Contents

Chapter 1

Review of Probability

Beginning of Jan.10, 2022

Some preliminaries first:

- Throughout this course, we will use Ω to denote the **universal set**.
- A **probability law** on ω is a function $\mathbb{P}:\Omega \to [0,1]$ satisfying the following axioms:
  (1) (Nonnegativity) $\mathbb{P}(A) \geqslant 0$ for all $A \subset X$[1].
  (2) (Countable additivity) For $\{A_i\}_{i\geqslant 1}$ with $A_i \cap A_j = \varnothing$ whenever $i \neq j$, $\mathbb{P}(\bigcup_{i\geqslant 1} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.
  (3) (Normalization) $\mathbb{P}(\Omega) = 1$.
- The following are direct consequences of the definition of a probability law:
  (1) If $A \subset B$ then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.
  (2) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
  (3) (Union bound) $\mathbb{P}(A \cup B) \leqslant \mathbb{P}(A) + \mathbb{P}(B)$ and more generally $\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) \leqslant \sum_{k=1}^{\infty} \mathbb{P}(A_k)$.
- Random variable definitions:
  (1) A **random variable** is a function $X:\Omega \to \mathbb{R}$ (or some different codomains). A **random vector** X is a function $X:\Omega \to \mathbb{R}^n$.
  (2) A **discrete random variable** is a random variable with finite or countable range.
  (3) A **probability density function** (PDF) is a function $f:\mathbb{R} \to [0,\infty)$ such that

  $$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \qquad \text{and } \int_a^b f(x)\,\mathrm{d}x \text{ exists for all } -\infty \leqslant a \leqslant b \leqslant \infty.$$

  (4) A random variable X is **continuous** if there exists a PDF f with

  $$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f(x)\,\mathrm{d}x \qquad \text{for all } -\infty \leqslant a \leqslant b \leqslant \infty.$$

  If so we say f is the PDF of X.

[1] For technical reasons we avoid measure theories and assume all $A \subset X$ are measurable.

(5) Let X be a random variable. We define the **cumulative distribution function** (CDF) to be $F : \mathbb{R} \to [0,1]$ by

$$F(x) := \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(t)\, \mathrm{d}t.$$

- Examples of some distributions:

(1) Bernoulli: let $0 < p < 1$ and define $\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p$ and $\mathbb{P} \equiv 0$ otherwise. "Flip one coin. Count the number of heads."

(2) Binomial: let $n \in \mathbb{N}$ and $0 < p < 1$. For $k \in \{0, \ldots, n\}$, define $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ and define $\mathbb{P} \equiv 0$ otherwise. Can be thought of the sum of n independent Bernoulli with parameter p. "Flip n coins. Count the number of heads."

(3) Geometric: let $0 < p < 1$ and define $\mathbb{P}(X = k) = (1-p)^{k-1}p$ for $k \in \mathbb{N}$ and 0 otherwise. "Flip a coin until heads shows up. Count the number of flips."

(4) Normal / Gaussian with mean μ and variance σ^2: the PDF is given by

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

(5) Poisson with parameter $\lambda > 0$:

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{for } k \in \mathbb{N}.$$

"Limit of binomial random variables subject to $\lim p_n = 0$ and $\lim np_n = \lambda$."

Definition: (1.17) Independent Sets

Let $\{A_i\}_{i \in I} \subset \Omega$ equipped with probability law Ω. We say $\{A_i\}$ are **independent** if, for all $S \subset I$ we have

$$\mathbb{P}(\bigcap_{i \in S} A_i) = \prod_{i \in S} \mathbb{P}(A_i).$$

Remark. This is *stronger* than pairwise independence, which only says $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for $i \neq j$. An example can be found here.

Beginning of Jan.12, 2022

Expected Value and Variance

Notation: given $A \subset \Omega$, we define the **indicator function** $1_A : \Omega \to \{0,1\}$ by

$$1_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Definition 1.0.1: (1.37) Expected Values

Let $\mathbb{P}$ be a probability law on Ω and let $X : \Omega \to [0, \infty]$. Define the **expected value** of X denoted $\mathbb{E}X$ to be

$$\mathbb{E}X := \int_0^\infty \mathbb{P}(X > t)\, \mathrm{d}t.$$

A simple application of Tonelli shows that if X is continuous then $\mathbb{E}X$ agrees with $\int_{-\infty}^{\infty} x f_X(x)\, \mathrm{d}x$ which we are more familiar with. If X is discrete, the analogous version is $\mathbb{E}X = \sum_{k\in\mathbb{R}} k\mathbb{P}(X = k)$.

In particular, if $X : \mathbb{R} \to \mathbb{R}$ and if $\mathbb{E}|X| < \infty$, then we can define

$$\mathbb{E}X := \mathbb{E}X^+ - \mathbb{E}X^-$$

where

$$X^+ := \max\{X, 0\} \qquad \text{and } X^- := \max\{-X, 0\}.$$

Remark. If $X : \Omega \to [0, \infty)$, then for positive integer n,

$$\mathbb{E}X^n = \int_0^\infty nt^{n-1}\mathbb{P}(X > t)\, \mathrm{d}t.$$

More generally, if $g : [0, \infty) \to [0, \infty)$ continuous differentiable with $g(0) = 0$, then

$$\mathbb{E}g(X) = \int_0^\infty g'(t)\mathbb{P}(X > t)\, \mathrm{d}t.$$

Proposition: (1.43) Linearity of $\mathbb{E}$

Let $X_1, \ldots, X_n$ be random variables. Then $\mathbb{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{E}X_i$.

Definition: (1.44) Variance

If $\mathbb{E}|X| < \infty$, define $\mathrm{var}(X) := \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$ to be the **variance** of X.

Remark. If $X : \Omega \to \mathbb{C}$ is complex valued, then if $\mathbb{E}|X| < \infty$, we can define

$$\mathbb{E}X := \mathbb{E}\mathfrak{Re}(X) + i\mathbb{E}\mathfrak{Im}(X)$$

and $\mathrm{var}(X) := \mathbb{E}(X - \mathbb{E}X)^2$ as before.

Joint Distributions

Definition: (1.47) Joint PDF

A **joint PDF** for two random variables is a function $f : \mathbb{R}^2 \to [0, \infty)$ with

$$\iint_{\mathbb{R}^2} f(x, y)\, \mathrm{d}x\mathrm{d}y = 1$$

and such that

$$\int_c^d \int_a^b f_{X,Y}(X, Y)\, \mathrm{d}x\mathrm{d}y$$

exists for all $[a, b] \times [c, d] \in \overline{\mathbb{R}}^2$.
We say X, Y are **jointly continuous** with joint PDF $f_{X,Y}$ if

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\, \mathrm{d}x\mathrm{d}y \qquad \text{for “all” } A \subset \mathbb{R}^2.$$

Definition: (1.48) Marginals

We define the **marginal PDF** f_X of X to be

$$f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y \qquad \text{for all } x \in \mathbb{R}.$$

Similarly, if $g : \mathbb{R}^2 \to \mathbb{R}$, we define

$$\mathbb{E}g(X, Y) := \iint_{\mathbb{R}^2} g(x, y) f_{X,Y}(x, y)\, \mathrm{d}x\mathrm{d}y.$$

Definition: (1.55) Independence of RVs

Let $X_1, \ldots, X_n$ be r random variables on Ω. We say they are **independent** if

$$\mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) = \prod_{i=1}^n \mathbb{P}(X_i \leqslant x_i) \qquad \text{for all } (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

In particular if $X_1, \ldots, X_n$ are continuous, then the definition is equivalent to saying

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) \qquad \text{for all } (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Proposition: (1.59, 1.60)

If $X_1, \ldots, X_n$ are independent and $\mathbb{E}X_i < \infty$, then

$$\mathrm{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{var}(X_i),$$

and

$$\mathbb{E}(\prod_{i=1}^n X_i) = \prod_{i=1}^n \mathbb{E}(X_i).$$

Conditional Probability

Let $A, B \subset \Omega$ with $\mathbb{P}(B) > 0$. We define

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

and read the **probability of A given B**.

For a fixed B, we define

$$\mathbb{E}(X \mid B) := \frac{\mathbb{E}X \cdot 1_B}{\mathbb{P}(B)}.$$

Proposition: Laws of Total Probability & Expectation

If $A \subset \Omega$ and $\{B_i\}$ partitions Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$$

and

$$\mathbb{E}X = \sum_{i=1}^{\infty} \mathbb{E}(X 1_{B_i}) = \sum_{i=1}^{\infty} \mathbb{E}(X \mid B_i)\mathbb{P}(B_i).$$

Definition: (1.75) Conditioning a RV

Let X, Y be continuous random variables with joint PDF $f_{X,Y}$. Fix $y \in \mathbb{R}$ with $f_Y(y) > 0$. Then for any $x \in \mathbb{R}$ we define the **conditional PDF** of X given $Y = y$ by

$$f_{X|Y}(x \mid y) := \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

The **conditional expectation** is given by

$$\mathbb{E}(X \mid Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, \mathrm{d}x.$$

Beginning of Jan.14, 2021

Theorem: (1.78) Total Expectation Theorem, Continuous

Let X, Y be continuous random variables and assume $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ be continuous. Then

$$\mathbb{E}X = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y)\, \mathrm{d}y.$$

Some Useful Inequalities

Theorem: (1.91) Jensen's Inequality

Let $\varphi : \mathbb{R} \to \mathbb{R}$. We say φ is **convex** if for all $x, y \in \mathbb{R}$ and $\lambda \in (0,1)$ we have

$$\varphi(\lambda x + (1-\lambda)y) \leqslant \lambda\varphi(x) + (1-\lambda)\varphi(y).$$

We say φ is **strictly convex** if the above inequality can be replaced by $<$.
Jensen's inequality states that if $\mathbb{E}|X| < \infty$ and $\mathbb{E}|\varphi(X)| < \infty$, and if φ is convex, then

$$\varphi(\mathbb{E}X) \leqslant \mathbb{E}\varphi(X).$$

Theorem: (1.92) Markov's Inequality

For all $t > 0$, we have

$$\mathbb{P}(|X| > t) \leqslant \frac{\mathbb{E}|X|}{t}.$$

Moreover, if $n \geqslant 1$ is a positive integer, then

$$\mathbb{P}(|X| \geqslant t) \leqslant \frac{\mathbb{E}|X|^n}{t^n}.$$

Theorem: (1.97) Chebyshev's Inequality

Using $n = 2$ in Markov's inequality applied to the random variable $X - \mathbb{E}X$, we have

$$\mathbb{P}(|X - \mathbb{E}X| \geqslant t) \leqslant \frac{\text{var}(X)}{t^2}$$

or equivalently

$$\mathbb{P}(|X - \mu| \geqslant t\sigma) \leqslant \frac{1}{t^2}.$$

Proposition: (1.107) Sum & Convolution

Let X, Y be continuous, independent random variables. Then

$$f_{X+Y}(t) = (f_X * f_Y)(t)$$

where $*$ denotes the convolution:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(s) f_Y(t-s)\, \mathrm{d}s.$$

Proof. We use independence and the fact that PDFs are derivatives of CDFs:

$$\mathbb{P}(X+Y \leqslant t) = \int_{\{x+y \leqslant t\}} f_{X,Y}(x,y)\, \mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{t-x} f_X(x) f_Y(y)\, \mathrm{d}y\, \mathrm{d}x = \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{t-x} f_Y(y)\, \mathrm{d}y\, \mathrm{d}x,$$

so

$$\begin{aligned}
f_{X+Y}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(X+Y \leqslant t) \\
&= \frac{\mathrm{d}t}{\mathrm{d}x}\int_{-\infty}^{\infty} f_X(x)\int_{-\infty}^{t-x} f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} f_X(x)\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{t-x} f_Y(y)\,\mathrm{d}y\,\mathrm{d}x = \int_{-\infty}^{\infty} f_X(x)f_Y(t-x)\,\mathrm{d}x.
\end{aligned}$$

Of course, we have assumed once again that it is well-defined to differentiate w.r.t the integral. □

Chapter 2

Modes of Convergence & the Limit Theorems

2.1 Modes of Convergence

Definition: (2.1) Almost Sure (a.s.) Convergence

We say $\{Y_n\}$ converges to Y **almost surely** if

$$\mathbb{P}(\lim_{n\to\infty} Y_n = Y) = 1$$

or equivalently

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1.$$

Definition: (2.2) Convergence in Probability

We say $\{Y_n\}$ converges to Y **in probability** if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| > \epsilon) = 0,$$

or equivalently

$$\lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega : |Y_n(\omega) - Y(\omega)| > \epsilon\}) = 0.$$

Definition: (2.3) Convergence in Distribution

We say $\{Y_n\}$ converges to Y **in distribution** in distribution if

$$\lim_{n\to\infty} \mathbb{P}(Y_n \leqslant t) = \mathbb{P}(Y \leqslant t)$$

for all $t \in \mathbb{R}$ such that $s \mapsto \mathbb{P}(Y \leqslant s)$ is continuous at $s = t$.

Remark. Since a Gaussian has continuous PDF, the CLT, to be stated right below, is indeed a statement about convergence in distribution.

Definition: (2.4) Convergence in L^p

Let $0 < p \leqslant \infty$. We say that $\{Y_n\}$ converges to Y **in L^p** if $\|Y\|_p < \infty$ and

$$\lim_{n\to\infty} \|Y_n - Y\|_p = 0,$$

where

$$\|Y\|_p := \begin{cases} (\mathbb{E}|Y|^p)^{1/p} & \text{if } 0 < p < \infty \\ \operatorname{ess\,sup}|X| = \inf\{c > 0 : \mathbb{P}(|X| \leqslant c\} = 1) & \text{if } p = \infty. \end{cases}$$

Remark.

$$\text{Convergence in distribution} \Longleftarrow \text{Convergence in probability} \Longleftarrow \begin{cases} \text{a.s. convergence} \\ \text{convergence in } L^p \end{cases}$$

The converses are all false.

2.2 The Limit Theorems

Theorem: (2.10) Weak Law of Large numbers, Weak LLN

Let $X_1, \ldots, X_n$ be i.i.d. (independent identically distributed) and assume that $\mu := \mathbb{E}X_1 < \infty$. Then X_n converges to $\mathbb{E}X_1$ in probability, i.e., for $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| > \epsilon\right) = 0.$$

Theorem: (2.11) Strong Law of Large Numbers, Strong LLN

Let $X_1, \ldots, X_n$ be i.i.d. with $\mu := \mathbb{E}X_1 < \infty$. Then $X_n \to \mu$ almost surely, i.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = \mu\right) = 1.$$

Beginning of Jan.19, 2021

Theorem: (2.13) Central Limit Theorem, CLT

Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}|X_1| < \infty$ and $0 < \operatorname{var}(X_1) < \infty$. Then for any $t \in \overline{\mathbb{R}}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant t\right) = \mathbb{P}(Z \leqslant t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-s^2/2}\, \mathrm{d}s,$$

where $\mu := \mathbb{E}X_1$ and $\sigma^2 := \mathrm{var}(x_1)$. In particular, each quotient $(X_1 + \ldots + X_n - n\mu)/(\sigma\sqrt{n})$ does have mean 0 and variance 1.

Theorem: (2.30) Berry-Esseén Theorem for CLT

Assume in addition that $\mathbb{E}|X_1|^3 < \infty$. Then

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant t \right) - \mathbb{P}(Z \leqslant t) \right| \leqslant \frac{\mathbb{E}|X_1|^3}{\sigma^3\sqrt{n}},$$

so in particular if $\mathbb{E}X_1 = 0$ and $\mathrm{var}(X_1) = 1$, we have

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}} \leqslant t \right) - \mathbb{P}(Z \leqslant t) \right| \leqslant \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}.$$

Chapter 3

Exponential Families

3.1 Exponential Families

A general question in statistics is to *fit a parameter to some given data*, for example, to find the unknown mean of a Gaussian sample.
An exponential family is some family of PDF or PMFs that depends on a parameter $w \in \mathbb{R}^k$ for some $k \geqslant 1$. More formally,

Definition: (3.1) Exponential Families

Let n, k be positive integers and let μ be a measure on $\mathbb{R}^n$. Let $t_1, \dots, t_k : \mathbb{R}^n \to \mathbb{R}$, and let $h : \mathbb{R}^n \to [0, \infty]$ not identically zero. For any $w = (w_1, \dots, w_k) \in \mathbb{R}^k$, define

$$a(w) := \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x)\Big) \,\mathrm{d}\mu(x).$$

The set $\{w \in \mathbb{R}^k : a(w) < \infty\}$ is called the **natural parameter space**. On this set, the functions

$$f_w(x) := h(x) \exp\left(\sum_{i=1}^{k} w_i t_i(x) - a(w)\right) \qquad \text{for all } x \in \mathbb{R}^n$$

satisfy

$$\begin{aligned}\int_{\mathbb{R}^n} f_w(x)\,\mathrm{d}x &= \int_{\mathbb{R}^n} h(x) \frac{\exp\big(\sum_{i=1}^{k} w_i t_i(x)\big)}{\int_{\mathbb{R}^n} h(x) \exp(\sum_{i=1}^{k} w_i t_i(x))\,\mathrm{d}\mu(x)} \,\mathrm{d}\mu(x)\\ &= \frac{\int_{\mathbb{R}^n} h(x)\exp()\,\mathrm{d}\mu(x)}{\int_{\mathbb{R}^n} h(x)\exp()\,\mathrm{d}\mu(x)} = 1.\end{aligned}$$

Informally, the f_w's can be interpreted as probability density functions with respect to the measure μ. Then, the set of functions $\{f_w : a(w) < \infty\}$ is called a ***k*-parameter exponential family in canonical form**. (*We interpret f_w as a PDF or PMF according to μ the measure.*)
More generally, let $\Theta \subset \mathbb{R}^k$ and let $w : \Theta \to \mathbb{R}^k$. We define a ***k*-parameter exponential family** to be the set of functions $\{f_\theta : \theta \in \Theta, a(w(\theta)) < \infty\}$ where

$$f_\theta(x) := h(x) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x) - a(w(\theta))\right) \qquad \text{for all } x \in \mathbb{R}^n.$$

Example: (3.3) Writing Gaussians as an Exponential Family. Consider Gaussians with mean $\mu < \infty$ and standard deviation $\sigma > 0$. Then the PDF is given by

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right). \tag{1}$$

If we write $\theta = (\theta_1, \theta_2) := (\mu, \sigma^2) \in \mathbb{R}^2$ and define

$$\begin{aligned} t_1(x) &:= x, \qquad t_2(x) := x^2, \\ w_1(\theta) &:= \frac{\theta_1}{\theta_2} = \frac{\mu}{\sigma^2}, \qquad w_2(\theta) := -\frac{1}{2\theta_2} = -\frac{1}{2\sigma^2}, \\ a(w(\theta)) &:= \frac{\theta_1^2}{2\theta_2} + \frac{1}{2}\log\theta_2 = \frac{\mu^2}{2\sigma^2} + \log\sigma, \end{aligned}$$

and $h(x) := 1/\sqrt{2\pi}$, then (1) becomes

$$h(x)\exp\left(w_1(\theta)t_1(x) + w_2(\theta)t_2(x) - a(w(\theta))\right) \qquad \text{for all } x \in \mathbb{R}.$$

Let $\Theta := \mathbb{R} \times (0, \infty)$, and for $\theta \in \Theta$ we define

$$f_\theta(x) := h(x)\exp\left(\sum_{i=1}^{2} w_i(\theta)t_i(x) - a(w(\theta))\right) \qquad \text{for all } x \in \mathbb{R}.$$

From this we see that $\{f_\theta : \theta \in \Theta\}$ is a two parameter exponential family and that the Gaussians can be expressed by an exponential family.

Beginning of Jan.21, 2022

We can also rewrite the Gaussian familty has a two parameter exponential family *in canonical form*:

$$w_1(\theta) = \frac{\mu}{\sigma^2} \qquad \text{and} \qquad w_2(\theta) = -\frac{1}{2\sigma^2}$$

so we try to rewrite $a(w)$ in terms of w_1, w_2 by

$$\begin{aligned} a(w) = \frac{\mu^2}{2\sigma^2} + \log\sigma &= -\left(\frac{\mu}{\sigma^2}\right)^2 \cdot \left(-\frac{1}{2\sigma^2}\right)^{-1} - \frac{1}{2}\log\left(-2 \cdot \frac{-1}{2\sigma^2}\right) \\ &= -\frac{w_1^2}{4w_2} - \frac{\log(-2w_2)}{2}. \end{aligned}$$

Originally we had the restriction $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, so this is equivalent to the constraint $\{(w_1, w_2) \in \mathbb{R}^2 : w_2 < 0\}$.

Example: (3.4) Location Family. Let X be a random variable with continuous density $f : \mathbb{R} \to [0, \infty)$. Let $\mu \in \mathbb{R}$. Then the densities $\{f(x+\mu)\}_{\mu \in \mathbb{R}}$ is called the **location family** of X. This may *or may not* be an exponential family.

An example: Gaussian densities with a fixed variance — shifting the pdf simply results in a new Gaussian pdf with shifted mean and same variance.

A non-example: if X is uniform on $[0, 1]$ then the location family $1_{[-\mu, 1-\mu]}$ do not form an exponential family.

Example: (3.6) Scale Family. Let X be a random variable. The densities $\{\sigma^{-1} f(x/\sigma)\}_{\sigma>0}$ are called the **scale family** of X. (Divide by $1/\sigma$ because we need to ensure the integral is 1.) This family may *or may not* be an exponential family.

Example: (3.7) Location and Scale Family. Combining the two examples above, $\{\sigma^{-1} f((x+\mu)/\sigma)\}$ is caled the **location and scale family** of X. Again, this may *or may not* be an exponential family.

3.2 Differential Identities

Sometimes exponential families make certain computations easier. One obvious example is via differentiation.

Let X be a standard Gaussian. Then its moment generating function (MGF) is

$$\mathbb{E}e^{tX} = e^{t^2/2} \qquad \text{for al } t \in \mathbb{R}.$$

Using this we have

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m}\Big|_{t=0} \mathbb{E}e^{tX} = \mathbb{E}X^m,$$

so for example

$$\mathbb{E}X^2 = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} e^{t^2/2} = 1.$$

We can do similar things for exponential families. If

$$a(w) = \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\, \mathrm{d}\mu(x),$$

and let W be the natural parameter space (i.e., where $a(w) < \infty$), then we claim that

Lemma: (3.8)

$a(w)$ is continuous and has continuous partial derivatives on the interior of W (i.e. where $a(\cdot)$ is finite). Moreover, the derivative can be obtained by differentiating under the integral sign.

Proof. We prove the existence of first order partial derivative with respect to w_1 and the rest follows by iteration. Let $e_1 := (1, 0, \ldots, 0) \in \mathbb{R}^k$. Exponential is analytic so it suffices to show that $\exp(a(w))$ has continuous partial derivative along e_1. The difference quotient is

$$\begin{aligned}\frac{\exp(a(w+\epsilon e_1)) - \exp(a(w))}{\epsilon} &= \frac{1}{\epsilon}\int_{\mathbb{R}^n} h(x)\left[\exp\Big(\epsilon t_1(x) + \sum_{i=1}^k w_i t_i(x)\Big) - \exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\right] \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^n} h(x) \frac{\exp(\epsilon t_1(x)) - 1}{\epsilon} \exp\Big(w_i t_i(x)\Big)\, \mathrm{d}\mu(x).\end{aligned}$$

By the MVT, for any $\alpha \in (0,1)$ and for all $\beta \in \mathbb{R}$,

$$|e^{\alpha\beta - 1}| \leqslant |\alpha\beta| e^{|\beta|} \leqslant |\alpha| e^{2|\beta|} \leqslant |\alpha|(e^{2\beta} + e^{-2\beta}). \tag{*}$$

Beginning of Jan.24, 2022

Therefore, for $\delta > 0$, $\alpha := \epsilon/\delta$ and $\beta := \delta t_1(x)$,

$$\left|h(x)\frac{\exp(\epsilon t_1(x)) - 1}{\epsilon}\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\right| \leqslant h(x)\left|\frac{\exp(\epsilon t_1(x)) - 1}{\epsilon}\right|\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big) \tag{1}$$

$$\leqslant \frac{1}{\delta}h(x)[\exp(2\delta t_1(x) + \exp(-2\delta t_1(x))]\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big). \tag{2}$$

Note that we have gotten rid of the dependence of ϵ.

If we define $X_\epsilon :=$ the LHS of (1) and $Y :=$ (2), then $|X_\epsilon| \leqslant Y$ for $0 < \epsilon < \delta < 1$.Letting $\epsilon \to 0$ and using DCT,

$$\begin{aligned}\frac{\partial}{\partial w_1}\exp(a(w)) &= \lim_{\epsilon\to 0}\int_{\mathbb{R}^n}\left|h(x)\frac{\exp(\epsilon t_1(x)) - 1}{\epsilon}\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\right| \mathrm{d}\mu(x)\\ &= \int_{\mathbb{R}^n}\lim_{\epsilon\to 0} h(x)\left|\frac{\exp(\epsilon t_1(x)) - 1}{\epsilon}\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\right| \mathrm{d}\mu(x)\\ &= \int_{\mathbb{R}^n} h(x)t_1(x)\exp\Big(\sum_{i=1}^k w_i t_i(x)\Big)\,\mathrm{d}\mu(x),\end{aligned}$$

where the dominance of an integrable function is given by the fact that w is in the interior of W, so there exists $\delta > 0$ such that

$$a(w + 2\delta e_1) < \infty \qquad \text{and} \qquad a(w - 2\delta e_1) < \infty.$$

□

Remark. We can rewrite the above formula, using definition of $e^{-a(w)}$, as

$$\exp(-a(w))\frac{\partial}{\partial w_1}\exp(a(w)) = \int_{\mathbb{R}^n} t_1(x)h(x)\exp\Big(\sum_{i=1}^k w_i t_i(x) - a(w)\Big)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^n} t_1(x)f_w(x)\,\mathrm{d}\mu(x).$$

That is, differentiating $a(w)$ gives moment information for the exponential family $\{f_w(x)\}$.

Since $f_w(x)$ can be thought of as a PDF with respect to the measure μ, i.e. $\int_{\mathbb{R}^n} t_i f_w(x)\,\mathrm{d}\mu(x) = 1$, for convenience we define

$$\mathbb{E}_\theta t_i := \int_{\mathbb{R}^n} t_i f_w(x)\,\mathrm{d}\mu(x).$$

Remark. We proved the lemma for canonical exponential families. For non-canonical exponential families, a similar argument holds:

$$e^{-a(w(\theta))}\frac{\partial}{\partial\theta_1}e^{a(w(\theta))} = e^{-a(w(\theta))}\sum_{i=1}^k\frac{\partial e^{a(w)}}{\partial w_i}\frac{\partial w_i}{\partial\theta_1} = \sum_{i=1}^k\frac{\partial w_i}{\partial\theta_1}\mathbb{E}_\theta t_i = \mathbb{E}_\theta\Big(\sum_{i=1}^k\frac{\partial w_i}{\partial\theta_1}t_i\Big).$$

We will often use this version of the **differential identity**.

We can take *more* derivatives of $a(w(\theta))$ and obtain more moment information.

Example: (3.13) Gaussian revisited. Recall that, for Gaussians with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, we have $k = 2, n = 1$, and we defined $\theta = (\theta_1, \theta_2) := (\mu, \sigma^2) \in \mathbb{R}^2$, $t_1(x) := x$, $t_2(x) := x^2$,

$$w_1(\theta) := \frac{\theta_1}{\theta_2} = \frac{\mu}{\sigma^2}, \qquad w_2(\theta) := -\frac{1}{2\theta_2} = -\frac{1}{2\sigma^2},$$

and finally

$$a(w(\theta)) := \frac{\theta_1^2}{2\theta_2} + \frac{\log\theta_2}{2} = \frac{\mu^2}{2\sigma^2} + \log\sigma.$$

Then,

$$\begin{aligned} e^{-a(w(\theta))}\frac{\partial}{\partial\theta_1}e^{a(w(\theta))} &= e^{-a(w(\theta))}\frac{\mathrm{d}}{\mathrm{d}\theta_1}\exp\left[\frac{\theta_1^2}{2\theta_2} + \frac{\log\theta_2}{2}\right] \\ &= (2\theta_1)/(2\theta_2) = \mu/\sigma^2, \end{aligned}$$

whereas the previous remark gives

$$\mathbb{E}_\theta\Big(\sum_{i=1}^2 \frac{\partial w_i}{\partial\theta_1}t_i\Big) = \mathbb{E}_\theta\Big(\frac{\partial w_1}{\partial\theta_1}t_1 + 0\Big)\mathbb{E}_\theta(x/\theta_2) = \mathbb{E}_\theta(x)/\sigma^2.$$

That is,

$$\mathbb{E}_\theta(x)/\sigma^2 = \mu/\sigma^2 \implies \mathbb{E}_\theta(x) = \mu.$$

In totality, we've shown that ***expected value of a Gaussian with mean μ is indeed μ!***

Beginning of Jan.26, 2022

Example: (3.15) Binomial (n, p) has expected value np. Since

$$\begin{aligned} \mathbb{P}(X = x) &= \binom{n}{x}p^x(1-p)^{n-x} = \binom{n}{x}(1-p)^n\left(\frac{p}{1-p}\right)^x \\ &= \binom{n}{x}\exp\left(x\log\left(\frac{p}{1-p}\right) - (-1)n\log(1-p)\right), \end{aligned}$$

we define a one-parameter exponential family using $h(x) := \binom{n}{x}$ on $\mathbb{N}$, $\theta := p$, $\Theta := (0,1)$,

$$t(x) := x, \qquad w(\theta) := \log(\theta/(1-\theta)), \qquad \text{and } a(w(\theta)) := -n\log(1-\theta).$$

In doing so we have $f_\theta(x) = h(x)\exp(w(\theta)t(x) - a(w(\theta))$, so the differential identity gives

$$e^{-a(w(\theta))}\frac{\mathrm{d}}{\mathrm{d}\theta}e^{a(w(\theta))} = \frac{\mathrm{d}}{\mathrm{d}\theta}a(w(\theta)) = \mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}w(\theta)t\right).$$

Therefore, $\dfrac{n}{1-\theta} = \dfrac{\mathbb{E}_\theta(x)}{\theta(1-\theta)}$ which, upon rearranging, leads to

$$\mathbb{E}_\theta(x) = \frac{n\theta(1-\theta)}{1-\theta} = n\theta = np,$$

i.e., ***the expected value of a Binomial (n, p) has expected np.*** How surprising.

Chapter 4

Random Samples

4.1 Random Samples of Gaussians

Definition: (4.1) Random Samples

A **random sample** of size n is a sequence $X_1, \dots, X_n$ of independent identically distributed (i.i.d.) (real-valued) random variables.

Definition: (4.2) Statistic

Let n, k be positive integers. Let $X_1, \dots, X_n$ be a random sample and let $f : \mathbb{R}^n \to \mathbb{R}^k$. A **statistic** is a random variable of form $Y := f(X_1, \dots, X_n)$ and its distribution is called a **sampling distribution**.
Most common examples include the **sample mean**

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the **sample variance**

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(*We divide by $n-1$ because this makes S^2 unbiased to estimate σ^2; this will be discussed later.*)

Proposition: (4.7)

Let $n \geqslant 2$ and let $X_1, \dots, X_n$ be a random sample from a *Gaussian* distribution with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then:

(1) $\overline{X}$ and S are independent,

(2) $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, and

(3) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. (A chi-squared random variable with n degrees of freedom, χ^2_n, has the PDF obtained from adding n independent squared standard Gaussians, i.e., $\chi^2_n \sim Z_1^2 + \dots + Z_n^2$.)

Proof of (1). WLOG assume $\mu = 0$ and $\sigma = 1$ since the claim is invariant under shifting and scaling.
We first show that $\overline{X}$ is independent of $X_2 - \overline{X}, ..., X_n - \overline{X}$ (i.e., *pairwise* independent between X_2 *and* any one of these). To see this, note that $(1, ..., 1) \in \mathbb{R}^n$ is orthogonal to the span of

$$e_2 - \frac{(1, ..., 1)}{n}, \ \dots, e_n - \frac{(1, ..., 1)}{n}$$

(where e_i has the i^{th} component 1 and zero for all other components).
Exercise 3.16 shows that if $X = (X_1, ..., X_n)$, then $\langle X, v_1 \rangle, \langle X, v_2 \rangle, ..., \langle X, v_n \rangle$ are independent (random variables) if and only if $v_1, ..., v_n$ are pairwise orthogonal (vectors). Hence the result above shows that

$$\langle X, (1, ..., 1) \rangle = X_1 + ... + X_n$$

is independent of the span of

$$\langle X, e_2 - (1, ..., 1)/n \rangle = X_2 - \overline{X}, \ \dots, \langle X, (e_n - (1, ..., 1)/n \rangle = X_n - \overline{X}.$$

It remains to notice that

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 = (X_1 - \overline{X})^2 + \sum_{i=2}^{n}(X_i - \overline{X})^2 \\ &= (n\overline{X} - \overline{X} - \sum_{i=2}^{n} X_i)^2 + \sum_{i=2}^{n}(X_i - \overline{X})^2 = \Big(\sum_{i=2}^{n}(X_i - \overline{X})\Big)^2 + \sum_{i=2}^{n}(X_i - \overline{X})^2.\end{aligned}$$

That is, S^2 can be written as a function of $X_2 - \overline{X}, ..., X_n - \overline{X}$ only, all of which are independent to $n\overline{X}$. This proves the claim. □

Beginning of Jan.28, 2022

Proof of (3). Notation-wise, redefine $\overline{X}_n : \sum_{i=1}^{n} X_i/n$ and $S_n^2 := \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/(n-1)$. We use induction on n.
In the case $n = 2$, we have

$$S_2^2 = (X_1 - (X_1 + X_2)/2)^2 + (X_2 - (X_1 + X_2)/2)^2 = \frac{(X_1 - X_2)^2}{4} + \frac{(X_2 - X_1)^2}{4} = \frac{(X_1 - X_2)^2}{2}$$

Since $X_1 - X_2$ is a Gaussian with mean 0 and variance $2\sigma^2$ (by independence), $(X_1 - X_2)/(\sqrt{2}\sigma)$ is a standard Gaussian. Therefore $S_2^2/\sigma^2 \sim \chi_1^2$. Base case complete.
Now we induct on n. Some *simple algebraic manipulation* shows that

$$nS_{n+1}^2 = (n-1)S_n^2 + \frac{n}{n+1}(X_{n+1} - \overline{X}_n)^2 \qquad \text{for all } n \geqslant 2.$$

From part (1), S_n is independent of $\overline{X}_n$; also, X_{n+1} is independent of S_n, which is a function of $X_1, ..., X_n$ only. Therefore S_n is independent of their difference squared, i.e., $(X_{n+1} - \overline{X}_n)^2$. By inductive hypothesis, $(n-1)S_n^2/\sigma^2$ is χ_n^2. Also, $(X_{n+1} - \overline{X}_n)^2$ is a Gaussian with mean 0 and variance $\sigma^2 + \sigma^2/n = \sigma^2 n/(n+1)$. Therefore,

$$\frac{nS_{n+1}^2}{\sigma^2} = \frac{(n-1)S_n^2}{\sigma^2} + \frac{n(X_{n+1} - \overline{X}_n)^2}{(n+1)\sigma^2} \sim \chi_n^2 + Z^2 \sim \chi_{n+1}^2,$$

which finishes the inductive step. □

4.2 Student's t-distrubution

Recall that if $X_1, X_2, \ldots$ are a random sample from a Gaussian random variable with known parameters μ, σ, then

$$\frac{X_1 + \ldots + X_n}{\sigma\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim Z.$$

In practice, however, σ and/or μ are often times *unknown*. In this case, we can replace σ by S and instead examine

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

where μ becomes the only unknown quantity. By examine μ and plugging in different values, we might be able to determine the actual μ. However, it is not immediately clear what distribution $(\overline{X} - \mu)/(S/\sqrt{n})$ follows, since it is no longer a Gaussian —

Proposition: (4.9) Student's t-distribution

Let X be a standard Gaussian. Let $Y \sim \chi_p^2$ and assume that X, Y are *independent*. Then $X/\sqrt{Y/p}$ has the **student's t-distribution** with p degrees of freedom, characterized by the PDF

$$f_{X/(\sqrt{Y/p})}(t) := \frac{\Gamma((p+1)/2)}{\sqrt{\pi p}\,\Gamma(p/2)}\left(1 + \frac{t^2}{p}\right)^{-(p+1)/2} \qquad \text{where } t \in \mathbb{R}.$$

Proof. For convenience let $Z := \sqrt{Y/p}$, and our goal is find the PDF of Z. We compute CDF and the differentiate:

$$\begin{aligned} f_Z(y) &= \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Z \leqslant y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \leqslant y^2 p) = \frac{\mathrm{d}}{\mathrm{d}y}\int_0^{y^2 p} f_{\chi_p^2}(x)\,\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\int_0^{y^2 p} \frac{x^{p/2-1}e^{-x/2}}{2^{p/2}\Gamma(p/2)}\,\mathrm{d}x = (2yp) f_{\chi_p^2}(y^2/p) \\ &= \frac{2yp}{2^{p/2}\Gamma(p/2)}(y^2 p)^{p/2-1}e^{-y^2 p/2} = \frac{p^{p/2}y^{p-1}e^{-y^2 p/2}}{2^{p/2-1}\Gamma(p/2)}. \end{aligned}$$

Now we compute the CDF of X/Z. Let $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by $(b, a/b) \mapsto (a, b)$. (By doing so, the region below with constraint $x \leqslant ty$ becomes $x/y \leqslant t$, which makes things simpler.) The Jacobian determinant is $|a|$ for all $(a, b) \in \mathbb{R}^2$. Then,

$$\begin{aligned} \mathbb{P}(X/Z \leqslant t) = \mathbb{P}(X \leqslant tZ) &= \int_{\{(x,y):x \leqslant ty, y > 0\}} f_X(x) f_Z(y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_{\{(a,b):b \leqslant t, a > 0\}} |a| f_X(ab) f_Z(a)\,\mathrm{d}a\mathrm{d}b \\ &= \int_{-\infty}^t \int_0^\infty |a| f_X(ab) f_Z(a)\,\mathrm{d}a\,\mathrm{d}b. \end{aligned}$$

Differentiating with respect to t gives

$$\begin{aligned} f_{X/Z}(t) = \int_0^\infty |a| f_X(at) f_Z(a)\,\mathrm{d}a &= \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2-1}\Gamma(p/2)}\int_0^\infty a^p e^{-(p+t^2)a^2/2}\,\mathrm{d}a \\ &= \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)}\int_0^\infty x^{(p-1)/2}e^{-(p+t^2)x/2}\,\mathrm{d}x. \end{aligned}$$

Recall that a Gamma distributed random variable has PDF 1, i.e.,

$$\frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-x/\beta}\,\mathrm{d}x = 1 \implies \int_0^\infty x^{\alpha-1}e^{-x/\beta}\,\mathrm{d}x = \beta^\alpha\Gamma(\alpha).$$

Substituting with $\alpha - 1 := (p-1)/2$ and $\beta := 2/(p+t^2)$, we have

$$\begin{aligned} f_{X/Z}(t) &= \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)}\beta^\alpha\Gamma(\alpha) \\ &= \frac{p^{p/2}}{\sqrt{2\pi}2^{p/2}\Gamma(p/2)}\Gamma((p+1)/2)\left(\frac{2}{p+t^2}\right)^{(p+1)/2} \\ &= \frac{p^{p/2}\Gamma((p+1)/2)}{\sqrt{\pi}2^{(p+1)/2}\Gamma(p/2)}\left(\frac{p(1+t^2/p)}{2}\right)^{-(p+1)/2} \\ &= \frac{\Gamma((p+1)/2)}{\sqrt{\pi p}\,\Gamma(p/2)}\left(1+\frac{t^2}{p}\right)^{-(p+1)/2} \end{aligned}$$

which concludes the proof. □

4.3 The Delta Method

Beginning of Jan.31, 2022

Recall that if $X_1, X_2, \dots$ are i.i.d. with mean μ and variance $\sigma^2 \in \mathbb{R}$, then the CLT states that

$$\frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}} = \sqrt{n}\left(\frac{X_1 + \dots + X_n}{n} - \mu\right)$$

converges in distribution to a mean zero Gaussian with variance σ^2. That is, we have a "good" way of estimating the mean μ. The next question is, what about functions of μ, for example $1/\mu$ or μ^2?

Theorem: (4.14) Delta Method

Let $\theta \in \mathbb{R}$. Let $Y_1, Y_2, \dots$ be random variables such that $\sqrt{n}(Y_n - \theta)$ converges *in distribution* to $\mathcal{N}(0, \sigma^2)$ (assume $\sigma^2 > 0$). Let $f : \mathbb{R} \to \mathbb{R}$ and assume $f'(\theta)$ exists. Then

$$\sqrt{n}(f(Y_n) - f(\theta))$$

converges *in distribution* to a mean zero Gaussian with variance $\sigma^2(f'(\theta))^2$ as $n \to \infty$.

Since $f(\theta)$ is just a constant, we have

$$\sigma^2(f'(\theta))' \approx \operatorname{var}(\sqrt{n}(f(Y_n) - f(\theta))) = n\operatorname{var}(f(Y_n));$$

that is, the Delta method an *approximation* $\operatorname{var}(f(Y_n)) \approx \dfrac{\sigma^2(f'(\theta))^2}{n}$ (convergence in distribution is strictly weaker than that in L^2 so this limits might not equal; approximations, however, still makes sense).

Proof. Suppose $f'(\theta)$ exists, i.e., $\lim_{y\to\theta}\dfrac{f(y) - f(\theta)}{y - \theta}$ exists. By definition there exists a sublinear $h : \mathbb{R} \to \mathbb{R}$ satisfying

$$f(y) = f(\theta) + f'(\theta)(y - \theta) + h(y - \theta).$$

(That is, h satisfies $\lim_{z\to 0} h(z)/z = 0$.) Some algebraic manipulation gives

$$\sqrt{n}(f(Y_n) - f(\theta)) = \sqrt{n}f'(\theta)(Y_n - \theta) + \sqrt{n}h(Y_n - \theta). \tag{1}$$

It remains to justify that the last term "doesn't matter" as $n \to \infty$.

By convergence in distribution, for all $s, t > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - \theta| > st/\sqrt{n}) = \frac{2}{\sqrt{2\pi}} \int_{st}^{\infty} e^{-y^2/(2\sigma^2)}\,\mathrm{d}y. \tag{2}$$

Therefore, splitting the case $\sqrt{n}|h(Y_n - \theta)| > t$ by whether $|Y_n - \theta|$ is small, we have

$$\begin{aligned}\mathbb{P}(\sqrt{n}|h(Y_n-\theta)| > t) &= \mathbb{P}(\sqrt{n}|h(Y_n-\theta)| > t, |Y_n - \theta| > st/\sqrt{n}) + \mathbb{P}(\sqrt{n}|h(Y_n-\theta)| > t, |Y_n - \theta| \leqslant st/\sqrt{n}) \\ &\leqslant \mathbb{P}(|Y_n - \theta| > st/\sqrt{n}) + \mathbb{P}(\sqrt{n}|h(Y_n-\theta)| > t, |Y_n - \theta| \leqslant st/\sqrt{n}). \end{aligned} \tag{3}$$

Let $n \to \infty$. The first term in (3) converges to $\frac{2}{\sqrt{2\pi}} \int_{st}^{\infty} e^{-y^2/(2\sigma^2)}\,\mathrm{d}y$ by (2). For the second term, since

$$\sqrt{n}|h(Y_n - \theta)| = \frac{|h(Y_n-\theta)|}{|Y_n - \theta|} \cdot \sqrt{n}|Y_n - \theta| \leqslant st \frac{|h(Y_n-\theta)|}{|Y_n-\theta|} \to 0,$$

the entire probability tends to 0. Therefore, for any $s, t > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}|h(Y_n-\theta)| > t) \leqslant \frac{2}{\sqrt{2\pi}} \int_{st}^{\infty} e^{-y^2/(2\sigma^2)}\,\mathrm{d}y. \tag{4}$$

Note that the LHS of (4) is independent of s, so we can let $s \to \infty$ for any fixed t and obtain

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}|h(Y_n-\theta)| > t) = 0, \tag{5}$$

i.e., $\sqrt{n}h(Y_n - \theta)$ converges to the zero constant random variable in probability.
By *Slutsky's Theorem* ($X_n \to X$ in probability and $Y_n \to$ a constant c in distribution together imply $X_n + Y_n \to X + c$ in distribution),

$$\sqrt{n}(f(Y_n) - f(\theta)) = \underbrace{\sqrt{n}h(Y_n-\theta)}_{\text{conv. in prob.}} + \underbrace{\sqrt{n}f'(\theta)(Y_n-\theta)}_{\text{con. in dist.}}$$

converges *in distribution* to a Gaussian random variable with mean 0 and variance $\sigma^2(f'(\theta))^2$. $\square$

Beginning of Feb.2, 2022

Example: (4.15). Let $\overline{X}_n$ be the sample mean for $X_1, \dots, X_n$. We assume $\mathrm{var}(X_1) < \infty$. Let $\mu := \mathbb{E}X_1 \neq 0$.
By CLT, $\sqrt{n}(\overline{X}_n - \mu)$ converges in distribution to a mean zero Gaussian with variance $\sigma^2 := \mathrm{var}(X_1)$.
If we let $f(x) := 1/x$ for nonzero x, then by the Delta method

$$\sqrt{n}(f(\overline{X}_n) - f(\mu)) = \sqrt{n}\left(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\right)$$

converges in distribution to a mean zero Gaussian with variance $\sigma^2(f'(\mu))^2 = \sigma^2/\mu^4$. Put informally, we have the approximation $\mathrm{var}(1/\overline{X}_n) \approx \sigma^2/(n\mu^4)$.
The last approximation is not rigorous – convergence in distribution does not necessarily imply converges in variance. In order to make this rigorous, we need to assume that there exist $\epsilon, c > 0$ such that

$$\mathbb{E}\left|\sqrt{n}\left(f(\overline{X}_n) - \frac{1}{\mu}\right)\right|^{2+\epsilon} \leqslant c$$

for all $c > 0$.

Theorem: (4.16) Convergence Theorem with Bounded Moment

Let $X_1, X_2, \ldots$ be random variables that converge in distribution X. Assume that there exist $0 < \epsilon, c < \infty$ such that $\mathbb{E}|X_n|^{1+\epsilon} \leqslant c$ for all $n \geqslant 1$. Then

$$\mathbb{E}X = \mathbb{E} \lim_{n\to\infty} X_n = \lim_{n\to\infty} \mathbb{E}X_n.$$

Remark. If $f'(\theta) = 0$ then the Delta method simply says that $\sqrt{n}(f(Y_n) - f(\theta))$ converges in distribution to the zero random variable. This kills the purpose of analyzing the variance alongside convergence. We fix this issue by introducing the second-order Delta method.

Theorem: (4.17) Second Order Delta Method

Let the above assumptions hold. Let $f'(\theta) = 0$ and $f''(\theta)$ exist and be nonzero. Then

$$n(f(Y_n) - f(\theta))$$

converges in distribution to $\sigma^2/2 \cdot f''(\theta)$ times χ_1^2. More generally, if $f'(\theta) = \cdots = f^{(m-1)}(\theta) = 0$ and if $f^{(m)}(\theta)$ exists and is nonzero, then

$$\sqrt{n^m}(f(Y_n) - f(\theta))$$

converges in distribution to $\sigma^2/m! \cdot f^{(m)}(\theta)$ times $(\mathcal{N}(0,1))^m$.

Chapter 5

Data Reduction

Question. How to find a parameter that fits data well using as little information as possible? One way is by using a sufficient statistic.

5.1 Sufficient Statistics

Definition: (5.1) Sufficient Statistic

Let $X = (X_1, ..., X_n)$ be a random sample from a distribution $f \in \{f_\theta : \theta \in \Theta\}$. Let $t : \mathbb{R}^n \to \mathbb{R}^k$ so that $Y := t(X_1, ..., X_n)$ is a statistic. We say Y is **sufficient** for θ if, for every $y \in \mathbb{R}^k$ and every $\theta \in \Theta$, the conditional distribution of $X = (X_1, ..., X_n)$ given $Y = y$ does *not* depend on θ. In other words, Y provides sufficient information to *estimate* θ from $X_1, ..., X_n$.

As we shall see from the next example, Y being sufficient does not mean Y allows us to *exactly* determine θ. All it says is that we have sufficient information to *guess* or *give a good estimate* for the unknown θ.

Beginning of Feb.4, 2022

Example: (5.5) Sufficient statistics always exist. Though trivial, the statistic $(X_1, ..., X_n)$ is always sufficient, for the distribution of $(X_1, ..., X_n) \mid (X_1, ..., X_n)$ clearly does not depend on θ.

We now look at two nontrivial, more succinct sufficient statistics, and later we will determine if there exists a sufficient statistic with "minimal amount of information", i.e., a "most useful" sufficient statistic.

Example: (5.2). Let $X_1, ..., X_n$ be i.i.d. Bernoulli distributions with parameter $\theta \in (0, 1)$. Then $Y := X_1 + ... + X_n$ is sufficient for θ.

Proof. Let $(x_1, ..., x_n) \in \{0, 1\}$ and let $0 \leqslant y \leqslant n$. Then Y is a binomial distribution with parameters (n, θ). Then

$$\mathbb{P}((X_1, ..., X_n) = (x_1, ..., x_n) \mid Y = y) = \begin{cases} 0 \text{ (trivial)} & \text{if } \sum x_i \neq y \\ \text{something nontrivial} & \text{if } \sum x_i = y. \end{cases}$$

For this reason, we assume that $y = x_1 + \ldots + x_n$. Then,

$$\begin{aligned}
\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \mid Y = y) &= \frac{\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n)), Y = y}{\mathbb{P}(Y = y)} \\
&= \frac{\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n))}{\mathbb{P}(Y = y)} \\
&= \frac{\prod_{i=1}^n \mathbb{P}(X_i = x_i)}{\mathbb{P}(Y = y)} = \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} \\
&= \frac{\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} = \binom{n}{y}^{-1},
\end{aligned}$$

indeed an expression not depending on θ.

Again, it is clear that Y alone cannot determine exactly what θ is; it however provides enough information for us to estimate θ.

Also, more formally, we should say Y_n is sufficient for θ given a random sample of size n. However, since dependency on n is clear, we tend to drop the cumbersome subscript and simply say Y is sufficient.

Example: (5.3). Let $X_1, \ldots, X_n$ be i.i.d. Gaussians with unknown $\mu \in \mathbb{R}$ and *known* $\sigma^2 > 0$. We claim that the sample mean $Y := (X_1 + \ldots + X_n)/n$ is sufficient for μ.

Proof. Let $(x_1, \ldots, x_n) \in \mathbb{R}$ and $y \in \mathbb{R}$. Like above, we can assume that $y = (x_1 + \ldots + x_n)/n$. Then Y is a Gaussian with mean μ and variance σ^2/n, and

$$\begin{aligned}
f_{X_1,\ldots,X_n|Y}(x_1, \ldots, x_n \mid y) &= \frac{f_{X_1,\ldots,X_n,Y}(x_1, \ldots, x_n, y)}{f_Y(y)} = \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{f_Y(y)} \\
&= \frac{\prod_{i=1}^n (\sigma\sqrt{2\pi})^{-1}\exp(-(x-\mu)^2/(2\sigma^2))}{\exp\left(-\frac{((x_1 + \ldots + x_n)/n - \mu)^2}{2\sigma^2/n}\right)/\sqrt{2\pi}\sigma/\sqrt{n}} \\
&= \frac{\sigma^{-n}(2\pi)^{-n/2}}{n^{1/2}\sigma^{-1}(2\pi)^{-1/2}} \frac{\exp(-(x_1^2 + \ldots + x_n^2)/(2\sigma^2) - n\mu^2/(2\sigma^2) + \sum x_i\mu/\sigma^2)}{\exp(-y^2n/(2\sigma^2) - n\mu^2/(2\sigma^2) + n\mu y/\sigma^2)} \\
&= \frac{\sigma^{-n}(2\pi)^{-n/2}}{n^{1/2}\sigma^{-1}(2\pi)^{1/2}} \frac{\exp((-\sum x_i^2)/(2\sigma^2))}{\exp(-y^2n/(2\sigma^2))}.
\end{aligned}$$

The last expression does not depend on μ, so Y is indeed sufficient for μ.

We now provide an "easy" way to find and/or identify sufficient statistics. Later on, we will further draw connections with exponential families, which would make things even nicer.

Theorem: (4.12) Factorization Theorem

Suppose $X_1, \ldots, X_n$ is a random sample from $\{f_\theta : \theta \in \Theta\}$. Suppose $Y = t(X_1, \ldots, X_n)$ is a statistic where $t : \mathbb{R}^n \to \mathbb{R}^k$. Then Y is sufficient for θ if and only if there exist $h : \mathbb{R}^n \to [0, \infty)$ and $g_\theta : \mathbb{R}^k \to [0, \infty)$ such that

$$f_\theta(x_1, \ldots, x_n) = f_\theta(x) = g_\theta(t(x)) \cdot h(x) \qquad \text{for all } \theta \in \Theta.$$

A technical remark: in the PMF case, we assume that $\bigcup_{\theta\in\Theta}\{x \in \mathbb{R}^n : f_\theta(x) > 0\}$ *is at most countable and require* *the above equation to hold on this set; in the PDF case, we require the above equality to hold almost everywhere.*

Beginning of Feb.8, 2022

Proof of Factorization Theorem, PMF Case. We first show that (sufficient) $\Rightarrow$ (factorization). Let $x \in \mathbb{R}^n$. Then

$$\begin{aligned} f_\theta(x) &= \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x \text{ and } Y = t(x)) \\ &= \mathbb{P}_\theta(Y = t(x))\mathbb{P}_\theta(X = x \mid Y = y) = \mathbb{P}_\theta(Y = t(x))\mathbb{P}(X = x). \end{aligned}$$

where the last step is by the sufficiency of Y. Thus we have obtained a factorization.
Conversely, suppose $f_\theta(x)$ admits a factorization $f_\theta(x) = g_\theta(t(x))h(x)$. Some definitions first: we define

$$r_\theta(z) := \mathbb{P}_\theta(t(X) = z) = \mathbb{P}_\theta(Y = z) \qquad \text{where } z \in \mathbb{R}^k,$$

$$\tilde{t}(t(x)) := \{y \in \mathbb{R}^n : t(y) = t(x)\} \qquad \text{where } x \in \mathbb{R}^n.$$

Now we expand the conditional probability:

$$\begin{aligned} \mathbb{P}_\theta(X = x \mid Y = t(x)) &= \frac{\mathbb{P}_\theta(X = x \text{ and } Y = t(x))}{\mathbb{P}_\theta(Y = t(x))} = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(Y = t(x))} \\ &= \frac{g_\theta(t(x)) \cdot h(x)}{\mathbb{P}_\theta(Y = t(x))} = \frac{g_\theta(t(x)) \cdot h(x)}{\sum_{z \in \tilde{t}t(x)} \mathbb{P}_\theta(X = z)} && \text{(total probability law)} \\ &= \frac{g_\theta(t(x)) \cdot h(x)}{\sum_{z \in \tilde{t}t(x)} g_\theta(t(z)) \cdot h(z)} && \text{(factorization assumption)} \\ &= \frac{g_\theta(t(x)) \cdot h(x)}{\sum_{z \in \tilde{t}t(x)} g_\theta(t(x)) \cdot h(z)} && (\text{since } z \in \tilde{t}t(x) \Rightarrow t(z) = t(x)) \\ &= \frac{g_\theta(t(x))}{g_\theta(t(x))} \frac{h(x)}{\sum_{z \in \tilde{t}t(x)} h(z)} = \frac{h(x)}{\sum_{z \in \tilde{t}t(x)} h(z)}, \end{aligned}$$

which is indeed independent of θ. ☐

We now move on to address the question of whether there exists a "more succinct" sufficient statistic, as mentioned before.

5.2 Minimal Sufficient Statistics

Suppose $t : \mathbb{R}^n \to \mathbb{R}^k$ and $Y = t(X_1, .., X_n)$ is sufficient for θ. Suppose $s : \mathbb{R}^n \to \mathbb{R}^m$ so $Z := s(X_1, ..., X_n)$ is another statistic. If there exists a function $\varphi : \mathbb{R}^m \to \mathbb{R}^k$ such that $\varphi \circ s = t$, i.e., $Y = \varphi(Z)$, then from the factorization above, Z is also sufficient, in the sense that

$$f_\theta(x) = g_\theta(t(x))h(x) = g_\theta(\varphi(s(x)))h(x) = (g \circ \varphi)_\theta(s(x))h(x).$$

That is, *if Y is sufficient and Y is a function of Z, then Z is automatically sufficient*.
Now we present the minimal sufficient statistics, as promised.

Definition: (5.6) Minimal Sufficient Statistic (MSS)

Suppose $X = (X_1, \dots, X_n)$ is a random sample of size n following a distribution in $\{f_\theta : \theta \in \Theta\}$. Let $Y = t(X_1, \dots, X_n)$ where $t : \mathbb{R}^n \to \mathbb{R}^k$ and assume Y is sufficient for θ. Then we say Y is **minimal sufficient** if, for every other sufficient $Z : \Omega \to \mathbb{R}^m$, there exists some function $r : \mathbb{R}^m \to \mathbb{R}^k$ such that $Y = r(Z)$.
Connecting to our introduction of MSS, this implies Y is the "most succint" sufficient statistic, as any other sufficient statistic requires more information.

Beginning of Feb.9, 2022

Example: (5.7). Let $X_1, \dots, X_n$ be a random Gaussian sample with (known) variance 1 but *unknown* mean $\mu \in \mathbb{R}$. We previous showed that the sample mean $\overline{X}$ is sufficient; in fact, it is minimal sufficient.
Connecting to another previous example, if we define $Y = t(X) := (X_1, \dots, X_n)$, then Y is trivially sufficient, since $\overline{X}$ can be expressed as the average of components of Y. Unless $n = 1$, it is not minimal sufficient — for $n \geqslant 2$, we cannot write $Y = (X_1, \dots, X_n)$ as a function of $\overline{X}$.
We will not prove that $\overline{X}$ is minimal sufficient; the proof is rather hard.

Theorem: (5.8) Characterization of Minimal Sufficiency

Let $X_1, \dots, X_n$ is a random sample with *joint* PDF/PMF from $\{f_\theta : \theta \in \Theta\}$. (If it is from a family of PMFs, assume the set $E := \bigcup_{\theta \in \Theta} \{x \in \mathbb{R}^n : f_\theta(x) > 0\}$ is at most countable.) Let $t : \mathbb{R}^n \to \mathbb{R}^m$ and $Y = t(X_1, \dots, X_n)$ be a statistic. If the following holds (a.e.) on $\mathbb{R}^n$ for PDFs or on E for PMFs, then Y is minimal sufficient:

There exists $c(x, y) \in \mathbb{R}$, dependent on x, y but *not* on θ, such that
$f_\theta(x) = c(x, y) f_\theta(y)$ for all $\theta \in \Theta$ if and only if $t(x) = t(y)$.

Proof. To avoid technical issues arising in measure theory, we again only consider the PMF case.
We first show that Y is sufficient. For any $z \in t(\mathbb{R}^n)$, let y_z be any element of $t^{-1}(z)$ so that $t(y_z) = z$. Then, for $x \in \mathbb{R}^n$, $t(y_{t(x)}) = t(x)$ so by assumption

$$f_\theta(x) = c(x, y_{t(x)}) f_\theta(y_{t(x)}).$$

Therefore, for all $z \in \mathbb{R}^m$ and all $x \in E$, if we define

$$g_\theta(z) := f_\theta(y_z) \qquad \text{and } h(x) := c(x, y_{t(x)}),$$

then we admit a factorization which completes the proof of sufficiency.

$$f_\theta(x) = g_\theta(t(x)) h(x),$$

Now we show that Y is minimal sufficient. Let Z be any other sufficient statistic with $Z = u(X_1, \dots, X_n)$. We need to show that t is a function of u.
By factorization theorem on Z, we can can write

$$f_\theta(x) = \tilde{g}_\theta(u(x)) \cdot \tilde{h}(x) \qquad \text{for all } \theta \in \Theta \text{ and all } x \in E.$$

Let $y \in \mathbb{R}^n$. WLOG assume $\tilde{h}(y) \neq 0$; otherwise $f_\theta(y) = 0$ for all θ, so by definition $y \notin E$ and we can simply ignore the case. Suppose for $u, y \in \mathbb{R}^n$ we have $u(x) = u(y)$. Then

$$f_\theta(x) = \tilde{g}_\theta(u(x)) \cdot \tilde{h}(x) = \tilde{g}_\theta(u(y)) \cdot \tilde{h}(x) = \tilde{g}_\theta(u(y)) \cdot \tilde{h}(y) \cdot \frac{\tilde{h}(x)}{\tilde{h}(y)}.$$

Using the converse of factorization theorem again,

$$f_\theta(x) = f_\theta(y) \frac{\tilde{h}(x)}{\tilde{h}(y)}, \qquad \text{for all } \theta \in \Theta.$$

Define $c(x, y) := \tilde{h}(x)/\tilde{h}(y)$, which is independent of θ indeed. We have shown that $f_\theta(x) = c(x, y) f_\theta(y)$ for all $\theta \in \Theta$. By the Theorem's assumption, this implies $t(x) = t(y)$. In other words, $u(x) = u(y)$ implies $t(x) = t(y)$. This implies that there exists a function φ with $t = \varphi \circ u$ (Exercise 5.9), which concludes the proof. $\square$

Example: (5.10) Exponential Families Gives MSS. Let $\{f_\theta : \theta \in \Theta\}$ be a k-parameter exponential family in canonical form

$$f_w(x) = h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x) - a(w(\theta)) \Big).$$

Let $X_1, \ldots, X_n$ be i.i.d. from f_w. Define

$$Y := t(X) := \sum_{i=1}^n (t_i(X_j), \ldots, t_k(X_j)).$$

Then Y is MSS for θ. **Upshot**: we can easily construct MSS from exponential families.
For example, if we sample from a Gaussian with unknown μ and $\sigma^2 > 0$, then $\overline{X}$ is minimal sufficient for θ and $(\overline{X}, S^2)$ is minimal sufficient for (μ, σ^2).

Existence and Uniqueness of MSS

Beginning of Feb.11, 2022

Observe that since $\overline{X}$ is MSS for μ where $X_1, \ldots, X_n$ are i.i.d. Gaussians wit known variance, then so is $c\overline{X}$ for any constant c. It turns out this uniqueness is "up to invertible transformations".

Remark: (5.11) Uniqueness of MSS up to Invertible Transformation. If $Y : \Omega \to \mathbb{R}^n, Z : \Omega \to \mathbb{R}^m$ are both MSS, then by definition there exist $r : \mathbb{R}^m \to \mathbb{R}^n$ with $Y = r(Z)$ and $s : \mathbb{R}^n \to \mathbb{R}^m$ with $Z = s(Y)$. Composing gives $r \circ s = \mathrm{id}_Y$ and $s \circ r = \mathrm{id}_Z$. Hence Y and Z are invertible images of each other.

Note that this also connects to the characterization of MSS in some sense. In particular, if Y is MSS, then the condition

$$f_\theta(x) = c(x, y) f_\theta(y) \iff t(x) = t(y)$$

should hold.

We now show existence of MSS.

Theorem: (5.12) Existence of MSS

Suppose $X_1, \ldots, X_n$ is a random sample of size n from $\{f_\theta : \theta \in \Theta\}$. *In the case of PMFs, assume* $\bigcup_{\theta\in\Theta} \{x \in \mathbb{R}^n : f_\theta(x) > 0\}$ *is countable.* Then there exists a MSS Y for θ.

***Proof for countable* Θ.** We label elements of $\{f_\theta : \theta \in \Theta\}$ as $\{f_n\}_{n\geqslant 1}$. We define an equivalence relation on $\mathbb{R}^{\mathbb{N}}$ by $x \sim y$ if x is a scalar multiple of y. Consider $t : \mathbb{R}^n \to \mathbb{R}^{\mathbb{N}}/\sim$ by

$$t(x) := (f_1(x), f_2(x), \ldots)$$

Define $Y := t(X_1, \ldots, X_n)$. We now check that such Y satisfies the condition in the MSS characterization theorem. On one hand, if $t(x) = t(y)$, then $f_k(x) = \alpha f_k(y)$ for some constant α that works for all k. Conversely, if for each θ, the corresponding $f_k(x)$ is some fixed α times $f_k(y)$, then again $t(x) = t(y)$ modulo $\sim$.
Therefore, the characterization theorem applies and Y, despite its weird appearance, is sufficient. □

From above, MSS sometimes might still contain "excess information". After all $(f_1(x), f_2(x), \ldots)$ is an infinite sequence. Though this is minimal sufficient, it is more interesting to come up with a way to get rid of the excess information of a statistic.

5.3 Ancillary Statistics

Definition: (5.14) Ancillary Statistic

Suppose $X_1, \ldots, X_n$ is a random sample of size n from $\{f_\theta : \theta \in \Theta\}$. A statistic $Y = t(X_1, \ldots, X_n)$ is **ancillary** for θ if the distribution of Y does not depend on θ.

Example: (5.15). Let $X_1, \ldots, X_n$ be a random sample from the location family for the **Cauchy distribution**. The joint PDF is

$$f_\theta(x) := \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2}, \qquad x \in \mathbb{R}^n, \theta \in \mathbb{R}.$$

The order statistics $(X_{(1)}, \ldots, X_{(n)})$, all put together, are minimal sufficient for θ. For sufficiency, we have

$$f_\theta(X) = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (X_i - \theta)^2} = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (X_{(i)} - \theta)^2} \cdot 1.$$

For minimal sufficiency, if $x, y \in \mathbb{R}^n$ are fixed, then

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{\prod_{i=1}^{n}(1 + (y_i - \theta)^2)}{\prod_{i=1}^{n}(1 + (x_i - \theta)^2)}$$

only when $t(x) = t(y)$. (Both top and bottom are polynomials of θ and their ratio is constant if and only if they share the same roots. Ordering them gives the same result, so $t(x) = t(y)$.) Then using the characterization theroem, we see $(X_{(1)}, \ldots, X_{(n)})$ is indeed MSS.
However, we began with a vector $(X_1, \ldots, X_n) \in \mathbb{R}^n$ and we ended up with another vector in $\mathbb{R}^n$. Something should be excess here.

For example, $X_{(n)} - X_{(1)}$ is ancillary for θ. If we let $Z_1, ..., Z_n$ be i.i.d. Cauchy random variables with pdf $\pi^{-1}1/(1+x^2)$, then $X_i = Z_i + \theta$ and $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$, which is indeed independent of θ. Because $(X_{(1)}, ..., X_{(n)})$ contains such ancillary statistic, it has "excess information" for θ.

5.4 Complete Statistics

Beginning of Feb.14, 2022

Continuing the above example, since $X_{(n)} - X_{(1)}$ is ancillary, its distribution does not rely on θ. Hence there exists a constant c such that, *for all* $\theta \in \Theta$,

$$\mathbb{E}_\theta (X_{(n)} - X_{(1)}) 1_{\{-1 \leqslant X_{(1)} \leqslant X_{(n)} \leqslant c\}} = c.$$

(The indicator function only serves to ensure that the above expression is well-defined, i.e., finite.)
Let $Y := (X_{(1)}, ..., X_{(n)})$ and let

$$f(x_1, ..., x_n) := (x_n - x_1) 1_{\{-1 \leqslant x_1, x_n \leqslant 1\}} - c \qquad \text{for } (x_1, ..., x_n) \in \mathbb{R}^n.$$

Then as stated above, $\mathbb{E}_\theta f(Y) = 0$ for all $\theta \in \Theta$ with $f(Y) \neq 0$. We claim that this implies Y contains extraneous information, and we turn the negation into a definition:

Definition: (5.16) Complete Statistic

Suppose $X_1, ..., X_n$ is a random sample with distribution from $\{f_\theta : \theta \in \Theta\}$. Let $t : \mathbb{R}^n \to \mathbb{R}^m$. We say a statistic $Y = t(X_1, ..., X_n)$ is **complete** for $\{f_\theta : \theta \in \Theta\}$ if, for any $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbb{E}_\theta f(Y) = 0$ for all $\theta \in \Theta$, we have $f(Y) = 0$.
(We implicitly assume $\mathbb{E}_\theta f(Y)$ is well-defined and $\mathbb{E}_\theta |f(Y)| < \infty$ for all $\theta \in \Theta$.)
Intuition: being complete means we have no excess information about θ.

Remark: Nonconstant Complete $\Rightarrow$ Not Ancillary. Let Y be nonconstant and complete. If Y is ancillary then there exists $c \in \mathbb{R}$ with $\mathbb{E}_\theta Y = c$ or $\mathbb{E}_\theta (Y - c) = 0$ for all $\theta \in \Theta$. By completeness this forces us to have $Y = c$, a contradiction.

Remark: Complete and Ancillary $\not\Rightarrow$ Sufficient. Consider a constant statistic.

Remark. We always have trivial complete statistics (like the constant one above), but unfortunately *complete sufficient* statistics might not exist. When they do, they are "good."

Example: (5.21) Binomial Revisited. Let $X = (X_1, ..., X_n)$ be a random sample from a Bernoulli distribution with parameter $0 < \theta < 1$. We showed that $Y = \sum_{i=1}^{n} X_i$ is sufficient for θ. We now show that Y is also complete.

Proof. Let $f : \mathbb{R} \to \mathbb{R}$ be such that $\mathbb{E}_\theta f(Y) = 0$ for all $\theta \in (0,1)$. Writing this explicitly,

$$0 = \mathbb{E}_\theta f(Y) = \sum_{k=0}^{n} f(k)\binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad \theta \in (0,1).$$

Since

$$0 = \sum_{k=0}^{n} f(k)\binom{n}{k}\alpha^k$$

where $\alpha := \theta/(1-\theta)$, we see the above is a polynomial that equals zero for all $\alpha > 0$. That is, the polynomial itself must be identically 0. Since binomial coefficients are not, we must have $f(k) = 0$ for $k \in \{0,1,\dots,n\}$, which completes our proof showing Y is complete.

Example: (5.22) Gaussians Revisited. Recall that if $X_1,\dots,X_n$ are i.i.d. Gaussians with known variance $\sigma^2 > 0$ and unknown $\mu \in \mathbb{R}$, then $Y = \overline{X}$ is (minimal) sufficient. We now claim that Y is also complete.
For simplicity we assume $n = \sigma = 1$ so Y is simply a standard Gaussian. Let $f : \mathbb{R} \to \mathbb{R}$ and assume $\mathbb{E}_\mu|f(Y)| < \infty$ for all μ. We further assume that

$$0 = \mathbb{E}_\mu f(Y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)\exp(-(t-\mu)^2/2)\,\mathrm{d}t, \qquad \text{for all } \mu \in \mathbb{R}.$$

Equivalently, after expansion and getting rid of the constants,

$$\int_{\mathbb{R}} f(t)\exp(-t^2/2)e^{t\mu}\,\mathrm{d}t = 0 \qquad \text{for all } \mu \in \mathbb{R}.$$

If $f \geqslant 0$ then clearly f needs to be identically 0. Otherwise we split f into positive and negative parts and will also obtain the result after some algebra.

Beginning of Feb.18, 2022

Theorem: (5.25) Bahadur's Theorem

If Y is complete and sufficient for $\{f_\theta : \theta \in \Theta\}$ then Y is minimal sufficient.
(For PMFs we assume $\bigcup_{\theta\in\Theta}\{x \in \mathbb{R}^n : f_\theta(x) > 0\}$ is countable.)

Proof. By a previous remark, there exists a MSS Z, so it suffices to show that there exists a function r with $Y = r(Z)$ (because any sufficient statistic is a function of Z, so Y is a composite function of that sufficient statistic).
Define $r(Z) := \mathbb{E}_\theta(Y \mid Z)$. We will show that $r(Z) = Y$. Since Z is MSS and Y sufficient, Z can be written as a function of Y, say $Z = u(Y)$. Therefore, using properties of conditionals,

$$\begin{aligned}
\mathbb{E}_\theta(r(u(Y))) &= \mathbb{E}_\theta(r(Z)) \\
&= \mathbb{E}_\theta[\mathbb{E}_\theta(Y \mid Z)] && \text{(definition of } r(Z)) \\
&= \mathbb{E}_\theta(Y). && \text{(total expected value)}
\end{aligned}$$

Therefore $\mathbb{E}_\theta(r(u(Y)) - Y) = 0$ for all $\theta \in \Theta$. By completeness this means $r(u(Y)) = Y$, i.e., $r(Z) = Y$. $\square$

Theorem: (5.27) Basu's Theorem

Let Y be complete and sufficient for $\{f_\theta : \theta \in \Theta\}$. If Z is ancillary for θ, then Y and Z are independent with respect to f_θ.
"Complete sufficient statistics are very nice since they do not contain ancillary data."

Proof. Let $Y : \Omega \to \mathbb{R}^k$ and $Z : \Omega \to \mathbb{R}^m$. Let $A \subset \mathbb{R}^k$ and $B \subset \mathbb{R}^m$. To show independence, we need to verify that

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{P}_\theta(Y \in A)\mathbb{P}_\theta(Z \in B) \qquad \text{for all } \theta \in \Theta.$$

That is,

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{E}_\theta 1_{Y\in A} 1_{Z\in B} = \mathbb{E}_\theta[\mathbb{E}_\theta(\theta(1_{Y\in A}1_{Z\in B}) \mid Y] = \mathbb{E}_\theta[1_{Y\in A}\mathbb{E}_\theta(1_{Z\in B} \mid Y)]$$

where the last = is by the tower property (i.e., $\mathbb{E}[\mathbb{E}(Xh(Y) \mid Y)] = h(Y)\mathbb{E}(X \mid Y)$). Since Y is sufficient, the conditional distribution does not depend on θ, so (check) $g(Y) := \mathbb{E}_\theta(1_{Z\in B} \mid Y)$ should not depend on θ. Therefore

$$\mathbb{E}_\theta g(Y) = \mathbb{E}_\theta[\mathbb{E}_\theta(1_{Z\in B} \mid Y)] = \mathbb{E}_\theta(1_{Z\in B}) = \mathbb{P}_\theta(Z \in B).$$

Since Z is ancillary we see $\mathbb{E}_\theta g(Y)$ does not depend on θ. Define this quantity to be c. Then

$$\mathbb{E}_\theta(g(Y) - c) = 0$$

for all $\theta \in \Theta$. By completeness this implies $g(Y) = c$, i.e., $g(Y)$ is constant. Therefore,

$$\mathbb{P}_\theta(Y \in A, Z \in B) = \mathbb{E}_\theta(1_{Y\in A} \cdot c) = \mathbb{E}_\theta(1_{Y\in A})\mathbb{P}_\theta(Z \in B) = \mathbb{P}_\theta(Y \in A)\mathbb{P}_\theta(Z \in B).$$

□

Chapter 6

Point Estimation

Goal in a nutshell: estimate some known $\theta \in \Theta$ using a function / statistic of a random sample $X_1, \ldots, X_n$. Such statistic $Y = t(X_1, \ldots, X_n)$ is called an **estimator** or **point estimator**. Unless otherwise specified, we assume $X_1, \ldots, X_n$ are i.i.d. from $\{f_\theta : \theta \in \Theta\}$. We also assume Y is a statistic of $X_1, \ldots, X_n$.

6.1 Evaluating Estimators; UMVU

Definition: (6.1) Likelihood Function

If $x \in \mathbb{R}^n$, then the function $\ell : \Theta \to [0, \infty)$ defined by $\ell(\theta) := f_\theta(x)$ is the **likelihood function**.

Beginning of Feb.23, 2022

Definition: (6.2) Unbiased Estimator

Let Y be an estimator for $g(\theta)$ where $g : \Theta \to \mathbb{R}^k$. We say Y is **unbiased** for $g(\theta)$ if

$$\mathbb{E}_\theta Y = g(\theta) \qquad \text{for all } \theta \in \Theta.$$

(Unbiased estimators always exist; for example consider the trivial constant statistic.)

For example, we have shown that the sample mean and variance are unbiased for a Gaussian's mean and variance, respectively.

However, it should be clear that just being unbiased doesn't necessarily guarantee a "good" estimator. For example, any statistic taking value $+r$ with probability $1/2$ and $-r$ with $1/2$ has expected value 0. If the quantity it estimates has expected value 0 then all such estimators are unbiased, but clearly as r gets large, this estimator gets "bad" since its distribution gets spread more widely. A workaround is to examing the **mean-squared error** (or L^2 norm):

$$\mathbb{E}_\theta (Y - g(\theta))^2.$$

For unbiased estimators, the above quantity equals $\mathrm{var}(Y)$.

Definition: (6.3) Uniformly Minimum Variance Unbiased Estimators, UMVU

Let $g : \Theta \to \mathbb{R}$. Assume Y is unbiased. We say Y is (an) **uniformly minimum variance unbiased** (estimator), **UMVU**, for $g(\theta)$ if for any other unbiased estimator Z for $g(\theta)$,

$$\operatorname{var}_\theta(Y) \leqslant \operatorname{var}_\theta(Z) \qquad \text{for all } \theta \in \Theta.$$

(UMVU might not exist a priori. See below.)

Definition: (6.4) Uniformly Minimum Risk Unbiased Estimators, UMRU

This generalizes the notion of UMVU. Suppose we are given a **loss function**

$$L : \Theta \times \mathbb{R}^k \to \mathbb{R}$$

(for example, consider $L(\theta, y) := (y - g(\theta))^2$, in which case the UMRU defined below is simply UMVU; also, we often assume that $L(\theta, y)$ is strictly convex in y) and we define the **risk function** to be

$$r(\theta, Y) = \mathbb{E}_\theta L(\theta, Y) \qquad \text{for all } \theta \in \Theta.$$

Again, assume Y is unbiased for $g(\theta)$. We say Y is (an) **uniformly minimum risk unbiased** (estimator), **UMRU**, for $g(\theta)$ if for any other unbiased estimator Z for $g(\theta)$,

$$r(\theta, Y) \leqslant r(\theta, Z) \qquad \text{for all } \theta \in \Theta.$$

Example: (6.5) UMVU might not exist. Suppose X is a binomial random variable with parameter n (known) and $\theta \in (0,1)$ (unknown), and we want to estimate $\theta/(1-\theta)$. It turns out there is *no unbiased estimator* for $g(\theta)$ (which implies there is no UMVU): for any estimator $Y = t(X)$,

$$\mathbb{E}_\theta Y = \mathbb{E}_\theta t(X) = \sum_{j=0}^{n} \binom{n}{i} t(i) \theta^i (1-\theta)^i,$$

a polynomial of θ, whereas $\theta/(1-\theta)$ is not.

6.2 Rao-Blackwell & Lehman-Scheffé

Theorem: (6.7) Rao-Blackwell Theorem

If $L(\theta, y)$ is convex in y, then *conditioning an unbiased on a sufficient one will only improve it*. More formally, if Z is sufficient for $\{f_\theta : \theta \in \Theta\}$ and Y unbiased for $g(\theta)$. Let $\theta \in \Theta$ with $r(\theta, Y) < \infty$ and such that $L(\theta, y)$ is convex in y. Then $W := \mathbb{E}_\theta(Y \mid Z)$ is unbiased and

$$r(\theta, W) \leqslant r(\theta, Y).$$

If in addition the risk function is strictly convex in y, then the inequality is strict unless $W = Y$.

Beginning of Feb.28, 2022

Proof. First note that since Z is sufficient, the distribution of W does not depend on θ, so W is indeed well-defined. Also, since Y is unbiased, so is W, since $\mathbb{E}_\theta W = \mathbb{E}_\theta \mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta Y$.
By definition $L(\theta, W) = L(\theta, \mathbb{E}_\theta(Y \mid Z))$. By Jensen's inequality we have

$$L(\theta, W) = L(\theta, \mathbb{E}_\theta(Y \mid Z)) \leqslant \mathbb{E}_\theta(L(\theta, Y) \mid Z). \tag{*}$$

Taking expectation on both sides again,

$$r(\theta, W) = \mathbb{E}_\theta L(\theta, W) \leqslant \mathbb{E}_\theta \mathbb{E}_\theta(L(\theta, Y) \mid Z) = \mathbb{E}_\theta L(\theta, Y) = r(\theta, Y).$$

Finally, if L is strictly convex, then the above inequality is strict unless (*) is attains equality; this happens when Y is a function of Z. If so, $W = \mathbb{E}_\theta(Y \mid Z) = Y$. □

Remark. We will later show that if Y is unbiased and Z is sufficient and *complete*, then the corresponding W automatically gives the UMRU.

Example: (6.12). Let $X_1, \ldots, X_n$ be i.i.d. with unknown mean $\mu \in \mathbb{R}$. Let $Y := t(X_1, \ldots, X_n) := X_1$, a bad yet unbiased estimator.
A bad example of Rao-Blackwell: condition Y on the trivially sufficient $(X_1, \ldots, X_n)$, which gives

$$W = \mathbb{E}(X_1 \mid X_1, \ldots, X_n) = \mathbb{E}(X_1 \mid X_1) = X_1.$$

A better example: we now condition Y on $\sum_{i=1}^n X_i$ (no guarantee if this is sufficient, but we condition it anyways). Then

$$\sum_{j=1}^n \mathbb{E}(X_j \mid \sum_{i=1}^n X_i) = n\mathbb{E}(X_1 \mid \sum_{i=1}^n X_i) \implies W := \mathbb{E}(X_1 \mid \sum_{i=1}^n X_i) = \frac{1}{n} \sum_{i=1}^n X_i,$$

so (whether or not) Rao-Blackwell gives a much better unbiased estimator.

Beginning of March 2, 2022

Example: Order statistics and sufficiency. If $X_1, \ldots, X_n$ are i.i.d. from $\{f_\theta : \theta \in \Theta\}$, then $(X_{(1)}, \ldots, X_{(n)})$ is always sufficient.
On the other hand, suppose also that $Y_1, \ldots, Y_n$ are i.i.d. from $\{g_\theta : \theta \in \Theta\}$. Suppose we want to estimate $\mathrm{var}(X_1, Y_1) = \mathbb{E}[(X_1 - \mathbb{E}X_1)(Y_1 - \mathbb{E}Y_1)]$. By reordering X_i into $X_{(1)}, \ldots, X_{(n)}$ and Y_i into $Y_{(1)}, \ldots, Y_{(n)}$ *separately*, there is no guarantee that X_i, Y_i still share the same index after using order statistics. Hence $X_{(1)}, \ldots, X_{(n)}, Y_{(1)}, \ldots, Y_{(n)}$ might not be sufficient for the covariance.

Theorem: (6.13) Lehmann-Scheffé

Conditioning an unbiased statistic on a complete sufficient one gives the UMRU/UMVU.
Let Z be a complete sufficient statistic for $\{f_\theta : \theta \in \Theta\}$, let Y be unbiased for $g(\theta)$, let $L(\theta, y)$ be convex in y, and define $W := \mathbb{E}_\theta(Y \mid Z)$. Then W is UMRU for $g(\theta)$.
Moreover, if $L(\theta, y)$ is strictly convex, then W is unique. (In particular, UMVU is unique.)

Proof. Since Y is unbiased, so is W. We first show that W does not depend on Y. (*Intuitively, given a strictly convex loss function, the unique UMRU should not depend on what Y on which we conditioned.*) Let Y' be another unbiased estimator for $g(\theta)$. Then

$$\mathbb{E}_\theta[\mathbb{E}_\theta(Y \mid Z) - \mathbb{E}_\theta(Y' \mid Z)] = \mathbb{E}_\theta Y - \mathbb{E}_\theta Y' = g(\theta) - g(\theta) = 0 \qquad \text{for all } \theta \in \Theta$$

so by completeness

$$\mathbb{E}_\theta(Y \mid Z) = \mathbb{E}_\theta(Y' \mid Z) \qquad \text{for all } \theta \in \Theta.$$

Therefore W does not depend on the choice of Y. Using Rao-Blackwell,

$$r(\theta, W) = r(\theta, \mathbb{E}_\theta(Y \mid Z)) = r(\theta, \mathbb{E}_\theta(Y' \mid Z)) \leqslant r(\theta, Y') \qquad \text{for all } \theta \in \Theta.$$

for all unbiased Y'. That is, W is a UMRU. Uniqueness when L is convex follows from Rao-Blackwell as well. □

Remark: (6.14). Here is a method to think backwards on obtaining a UMVU via Lehmann-Scheffé.
Let $Z : \Omega \to \mathbb{R}^k$ be complete sufficient for $\{f_\theta : \theta \in \Theta\}$. Let $h : \mathbb{R}^k \to \mathbb{R}^m$ and let $g(\theta) := \mathbb{E}_\theta h(Z)$. Then $W := \mathbb{E}_\theta(h(Z) \mid Z) = h(Z)$ is unbiased for $g(\theta)$. That is, $h(Z)$ is UMVU for $g(\theta)$.
If we can guess or solve a function h such that $g(\theta) = \mathbb{E}_\theta h(Z)$, then we are done.

Beginning of March 4, 2022

Example: (6.15) Gaussian and UMVU (backward thinking). Suppose we are sampling from a Gaussian with unknown $\mu \in \mathbb{R}$ and unknown $\sigma^2 > 0$. We take it for granted that $(\overline{X}, S^2)$ is complete for (μ, σ^2). So $\overline{X}$ is UMVU for μ:

$$h(x, y) := x \text{ and } g(\mu, \sigma^2) := \mu \implies g(\mu, \sigma^2) = \mathbb{E}_\theta h(Z).$$

Similarly, S^2 is UMVU for σ^2:

$$h(x, y) := y \text{ and } g(\mu, \sigma^2) := \sigma^2 \implies g(\mu, \sigma^2) = \mathbb{E}_\theta h(Z).$$

Finally, to find the UMVU for μ^2, we try to express it in terms of $\overline{X}$ and S^2:

$$\mathbb{E}\overline{X}^2 = \operatorname{var}(\overline{X}) + (\mathbb{E}\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

so

$$\mu^2 = \mathbb{E}(\overline{X}^2 - S^2/n).$$

That is, $\overline{X}^2 - S^2/n$ is UMVU for μ^2.

Example: (6.16) Binomial and UMVU (backward thinking). Consider a binomial random variable with parametrs n and $\theta \in (0,1)$. Suppose we want to estimate $g(\theta) := \theta(1-\theta)$, the variance of X. Using "backward thinking", we want to find $h : \mathbb{R} \to \mathbb{R}$ such that

$$\theta(1-\theta) = \mathbb{E}_\theta h(X) = \sum_{j=0}^{n} h(j)\binom{n}{j}\theta^j(1-\theta)^{n-j}.$$

Let $a := \theta/(1-\theta)$ so

$$\sum_{j=0}^{n} h(j)\binom{n}{j}a^j = (1-\theta)^{-n}\mathbb{E}_\theta h(X) = \theta(1-\theta)^{1-n}. \tag{1}$$

Since $\theta = a/(1+a)$ and so $1-\theta = 1/(1+a)$, binomial theorem gives

$$(1-\theta)^{-n}\mathbb{E}_\theta h(X) = (1+a)^{-1}a(1+a)^{n-1} = a(1+a)^{n-2} = a\sum_{j=0}^{n-2}\binom{n-2}{j}a^j = \sum_{j=1}^{n-1}\binom{n-2}{j-1}a^j. \tag{2}$$

Comparing the LHS of (1) and the RHS of (2) we see that the polynomials are equal on $(0,1)$, so their coefficients must be identical. Therefore

$$h(j) = \binom{n-2}{j-1}\binom{n}{j}^{-1} = \frac{(n-2)!}{(j-1)!(n-j-1)!}\frac{j!(n-j)!}{n!} = \frac{(n-j)j}{n(n-1)},$$

i.e., the UMVU for $\theta(1-\theta)$ is $\dfrac{X(n-X)}{n(n-1)}$ (assuming $n \geqslant 2$).

Example: (6.17) Bernoulli and UMVU (Lehman-Scheffé). Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli with $\theta \in (0,1)$. We have shown previosuly that $Z := \sum_{i=1}^{n} X_i$ is complete and sufficient and $\overline{X}$ is unbiased for θ. Therefore $\overline{X}$ is UMVU for θ.

Suppose we want to estimate θ^2. Since $Y := X_1X_2$ is unbiased, $\mathbb{E}(Y \mid Z)$ will be the UMVU.

Let $2 \leqslant z \leqslant n$. Since $Y = 1$ if and only if $X_1 = X_2 = 1$,

$$\begin{aligned}
\mathbb{E}_\theta(Y \mid Z = z) &= \mathbb{E}_\theta(1_{X_1=X_2=1} \mid Z = z) = \mathbb{P}_\theta(X_1 = X_2 = 1 \mid Z = z) \\
&= \mathbb{P}_\theta\Big(X_1 = X_2 = 1 \;\Big|\; \sum_{i=1}^{n} X_i = z\Big) = \frac{\mathbb{P}_\theta(X_1 = X_2 = 1, \sum_{i=1}^n X_i = z)}{\mathbb{P}_\theta(\sum_{i=1}^n X_i = z)} \\
&= \frac{\mathbb{P}_\theta(X_1 = X_2 = 1, \sum_{i=3}^n X_i = z-2)}{\mathbb{P}_\theta(\sum_{i=1}^n X_i = z)} \\
&= \frac{\theta^2\binom{n-2}{z-2}\theta^{z-2}(1-\theta)^{n-z}}{\binom{n}{z}\theta^z(1-\theta)^{n-z}} = \binom{n-2}{z-2}\binom{n}{z}^{-1} \\
&= \frac{(n-2)!}{(z-2)!(n-z)!}\frac{z!(n-z)!}{n!} = \frac{z(z-1)}{n(n-1)}.
\end{aligned}$$

We check that the cases $z = 1, z = 2$ still satisfy this relation. Hence the UMVU for θ^2 is $\mathbb{E}_\theta(Y \mid Z) = \dfrac{Z(Z-1)}{n(n-1)}$.

Beginning of March 7, 2022

One More Remark on UMVU

Question. if W_1 is UMVU for $g_1(\theta)$ and W_2 UMVU for $g_2(\theta)$, does it follow that $W_1 + W_2$ is UMVU for $g_1(\theta) + g_2(\theta)$? By Lehman-Scheffé, if Y is unbiased for $g_1(\theta)$ and Y_2 unbiased for $g_2(\theta)$, and if Z is complete and sufficient, then by uniqueness $W_i = \mathbb{E}_\theta(Y_i \mid Z)$, and by linearity

$$W_1 + W_2 = \mathbb{E}_\theta(Y_1 + Y_2 \mid Z)$$

is the UMVU for $g_1(\theta) + g_2(\theta)$. But what if we don't assume the existence of a complete sufficient Z a priori? The answer is still yes:

Theorem: (6.18) Alternate Characterization of UMVU

Let $\{f_\theta : \theta \in \Theta\}$ be a family of distributions and let W be unbiased of $g(\theta)$. Let $L_2(\Omega)$ be the set of statistics with finite second moment. then $W \in L_2(\Omega)$ is UMVU for $g(\theta)$ if and only if $\mathbb{E}_\theta(WU) = 0$ for all $\theta \in \Theta$ and all $U \in L_2(\Omega)$ with $\mathbb{E}_\theta U = 0$.

Remark. For the W_1, W_2 example above, this theorem gives that $\mathbb{E}_\theta(W_1 U) = \mathbb{E}_\theta(W_2 U) = 0$ for all $U \in L_2(\Omega)$ with $\mathbb{E}_\theta U = 0$. Then $W_1 + W_2$ is unbiased with $\mathbb{E}_\theta((W_1 + W_2)U) = 0$.

Proof. We first assume that W is UMVU for $g(\theta)$. Let U be unbiased for 0. Let $s \in \mathbb{R}$ and consider $W + sU$, an unbiased estimator for $g(\theta)$ again. Then

$$\mathrm{var}_\theta(W) \leqslant \mathrm{var}_\theta(W + sU) = \mathrm{var}_\theta(W) + 2s\mathbb{E}_\theta(W - \mathbb{E}_\theta W)U + s^2\,\mathrm{var}_\theta(U).$$

The minimum value occurs at $s = 0$ if and only if the derivative vanishes at $s = 0$. That is, $\mathbb{E}_\theta WU = \mathbb{E}_\theta(W - \mathbb{E}_\theta W)U = 0$.

Conversely, assume $\mathbb{E}_\theta(WU) = 0$ for all $U \in L_2(\Omega)$ unbiased for 0. If Y is unbiased, then $U := Y - W$ is unbiased for 0. Comparing the variance of Y with $W + U$ we have

$$\mathrm{var}_\theta(Y) = \mathrm{var}_\theta(U + W) = \ldots = \mathrm{var}_\theta(U) + \mathrm{var}_\theta(W) \geqslant \mathrm{var}_\theta(U).$$

□

6.3 Fisher Information & Cramér-Rao

In this section we assume $\Theta \subset \mathbb{R}$ unless otherwise specified.

Definition: (6.19) Fisher Information

Let $\{f_\theta : \theta \in \Theta\}$ be a family of multivariate PDFs or PMFs. Let X be a random vector with distribution f_θ. The **Fisher information** of the family is defined to be

$$I(\theta) = I_X(\theta) := \mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)^2 \qquad \text{for all } \theta \in \Theta$$

if this quantity exists and is finite. We also implicitly assume that $\{x \in \mathbb{R} : f_\theta(x) > 0\}$ does not depend on θ.

Beginning of March 9, 2022

Example: (6.20) Gaussians & Fisher. Let $\sigma > 0$. Let $f_\theta(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$ for all $x \in \mathbb{R}$, $\theta \in \mathbb{R}$. Then we have

$$\log f_\theta(x) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \cdot -\frac{(x-\theta)^2}{2\sigma^2}$$

so

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) = \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{-(X-\theta)^2}{2\sigma^2},$$

and so

$$I(\theta) = \mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{-(X-\theta)^2}{2\sigma^2}\right)^2 = \mathbb{E}_\theta\left(\frac{X-\theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^4}\operatorname{var}(X-\theta) = \frac{1}{\sigma^2}.$$

In general, $I(\theta)$ depends on θ, but in this case it does not. Here, when σ is small, f_θ looks like a sharp bump rather than a flat curve. A smaller σ corresponds to a larger $I(\theta)$ which gives us more information about where and how the random variable is distributed. Later we will establish the Cramér-Rao bound and draw connection between Fisher information and UMVU.

We now provide two alternate forms for the Fisher information which might be useful sometimes:

Remark. Without the square,

$$\mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right) = \int_{\mathbb{R}^n} \frac{\mathrm{d}/\mathrm{d}\theta f_\theta(x)}{f_\theta(x)} f_\theta(x)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathbb{R}^n} f_\theta(x)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta}(1) = 0.$$

Therefore, treating $\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)$ as a random variable,

$$I(\theta) = \mathbb{E}_\theta(\dots)^2 = \operatorname{var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right).$$

Remark. Alternatively,

$$\begin{aligned}
\mathbb{E}_\theta\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f_\theta(X)\right) &= \int_{\mathbb{R}^n} \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\mathrm{d}/\mathrm{d}\theta f_\theta(x)}{f_\theta(x)} f_\theta(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^n} \frac{f_\theta(x)\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f_\theta(x) - (\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x))^2}{(f_\theta(x))^2} f_\theta(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^n} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f_\theta(x) - \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(x)\right)^2 f_\theta(x)\,\mathrm{d}x \\
&= \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(1) - \int_{\mathbb{R}^n}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(x)\right)^2 f_\theta(x)\,\mathrm{d}x = 0 - I(\theta) = -I(\theta).
\end{aligned}$$

Proposition: (6.21)

Let X, Y be independent where their distributions are from $\{f_\theta : \theta \in \Theta\}$ and $\{g_\theta : \theta \in \Theta\}$ respectively (not necessarily the same distribution, but same parameter space). Then

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$

Proof. Using the variance expression,

$$\begin{aligned} I_{(X,Y)}(\theta) &\overset{*}{=} \operatorname{var}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f_\theta(X)g_\theta(Y))\right) = \operatorname{var}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(X) + \log g_\theta(X)\right) \\ &\overset{*}{=} \operatorname{var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right) + \operatorname{var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log g_\theta(X)\right) = I_X(\theta) + I_Y(\theta). \end{aligned}$$

(The starred equations are because of independence.) □

Theorem: (6.23) Cramér-Rao / Information Inequality

Let $X : \Omega \to \mathbb{R}^n$ be a random variable with distribution from $\{f_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$. Let $Y := t(X)$ be a statistic. For $\theta \in \Theta$, define $g(\theta) := \mathbb{E}_\theta Y$. Then

$$\operatorname{var}_\theta(Y) \geqslant \frac{|g'(\theta)|^2}{I_X(\theta)} \qquad \text{for all } \theta \in \Theta.$$

In particular if Y is *unbiased* then $g(\theta) = \theta$ and $g'(\theta) = 1$, so

$$\operatorname{var}_\theta(Y) \geqslant \frac{1}{I_X(\theta)} \qquad \text{for all } \theta \in \Theta.$$

In both cases, "=" happens only when $\dfrac{\mathrm{d}/\mathrm{d}\theta(\log f_\theta(X))}{Y - \mathbb{E}_\theta Y} \in \mathbb{R}$ for some $\theta \in \Theta$.

This theorem provides a lower bound on the variance of unbiased estimators of θ — in general, we cannot get estimators with arbitrarily small variance.

Remark. If $X_1, \dots, X_n$ are i.i.d. and $X = (X_1, \dots, X_n)$, then (by last proposition) $I_X(\theta) = nI_{X_1}(\theta)$. If $\mathbb{E}_\theta Y = \theta$, then $\operatorname{var}_\theta(Y) \geqslant 1/(nI_{X_1}(\theta))$ for all $\theta \in \Theta$.

Proof. Define $g(\theta)$, Y, and t accordingly. If X is continuous (similar for discrete),

$$\begin{aligned} |g'(\theta)| &= \left|\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathbb{R}^n} f_\theta(x)t(x)\,\mathrm{d}x\right| = \left|\int_{\mathbb{R}^n}\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)t(x)\,\mathrm{d}x\right| \\ &\underset{*}{=} \left|\int_{\mathbb{R}^n}\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(x))\,t(x)f_\theta(x)\,\mathrm{d}x\right| \\ &\overset{*}{=} \left|\operatorname{cov}(\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(X)), t(X))\right| \\ &\leqslant \left(\operatorname{var}_\theta(\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(X)))\right)^{1/2} \operatorname{var}_\theta(t(X))^{1/2} \\ &= \sqrt{I_X(\theta)}\sqrt{\operatorname{var}_\theta Y}. \end{aligned}$$

For $\underset{*}{=}$: $\dfrac{\mathrm{d}}{\mathrm{d}\theta}(\log f_\theta(x)) = \dfrac{1}{f_\theta(x)}\dfrac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)$ [note that $t(x)$ is treated as a constant when doing $\mathrm{d}/\mathrm{d}\theta$], and for $\overset{*}{=}$: if $\mathbb{E}W = 0$, then $\operatorname{cov}(W, Z) = \mathbb{E}[(W - \mathbb{E}W)(Z - \mathbb{E}Z)] = \mathbb{E}[W(Z - \mathbb{E}Z)] = \mathbb{E}(WZ)$.

Note that equality in Cramér-Rao happens if and only if the Cauchy-Schwarz step is attained, i.e., when

$$\frac{\mathrm{d}/\mathrm{d}\theta(\log f_\theta(X)) - \mathbb{E}(\ldots)}{t(X) - \mathbb{E}(t_\theta(X))} = \frac{\mathrm{d}/\mathrm{d}\theta(\log f_\theta(X))}{Y - \mathbb{E}_\theta Y} \text{ is a constant.}$$

□

Example: (6.24). Let $f_\theta(x) := \theta x^{\theta-1}\chi_{(0,1)}(x)$ for $x \in \mathbb{R}$ and $\theta > 0$. Then for $x \in (0,1)$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(x) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log(\theta x^{\theta-1}) = \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\log\theta + (\theta-1)\log x\right] = \frac{1}{\theta} + \log x.$$

Then if $X_1, \ldots, X_n$ are i.i.d., for $(x_1, \ldots, x_n) \in (0,1)^n$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^n f_\theta(x_i) = \sum_{i=1}^n(\theta^{-1} + \log x_i) = n\left(\frac{1}{\theta} + \frac{1}{n}\log\sum_{i=1}^n x_i\right).$$

By Cramér-Rao, any multiple of $\frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^n f_\theta(X_i)$ (plus a constant) is UMVU for $\mathbb{E}_\theta Y$.

For example, since $\mathbb{E}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log\prod_{i=1}^n f_\theta(X_i)) = 0$, we know $\mathbb{E}\sum_{i=1}^n \log X_i = -n/\theta$. Hence if we define $Y := -\frac{1}{n}\log\prod_{i=1}^n X_i$, its expected valve is $1/\theta$, and we claim that this is UMVU of its expectation.

6.4 Bayes Estimation

Beginning of March 21, 2022

In **Bayes estimation**, the unknown $\theta \in \Theta$ *itself* is regarded as random variable Ψ; the distribution of Ψ represents our **prior** knowledge about its probable values. Given $\Psi = \theta$, the condition distribution of $X \mid \Psi = \theta$ is assumed to be $\{f_\theta : \theta \in \Theta\}$.

Suppose $t : \mathbb{R}^n \to \mathbb{R}^k$, $y = t(X)$, and we have a loss function $L : \Theta \times \mathbb{R}^k \to \mathbb{R}$. Let $g : \Theta \to \mathbb{R}^k$.

Definition: (6.26) Bayes Estimator

A **Bayes estimator** for $g(\theta)$ w.r.t. Ψ is one such that

$$\mathbb{E}L(g(\Psi), Y) \leqslant \mathbb{E}L(g(\Psi), Z) \qquad \text{for all estimators } Z.$$

Proposition: (6.27) Minimizing Conditional Risk ⇒ Bayes

In order to find a Bayes estimator, it suffices to minimize the conditional risk.

Suppose there exists $t : \mathbb{R}^n \to \mathbb{R}$ such that, for almost every $x \in \mathbb{R}^n$, $Y := t(X)$ minimizes the conditional risk

$$\mathbb{E}(L(g(\Psi), Z) \mid X = x)$$

over all estimators Z. Then $t(X)$ is Bayes for $g(\theta)$ w.r.t. Ψ.

Proof. Total expectation. If

$$\mathbb{E}(L(g(\Psi), Z) \mid X = x) \leqslant \mathbb{E}(L(g(\Psi), Z) \mid X = x)$$

for (almost) all x, then taking the expectation again preserves $\leqslant$. *The probability measure is induced by the marginal*

$$\mathbb{P}(X \in A) := \int_{\Omega} \mathbb{P}_{\theta}(X \in A) \,\mathrm{d}\Psi(\theta).$$

The distribution of $t(X)$ can depend on the distribution of Ψ. □

Example: (6.29). Let $n = 1$, $g(\theta) := \theta$, and $L(\Psi, Y) := (\Psi - Y)^2$. The conditional stated above is minimized when $t(x) = E(\Psi \mid X = x)$, since

$$\begin{aligned}\mathbb{E}((\Psi - t(X)^2 \mid X = x)) &= \mathbb{E}(\Psi^2 - 2\Psi t(x) + t(x)^2 \mid X = x)\\ &= \mathbb{E}(\Psi^2 \mid X = x) - 2t(x)\mathbb{E}(\Psi \mid X = x) + t(x)^2.\end{aligned}$$

Therefore $\mathbb{E}(\Psi \mid X)$ is Bayes for θ with respect to Ψ.
Given $\Psi = \theta > 0$, suppose X us uniform on $[0, \theta]$ and assume that Ψ has a gamma distribution with $\alpha = 2, \beta = 1$ so its distribution is θe^{θ} for $\theta > 0$. Then

$$f_{\Psi,X}(\theta, x) = f_{X|\Psi=\theta}(x \mid \theta) f_{\Psi}(\theta) = e^{-\theta} 1_{x\in(0,\theta)}$$

and the marginal of X is

$$f_X(x) = 1_{x>0} \int_{-\infty}^{\infty} f_{\Psi,X}(\theta, x) \,\mathrm{d}\theta = 1_{x>0} \int_{x}^{\infty} e^{-\theta} \,\mathrm{d}\theta = 1_{x>0} \cdot e^{x}.$$

Therefore

$$f_{\Psi|X=x}(\theta \mid x) = \frac{f_{\Psi,X}(\theta, x)}{f_X(x)} = \frac{e^{-\theta} \cdot 1_{x\in(0,\theta)}}{e^{-x} \cdot 1_{x>0}} = e^{x-\theta} \cdot 1_{x\in(0,\theta)}$$

and so

$$\mathbb{E}(\Psi \mid X = x) = \int_{-\infty}^{\infty} \theta f_{\Psi|X=x}(\theta \mid x) \,\mathrm{d}\theta = \int_{x}^{\infty} \theta e^{x-\theta} \,\mathrm{d}\theta = e^{x}((x+1)e^{-x}) = x + 1,$$

which says that the Bayes estimator for the **mean squared error** (MSE) $L(\Psi, Y) = (\Psi - Y)^2$ is in this case $t(X) = X + 1$.
In contrast, the UMVU for θ is $(1 + 1/n)X_{(n)}$ and in this case $2X$.

Beginning of March 23, 2022

6.5 Method of Moments

Definition: (6.30) Consistency

Let $\{f_\theta : \theta \in \Theta\}$ be a family of distributions and let $Y_1, Y_2, \ldots$ be a sequence of estimators for $g(\theta)$. We say $Y_1, Y_2, \ldots$ is **consistent** for $g_9\theta)$ if, for any $\theta \in \Theta$, $Y_1, Y_2, \ldots$ converges in probability to the constant value $g_9\theta)$.

Remark. If $h : \mathbb{R}$ is continuous, and if $Y_1, Y_2, \ldots$ converges in probability to $c \in \mathbb{R}$, then $h(Y_1), h(Y_2), \ldots$ converges in probability to $h(c)$.

Example: (6.31). Let $X_1, \ldots, X_n$ be a sample of size n with distribution f_θ. The WLLN states that the sample mean is consistent when $\mathbb{E}_\theta |X_1| < \infty$ for all $\theta \in \Theta$. The same holds for the j^{th} moment given that $\mathbb{E}_\theta |X_1|^j < \infty$ for all $\theta \in \Theta$. If we define

$$\mu_j(\theta) := \mathbb{E}X_1^j \qquad \text{and } M_j(\theta) := \frac{1}{n}\sum_{i=1}^n X_i^j$$

then $M_j(\theta)$ converges in probability to $\mu_j(\theta)$. This gives rise to the Method of Moments.

Definition: (6.32) Methods of Moments

Suppose we want to estimate $g(\theta)$ and suppose there exists $h : \mathbb{R}^j \to \mathbb{R}^k$ such that

$$g(\theta) = h(\mu_1, \ldots, \mu_j).$$

Then the estimator $h(M_1, \ldots, M_j)$ is called the **method of moments** estimator for $g(\theta)$.

Example: (6.33). Let $g(\theta)$ be the variance. We know $\operatorname{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$. Then the MoM for $g(\theta)$ is $M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$.

Example: Consistent but Biased Estimator. Following the previous example, define

$$Y_n := \sqrt{\sum_{i=1}^n X_i^2/n - (\sum_{i=1}^n X_i/n)^2}.$$

Since $(a, b) \mapsto \sqrt{a - b^2}$ is continuous, and since $\sum_{i=1}^n X_i^2/n$ and $\sum_{i=1}^n X_i/n$ converge to $\mathbb{E}X^2$ and $\mathbb{E}X$ respectively, we claim that $Y_n \to \sqrt{\mathbb{E}X^2 - (\mathbb{E}X)^2}$ as $n \to \infty$. This implies that Y_n is *consistent*.
However, Y_n *is* biased! Take $n = 1$ and X the uniform distribution on $[0, 1]$. Then

$$\mathbb{E}X = \frac{1}{2}, \mathbb{E}X^2 = \frac{1}{3}, \operatorname{var}(X) = \frac{1}{12}, \text{ and } \sigma = \frac{1}{2\sqrt{3}}.$$

On the other hand,

$$\mathbb{E}\sqrt{X^2 - X^2} = 0.$$

Therefore Y_n is *consistent but biased*.

Beginning of March 25, 2022

Example: (6.34). Let $X_1, \ldots, X_n$ be a random sample of size n from $[0, \theta]$ where $\theta > 0$ is unknown. Previously we mentioned that $(1 + 1/n)X_{(n)}$ is UMVU for θ. Ont he other hand, $\mathbb{E}_\theta X_1 = \theta/2$ so the MoM estimator is $2/n \cdot \sum_{i=1}^{n} X_i$. The variance of this estimator is

$$\frac{4}{n^2}\sum_{i=1}^{n}\operatorname{var}(X_i) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

The variance for the UMVU is

$$\begin{aligned}\operatorname{var}((1+1/n)X_{(n)}) &= \left(\frac{n+1}{n}\right)^2 \operatorname{var}(X_{(n)}) = \frac{(n+1)^2}{n^2}\mathbb{E}X_{(n)}^2 - \theta^2 \\ &= \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbb{P}(X_{(n)} > t)\,\mathrm{d}t - \theta^2 = \ldots = \frac{\theta^2}{n(n+2)}.\end{aligned}$$

From this we see that MoM might not be too good in terms of variance, in addition to its possibility of not being biased.

Example: (6.35). Suppose we have a binomial random variable with known parameters n, p where $0 < p < 1$. Then $\mathbb{E}X_1 = np$ and $\mathbb{E}X_1^2 = np(1-p) + n^2p^2$. Some algebra shows that $n = M_1/N$, where

$$N := \frac{M_1^2}{M_1 - (M_2 - M_1^2)}.$$

6.6 Maximum Likelihood Esimation

Beginning of March 28, 2022

Definition: (6.36) Maxiimum Likelihood Estimator, MLE

Let $X_1, \ldots, X_n$ be a random sample from f_θ where $\theta \in \Theta$. If $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we define the **likelihood function** $\ell : \Theta \to [0, \infty)$ to be

$$\ell(\theta) := \prod_{i=1}^{n} f_\theta(x_i).$$

The **maximum likelihood estimator**, MLE, Y, is the estimator maximizing the likelihood.

Remark. MLE might not exist. Even if it exists, it might not be unique and can in fact have uncountably many.

For the nonexistent one: let $f_\theta(x) := \theta \cdot 1_{[0,1/\theta]}(x)$ where $\theta \in \mathbb{N}$. Then $\ell(\theta) = \theta$ has no maximum over $\theta \in \mathbb{N}$. However, note that if f_θ is continuous and Θ compact, then MLE at least exists.

For the uncountable one, let $f_\theta(x_1) := 1_{[\theta,\theta+1]}(x_1)$ for x_1 and unknown $\theta \in \mathbb{R}$. Then

$$\ell(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} 1_{[\theta \leqslant x_{(1)} \leqslant x_{(n)} \leqslant \theta+1}.$$

If $x_1 = \ldots = x_n = 0$, then

$$\ell(\theta) = 1_{\theta \in [-1,0]}.$$

That is, any $\theta \in [-1, 0]$ works as a MLE in this case.

Remark. We will show later that under certain conditions MLE is consistent and will have the optimal variance as $n \to \infty$.

Definition: (6.40) Log Concavity ⇒ Uniqueness of MLE If It Exists

If each function $\theta \mapsto f_\theta(x_i)$ is strictly log-concave, then for $x_1, \ldots, x_n \in \mathbb{R}$, then likelihood function has at most maximum value.
Note that this does not guarantee existence — for example e^{-x} is log-concave but does not have maximum on $\mathbb{R}$.

Beginning of March 30, 2022

Example: (6.45 MLE and Gaussian). Consider a Gaussian with unknown $\mu \in \mathbb{R}$ and unknown $\sigma^2 > 0$ so $\theta = (\mu, \sigma)$. Suppose we want to find the MLE for the pair (μ, θ). Here we maximize $\log \ell(\theta)$:

$$\log \ell(\theta) = \log \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \sum_{i=1}^n \left[-\log\sigma - \frac{\log 2\pi}{2} - \frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

Computing its partials,

$$\frac{\partial}{\partial\mu}\log\ell(\theta) = \frac{x_i-\mu}{\sigma^2} \qquad \frac{\partial}{\partial\sigma}\log\ell(\theta) = \sum_{i=1}^n -\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}.$$

Setting them to 0, we obtain

$$\mu = \frac{1}{n}\sum_{i=1}^n x_i \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2.$$

(Note that we did not get $1/(n-1)$ for σ^2, but nevertheless this is still pretty good.)
Now that we found a critical point, we need to verify that it is a maximum. Write $\alpha := 1/\sigma^2$. Then

$$\log\ell(\theta) = \frac{1}{2}(\sum_{i=1}^n \log\alpha - \log 2\pi - \alpha(x_i-\mu)^2)$$

For fixed α, $\log\ell(\theta)$ is strictly concave function of μ; likewise, fixing μ, $\log\ell(\theta)$ is a strictly concave function of α (alternatively, do first derivative test on σ), so the critical point must have been a global maximum. We have therefore found *the* (only) MLE:

$$\theta = (\mu, \sigma^2) = \left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{i=1}^n X_i)^2\right).$$

Note that such MLE is biased for σ^2 but asymptotically unbiased.

Beginning of April 8, 2022

Theorem: (6.52) Consistency of MLE

Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^n$ be i.i.d. with pdf f_θ. Suppose Θ is compact and $f_\theta(x_1)$ is a continuous function for θ for a.e. $x_1 \in \mathbb{R}$. Assume $\mathbb{E}_\theta \sup_{\theta' \in \Theta} |\log f_{\theta'}(X_1)| < \infty$ and $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$ for all $\theta' \neq \theta$. Then the MLE Y_n of θ converges in probability to the constant function θ with respect to $\mathbb{P}_\theta$.

***Proof for finiteh* Θ.** Fix $\theta \in \Theta$. For $\theta' \in \Theta$ and $n \geqslant 1$, let

$$\ell_n(\theta') := \frac{1}{n}\sum_{i=1}^{n} \log f_{\theta'}(X_i).$$

Note that each $\log f_{\theta'}(X_i)$ is a random variable with finite expectation, so by WLLN, $\ell_n(\theta')$ converges in probability with respect to $\mathbb{P}_\theta$ to the constant $\mu(\theta') := \mathbb{E}_\theta \log f_{\theta'}(X_1)$.

Enumerate Θ as $\{\theta, \theta_1, ..., \theta_k\}$. Since $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$ for all $\theta' \neq \theta$, we have by information inequality that $I(\theta, \theta') = \mu(\theta) - \mu(\theta') > 0$.

For $n \geqslant 1$, define

$$\Omega \supset A_n := \{\ell_n(\theta) > \ell_n(\theta_j) \text{ for all } 1 \leqslant j \leqslant k\}$$

Then $\lim\limits_{n\to\infty} \mathbb{P}_\theta(A_n) = 1$ because $\ell_n(\theta) \to \mu(\theta)$ in probability and $\ell_n(\theta_j) \to \mu(\theta') < \mu(\theta)$ in probability for each j and there are only finitely many j's. (For infinite case the proof needs to be modified). By convergence in probability,

$$\lim_{n\to\infty} \mathbb{P}_\theta(|\ell_n(\theta) - \mu(\theta)| > \epsilon) = \lim_{n\to\infty} \mathbb{P}_\theta(|\ell_n(\theta') - \mu(\theta')| > \epsilon_0 = 0.$$

Using triangle inequality,

$$|\ell_n(\theta) - \ell_n(\theta_j)| = |\ell_n(\theta) - \mu(\theta) + \mu(\theta) - \mu(\theta_j) + \mu(\theta_j) - \ell_n(\theta_j)|$$

where the first two terms are $< \epsilon$, last two $< \epsilon$, and the middle two can be $> 3\epsilon$ for small ϵ. Then the entire thing $> \epsilon$. Taking maximum index over all j's again,

$$\lim_{n\to\infty} \mathbb{P}_\theta(|\ell_n(\theta) - \ell_n(\theta_j)| > \epsilon \text{ for all } 1 \leqslant j \leqslant k) = \lim_{n\to\infty} \mathbb{P}_\theta(A_n) = 1.$$

On each A_n, the MLE Y_n is well-defined and unique with $Y_n = \theta$, so $\{Y_n = \theta\}^c \subset A_n^c$. Using $\lim\limits_{n\to\infty} \mathbb{P}(A_n) = 1$ we have

$$\lim_{n\to\infty} \mathbb{P}_\theta(|Y_n - \theta| > \epsilon) \leqslant \lim_{n\to\infty} \mathbb{P}_\theta(A_n^c) = 0.$$

$\square$

Beginning of April 11, 2022

We now give a powerful theorem on the asymptotic variance of MLE and claim that it achieves it asymptotically achieve the Cramér-Rao lower bound.

Theorem: (6.53) Limiting Distribution of MLE

(Think of this as an analogue to the CLT/Delta.) Let $\{f_\theta : \theta \in \Theta\}$ be a family of PDFs with $f_\theta : \mathbb{R} \to [0, \infty)$ for all θ. Let $X_1, X_2, \ldots$ be i.i.d. with distribution f_θ. Assume that

(1) The set $A := \{x \in \mathbb{R} : f_\theta(x) > 0\}$ is independent of θ,

(2) For every $x \in A$, $\partial^2 f_\theta(x)/\partial\theta^2$ exists and is continuous in θ,

(3) The Fisher information $I_{X_1}(\theta)$ exists and is finite with

$$\mathbb{E}_\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X_1) = 0 \qquad \text{and} \qquad I_{X_1}(\theta) = -\mathbb{E}_\theta \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f_\theta(X_1) > 0,$$

(4) For every θ in the interior of Θ, there exists $\delta > 0$ such that

$$\mathbb{E}_\theta \sup_{\theta' \in \Theta} \left| 1_{[\theta-\delta,\theta+\delta]} \frac{\mathrm{d}^2}{\mathrm{d}[\theta']^2} \log f_{\theta'}(X_1) \right| < \infty,$$

and

(5) The MLE Y_n of θ is consistent.

Then, for any θ in the interior of Θ, as $n \to \infty$, $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian with variance $1/I_{X_1}(\theta)$ w.r.t. $\mathbb{P}_\theta$.

Proof. We assume Θ is finite for simplicity (in which case (4) is trivial). Fix $\theta \in \Theta$.
Define the log-likelihood to be

$$\ell_n(\theta') := \frac{1}{n} \sum_{i=1}^{n} \log f_{\theta'}(X_i).$$

Assuming Θ is finite, let $\epsilon > 0$ be small so that $[\theta - \epsilon, \theta + \epsilon] \cap \Theta = \{\theta\}$. Let A_n be the event where $Y_n = \theta$, and by (5) we have $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$. Since Y_n is MLE, we have $\ell_n'(Y_n) = 0$ on Y_n (assuming the notion of derivative works in a finite domain, thought in actuality it doesn't). Taylor expansion gives

$$0 = \ell_n'(Y_n) = \ell_n'(\theta) + \ell_n''(Z_n)(Y_n - \theta) \qquad \text{if } A_n \text{ occurs},$$

for some Z_n always lying between θ and Y_n. Therefore

$$\sqrt{n}(Y_n - \theta) = \frac{\sqrt{n}\ell_n'(\theta)}{-\ell_n''(Z_n)} \qquad \text{if } A_n \text{ occurs.} \tag{*}$$

By (3), each term in $\ell_n'(\theta)$ has mean zero and variance $I_{X_1}(\theta)$, so $\sqrt{n}\ell_n'(\theta)$ converges in distribution to a mean zero Gaussian with variance $I_{X_1}(\theta)$ by CLT.
For the denominator, first note that by (5), Y_n converges to θ, the constant. Then by (4) and WLLN, $\ell_n''(\theta)$ converges in probability to $\mathbb{E}_\theta \ell_n''(\theta)$, where $\ell_n''(\theta)$ is simply a fixed value. Therefore the denominator converges in probability to $\mathbb{E}_\theta \ell_n''(\theta) = -I_{X_1}(\theta)$. Therefore, (*) implies that $\sqrt{n}(Y_n - \theta)$ converges to a Gaussian with mean 0 and variance $1/I_{X_1}(\theta)$, as claimed. □

Beginning of April 13, 2022

6.7 EM Algorithm

Let $X : \Omega \to \mathbb{R}^n$ be a random variable. Let $h : \mathbb{R}^n \to \mathbb{R}^m$ be non-invertible and let $Y := t(X)$. Sometimes we want to ideally observe the sample X but in really we only have access to Y.
Suppose X has a distribution from $\{f_\theta : \theta \in \Theta\}$. To find the MLE of θ, we want to maximize

$$\log \ell(\theta) = \log f_\theta(X).$$

Yet, since X cannot be directly observed we cannot maximize the above. Instead, we try to approximate the maximum value by conditioning on Y.

Definition 6.7.1: Expectation-Maximization Algorithm

Initialize $\theta_0 \in \Theta$. Fix $k \geqslant 1$. For $1 \leqslant j \leqslant k$, repeat the following procedure:

(1) (Expectation) Given θ_{j-1}, let $\varphi_j(\theta) := \mathbb{E}_{\theta_{j-1}}(\log f_\theta(X) \mid Y)$, and

(2) (Maximization) Define $\theta_j := \operatorname{argmax} \varphi_j(\theta)$.

Beginning of April 15, 2022

A few examples:

(1) If $Y = X$ the whole sample then Y is sufficient. We have $\varphi_1(\theta) = \log f_\theta(X)$ so we get MLE in one run.

(2) If Y is constant, $\varphi_1(\theta) = \mathbb{E}_{\theta 0} \log f_\theta(X)$. We get $\theta = \theta_0$ in one run according to the likelihood inequality, and we keep getting this result iteratively.

(3) Let $t(x_1, \ldots, x_n) = (x_1, \ldots, x_m)$ where $m < n$. Then

$$\begin{aligned}
\varphi_j(\theta) &= \mathbb{E}_{\theta_{j-1}}\Big(\sum_{i=1}^n \log f_\theta(X_i) \mid (X_1, \ldots, X_m)\Big) \\
&= \mathbb{E}_{\theta_{j-1}}\Big(\sum_{i=1}^m \log f_\theta(X_i) \mid (X_1, \ldots, X_m)\Big) + \mathbb{E}_{\theta_{j-1}}\Big(\sum_{i=m+1}^n \log f_\theta(X_i) \mid (X_1, \ldots, X_m)\Big) \\
&= \sum_{i=1}^m \log f_\theta(X_i) + \mathbb{E}_{\theta_{j-1}} \sum_{i=m+1}^n \log f_\theta(X_i).
\end{aligned}$$

We now provide a “measure of progress” of the EM algorithm.

Proposition: (6.58)

Suppose X has density f_θ and $Y := t(X)$ has density h_θ. We denote $g_\theta(x \mid y) := f_{X|Y}(x \mid y)$. Then for any $\theta \in \Theta$,

$$\log h_\theta(Y) - \log h_{\theta_{j-1}}(Y) \geqslant \varphi_j(\theta) - \varphi_j(\theta_{j-1})$$

with equality only when $g_\theta(X \mid y) = g_{\theta_{j-1}}(X \mid y)$ a.s. w.r.t. $\mathbb{P}_{\theta_{j-1}}$ for fixed y.

Proof. Since $f_{X,Y}(x,y) = f_{X|Y}(x \mid y) f_Y(y)$, we have

$$\log f_Y(y) = \log f_{X,Y}(x,y) - \log f_{X|Y}(x \mid y).$$

Since $Y = t(X)$, we have $f_{X,Y}(x,y) = f_X(x) 1_{y=t(x)}$. Hence, when $y = t(x)$,

$$\log f_Y(y) = \log f_X(x) - \log f_{X|Y}(x \mid y) = \log f_\theta(x) - \log f_{X|Y}(x \mid y).$$

That is,

$$\log h_\theta(y) = \log f_\theta(x) - \log g_\theta(x \mid y).$$

Multiplying by $h_{\theta_{j-1}}(x \mid y)$ and integrating in x, we have

$$\mathbb{E}_{\theta_{j-1}}(\log h_\theta(Y) \mid Y = y) = \mathbb{E}_{\theta_{j-1}}(\log f_\theta(X) \mid Y = y) - \mathbb{E}_{\theta_{j-1}}(\log g_\theta(X \mid y) \mid Y = y) \qquad \text{for all } \theta \in \Theta.$$

Since the above holds for any θ, in particular we can set $\theta := \theta_{j-1}$. Note that the first term is simply $\log h_\theta(y)$. Subtracting gvies

$$\begin{aligned}\log h_\theta(y) - \log h_{\theta_{j-1}}(y) &= \mathbb{E}_{\theta_{j-1}}(\log f_\theta(X) \mid Y = y) - \mathbb{E}_{\theta_{j-1}}(\log f_{\theta_{j-1}}(X) \mid Y = y) \\ &\quad - \mathbb{E}_{\theta_{j-1}}(\log g_\theta(X \mid y) \mid Y = y) + \mathbb{E}_{\theta_{j-1}}(\log g_{\theta_{j-1}}(X \mid y) \mid Y = y).\end{aligned}$$

By likelihood inequality, the sum of the last two terms should be positive, and we recover our claim. □

Proposition: (6.59) EM Algorithm Improvement

Let $\theta_1, \ldots, \theta_k$ be an output of the EM algorithm. Then for all $1 \leqslant j \leqslant k$,

$$\log h_{\theta_j}(Y) \geqslant \log h_{\theta_{j-1}}(Y).$$

Moreover, equality occurs only when $g_{\theta_j}(X \mid y) = g_{\theta_{j-1}}(X \mid y)$ a.e. w.r.t. $\mathbb{P}_{\theta_{j-1}}$ for fixed y or when $\theta_j = \theta_{j-1}$.

Chapter 7

Resampling & Bias Reduction

Idea. For a fixed sample size n, there are ways to reduce the bias of an estimator on n samples by re-sampling from the n samples given.

7.1 Jackknife Resampling

Definition: (7.1) Jackknife Estimator

Let $X_1, X_2, \dots : \Omega \to \mathbb{R}$ be i.i.d. with distribution $f_\theta : \mathbb{R}^n \to [0, \infty)$. Suppose $Y_1, Y_2, \dots$ are estimators for θ so that $Y_n = t_n(X_1, \dots, X_n)$. For $n \geqslant 1$, we define the **jackknife estimator** of Y_n to be

$$Z_n := nY_n - \frac{n-1}{n} \sum_{i=1}^{n} t_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

Proposition: (7.2) Jackknife Reduces Bias

Suppose there exist $a, b \in \mathbb{R}$ such that

$$\mathbb{E}Y_n = \theta + \frac{a}{n} + \frac{b}{n^2} + \mathcal{O}(1/n^3).$$

Then

$$\mathbb{E}Z_n = \theta + \mathcal{O}(1/n^2)$$

and if $b = 0$ and $\mathcal{O}(1/n^3) = 0$ then Z_n is unbiased.

Proof.

$$\begin{aligned}
\mathbb{E}Z_n &= n\theta + a + \frac{b}{n} + \mathcal{O}(1/n^2) - \frac{n-1}{n} \sum_{i=1}^{n} \mathbb{E}t_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \\
&= n\theta + a + \frac{b}{n} + \mathcal{O}(1/n^2) - \frac{n-1}{n} \sum_{i=1}^{n} \left(\theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + \mathcal{O}(1/n^3) \right) \\
&= \theta + \frac{b}{n} - \frac{b}{n-1} + \mathcal{O}(1/n^2) = \theta + \mathcal{O}(1/n^2).
\end{aligned}$$

□

Example: (7.3) Jackknife and Sample Mean. The jackknife estimator of the sample mean is the sample mean:

$$\sum_{i=1}^{n} X_i - \frac{n-1}{n} \sum_{i=1}^{n} \frac{1}{n-1} \sum_{j \neq i} X_j = \sum_{i=1}^{n} X_i - \frac{n-1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Example: (7.4) Jackknife and Bernoulli. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli with parameter $\theta \in (0,1)$. Then the MLE for θ is the sample mean so that for θ^2 is simply sample mean squared $Y_n := \left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)^2$. Then

$$\mathbb{E}Y_n = \frac{1}{n^2}(n\theta + n(n-1)\theta^2) = \theta^2 + \frac{\theta - \theta^2}{n}$$

so the corresponding jackknife estimator is unbiased for θ^2.

Chapter 8

Concentration of Measure

Beginning of April 22, 2022

Theorem: (8.1) Hoeffding Inequality

Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Let $a_1, a_2, \dots \in \mathbb{R}$. Then for $n \geqslant 1$ and $t \geqslant 0$,

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right) \qquad \text{and therefore} \qquad \mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right).$$

Proof. We may assume $\sum_{i=1}^n a_i^2 = 1$. Let $\alpha > 0$. Then

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geqslant t\right) &= \mathbb{P}\left(\exp\left(\alpha \sum_{i=1}^n a_i X_i\right) \geqslant e^{\alpha t}\right) \\
&\leqslant e^{-\alpha t}\mathbb{E}\exp\left(\alpha \sum_{i=1}^n a_i X_i\right) = e^{-\alpha t}\mathbb{E}\prod_{i=1}^n e^{\alpha a_i X_i} = e^{-\alpha t}\prod_{i=1}^n \mathbb{E}e^{\alpha a_i X_i} \\
&= e^{-\alpha t}\prod_{i=1}^n \frac{e^{\alpha a_i} + e^{-\alpha a_i}}{2} = e^{-\alpha t}\prod_{i=1}^n \cosh(\alpha a_i) \\
&\leqslant e^{-\alpha t}\prod_{i=1}^n e^{\alpha^2 a_i^2/2} = e^{-\alpha t + \alpha^2/2}.
\end{aligned}$$

The LHS is independent of α. Letting $\alpha := t$ we have $\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geqslant t\right) \leqslant e^{-t^2 + t^2/2} = e^{-t^2/2}$. □

Theorem: (8.3) Chernoff Inequality

Let $0 < p < 1$ and let $X_1, X_2, \dots$ be i.i.d. Bernoulli. Then for $n \geqslant 1$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geqslant t\right) \leqslant e^{-np}\left(\frac{ep}{t}\right)^{tn} \qquad \text{for } t \geqslant p.$$

Theorem: (8.5) Concentration of Measure for Gaussians

Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz with constant 1, i.e., $|f(x) - f(y)| \leqslant \|x - y\|$. Let $X = (X_1, \dots, X_n)$ be a mean zero Gaussian random vector with identity covariance matrix (or i.i.d. standard Gaussians). Then for $t > 0$,

$$\mathbb{P}(x \in \mathbb{R}^n : |f(x) - \mathbb{E}f(X)| \geqslant t) \leqslant 2e^{-2t^2/\pi^2}.$$

Beginning of April 25, 2022

Proof. We assume all partial derivatives of f exist and are continuous. Let $Y = (Y_1, \dots, Y_n)$ be another mean zero Gaussian vector with identity covariance matrix and X and Y are independent. Then, for $\theta \in [0, \pi/2]$ define

$$Z_\theta := X \sin\theta + Y \cos\theta.$$

We have

$$\frac{\mathrm{d}}{\mathrm{d}\theta} Z_\theta = X\cos\theta - Y\sin\theta.$$

Note that $X_1 \sin\theta + Y_1\cos\theta$ is a Gaussian with mean zero and variance 1, and so is $X_1\cos\theta - Y_1\sin\theta$. But then their covariance is

$$\begin{aligned}\mathbb{E}(X_1\sin\theta + Y_1\cos\theta)(X_1\cos\theta - Y_1\sin\theta) &= \mathbb{E}X_1^2\sin\theta\cos\theta - \mathbb{E}Y_1^2\sin\theta\cos\theta - \mathbb{E}X_1Y_1\sin^2\theta + \mathbb{E}X_1Y_1\cos^2\theta \\ &= \mathbb{E}X_1^2\sin\theta\cos\theta - \mathbb{E}Y_1^2\sin\theta\cos\theta - 0 + 0 = 0.\end{aligned}$$

Jointly uncorrelated Gaussians are independent so Z_θ and $\frac{\mathrm{d}}{\mathrm{d}\theta}Z_\theta$ are. Note that $Z_0 = Y$ and $Z_{\pi/2} = X$. Also, since $(\sin\theta, \cos\theta)$ and $(\cos\theta, -\sin\theta)$ are orthogonal, $(Z, \mathrm{d}Z_\theta/\mathrm{d}\theta$ have the same joint distribution as X and Y.

Let $\varphi : \mathbb{R} \to [0, \infty)$ be convex. Then,

$$\begin{aligned}
\mathbb{E}\varphi[f(X) - \mathbb{E}f(Y)] &\leqslant \mathbb{E}\varphi(f(X) - f(Y)) && \text{(Jensen)}\\
&= \mathbb{E}\varphi\left(\int_0^{\pi/2} \frac{\mathrm{d}}{\mathrm{d}\theta} f(Z_\theta)\,\mathrm{d}\theta\right) && \text{(FTC)}\\
&= \mathbb{E}\varphi\left(\int_0^{\pi/2} \left\langle \nabla f(Z_\theta), \frac{\mathrm{d}}{\mathrm{d}\theta}Z_\theta\right\rangle \mathrm{d}\theta\right)\\
&= \mathbb{E}\varphi\left(\frac{1}{\pi/2}\int_0^{\pi/2} \frac{\pi}{2}\left\langle \nabla f(Z_\theta), \frac{\mathrm{d}}{\mathrm{d}\theta}Z_\theta\right\rangle \mathrm{d}\theta\right)\\
&\leqslant \frac{1}{\pi/2}\mathbb{E}\int_0^{\pi/2} \varphi\left(\frac{\pi}{2}\left\langle \nabla f(Z_\theta), \frac{\mathrm{d}}{\mathrm{d}\theta}Z_\theta\right\rangle\right) \mathrm{d}\theta && \text{(Jensen again)}\\
&= \frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}\varphi\left(\frac{\pi}{2}\nabla f(Z_\theta), \frac{\mathrm{d}}{\mathrm{d}\theta}Z_\theta\right) \mathrm{d}\theta && \text{(Fubini)}\\
&= \frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}\varphi\left(\frac{\pi}{2}\langle\nabla f(X), Y\rangle\right) \mathrm{d}\theta && (\langle Z, \mathrm{d}Z_\theta/\mathrm{d}\theta\rangle \sim (X, Y))\\
&= \frac{1}{\pi/2}\frac{\pi}{2}\mathbb{E}\varphi\left(\frac{\pi}{2}\langle\nabla f(X), Y\rangle\right) = \mathbb{E}\varphi\left(\frac{\pi}{2}\langle\nabla f(X), Y\rangle\right).
\end{aligned}$$

Let $\alpha \in \mathbb{R}$ and $\varphi(x) := e^{\alpha x}$ for $x \in \mathbb{R}$. Then

$$\begin{aligned}
\mathbb{E}\exp(\alpha[f(X) - \mathbb{E}f(Y)]) &\leqslant \mathbb{E}\exp\left(\alpha\frac{\pi}{2}\sum_{i=1}^n \frac{\partial f(X)}{\partial x_i}\cdot Y_i\right)\\
&= \mathbb{E}_X\prod_{i=1}^n \mathbb{E}_Y\exp\left(\alpha\frac{\pi}{2}\frac{\partial f(X)}{\partial x_i}\cdot Y_i\right)
\end{aligned}$$

where we can split the expectation of product into product of expected value because the Y_i's are independent (we don't care about the behavior of X_i's in this step).
By the property of MGF, for all $s \in \mathbb{R}$ and all $1 \leqslant i \leqslant n$,

$$\mathbb{E}_Y \exp(sY_i) = e^{s^2/2}.$$

Continuing the inequality above with $s := \alpha \frac{\pi}{2} \frac{\partial f(X)}{\partial x_i}$, we have

$$\mathbb{E} \exp(\alpha[f(X) - \mathbb{E} f(Y)]) \leqslant \mathbb{E} \exp\left(\alpha^2 \frac{\pi^2}{8} \sum_{i=1}^{n} \left(\frac{\partial f(X)}{\partial x_i}\right)^2\right).$$

Since f is 1-Lipschitz, $\|\nabla f(x)\| \leqslant 1$, so we further bound the quantity by $\exp(\alpha^2\pi^2/8)$. Then,

$$\begin{aligned}\mathbb{P}(f(X) - \mathbb{E} f(Y) > t) &= \mathbb{P}(\exp(\alpha[f(X) - \mathbb{E} f(Y)]) > e^{\alpha t}) \\ &\leqslant e^{-\alpha t} \exp(\alpha^2\pi^2/8) = \exp(-\alpha t + \alpha^2\pi^2/8).\end{aligned}$$

Like in Hoeffding, the LHS is independent of α. The RHS is minimized when $\alpha = 4t/\pi^2$, and when so we obtain

$$\mathbb{P}(f(X) - \mathbb{E} f(Y) > t) \leqslant \exp(-2t^2/\pi^2).$$

A symmetric argument to $\mathbb{P}(f(X) - \mathbb{E} f(Y) < -t)$, giving

$$\mathbb{P}(|f(X) - \mathbb{E} f(Y)| > t) \leqslant 2 \exp(-2t^2/\pi^2).$$

$\square$