - \mu| > \epsilon) = 0 \implies \bar{X_n} \xrightarrow{P} \mu$$

Thus, regardless of distribution, the sample mean is a consistent estimator of the population mean. The WLLN follows directly from Chebyshev's inequality.

4.1.3 Convergence Theorems

Theorem 18 (Continuous Mapping Theorem)**.** Let g be a continous function:

(a) If a sequence of RVs, $X_n \xrightarrow{D} X$, then $g(X_n) \xrightarrow{D} g(X)$

(b) If a sequence of RVs, $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$

Theorem 19 (Slutsky's Theorem)**.** Let X_n, Y_n be two sequences of RVs such that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$, where X is a RV and c is a constant. Then:

(a) $X_n + Y_n \xrightarrow{D} X + c$

(b) $X_n Y_n \xrightarrow{D} cX$

(c) $X_n / Y_n \xrightarrow{D} X/c \quad$ (if $c \neq 0$)

4.2 Central Limit Theorem

Theorem 20 (Central Limit Thoerem)**.** Let $X_1, \ldots, X_n$ be IID RVs with $E[X_1] = \mu$ and $\text{Var}[X_1] = \sigma^2 < \infty$. Let $\bar{X_n}$ be the sample average. Then:

$$\sqrt{n}\left(\frac{\bar{X_n} - \mu}{\sigma}\right) \xrightarrow{D} N(0,1)$$

4.3 Concentration inequality

Definition 29 (Concentration inequality)**.** In general, a concentration inequality describes the function that bounds the probability of the absolute difference between X_n and $E(X_n)$:

$$P(|X_n - E(X_n)| \geq \epsilon) \leq \phi_n(\epsilon)$$

where $\phi_n(\epsilon) \to 0$ is the concentration inequality. The *convergence rate* with respect to n is an important property describing how fast X_n converges to its mean. Note that Chebyshev's inequality gives a general concentration inequality that is polynomial with n, but with additional assumptions we can obtain better convergence rates.

Theorem 21 (Hoeffding's inequality)**.** Let $X_1, \ldots, X_n$ be IID RVs s.t. $a \leq X_1 \leq b$ and let $\bar{X}_n$ be the sample average. Then for any $\epsilon > 0$:

$$P(|\bar{X}_n - E(\bar{X}_n)| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

Consider the case of estimating a high dimensional proportion (i.e., proportion who replied Yes to d survey questions). The vector $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_d)$ is a good estimator, but can we estimate every proportion accurately?

$$\begin{aligned} P(||\hat{\pi} - \pi||_{\max} > \epsilon) &= P\left(\max_{\{j=1,\dots,d\}} |\hat{\pi}_j - \pi_j| > \epsilon\right) \qquad \text{definition of max norm} \\ &\leq \sum_{j=1}^{d} P(|\hat{\pi}_j - \pi_j| > \epsilon) \\ &= dP(|\hat{\pi}_j - \pi_j| > \epsilon) \leq 2de^{-2n\epsilon^2} \end{aligned}$$

Which converges in probability to π (i.e., is a consistent estimator of π) as long as $2de^{-2n\epsilon^2} \to 0$, which holds when $\frac{\log(d)}{n} \to 0$.

Definition 30 (Gaussian concentration)**.** Given $X_1, \dots, X_n \sim N(0, \sigma^2)$ and $\bar{X}_n$ be the sample mean, we know $\bar{X}_n \sim N(0, \sigma^2/n)$. Then:

$$\begin{aligned} P(\bar{X}_n > \epsilon) &= P(e^{t\bar{X}_n} > e^{t\epsilon}) \\ &\leq \frac{E(e^{t\bar{X}_n})}{e^{t\epsilon}} \quad \text{By Markov's Inequal} \\ &\leq e^{\frac{1}{2n}\sigma^2 t^2 - t\epsilon} \quad \text{By MGF Gaussian} \\ &\leq e^{-\frac{n\epsilon^2}{2\sigma^2}} \quad \text{By finding maximum of quad equation wrt } s \\ &\implies P(|\bar{X}_n| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}} \\ &\implies \boxed{P(|\bar{X}_n - E(X_1)| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}} \end{aligned}$$

Note the Gaussian concentration has convergence rate that is exponential wrt n.

Note: for other RVs whose MGFs satisfies $M_X(t) \leq e^{1/2t^2\sigma^2}$ for all $t > 0$, this concentration inequality holds. These types of random variables are called *sub-Gaussian*. Note that bounded RVs are sub-Gaussian!

Definition 31 (Concentration of a maximum)**.** Suppose $X_1, \dots, X_n \sim N(0, \sigma^2)$ and $Z_n = \max\{|X_1|, \dots, |X_n|\}$. We know from the Gaussian concentration (just replace $\bar{X}$ with X_i and set $n = 1$):

$$\begin{aligned} P(|X_i| > \epsilon) &\leq 2e^{-\frac{\epsilon^2}{2\sigma^2}} \\ &\implies P(Z_n) > \epsilon = P(\max\{|X_1|, \dots, |X_n|\}) \\ &\leq \sum_{i=1}^{n} P(|X_i| > \epsilon) \\ &\leq 2ne^{-\frac{\epsilon^2}{2\sigma^2}} \end{aligned}$$

As long as the concentration inequality $2ne^{-\frac{\epsilon^2}{2\sigma^2}} \to \delta$ for some $\delta \in (0, 1)$, we can bound how fast the maximum diverges. Turns out $\gamma_n = \sigma\sqrt{2logn}$ is teh choice of sequence such that Z_n/γ_n does not diverge in probability.

5 Conditional distribution and conditional expectation

5.1 Conditional distribution

Definition 32 (Conditional distribution)**.** The conditional distribution of $X|Y$, denoted $p_{Y|X}(y|x)$ can be written in the following four ways:

1. If X, Y continuous: $p_{Y|X}(y|x) = \frac{\frac{\partial^2 F(x,y)}{\partial x \partial y}}{\int_{-\infty}^{\infty} p_{XY}(x,y)dy} = \frac{p_{XY}(x,y)}{p_X(x)}$
2. If X, Y discrete: $p_{Y|X}(y|x) = \frac{P(X=x,Y=y)}{P(X=x)} = \frac{p_{XY}(x,y)}{p_X(x)}$
3. If X discrete, Y continuous: $p_{Y|X}(y|x) = \frac{d}{dy}P(Y \leq y|X = x) = \frac{\frac{d}{dy}P(Y \leq y)}{P(X=x)} = \frac{p_{XY}(x,y)}{p_X(x)}$
4. If X continuous, Y discrete: we choose C s.t. $\{(X,Y) \in C\} = \{Y = y\}$. $p_{Y|X}(y|x) = P(Y = y|X = x) = P((X,Y) \in C|X = x) = \frac{1}{p_X(x)}\frac{d}{dx}P((X,Y) \in C, X \leq x = \frac{1}{p_X(x)}\frac{d}{dx}P(Y = y, X \leq x) = \frac{p_{XY(x,y)}}{p_X(x)}$

Strategy 3 (Calculating conditional distributions)**.** A tried and true strategy for finding $p_{Y|X}(y|x)$ is to take the joint pdf over the marginal of X: $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$.

However, note that $P_{Y|X}(y|x)$ is a function of y, so you can consider the x-terms as constants and consider only components involving y.

For example, consider discrete X and continuous Y s.t. $p_{X,Y}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!}$. Then:

$$\begin{aligned}
p_{X|Y}(x|y) &\propto p_{X,Y}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} \\
&\propto \frac{y^x}{x!} \\
&= \text{Pois}(y) \quad \text{(By recognizing the kernel of Poisson pdf)} \\
p_{Y|X}(x|y) &\propto p_{X,Y}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} \\
&\propto y^x e^{-(\lambda+1)y} \\
&= \text{Gamma}(x+1, \lambda+1) \quad \text{(By recognizing the kernel of Gamma pdf)}
\end{aligned}$$

5.2 Conditional Expectation

Definition 33 (Conditional Expectation)**.** The conditional expectation of Y given X is:

$$E(Y|X = x) = \begin{cases} \int yp(y|x)dy, & \text{if } Y \text{ is cont} \\ \sum_y yp(y|x), & \text{if } X \text{ is cont} \end{cases}$$

Theorem 22 (Law of Total Expectation)**.** In the more specific form:

$$E[Y] = E[E[Y|X]]$$

In the more general version:

$$E[g(X,Y)] = E[E[g(X,Y)|X]]$$

Property 16 (Properties of conditional expectation)**.** Conditional expectation has the following few properties:

1. If $X \perp Y$, $E(X|Y=y) = E(X)$
2. Suppose $g(x,y) = q(x)h(y)$, then $E[q(X)h(Y)] = E[q(X)E[h(Y)|X]]$
3. $\text{Cov}(g(X), q(Y)) = \text{Cov}(g(X), E(q(Y)|X))$ (Substitute $w(x) = E(q(y)|x)$)

Theorem 23 (Law of Total Variance)**.**

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}(E[Y|X])$$

Definition 34 (Cool example: Inverse Probability weighting)**.** Suppose we're interested in estimating Y given X, however, we don't always observe Y. $R = \begin{cases} 1 & \text{when X,Y observed} \\ 0 & \text{when Y not observed} \end{cases}$. Assume $R \perp Y|X$, meaning given some value of X, missingness is independent of Y. Thus, $P(R=1|X,Y) = P(R=1|X) = \pi(X)$ is a function of X. The *inverse probability weighting quantity*:

$$W = \frac{RY}{\pi(X)}$$

has the same mean as Y. Thus:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\pi(X_i)}$$

Is the inverse probability weighting estimator which is an unbiased estimator. Note: you can think about the inverse probability weighted estimator as the mean uncensored income *weighted* by the probability someone responded given their age in our dataset.

6 Correlation, Prediction, Regression

6.1 Correlation

Definition 35 (Pearson's Correlation)**.** Correlation measures the *linear* relationship between two variables.

$$\rho_{XY} = \text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Property 17 (Properties of Pearson's Correlation)**.** Correlation has the following properties

1. $\text{Cor}(X,Y) = \text{Cor}(Y,X)$
2. Location-scale: $\text{Cor}(aX+b, cY+d) = \text{sign}(ac)\text{Cor}(X,Y)$
3. $-1 \leq \text{Cor}(X,Y) \leq 1$ and Cor $= \pm 1$ iff $X = aY + b$ for constants a, b.

6.2 Regression function & MSE prediction

Definition 36 (Mean-square error)**.** To measure how good a predictor $g(X)$ is of Y, we often use the *mean-square error (MSE)*:

$$\begin{aligned} R(g) &= E((Y - g(X))^2) \\ &= E[\text{Var}(Y|X)] + E[(E[Y|X] - g(X))^2] \end{aligned}$$

MSE is the expected squared deviations from the target Y.

Strategy 4 (MSE wrt constant)**.** Note that when the MSE of Y relative to a constant/fixed number/target:

$$E[(Y-c)^2] = \text{Var}[Y] + (E[Y] - c)^2$$

Definition 37 (Regression function)**.** The *regression function/best predictor* is the $g(X)$ that minimizes the MSE, where $g(X) = E[Y|X]$ (see Definition 36 above).

Then, Y can be considered as:

$$Y = \underbrace{E[Y|X]}_{\text{best predictor}} + \underbrace{(Y - E[Y|X])}_{\text{residuals}}$$

Property 18 (Properties of regression function)**.**

1. Unbiased: $E[\text{best predictor}] = E[E[Y|X]] = E[Y]$ and $E[\text{residuals}] = 0$
2. Uncorrelated: $\text{Cov}(E[Y|X], (Y - E[Y|X])) = 0$
3. Residual variance: $\text{Var}(Y - E[Y|X]) = E[\text{Var}(Y|X)]$
4. Variance decomposition: $\text{Var}(Y) = \underbrace{\text{Var}(E[Y|X])}_{\text{Var(best predictor)}} + \underbrace{E[\text{Var}(Y|X)]}_{\text{average Var(residuals)}}$

6.3 Linear regression

Definition 38 (Best linear predictor)**.** We often restrict our search for a best predictor to a simple class of functions, for example, linear function: $Y = \alpha_\beta X$. Then we choose α, β that minimize the MSE:

$$\begin{aligned}
\alpha^*, \beta^* &= \text{argmin}_{\alpha,\beta} E((Y - \alpha - \beta X)^2) = \text{argmin}_{\alpha,\beta} R(\alpha, \beta) \\
\beta^* &= \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \\
\alpha^* &= E[Y] - \beta^* E[X] \quad \text{(By solving the gradient equation } \frac{\partial}{\alpha}R = 0 \text{ and } \frac{\partial}{\beta}R = 0
\end{aligned}$$

Therefore, the best linear predictor, $m^*(x)$ is:

$$\begin{aligned}
m^*(x) &= \alpha^* + \beta^* x \\
&= E[Y] + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(x - E[X]) \\
&= \mu_y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mu_x)
\end{aligned}$$

Property 19 (MSE under best linear predictor)**.** The MSE under the best linear predictor is:

$$\begin{aligned}
R(\alpha^*, \beta^*) &= E((Y - \alpha^* - \beta^* x)^2) \\
&= E(((Y - \mu_y) + (-\rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mu_x)))^2) \\
&= \sigma_Y^2(1 - \rho_{XY}^2)
\end{aligned}$$

Definition 39 (Multivariate linear regression)**.** Consider the data vector $Z = (X_1, \ldots, X_p)$ and coefficient vector $\gamma = (\alpha, \beta_1, \ldots, \beta_p)$. Then MSE:

$$\begin{aligned}
R(\gamma) = R(\alpha, \beta) &= E((Y - \gamma^T Z)^2) = E((Y - \alpha - \beta_1 X_1 - \ldots \beta_p X_p)^2) \\
&= E[Y^T Y] - 2\gamma^T E[ZY] + \gamma^T E[ZZ^T]\gamma \\
\frac{\partial}{\partial\gamma}R(\gamma) = 0 &\implies 0 = -2E[ZY] + 2E[ZZ^T]\gamma \implies \gamma^* = E[ZZ^T]^{-1}E[ZY]
\end{aligned}$$

Property 20 (Correctness vs misspecified)**.**

1. When the linear model is misspecified (incorrect), the best linear predictor will change as the distribution of the covariate changes... i.e., the model will be sensitive to changes in covariates

2. When the linear model is correct (i.e., Y is a linear function of Z), $Y = \bar{\gamma} Z + \epsilon$ for some $\bar{\gamma} \in \mathbb{R}^{p+1}$ and $\epsilon \perp Z$ and $E[\epsilon|Z] = 0$, the least square coefficient is the same as the true coefficient: $\gamma^* = \bar{\gamma}$.

6.4 Binary classification

Definition 40 (Binary Classifier)**.** $c(x)$ is a classifier if $c : \underbrace{\mathcal{X}}_{\text{support of } X\text{, usually } \mathbb{R}} \to \underbrace{Y}_{\{0,1\}}$

Definition 41 (0-1 loss)**.** The 0-1 loss is a function used to measure the success/accuracy of a binary classifier.

$$L(c(X), Y) = \begin{cases} 0 & \text{if } c(X) = Y \\ 1 & \text{if } c(X) \neq Y \end{cases}$$

Definition 42 (Bayes classifier)**.** The Bayes classifier (c^*) is the classifier that minimizes:

$$\begin{aligned} c^*(x) &= \underset{c}{\operatorname{argmin}}\ R(c) = \underset{c}{\operatorname{argmin}}\ E[L(c(x), y)] \\ &= \underset{c}{\operatorname{argmin}}\ E[E[L(c(x), y)]|X] \\ &= \underset{c}{\operatorname{argmin}}\ L(c(x), Y = 1)P(Y = 1|X) + L(c(x), Y = 0)P(Y = 0|X) \\ &= \underset{c}{\operatorname{argmin}}\ \mathbb{I}(c(x) = 0)P(Y = 1|X) + \mathbb{I}(c(x) = 1)P(Y = 0|X) \\ c^*(x) &= \begin{cases} 0 & \text{if } P(Y = 1|X) < P(Y = 0|X) \\ 1 & \text{if } P(Y = 1|X) > P(Y = 0|X) \end{cases} \end{aligned}$$

The classifier is the optimal classifier. It is called Bayes b/c it uses conditional probabilities to make its choice.

7 Estimators

In the real world, we observe data, $X_1, \dots, X_n$ a random, IID sample from a population. Assuming they were generated from a parametric model, we can use $X_1, \dots, X_n$ to estimate/learn the parameters of the model.

Definition 43 (Estimator)**.** An estimator is a statistic $W(X_1, \dots, X_n)$ such that W can be used to estimate θ, the parameter for a parametric model.

7.1 Method of Moments estimator

Strategy 5 (Method of Moments estimation)**.** The moments of a parametric model are determined by the underlying parameter. Suppose we have k parameters in the model: $\theta \in \mathbb{R}^k$.

We can take all the theoretical moments:

$$m_j(\theta) = \int x^j p(x;\theta)dx$$

And equate them to the sample moments (which are very easily calculated):

$$m_1(\theta) = \frac{1}{n}\sum_{i=1}^n X_i$$
$$m_2(\theta) = \frac{1}{n}\sum_{i=1}^n X_i^2$$
$$\dots$$
$$m_k(\theta) = \frac{1}{n}\sum_{i=1}^n X_i^k$$

The resulting θ that solves the equations above is called the **method of moments estimator** and is termed $\hat{\theta}_{\text{MoM}}$.

7.2 Maximum Likelihood estimation

Definition 44 (Likelihood function, Log Likelihood, Score Function, & MLE)**.** If we consider the data, X, is fixed and we want to choose θ that is most likely to generate X, we define the **likelihood function**:

$$L(\theta|X_1, \dots, X_n) = p(X_1, \dots, X_n; \theta)$$

When IID, the likelihood function can be written as:

$$L(\theta|X_1, \dots, X_n) = \prod_{i=1}^n L(\theta|X_i) = \prod_{i=1}^n p(X_i|\theta)$$

The **log likelihood** is a useful quantity to work with in finding the MLE:

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^n \log p(X_i;\theta)$$

The **score function** is the gradient of the log-likelihood, which when set to 0, is satisfied by the MLE:

$$S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ell(\theta|X_i) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p(X_i;\theta)$$

The **maximum likelihood estimator** is defined as:

$$\hat{\theta}_{MLE} = \text{argmax}_\theta L(\theta|X)$$

Definition 45 (The Fisher Information)**.** The **Fisher information (matrix)** describes the curvature of the log-likelihood surface.

$$I_n(\theta) = -\mathbb{E}[\nabla_\theta\nabla_\theta\ell_n(\theta|X_1)] = nI_1(\theta) = n\cdot -\mathbb{E}[\nabla_\theta\nabla_\theta p(X_1;\theta)]$$

Near the MLE, low fisher information implies a "blunt/shallow" maximum, meaning that there are other nearby points with similar log-likelihoods. Low fisher information implies high variance of the estimator.

Definition 46 (Cramer-Rao Lower Bound)**.** The **Cramer-Rao lower bound** provides a lower bound on the variance of an estimator:

$$\text{Var}(\hat{\theta}(Y)) \geq \frac{\left(\frac{\partial}{\partial\hat{\theta}}\mathbb{E}[\hat{\theta}(Y)]\right)^2}{I(\theta)}$$

For an unbiased estimator, we can write the inequality as:

$$\text{Var}(\hat{\theta}(Y)) \geq \frac{1}{I(\theta)}$$

If an estimator achieves it's Cramer-Rao lower bound, it is known as *efficient* (lowest possible variance).

Property 21 (Properties of the MLE)**.**

1. **Asymptotic Efficiency**: the MLE is asymptotically efficient, i.e., as $n\to\infty$, $\text{Var}(\hat{\theta}) = I^{-1}(\theta)$.
2. **Consistency**: the MLE is always consistent, i.e., $\lim_{n\to\infty} P(|\hat{\theta}-\theta| < \epsilon) = 1 \implies \hat{\theta}\xrightarrow{P}\theta$.
3. **Asymptotically unbiased**: the MLE is always asymptotically unbiased: $\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_n] = \theta$
4. **Asymptotic Normality**: the MLE is always asymptotically normal: $\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{D}N(0, I^{-1}(\theta))$

Strategy 6 (Maximum Likelihood Estimation)**.** A common trick to find the MLE is to set the gradient of the log-likelihood (score function) equal to 0 and find the parameter estimates that generate maxima. Always remember to evaluate the second derivative to see if we have a maximum!

7.3 Bayesian estimation

Definition 47 (Prior, Posterior, Conjugate Prior)**.** Bayesian inference does not assume a probability model to be true, meaning that there is NO true parameter. Rather, the paradigm just considers probability models as useful mathematical tools for analyzing data.

Bayesian inference focuses on the distribution of θ after observing $X_1, \dots, X_n$, i.e., the **posterior distribution**:

$$\pi(\theta|X_1, \dots, X_n) = \frac{p(X_1, \dots, X_n, \theta)}{p(X_1, \dots, X_n)} \propto \underbrace{p(X_1, \dots, X_n|\theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}}$$

Thus, the posterior is proportional to the likelihood times the **prior distribution** which reflects our belief about the value of θ. A **conjugate prior** is a prior distribution that produces a posterior distribution from the same family as the prior.

Definition 48 (Posterior Mean and MAP estimates)**.** The **posterior mean** is the mean of the posterior distribution, a common estimator θ:

$$\hat{\theta}_\pi = \mathbb{E}(\theta|X_1, \dots, X_n) = \int \theta\pi(\theta|X_1, \dots, X_n)d\theta$$

The **Maximum a posteriori (MAP) estimate**, similarly to the MLE, chooses the value of θ that is most likely:

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ \pi(\theta|X_1, \dots, X_n)$$

Property 22 (Decomposition of posterior mean)**.** In some cases, the posterior mean can be shown to be a weighted average of the MLE and prior mean. For example, for $Y \sim \text{Bin}(N, \theta)$ ad $\theta \sim \text{Beta}(\alpha, \beta)$, we can calculate $\pi(\theta|Y) \sim \text{Beta}(Y + \alpha, N - Y + \beta)$. Then the posterior mean is:

$$\begin{aligned}\hat{\theta}_\pi &= \frac{Y}{N} \times \frac{N}{N + \alpha + \beta} + \frac{\alpha}{\alpha + \beta} \times \frac{\alpha + \beta}{N + \alpha + \beta} \\ &= \hat{\theta}_{MLE} \times W + \text{Prior mean} \times (1 - W)\end{aligned}$$

And as the sample size $N \to \infty$, i.e., as we get more data, the MLE dominates. When N is small, the prior mean contributes more.

7.4 Empirical Risk Minimization

Definition 49 (Loss function, Risk function, and Empirical risk minimizer)**.** A *loss function* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a function that measures the quality of prediction. Consider a prediction model $f_\beta(X) = X^T\beta$

We can define the *risk function* like so:

$$R(\beta) = \mathbb{E}[L(Y, f_\beta(X))]$$

The *empirical risk* is the estimated/computable version of the risk function:

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} L(Y, f_\beta(X))$$

The *empirical risk minimizer* is defined as the choice of β that minimizes the empirical risk:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ \hat{R}(\beta)$$

7.4.1 M-estimation

Definition 50 (M-estimation)**.** M-estimation finds an estimator by maximizing an empirical objective function:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \omega(\theta; X_i)$$

When we choose the log-likelihood as the objective function, the M-estimator is the MLE.

8 Multinomial distribution

8.1 Multinomial distribution, Properties, MLE

Definition 51 (Multinomial distribution)**.** The multinomial distribution is useful for characterizing categorical variables. If X has k categories with $(p_1, \dots, p_k)$ describing the category-wise probabilities and with the constraint that $\sum_j p_j = 1$, then we describe $X = (X_1, \dots, X_k) \sim M_k(n; p_1, \dots, p_k)$ where:

$$p(X = x) = p(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

And:

$$\textbf{MGF:} \qquad M_X(s) = \mathbb{E}\left[e^{s^T X}\right] = \mathbb{E}\left[e^{s^T Y_1}\right]^n = \left(\sum_{j=1}^k p_j e^{s_j}\right)^n$$

Property 23 (Properties of multinomial)**.** The multinomial has some nice properties:

(a) **Sum of independent multinomials:** If $X \sim M_k(n; p_1, \dots, p_k)$ and $V \sim M_K(m; p_1, \dots, p_k)$, then:

$$X + V \sim M_k(n + m; p_1, \dots, p_k)$$

(b) **Sum of IID draws:** If $X \sim M_k(n; p_1, \dots, p_k)$, then

$$X = \sum_{i=1}^n Y_i$$

Where $Y_i \overset{iid}{\sim} M_k(1; p_1, \dots, p_k)$. In other words, a multinomial RV with sample size n can be interpreted as n single draws from the multinomial dist with same params.

(c) **"Block" decomposition produces conditionally independent multinomial RVs**: Suppose we partition $X = (X_1, \dots, X_k)$ into r blocks.

$$\underbrace{(X_1, \dots, X_{k_1})}_{B_1}, \dots, \underbrace{(X_{k_{r-1}}, \dots, X_k)}_{B_r}$$

Then $B_1, \dots, B_r$ are conditionally independent given $S_1, \dots, S_r$ where S_i is the block-specific sum for B_i. Implying:

$$B_j | S_j \sim M_{k_j, \dots, k_{j-1}}\left(S_j; \frac{p_{k_{j-1}+1}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}, \dots, \frac{p_{k_j}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}\right)$$

(d) **Conditional distribution of $X_i | X_j$:** this is equivalent to the blocking case when $r = 2$:

$$(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_k) | X_j \sim M_{k-1}(n - X_j; \frac{p_1}{1 - p_j}, \dots, \frac{p_k}{1 - p_j})$$

So the marginal distribution is:

$$X_i | X_j \sim \text{Binomial}(n - X_j, \frac{p_i}{1 - p_j})$$

(e) **Negative correlation between entries**: Suppose $X = (X_1, \dots, X_K) \sim M_k(n; p_1, \dots, p_k)$. Then $\text{Cov}(X_i, X_j) < 0$ because if X_i is really large, X_j can't be really large b/c they must sum to n. More technically:

$$\begin{aligned}
\text{Cov}(X_i, X_j) &= \mathbb{E}[\underbrace{\text{Cov}(X_i, X_j|X_j)}_{=0}] + \text{Cov}(\mathbb{E}[X_i|X_j], \mathbb{E}[X_j|X_j]) \\
&= \text{Cov}(\mathbb{E}[X_i|X_j], X_j) \\
&= \text{Cov}\left((n - X_j)\frac{p_i}{1-p_j}, X_j\right) \\
&= -\frac{p_i}{1-p_j}\text{Var}(X_j) \\
&= -np_ip_j
\end{aligned}$$

Strategy 7 (Constrained optimization with Lagrange multipliers)**.** In the case of a multinomial distribution, we are unable to simply set the gradient of the log-likelihood equal to 0 and solve for $p_1, \dots, p_k$. We need to account for the constraint $\sum_{i=1}^k p_i = 1$.

Optimizing the log-likelihood over the constraint region is akin to having the log-likelihood tangentially intersect the constrained space. This implies that the gradients of the log-likelihood and constrained spaces are scalar multiples of each other:

$$\begin{aligned}
&\nabla \ell_n(p_1, \dots, p_k|X) = \lambda \nabla g(p_1, \dots, p_k) \\
&\implies 0 = \nabla \ell_n(p_1, \dots, p_k|X) - \lambda \nabla g(p_1, \dots, p_k)
\end{aligned}$$

In the multinomial case:

$$\begin{aligned}
&0 = \nabla \ell_n(p_1, \dots, p_k|X) - \lambda \nabla g(p_1, \dots, p_k) \\
&= \nabla \sum_{j=1}^k X_j \log p_j + \lambda \nabla \underbrace{\left(1 - \sum_{j=1}^k p_j\right)}_{\text{Constraint } = 0} \\
&\frac{\partial F}{\partial p_j} = \frac{X_j}{p_j} - \lambda = 0 \\
&\implies X_j = \hat{\lambda}\, \hat{p}_{MLE,j}
\end{aligned}$$

And since $n = \sum X_j = \hat{\lambda} \sum p_j = \hat{\lambda}$, $\boxed{\hat{p}_{MLE,j} = \frac{X_j}{n}}$

See here for more discussion on Lagrange multipliers.

8.2 Dirichlet Distribution & Connections to Multinomial

Definition 52 (Dirichlet distribution)**.** The **Dirichlet distribution** models random vectors with length k and non-negative elements that sum to 1. In other words, it generates a random probability vector. In other words, the dirichlet distribution can be considered a generalization of the beta distribution.

It is characterized by pdf:

$$p(X_1, \dots, x_k; \alpha_1, \dots, \alpha_k) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

And has important values:

$$\textbf{Mean:} \qquad \mathbb{E}(X_i) = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j}$$

$$\textbf{Mode:} \qquad \text{Mode}(X_i) = \frac{\alpha_i - 1}{\sum_{j=1}^{k} \alpha_j - k}$$

Property 24 (Bayesian inference with multinomial likelihood and Dirichlet prior)**.** Dirichlet distributions are often used as priors for multinomial likelihoods.

$$\begin{aligned} \pi(p|X) &\propto \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \times \frac{1}{\beta(\alpha)} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \\ &\propto p_1^{x_1 + \alpha_1 - 1} \dots p_k^{x_k + \alpha_k - 1} \\ &\sim \text{Dirch}(x_1 + \alpha_1, \dots, x_k + \alpha_k) \end{aligned}$$

Thus, the posterior mean is:

$$\hat{p}_{\pi,i} = \frac{x_i + \alpha_i}{\sum_{j=1}^{k} x_j + \alpha_j}$$

And the α_j (prior parameters) can be viewed as the pseudocount of the category j before collecting the data.

9 Linear, MVN Dist, Chi-Square Dist

9.1 Linear Algebra

Definition 53 (Basic definitions)**.**

(a) **Rank**: dimension of the columnspace of a matrix (the space R^n that is spanned by vectors that are the columns of your matrix)

(b) **Inverse/regular**: a matrix A is invertible if there exists a matrix A^{-1} s.t. $AA^{-1} = I_n$. Note TFAE:

  (i) A is invertible
  (ii) A is full rank
  (iii) $\det(A) \neq 0$

  Also note the following properties

  (i) If $n \times n$ matrices A and B are invertible, AB is invertible with inverse $(AB)^{-1} = B^{-1}A^{-1}$
  (ii) For a diagnonal matrix $D = \text{Diag}(d_1, \ldots, d_n)$, its inverse $D^{-1} = \text{Diag}(d_1^{-1}, \ldots, d_n^{-1})$

(c) **Transpose**: for an $m \times n$ matrix A, it's transpose, A^T is an $n \times m$ matrix s.t. $[A^T]_{ij} = A_{ji}$. Also note the following properties:

  (i) $(A + B)^T = A^T + B^T$
  (ii) $(AB)^T = B^T A^T$
  (iii) $(A^{-1})^T = (A^T)^{-1}$

(d) **Trace**: trace is the sum of diagonal entries. Also note the following properties:

  (i) $\text{Tr}(aA + bB) = a\text{Tr}(A) + b\text{Tr}(B)$
  (ii) $\text{Tr}(A) = \text{Tr}(A^T)$
  (iii) $\text{Tr}(AB) = \text{Tr}(BA)$
  (iv) Cylic properties: $\text{Tr}(ABC) = \text{Tr}(BCA) = \text{Tr}(CAB)$

(e) **Determinant**: For an $n \times n$ matrix, its determinant is:

$$\det(A) = \sum_{\pi} \epsilon(\pi) \prod_{i=1}^{n} A_{i\pi(i)}$$

  Where π is all possible permutations of $\{1, 2, \ldots, n\}$ and $\epsilon(\pi) = \pm 1$ depending if it is an even or odd permutation. Note the following properties:

  (i) $\det(AB) = \det(A) \cdot \det(B)$ when both square
  (ii) $\det(A)^{-1} = \det(A^{-1})$
  (iii) $\det(A^T) = \det(A)$
  (iv) $\det(A) = \prod_{i=1}^{n} A_{ii}$ if A is triangular

(f) **Orthogonal matrix**: An $n \times n$ matrix is orthogonal if $A^T A = I_n$. In other words, its column vectors from an orthonormal basis of $\mathbb{R}^n$ (i.e., they span $\mathbb{R}^n$, are mutually perpendicular, and all have norm $= 1$). Note that for an orthogonal matrix, $A^T = A^{-1}$.

(g) **Eigenvalues and eigenvectors**: eigenvalues are the roots of $\lambda_1, \ldots, \lambda_n$ to the following equation:

$$\det(A - \lambda I_n) = 0$$

  For each λ_j, there exists an eigenvector u_j s.t. $(A - \lambda_j I_n)u_j = 0$.

Definition 54 (Symmetric Matrices)**.** A matrix A is symmetric if $A = A^T$. The following are properties of symmetric matrices:

(i) **Real eigenvalues/eigenvectors**

(ii) **Orthogonal eigenvectors**: for $\lambda_j \neq \lambda_k$, $u_j \perp u_k$, i.e., u_j is orthogonal to u_k or $u_j^T u_k = 0$

(iii) **Spectral decomposition**: a symmetric matrix A can be factorized according to its eigenvalues and eigenvectors: let $\lambda_1, \dots, \lambda_n$ be the eigenvalues and $u_1, \dots, u_n$ be the eigenvectors. Let $\Lambda = \text{Diag}(\lambda_1, \dots, \lambda_n)$ and $U = [u_1, \dots, u_n]$. Then:

$$A = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$

(iv) **Trace**: $\text{Tr}(A) = \sum_{i=1}^n \lambda_i$

(v) **Determinant**: $\det(A) = \prod_{i=1}^n \lambda_i$

Definition 55 (Positive definite/Positive semi-definite matrices)**.** A symmetric matrix is positive semi-definite matrix if

$$x^T A x \geq 0$$

For all $x \in \mathbb{R}^n$. A matrix is positive definite matrix if

$$x^T A x > 0$$

For all $x \in \mathbb{R}^n$ and $x^T x > 0$. Here are some properties of PD and PSD matrices:

(i) All PSD/PD matrices are symmetric

(ii) I_n is PD

(iii) A diagonal matrix is PD/PSD if $D_{ii} > 0$ or $D_{ii} \geq 0$ for all i respectively.

(iv) If $S \in \mathbb{R}^{n \times n}$ is PSD and $A \in \mathbb{R}^{m \times n}$, then ASA^T is PSD.

(v) If $S \in \mathbb{R}^{n \times n}$ is PD and $A \in \mathbb{R}^{m \times n}$ has $\text{rank}(A) = m \leq n$, then ASA^T is PD.

(vi) AA^T is PSD

(vii) AA^T is PD if $\text{rank}(A) = m \leq n$

(viii) A is PD $\implies$ A is full rank $\implies$ A has an inverse $\implies$ A^{-1} is PD.

(ix) A symmetric matrix is PSD/PD if all its eigenvalues $\lambda \geq 0$ or $\lambda > 0$ respectively.

(x) The square root of a PD matrix is $C = U\sqrt{\Lambda}U^T$ where $U\lambda U^T$ is the spectral decomposition of A.

Property 25 (Block decompositions of PD matrices)**.** Suppose that $A \in \mathbb{R}^{n \times n}$ PD matrix and suppose we can decompose it into 4 submatrices:

$$A = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

Then the following properties hold:

(i) S_{11} and S_{22} are PD

(ii) $S_{11,2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$ is PD

(iii) $S_{22,1} = S_{22} - S_{21}S_{11}^{-1}S_{12}$ is PD

(iv) For any vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^n$, then

$$xA^{-1}x = (x_1 - S_{12}S_{22}^{-1}x_2)S_{11,2}^{-1}(x_1 - S_{12}S_{22}^{-1}x_2) + x_2S_{22}^{-1}x_2$$

Definition 56 (Projection Matrix)**.** A matrix P is a projection matrix if it is symmetric and idenpotent ($A^2 = A$). The following properties hold:

(i) A is a projection matrix iff there exists orthogonal matrix U s.t.

$$A = U \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} U^T$$

(ii) If we partition $U = [U_1, U_2]$ where $U_1 \in \mathbb{R}^{n\times m}$ and $U_2 \in \mathbb{R}^{n\times(n-m)}$, then $P = U_1U_1^T$, implying that P projects any vector in $\mathbb{R}^n$ into the column space of U_1 and is orthogonal to the columnspace of U_2.

(iii) $\text{rank}(P) = m$

(iv) $I_n - P$ is also a projection matrix that projects any vector in $\mathbb{R}^n$ to the space orthogonal to the columnspace of U_1.

9.2 Transformations of multiple RVs and the Jacobian Method

Strategy 8 (First principles)**.** Given $U = f(X,Y)$ and $V = g(X,Y)$, we can derive the joint CDF of U, V using first principles:

$$\begin{aligned} F_{U,V}(u,v) &= P(U \leq u, V \leq v) \\ &= P(f(X,Y) \leq u, g(X,Y) \leq v) \\ &= \int_{R(U,V)} p_{XY}(x,y)dxdy \end{aligned}$$

Where $R(U,V) = \{(x,y) : f(x,y) \leq ug(x,y) \leq v\}$

Strategy 9 (Jacobian Method)**.** Consider x and $y \in \mathbb{R}^n$ and a 1-1, onto mapping between them T. The Jacobian matrix is defined as:

$$J_T(x) = \left(\frac{\partial T(x)}{x}\right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_n}{\partial x_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \dots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

The Jacobian is the absolute value of the determinant of the Jacobian matrix: $|\det(J_T(x))|$

Suppose we know how (X, Y) jointly vary and we have $U = T_1(X, Y)$ and $V = T_2(X, Y)$ where T is 1-1 and onto. Then:

$$p_{U,V}(y) = p_{X,Y}(T^{-1}(x, y)) \left|\frac{\partial x}{\partial y}\right|$$

Where $|\frac{\partial x}{\partial y}|$ is the absolute value of the determinant of the Jacobian under the inverse mapping:

$$\begin{aligned} \left|\frac{\partial(x, y)}{\partial(u, w)}\right| &= \left|\det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial w} & \frac{\partial y}{\partial w} \end{pmatrix}\right| \\ = \left|\frac{\partial T^{-1}(u, w)}{\partial(u, w)}\right| &= \left|\det\begin{pmatrix} \frac{\partial T_1^{-1}(u,w)}{\partial u} & \frac{\partial T_2^{-1}(u,w)}{\partial u} \\ \frac{\partial T_1^{-1}(u,w)}{\partial w} & \frac{\partial T_2^{-1}(u,w)}{\partial w} \end{pmatrix}\right| \end{aligned}$$

9.3 Covariance matrix

Definition 57 (Covariance matrix)**.** The covariance matrix a vector of random variables $X = (X_1, \dots, X_n)$ is:

$$\begin{aligned} \mathrm{Cov}(X) &= \mathbb{E}[(x - \mathbb{E}(X))(X - \mathbb{E}(X))^T] \\ &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Var}(X_n) \end{pmatrix} \end{aligned}$$

Note the following properties of the covariance matrix:

(i) $\mathrm{Cov}(X) = \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X)^T$

(ii) For $A \in \mathbb{R}^{m \times n}$ (fixed) and $b \in \mathbb{R}^m$:

$$\mathrm{Cov}(AX + b) = A\mathrm{Cov}(X)A^T$$

(iii) For $a \in \mathbb{R}^n$, $\mathrm{Var}(a^T X) = a^T \mathrm{Cov}(X) a$

(iv) The covariance matrix is PSD

(v) The covariance matrix is PD if the only vector $a \in \mathbb{R}^n$ s.t. $\mathrm{Var}(a^T X) = 0$ is $a = 0$.

9.4 Multivariate Normal Distribution

Definition 58 (MVN)**.** Consider $Z = (Z_1, \dots, Z_n)$ where $Z_1, \dots Z_n \overset{iid}{\sim} N(0,1)$. Now consider $X = AZ + \mu$ where $A \in \mathbb{R}^{n \times n}$ is an invertible (fixed) square matrix and $\mu \in \mathbb{R}^n$ is a fixed vector. Then:

$$X \sim N(\mu, \Sigma)$$
$$p_X(x_1, \dots, x_n) = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Where $\Sigma = \text{Cov}(X) = AA^T$.

Note: the multivariate normal distribution $X \sim MVN(\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{n \times n}$ has the following MGF:

$$M_X(t) = e^{t^T \mu + \frac{1}{2} t^T \sigma t}$$

Property 26 (Properties of MVN)**.** The multivariate normal distribution has the following properties:

(i) **Linearity**: a linear transformation of a multivariate normal distribution is still normal:

$$Y = CX + b \sim N(C\mu + b, C\Sigma C^T)$$

(i) **Independence $\iff$ uncorrelation**:

$$X_i \perp X_j \iff \text{Cov}(X_i, X_j) = \Sigma_{ij} = 0$$

(i) **Marginal is normal**: Suppose $X = (X_1, X_2)$, where $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$. $\mu = (\mu_1, \mu_2)$, and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, then:

$$X_i \sim N(\mu_i, \Sigma_{ii})$$

(i) **Conditional is normal**:

$$X_1 | X_2 \sim N_{n_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11,2})$$

Where $\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

9.5 Chi-Square Dist

Definition 59 (Chi-Square Dist)**.** Let $X = (X_1, \dots, X_n)^T \sim MVN(0, I_n)$. Then the random variable:

$$W_n = X^T X = \sum_{i=1}^n X_i^2 = ||X||^2 \sim \chi_n^2$$

Note the following properties of a Chi-square RV:

$$\begin{aligned} \textbf{Mean}: &\quad \mathbb{E}(W_n) = n \\ \textbf{Variance}: &\quad \text{Var}(W_n) = 2n \end{aligned}$$

Property 27 (Normalizing Gaussian vector)**.** Suppose $Y \sim N(\mu, \Sigma)$:

$$\begin{aligned} Z = \Sigma^{-\frac{1}{2}}(Y-\mu) &\sim N(0, I_n) \\ Z^T Z = (Y-\mu)^T \Sigma^{-1}(Y-\mu) &\sim \chi^2_n \end{aligned}$$

Property 28 (Chi-Square connections to Gamma, Normal)**.** The Chi-square distribution is just a special kind of Gamma distribution:

$$\chi^2_p \overset{D}{\iff} \text{Gamma}(\alpha = \frac{p}{2}, \gamma = 2)$$

And if $X \sim N(0, \sigma^2)$ is a 0-centered, nonstandard normal:

$$\begin{aligned} &\frac{X^2}{\sigma^2} \sim \chi^2_1 \\ &X^2 = \sigma^2 \chi^2_1 \sim \text{Gamma}(\frac{1}{2}, 2\sigma^2) \end{aligned}$$

Suppose we have $X_1, \ldots, X_n \overset{iid}{\sim} N(0, \sigma^2)$:

$$\begin{aligned} &\sum_{i=1}^n \frac{X_i^2}{\sigma^2} \sim \chi^2_n \\ &\sum_{i=1}^n X_i^2 = \sigma^2 \chi^2_n \sim \text{Gamma}\left(\frac{n}{2}, 2\sigma^2\right) \end{aligned}$$

Property 29 (Projection property)**.** Let $X \sim N(\mu, I_n)$ be multivariate normal vector and P be a projection matrix with $\text{rank}(P) = m < n$. Then

$$(X-\mu)^T P (X-\mu) \sim \chi^2_m$$

Property 30 (iid normals)**.** Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ be the sample variance. The following results hold:

(i) $\bar{X}_n \perp S_n^2$

(ii) $\bar{X}_n \sim N(\mu, \sigma^2/n)$

(iii) $(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

10 Order statistics

Definition 60 (Order statistics)**.** Let $X_1, \dots, X_n \overset{iid}{\sim}$ a continuous distribution. The order statistics $Y_1 < \dots < Y_n$, also denoted $X_{(1)}, \dots X_{(n)}$, are the ordered version of the n random variables. Thus, the mapping:

$$(X_1, \dots, X_n) \to (Y_1, \dots, Y_n)$$

is n!-to-1, because any permutation of $(X_1, \dots, X_n)$ yields the same set of order statistics.

Property 31 (Marginal Dist and Joint Distribution)**.** The following are properties of order statistics:

1. **Distribution of** Y_j:

$$\begin{aligned} p_{Y_j}(y)dy &\approx P(y \le Y_j \le y + dy) \\ &\iff P((j-1)\ X_i\text{'s are below } y,\ (n-j)\ X_i \text{ are above } y + dy, \text{ one } X_i \text{ within } [y, y+dy]) \\ &\iff \binom{n}{j-1} F_X(y)^{j-1} \binom{n-j+1}{n-j} (1 - F_X(y+dy))^{n-j}\, P(y \le Y_j \le y + dy) \\ &= \binom{n}{j-1} F_X(y)^{j-1} \binom{n-j+1}{n-j} (1 - F_X(y))^{n-j}\, p_X(y)dy \end{aligned}$$

Dividing both sides by dy, we obtain:

$$\boxed{p_{Y_j}(y) = \frac{n!}{(j-1)!(n-j)!} F_X(y)^{j-1} (1 - F_X(y))^{n-j}\, p_X(y)}$$

2. **Distribution of** Y_j, Y_ℓ:

$$\begin{aligned} p_{Y_j, Y_\ell}(y, z)dydz &\approx P(y \le Y_j \le y + dy, z \le Y_\ell \le z + dz) \\ &\iff P(A) \end{aligned}$$

Where A is the event that:

(i) $(j-1)$ X_i's below y

(ii) One X_i between $[y, y + dy]$

(iii) $(\ell - j - 1)$ X_i's between $(y + dy, z)$

(iv) One X_i between $[z, z + dz]$

(v) Remaining $(n - \ell)$ X_i's above $z + dz$

Writing out $P(A)$ and dividing both sides by $dydz$, we obtain:

$$\boxed{p_{Y_j, Y_\ell}(y, z) = \frac{n!}{(j-1)!(\ell-j-1)!(n-\ell)!} F_x(y)^{j-1} p_X(y)(F_X(z) - F_X(y))^{(\ell-j-1)} p_X(z)(1 - F_X(z))^{(n-\ell)}}$$

3. **Distribution of** $Y_1, \dots, Y_n$: Applying the procedure outlined above:

$$\boxed{p_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = n! p_X(y_1) \dots p_X(y_n)}$$

Property 32 (Order statistics of Standard Uniform)**.** $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}[0,1]$. Then the pdf of Y_j is:

$$\begin{aligned} p_{Y_j}(y) &= \frac{n!}{(j-1)!(n-j)!} y^{j-1}(1-y)^{n-j} \\ \implies Y_j &\sim \text{Beta}(j, n-j+1) \end{aligned}$$

So the variance of Y_j:

$$\text{Var}(Y_j) = \frac{j(n-j+1)}{(n+1)^2(n+2)}$$

Which, when n is odd, is maximized at $j = \frac{n+1}{2}$, i.e., the median:

$$\text{Var}(Y_{\frac{n+1}{2}}) = \frac{1}{4(n+2)} = \mathcal{O}(n^{-1})$$

While the variance of the max and min are lowest:

$$\text{Var}(Y_1) = \text{Var}(Y_n) = \frac{n}{(n+1)^2(n+2)} = \mathcal{O}(n^{-2})$$

Definition 61 (Spacing between order statistics of standard uniform)**.** Let $W_1, \ldots, W_{n+1}$ be the spacing between consecutive order statistics:

$$\begin{aligned} W_1 &= Y_1 - 0 \\ W_2 &= Y_2 - Y_1 \\ &\vdots \\ W_n &= Y_n - Y_{n-1} \\ W_{n+1} &= 1 - Y_n \end{aligned}$$

Where $W_i \in [0,1]$, $\sum_{i=1}^{n+1} W_i = 1$, and we can define the original order statistics as the partial sum of the gaps: $Y_j = \sum_{i=1}^{j} W_i$.

We know the pdf of $Y_1, \ldots, Y_n$ is:

$$\begin{aligned} p_{Y_1,\ldots,Y_n} &= n! \\ \implies p_{W_1,\ldots,W_n} &= p_{Y_1,\ldots,Y_n}(w_1, \ldots, w_n) \left| \det\left(\frac{dY}{dW}\right) \right| \\ p_{W_1,\ldots,W_n} &= n! \end{aligned}$$

Because the Jacobian of the inverse mapping is upper triangular with 1's on the diagonals.

Exchangeable: Since the distribution of $W_1, \ldots, W_n$ is invariant under any permutation (they are *exchangeable*), the marginal distribution of W_i is the same as the marginal of W_j for all j. Since $W_1 = Y_1 \sim \text{Beta}(1, n)$, then

$$W_j \sim \text{Beta}(1, n)$$

Note the distribution of $W_i, W_j \overset{D}{\sim} W_1, W_2$ by exchangeability property. So:

$$\begin{aligned}
\text{Cov}(W_i, W_j) &= \text{Cov}(W_1, W_2) \\
&= \frac{1}{2}\left(\text{Var}(W_1 + W_2) - \text{Var}(W_1) - \text{Var}(W_2)\right) \\
&= \frac{1}{2}\left(\text{Var}(Y_2) - \text{Var}(Y_2)\right) \\
&= \frac{-1}{(n+1)^2(n+2)} < 0 \qquad \text{(By plugging in the variance formula from above)}
\end{aligned}$$

11 Functionals & Bootstrap

11.1 EDF

Definition 62 (Empirical Distribution Function)**.** The EDF is an estimator of the CDF. Recall the CDF of X: $F_X(x) = \mathbb{P}(X \leq x)$. Given a sample, $X_1, \ldots, X_n$, our estimator of the CDF is:

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

The *empirical distribution function* essentially places a point mass of $\frac{1}{n}$ on each observed X_i.

Property 33 (Unbiasedness, consistency, and asymptotic normality of EDF)**.** Let $Y_i = I(X_i \leq x) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x \end{cases}$. Then $Y_i \sim \text{Bernoulli}(F(x))$. This implies:

$$\mathbb{E}(I(X_i \leq x)) = \mathbb{E}(Y_i) = F(x)$$
$$\text{Var}(I(X_i \leq x)) = \text{Var}(Y_i) = F(x)(1 - F(x))$$

Now for $\hat{F}_n(x)$:

$$\begin{aligned}\mathbb{E}(\hat{F}_n(x)) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right) \\ &= \mathbb{E}(I(X_1 \leq x)) = F(x) \\ \text{Var}(\hat{F}_n(x)) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) = \frac{F(x)(1-F(x))}{n}\end{aligned}$$

Since $\mathbb{E}(\hat{F}_n(x)) = F(x)$ for any x, $\hat{F}_n(x)$ is an **unbiased** estimator of F. Since it is unbiased and $\text{Var}(\mathbb{E}(\hat{F}_n(x))) \to 0$ as $n \to \infty$, $\hat{F}_n(x)$ is also a **consistent** estimator of F. Also:

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{D} N(0, F(x)(1 - F(x)))$$

Meaning $\hat{F}_n(x)$ is **asymptotically normal** around $F(x)$.

11.2 Statistical Functionals & The Plug In Principle

Definition 63 (Functionals and Plug-In Principle)**.** A **statistical functional** is a function of a function. One can consider any parameter of interest, θ as a functional of the population CDF: $\theta(F)$.

Plug-in Principle: The power of statistical functionals is that we phrase parameters of interest in terms of the CDF. Since we have an unbiased and consistent estimate of the CDF (the EDF), we can obtain estimates of the parameters of interest by simply *"plugging-in"* the EDF for the CDF in the functional formula.

Property 34 (Some useful functionals and their plugin estimators)**.**

1. **Mean**: $\mu = T_{\text{Mean}}(F) = \int x dF(x)$ is the mean functional. The Plug-in estimate is as follows:

$$\begin{aligned}\hat{\mu} &= T_{\text{Mean}}(\hat{F}_n)\\ &= \int x d\hat{F}_n(x)\\ &= \sum_{X_1,\dots,X_n} x \underbrace{\hat{p}(x)}_{\frac{1}{n}}\\ &= \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n\end{aligned}$$

2. **Variance**: $\sigma^2 = T_{\text{Var}}(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$ is the variance functional. The Plug-in estimate is as follows:

$$\begin{aligned}\hat{\sigma_2} &= T_{\text{Var}}(\hat{F}_n)\\ &= \int x^2 d\hat{F}_n - \left(\int x d\hat{F}_n\right)^2\\ &= \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2\\ &= \frac{1}{n}\sum (X_i - \bar{X}_n)^2\end{aligned}$$

3. α**-quantile**: $\theta_\alpha = T_\alpha(F) = F^{-1}(\alpha)$ is the α-quantile functional. The Plug-in estimate is as follows:

$$\begin{aligned}\hat{\theta_\alpha} &= T_{\text{Cov}}(\hat{F}_n)\\ &= \hat{F}_n^{-1}(\alpha) \qquad \text{(The } \alpha\text{-sample quantile)}\end{aligned}$$

4. **Covariance**: Suppose $(X_1, Y_1), \dots, (X_n, Y_n) \overset{iid}{\sim} F(x, y)$. $\theta_{\text{Cov}}(F) = T(F) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \int xy dF(x, y) - \int x dF(x, y) \int y dF(x, y)$ is the covariance functional. The Plug-in estimate is as follows:

$$\begin{aligned}\theta_{\text{Cov}}(\hat{F}_n) &= T_{\text{Cov}}(\hat{F}_n)\\ &= \frac{1}{n}\sum X_i Y_i - \bar{X}_n \bar{Y}_n\\ &= \frac{1}{n}\sum (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)\end{aligned}$$

5. **Correlation**: $\theta_{\text{Cor}}(F) = T(F) = \frac{T_1(F)}{\sqrt{T_2(F)T_3(F)}}$, where $T_1(F)$ is the covariance functional and $T_2(F)$, $T_3(F)$ are the variance functionals with respect to X and Y respectively. We define the functional as:

$$T(F) = \frac{\int xy dF(x, y) - (\int x dF(x, y))(\int y dF(x, y))}{\sqrt{\left(\int x^2 dF(x) - \left(\int x dF(x, y)\right)^2\right)\left(\int y^2 dF(y) - \left(\int y dF(x, y)\right)^2\right)}}$$

The plug-in estimator is as follows:

$$T(\hat{F}_n) = \frac{\frac{1}{n}\sum (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\frac{1}{n}\sum (X_i - \bar{X}_n)^2 \frac{1}{n}\sum (Y_i - \bar{Y}_n)^2}}$$

6. **Linear regression**: Recall that the risk (prediction error) of a linear regression is $R(\alpha,\beta) = \mathbb{E}[(Y - \alpha - \beta X)^2]$ which is writeable in the functional form: $\theta_{\alpha,\beta} = \int (y-\alpha-\beta x)dF(x,y)$. The least squares functional is: $\theta_{LSE} = \underset{\alpha,\beta}{\text{argmin}} \int (y-\alpha-\beta x)dF(x,y)$. Our plug-in estimators are as follows:

$$\theta_{\alpha,\beta}(\hat{F}_n) = \frac{1}{n}\sum (Y_i - \alpha - \beta X_i)^2$$
$$\theta_{LSE}(\hat{F}_n) = \underset{\alpha,\beta}{\text{argmin}} \frac{1}{n}\sum (Y_i - \alpha - \beta X_i)^2$$

Which correspond to the *empirical risk* and *empirical risk minimizer* respectively.

7. **MLE**: The statistical functional of the MLE can be written as:

$$\theta^* = T_{MLE}(F) = \underset{\theta}{\text{argmax}} \int \log p(x;\theta) dF(x)$$

We obtain this expression:

$$\hat{\theta} = T_{MLE}(F) = \underset{\theta}{\text{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log p(x_i;\theta)$$

11.3 Delta Method

Definition 64 (Delta Method)**.** Assume we have a sequence of random variables $Y_1, \ldots, Y_n$ such that

$$\sqrt{n}(Y_n - y_0) \xrightarrow{D} N(0, \sigma_Y^2)$$

If a function f is differentiable at y_0, then using the first order Taylor expansion:

$$\begin{aligned}\sqrt{n}(f(Y_n) - f(y_0)) &= \sqrt{n}\left((Y_n - y_0)f'(y_0) + \text{higher order terms}\right)\\ &\approx \underbrace{f'(y_0)}_{\text{fixed}}\ \underbrace{\sqrt{n}(Y_n - y_0)}_{\text{Converges in dist}}\\ &\xrightarrow{D} N(0, \sigma_Y^2 |f'(y_0)|^2), \qquad \text{Var}(f(Y_n)) \approx \frac{1}{n}|f'(y_0)|^2 \sigma_Y^2\end{aligned}$$

11.4 Linear functionals, Influence Functions, and Nonlinear functionals

Definition 65 (Linear functional)**.** A linear functional is of the form:

$$T_\omega(F) = \int \omega(x) dF(x) \qquad (= \mathbb{E}(\omega(x)))$$

Where ω is a function.

Definition 66 (Influence function)**.** The influence function of a linear statistical functional is:

$$\begin{aligned} L_F(x) &= \omega(x) - T_\omega(F) \\ &= \omega(x) - \int \omega(x) dF(x) \end{aligned}$$

Property 35 (Properties of Influence function of linear functional)**.** Suppose T_ω is a linear functional with influence function $L_F(x)$. Then:

$$\sqrt{n}\left(T_\omega(\hat{F_n}) - T_\omega(F)\right) \xrightarrow{D} N\left(0, \int L_f^2(x) dF(x)\right)$$

Thus:

$$\begin{aligned} \mathbb{E}(L_F(x)) &= 0 \\ \text{Var}(T_\omega(F)) &= \int L_f^2(x) dF(x) \end{aligned}$$

And a consistent estimator for the $\text{Var}(T_\omega(F))$ is

$$\text{Var}(T_\omega(\hat{F_n})) = \frac{1}{n} \sum L_F^2(X_i)$$

Property 36 (Influence function of non-linear functional)**.** Consider the median functional (which is non-linear): $T_{\text{med}}(F) = F^{-1}(0.5)$. To analyze the properties of this functional, we need a different notion of influence function. The *influence function of a general statistical functional* T_{target} is:

$$L_F(x) = \lim_{\epsilon \to 0} \frac{T_{\text{target}}((1-\epsilon)F + \epsilon\delta_x) - T_{\text{target}}(F)}{\epsilon}$$

Where δ_x denotes a point mass at x. This expression is a special kind of derivative that perturbs the CDF by adding a point mass at x.

11.5 Bootstrap

Definition 67 (The Bootstrap algorithm)**.** Bootstrapping is a nonparametric method for assessing the uncertainty of an estimate. The process is quite simple. Given data $X_1, \ldots, X_n$ and a statistic of interest M_n from the data:

1. Generate a **bootstrap sample**, $X_1^*, \ldots, X_n^*$, by sampling *with replacement* from the n data points.
2. Calculate the bootstrap sample statistic, $M_n^{*(1)}$ from the bootstrap sample.
3. Repeat steps (1), and (2) B times, yielding $M_n^{*(1)}, \ldots, M_n^{*(B)}$

Once we have our distribution of bootstrap statistics, $M_n^{*(1)}, \ldots, M_n^{*(B)}$:

1. **Bootstrap estimate of variance**: $\hat{\text{V}}\text{ar}_B(M_n)$ is an estimate of $\text{Var}(M_n)$ where

$$\widehat{\text{Var}}_B(M_n) = \frac{1}{B-1}\sum_{i=1}^{B}\left(M_n^{*(i)} - \bar{M}_B^*\right)^2$$

2. **Bootstrap estimate of the MSE**:

$$\widehat{\text{MSE}}(M_n) = \frac{1}{B}\sum_{i=1}^{B}\left(M_n^{*(i)} - M_n\right)^2$$

3. **Bootstrap CI**: we can construct a $1-\alpha$ CI:

$$M_n \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}_B(M_n)}$$

Property 37 (When and why the Bootstrap works)**.** Consider the CDF of the median, M_n, which is determined by the data distribution, F, and the sample size, n:

$$F_{M_n}(x) = \Psi(x; F, n)$$

When we sample with replacement from our data, $X_1, \ldots, X_n$, each element has a $\frac{1}{n}$ chance of being selected at each draw. Thus, we are sampling from the EDF, $\hat{F}_n$! Thus, the CDF of the bootstrap sample median is:

$$F_{M_n^{*(i)}}(x) = \Psi(x; \hat{F}_n, n)$$

Since $\hat{F}_n$ is a good approximation of F under sufficiently large samples, then as long as Ψ is smooth:

$$\hat{F}_n \approx F \implies F_{M_n^{*(i)}}(x) \approx F_{M_n}(x)$$

Meaning the CDF of the bootstrap median approximates the CDF of the true median. And then the variance of the bootstrap median approximates the variance of the median itself.

12 General Strategies

12.1 Finding distribution of random variables

1. Use first principles and find the CDF: $P(X \leq x...$
2. Perform a transformation from a known random variable (single variable or multivariable w/ Jacobian method)
3. Use MGFs. This is especially useful when X, Y, Z are independent and we want the distribution of $X + Y + Z$
4. Use definition of known distributions (e.g., counting blue balls drawn with replacement from bag filled with blue/green balls – binomial dist)
5. Are these ordered statistics?
6. Use properties of known distributions (e.g., sum of independent Poisson RVs is Poisson / Sum of independent exponentials is gamma).

12.2 Showing independence

1. Show that the joint density is the product of the marginals
2. Show that $P(X|Y = y)$ does not depend on y (be careful with support)
3. Use properties of normal distribution (e.g., for X MVN, independence of $X_i, X_j \iff X_i, X_j$ are uncorrelated, or $S_n^2 \perp \bar{X}_n$)

12.3 Computing expectations, variances, and covariances

1. Cite $\mathbb{E}(X), \text{Var}(X)$ of known distributions
2. Use definition
3. Use iterated law of total expectation, variance, or covariance tricks

12.4 Convergence of RVs

1. Employ Hoeffding, Markov, Chebyshev inequalities as necessary. Chebyshev is very useful for convergence in probability when Variance is shrinking with n.
2. Strong/Weak LLN: use when you have sample means and WTS convergence in probability.
3. CLT: use when we want to show convergence in distribution when format looks like $\sqrt{n}(\bar{X}_n - \cdot) \xrightarrow{D} N(\cdot, \cdot)$
4. CLT followed by delta method
5. Slutsky's theorem (for convergence in distribution) or Continuous mapping theorem (for either convergence in distribution or probability)
6. MGFs: convergence of MGFs $\iff$ convergence in distribution. If you only have access to the conditional MGF, use law of total expectation!

12.5 Finding MLE

1. Find the log-likelihood and maximize it
2. If log-likelihood is not differentiable, are there other ways to maximize this? Consider the likelihood maximization more broadly.
3. Use invariance property of MLE

12.6 Different definitions of e

1. Limit definition: $\lim\limits_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$
2. Power series: $\sum_{n\geq 0}\frac{x}{n!} = e^x$

12.7 Taylor expansion

1. The Taylor expansion about an infinitely differentiable function at value a is:

$$f(a) + \frac{f^{'(a)}}{1!}(x-a) + \frac{f^{''}(a)}{2!}(x-a)^2 + \ldots = \sum_{n=0}^{\infty}\frac{f^{(n)}(a)}{n!}(x-a)^b$$

12.8 Useful identities

1. Markov Inequality: $P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$.
2. Chebyshev Inequality: $P(|X - E(X)| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}$
3. Jensen's inequlaity: $\phi(E(X)) \leq E(\phi(X))$ where ϕ is convex function
4. Geometric series:
   - Finite: for $r \neq 1$, $\sum_{i=0}^{n} ar^i = a(\frac{1-r^{n+1}}{1-r}\Big)$
   - Infinite: if $r < 1$, $\sum_{i=0}^{\infty} ar^i = \frac{a}{1-r}$
5. Exponential sums
   - Finite: $\sum_{n=0}^{N-1} p^i e^{inx} = \frac{1-e^{iNx}}{1-e^{ix}}$
   - Infinite: $\sum_{n=0}^{\infty} p^n e^{inx} = \frac{1}{pe^{ix}-1}$
6. Calculus tricks:
   - Quotient rule: if $f(x) = g(x)/h(x)$, $f'(x) = \frac{g'(x)h(x)-g(x)h'(x)}{h(x)^2}$
   - Integration by parts: $\int udv = uv - \int vdu$
   - Integrating over two random variables: bounds of interior integral can be in terms of other RV, exterior bounds must be numeric.