Probability and Mathematical Statistics

NMSA202 – Lecture Notes

Michal Pešta

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University
Prague

May 3, 2024

Overture

- "*Lasciate ogni speranza o voi ch'entrate.*"
  (Abandon all hope, ye who enter.)

  [Dante Alighieri; Divina Commedia, Inferno]

- "*Und die Pforte ist enge, und der Weg ist schmal, der zum Leben führet; und wenig ist ihrer, die ihn finden.*"
  (How strait is the gate, and narrow the way, that leadeth to life, and there be few that find it.)

  [Matthaeus 7:14; Lutherbibel]

Strait Gate and Narrow Way

- Need of stochastic thinking: 95% & 95% = ?

- Three ordinary issues outside our long path:
  - Derivative should have been defined as log-derivative, but ... Physics
  - Central Limit Theorem is only the Holy pre-Grail, but a non-differentiable continuous function to which everything converges is the Holy Grail
  - Anyone uses neural networks, but just a few knows how to use them

Introductory Course

- $\neq$ Elementary Lecture
- $>$ Preliminary Course
- $<$ Comprehensive Course

Overview

Agenda

Introduction and Motivation

0.1 Literature
0.2 Structure
0.3 Data Science
0.4 Data Mining
0.5 Machine Learning

Literature

- Main source [Wasserman, 2013]
- Czech partial alternative [Dupač and Hušková, 2013]
- Additional and extending material
  [Casella and Berger, 2001, Chung, 2001, Resnick, 2013, Rosenthal, 2006, Ross, 2020]

Blocks of Highlighted Text

Some important text will be highlighted, because it's important.

Definition 0.1 (Name of the definition)

Definitions are in red.

Example 0.2 (Name of the example)

Examples are in green.

Theorem 0.3 (Name of the theorem)

Theorems are in italics and blue.

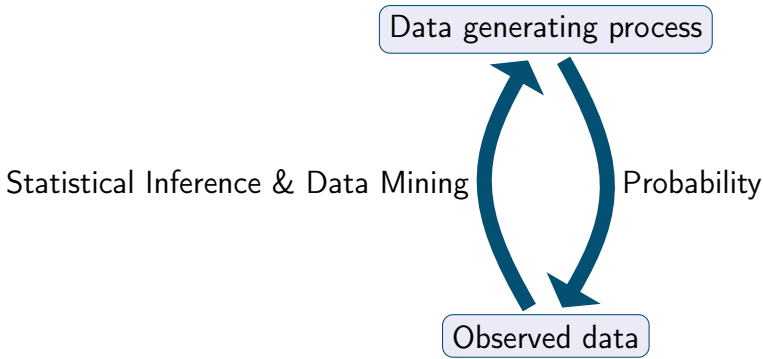
Data Science

= Mathematical Statistics + Algorithmic Computing

- Mathematical Statistics = Justified Statistics ... not just "How?", but also "Why?"

Data Mining

+ Statistical Inference = Stochastic Modeling



Machine Learning

$\subsetneq$ Statistical Learning

Mathematical Statistics	**Machine Learning**
estimation	learning
classification	supervised learning
clustering	unsupervised learning
data	training sample
covariates	features
classifier	hypothesis
hypothesis	—
confidence interval	—

Agenda

1. Probability and Random Events

1.1 Measurable Space
1.2 Probability Space
1.3 Independent Events
1.4 Conditional Probability
1.5 Bayes' Theorem

Measurable Space

The sample space Ω is the set of possible outcomes of an experiment. Points ω in Ω are called sample outcomes, realizations, or elements. Subsets of Ω are called (random) events.

Example 1.1 (Tossing a coin twice)

If we toss a coin twice, then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$.

Definition 1.2 (Measurable space)

Let $\Omega \neq \emptyset$ be some set and let 2^Ω represent its power set. A subset $\mathcal{A} \subseteq 2^\Omega$ is called a σ-algebra or σ-field iff it satisfies

(i) $\emptyset \in \mathcal{A}$;

(ii) if $A \in \mathcal{A}$, then $A^\complement \in \mathcal{A}$;

(iii) if $A_1, A_2, \ldots \in \mathcal{A}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Then, the tuple $(\Omega, \mathcal{A})$ is called a *measurable space*.

Probability Space

We will assign a real number $\mathbb{P}(A)$ to every event $A \in \mathcal{A}$, called the probability of A.

Example 1.3 (Two coin tosses)

Let H_1 be the event that heads occurs on toss 1 and let H_2 be the event that heads occurs on toss 2. If all outcomes are equally likely, then the probability that at least one head occurs (i.e., $H_1 \cup H_2$) is 3/4.

Definition 1.4 (Probability space)

Let $(\Omega, \mathcal{A})$ be a measurable space. A mapping (a set function) $\mathbb{P} : \mathcal{A} \to [0, 1]$ is a probability distribution or a probability measure iff it satisfies

(i) $\mathbb{P}(\Omega) = 1$;

(ii) if $A_1, A_2, \ldots \in \mathcal{A}$ are (pairwise) disjoint, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Then, the triple $(\Omega, \mathcal{A}, \mathbb{P})$ is called a *probability space*.

Basic Properties of Probability

Definition 1.4 clearly implies: $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(A^{\complement}) = 1 - \mathbb{P}(A)$, $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$, $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

Lemma 1.5 (Probability of Union)

For any events $A, B \in \mathcal{A}$,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Theorem 1.6 (Continuity of Probabilities)

Either $A_n \uparrow A$ or $A_n \downarrow A$,

$$\mathbb{P}(A_n) \to \mathbb{P}(A).$$

Finite Sample Spaces

Suppose that the sample space $\Omega = \{\omega_1, \ldots, \omega_n\}$ is finite.

Example 1.7 (Toss a die twice)

Ω has 36 elements: $\Omega = \{(i, j) : i, j \in \{1, \ldots, 6\}\}$. If each outcome is equally likely, then $\mathbb{P}(A) = |A|/36$, where $|A|$ denotes the number of elements in A. The probability that the sum of the dice is 11 is $2/36$, since there are two outcomes that correspond to this event.

If Ω is finite and if each outcome is equally likely, then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|},$$

which is called the uniform probability distribution.

Independence I

Definition 1.8 (Independent Events)

Two events A and B are independent if

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

and we write $A \perp\!\!\!\perp B$. A set of events $\{A_i : i \in I\}$ is independent if

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

for every finite subset $J \subseteq I$. If A and B are not independent, we write $A \not\perp\!\!\!\perp B$.

- Independence can sometimes be assumed (believed in) or sometimes derived (proved)
- Disjoint events with positive probability are not independent

Independent Experiments

Example 1.9 (Toss a fair coin 10 times)

Let $A =$ "at least one head". Let T_j be the event that tails occurs on the jth toss. Then

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^{\mathsf{C}}) \\
&= 1 - \mathbb{P}(\text{all tails}) \\
&= 1 - \mathbb{P}(T_1 \cap \ldots \cap T_{10}) && \text{using independence} \\
&= 1 - \mathbb{P}(T_1) \ldots \mathbb{P}(T_{10}) \\
&= 1 - (1/2)^{10} \approx .999.
\end{aligned}$$

Conditional Probability

Definition 1.10 (Conditional Probability)

If $\mathbb{P}(B) > 0$, then the conditional probability of A given B is

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

- Think of $\mathbb{P}(A|B)$ as the fraction of times A occurs among those in which B occurs. For any fixed B such that $\mathbb{P}(B) > 0$, $\mathbb{P}(\cdot|B)$ is a probability.
- In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.
- A and B are independent iff $\mathbb{P}(A|B) = \mathbb{P}(A)$, given $\mathbb{P}(B) > 0$.
- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$, given $\mathbb{P}(A)\mathbb{P}(B) > 0$.

Sensitivity and Specificity I

Example 1.11 (A medical test for a disease D has outcomes + and −)

The probabilities are:

	D	D^C
+	.009	.099
−	.001	.891

From the definition of conditional probability,

$$\mathbb{P}(+|D) = \frac{\mathbb{P}(+ \cap D)}{\mathbb{P}(D)} = \frac{.009}{.009 + .001} = .9, \; \mathbb{P}(-|D^C) = \frac{\mathbb{P}(- \cap D^C)}{\mathbb{P}(D^C)} = \frac{.891}{.891 + .099} \approx .9.$$

Apparently, the test is fairly accurate. Sick people yield a positive 90% of the time and healthy people yield a negative about 90% of the time.

Sensitivity and Specificity II

Example 1.12 (A medical test for a disease D has outcomes + and − (con't))

Suppose you go for a test and get a positive. What is the probability you have the disease? Most people answer .90. The correct answer is

$$\mathbb{P}(D|+) = \frac{\mathbb{P}(D \cap +)}{\mathbb{P}(+)} = \frac{.009}{.009 + .099} \approx .08.$$

Don't trust your intuition.

However, don't trust a *black box* neither: A randomly selected person is considered (not a symptomatic one).

Law of Total Probability and Bayes' Theorem

Theorem 1.13 (The Law of Total Probability)

Let $A_1, A_2, \ldots$ be a disjoint countable partition of Ω such that $\mathbb{P}(A_i) > 0$ for each $i \in \mathbb{N}$. Then, for any event B,

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

$\mathbb{P}(A_i)$ is the prior probability of A_i and $\mathbb{P}(A_i|B)$ is the posterior probability of A_i

Theorem 1.14 (Bayes' Theorem)

Let $A_1, A_2, \ldots$ be a disjoint countable partition of Ω such that $\mathbb{P}(A_i) > 0$ for each $i \in \mathbb{N}$. If $\mathbb{P}(B) > 0$, then, for each i,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

Email Filtering

Example 1.15 (Three categories of emails)

A_1="spam", A_2="low priority" and A_3="high priority". From previous experience, I find that $\mathbb{P}(A_1) = .7$, $\mathbb{P}(A_2) = .2$ and $\mathbb{P}(A_3) = .1$. Of course, $.7 + .2 + .1 = 1$. Let B be the event that the email contains the word "free". From previous experience, $\mathbb{P}(B|A_1) = .9$, $\mathbb{P}(B|A_2) = .01$, $\mathbb{P}(B|A_3) = .01$. (Note: $.9 + .01 + .01 \neq 1$.) I receive an email with the word "free". What is the probability that it is spam? Bayes' theorem yields,

$$\mathbb{P}(A_1|B) = \frac{.9 \times .7}{.9 \times .7 + .01 \times .2 + .01 \times .1} = .995.$$

Agenda

2. Random Variables

Measurable Mapping

Definition 2.1 (Random Variable)

Let $(\Omega, \mathcal{A})$ be a measurable space. A random variable is a measurable mapping that assigns a real number $X(\omega)$ to each outcome ω. It means that

$$X : \Omega \to \mathbb{R} \quad \& \quad \{\omega \in \Omega : \ X(\omega) \leq x\} \in \mathcal{A}, \forall x \in \mathbb{R}.$$

$$[X \in A] \equiv X^{-1}(A) := \{\omega \in \Omega : \ X(\omega) \in A\}$$

Example 2.2 (Flip a coin ten times)

Let $X(\omega)$ be the number of heads in the sequence ω. If $\omega = HHTHHTHHTT$, then $X(\omega) = 6$.

Distribution Function

Definition 2.3 (Cumulative Distribution Function)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. The cumulative distribution function (CDF) of X is the function $F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

Theorem 2.4

Let X have CDF F and let Y have CDF G. If $F(x) = G(x)$ for all $x \in \mathbb{R}$, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all (measurable) $A \in \mathcal{B}(\mathbb{R})$.

Flip a Coin Twice

Example 2.5

Flip a fair coin twice and let X be the number of heads. Then $\mathbb{P}(X = 0) = \mathbb{P}(X = 2) = 1/4$ and $\mathbb{P}(X = 1) = 1/2$. The distribution function is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2. \end{cases}$$

Notice that the function is right continuous, non-decreasing, and that it is defined for all x, even though the random variable only takes values 0, 1, and 2.

Properties of the CDF

Theorem 2.6 (Three Basic Properties)

If F is a CDF of a random variable X, then:

(i) *F is non-decreasing: $x_1 < x_2$ implies that $F(x_1) \leq F(x_2)$.*

(ii) *F is normalized:*

$$\lim_{x \downarrow -\infty} F(x) = 0$$

and

$$\lim_{x \uparrow +\infty} F(x) = 1.$$

(iii) *F is right-continuous: $F(x) = F(x^+)$ for all x, where*

$$F(x^+) = \lim_{y \downarrow x+} F(y).$$

Proving the other direction – namely, that if a function F mapping the real line to $[0, 1]$ satisfies (i), (ii), and (iii), then it is a CDF for some random variable – uses some deep tools in analysis.

Probability Mass Function

Definition 2.7 (Probability Mass Function)

X is *discrete* if it takes countably many values $\{x_1, x_2, \ldots\}$. We define the *probability function* or *probability mass function* (PMF) for X by $f_X(x) = \mathbb{P}(X = x)$.

Example 2.8

The probability function for Example 2.5 is

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise.} \end{cases}$$

Probability Density Function

Definition 2.9 (Probability Density Function)

A random variable X is *continuous* if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$ and for every $a \leq b$,

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)\mathrm{d}x.$$

The function f_X is called the *probability density function* (PDF). We have that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

and $f_X(x) = F_X'(x)$ at all points x at which F_X is differentiable.

Consequences of the CDF Definition

Lemma 2.10 (Consequences of the CDF Definition)

Let F be the CDF for a random variable X. Then:

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$ *where* $F(x^-) = \lim_{y \uparrow x^-} F(y)$;
2. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$;
3. $\mathbb{P}(X > x) = 1 - F(x)$;
4. *If X is continuous, then*

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b). \end{aligned}$$

Quantile Function

Definition 2.11 (Quantile)

Let X be a random variable with CDF F. The *inverse CDF* or *quantile function* is defined by

$$F^{-1}(q) = \inf\left\{x : F(x) > q\right\}$$

for $q \in (0,1)$. If F is strictly increasing and continuous, then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$.

- $F^{-1}(1/4)$ is the first quartile, $F^{-1}(1/2)$ is the median (or second quartile), and $F^{-1}(3/4)$ the third quartile.
- Two random variables X and Y are equal in distribution – written $X \overset{d}{=} Y$ – iff $F_X(x) = F_Y(x)$ for all x. This does not mean that $X = Y$.

Examples – Discrete Random Variables I

Example 2.12 (Point Mass Distribution)

X has a point mass distribution at a iff $\mathbb{P}[X = x] = \mathbb{1}\{x = a\}$, $x \in \mathbb{R}$. Written $X \sim \delta_a$ (Dirac measure at a). Then, $F_X(x) = \mathbb{1}\{x \geq a\}$.

Example 2.13 (Discrete Uniform Distribution)

X has a discrete uniform distribution on $\{1, \ldots, k\}$ iff $f_X(x) = \begin{cases} 1/k, & \text{for } x = 1, \ldots, k; \\ 0, & \text{otherwise.} \end{cases}$

Example 2.14 (Bernoulli Distribution)

X has a Bernoulli (or alternative; 0–1) distribution with parameter $p \in (0, 1)$ iff $f_X(x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$. Written $X \sim \text{Be}(p)$.

Examples – Discrete Random Variables II

Example 2.15 (Binomial Distribution)

X has a binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$ iff $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{1}\{x \in \{0, \dots, n\}\}$. Written $X \sim \text{Bi}(n, p)$.

Example 2.16 (Geometric Distribution)

X has a geometric distribution with parameter $p \in (0,1)$ iff $f_X(x) = p(1-p)^x$ for $x \in \mathbb{N}_0$. Written $X \sim \text{Ge}(p)$.

Example 2.17 (Poisson Distribution)

X has a Poisson distribution with parameter $\lambda > 0$ iff $f_X(x) = \exp\{-\lambda\}\lambda^x / x!$ for $x \in \mathbb{N}_0$. Written $X \sim \text{Po}(\lambda)$.

$X \sim \text{Po}(\lambda_X) \perp\!\!\!\perp Y \sim \text{Po}(\lambda_Y) \Rightarrow X + Y \sim \text{Po}(\lambda_X + \lambda_Y)$

... later & easy; not valid w/o $\perp\!\!\!\perp$ (try counterEx); $\Leftarrow$ non-trivial (Raikov's theorem)

Examples – Continuous Random Variables I

Example 2.18 (Uniform Distribution)

X has a uniform distribution on interval $[a, b]$ iff $f_X(x) = (b-a)^{-1}\mathbb{1}\{x \in [a, b]\}$.
Written $X \sim \mathsf{U}(a, b)$.

Example 2.19 (Normal or Gaussian Distribution)

X has a normal (or Gaussian) distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ iff $f_X(x) \equiv \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$, $x \in \mathbb{R}$. Written $X \sim \mathsf{N}(\mu, \sigma^2)$.

- $X \sim \mathsf{N}(\mu, \sigma^2) \Rightarrow Z = (X-\mu)/\sigma \sim \mathsf{N}(0, 1)$ (standard normal)
- $X \sim \mathsf{N}(\mu_X, \sigma_X^2) \perp\!\!\!\perp Y \sim \mathsf{N}(\mu_Y, \sigma_Y^2) \Rightarrow X + Y \sim \mathsf{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
  ... later & easy; not valid w/o $\perp\!\!\!\perp$; $\Leftarrow$ non-trivial (Cramér's decomposition theorem)

Examples – Continuous Random Variables II

- CDF of $\mathrm{N}(\mu, \sigma^2)$: $\Phi(x) = \int_{\infty}^{x} \phi(t)\mathrm{d}t$ ... no closed-form expression

Example 2.20

Suppose that $X \sim \mathrm{N}(3, 5)$. Find $\mathbb{P}[X > 1]$. The solution is

$$\mathbb{P}[X > 1] = 1 - \mathbb{P}[X < 1] = 1 - \mathbb{P}\left[Z < \frac{1-3}{\sqrt{5}}\right] = 1 - \Phi(-0.8944) = 0.81.$$

Now find $q = \Phi^{-1}(0.2)$. We solve this by writing

$$0.2 = \mathbb{P}[X < q] = \mathbb{P}\left[Z < \frac{q-\mu}{\sigma}\right] = \Phi\left[\frac{q-\mu}{\sigma}\right].$$

From the Normal table, $\Phi(-0.8416) = 0.2$. Therefore, $-0.8416 = \frac{q-\mu}{\sigma} = \frac{q-3}{\sqrt{5}}$ and, hence, $q = 3 - 0.8416\sqrt{5} = 1.1181$.

Examples – Continuous Random Variables III

Example 2.21 (Exponential Distribution)

X has an exponential distribution with parameter $\beta > 0$ iff $f_X(x) = \beta^{-1}\exp\{-x/\beta\}\mathbb{1}\{x > 0\}$. Written $X \sim \text{Exp}(\beta)$.

Example 2.22 (Gamma Distribution)

X has a Gamma distribution with parameters $\alpha, \beta > 0$ iff $f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1}\exp\{-x/\beta\}\mathbb{1}\{x > 0\}$, where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\exp\{-t\}\mathrm{d}t$ is the Gamma function. Written $X \sim \text{Gamma}(\alpha, \beta)$.

- $X \sim \text{Exp}(\beta) \Rightarrow X \sim \text{Gamma}(1, \beta)$
- $X \sim \text{Gamma}(\alpha_X, \beta) \perp\!\!\!\perp Y \sim \text{Gamma}(\alpha_Y, \beta) \Rightarrow X + Y \sim \text{Gamma}(\alpha_X + \alpha_Y, \beta)$

Examples – Continuous Random Variables IV

Example 2.23 (Beta Distribution)

X has a Beta distribution with parameters $\alpha, \beta > 0$ iff
$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}\{x \in (0,1)\}$. Written $X \sim \text{Beta}(\alpha, \beta)$.

Example 2.24 (χ^2-Distribution)

X has a χ^2-distribution with p degrees of freedom iff
$f_X(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1}\exp\{-x/2\}\mathbb{1}\{x > 0\}$. Written $X \sim \chi^2_p$.

$X_1, \ldots, X_p \overset{\perp\!\!\!\perp}{\sim} \text{N}(0,1) \Rightarrow \sum_{i=1}^{p} X_i^2 \sim \chi^2_p$

Examples – Continuous Random Variables V

Example 2.25 (Student's t-Distribution)

X has Student's t-distribution with ν degrees of freedom iff $f_X(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{(1+x^2/\nu)^{(\nu+1)/2}}$. Written $X \sim t_\nu$.

Example 2.26 (Cauchy Distribution)

The Cauchy distribution is a special case of the t-distribution corresponding to $\nu = 1$. Written $X \sim$ Cauchy.

- $X \sim$ Cauchy $\Rightarrow f_X(x) = \frac{1}{\pi(1+x^2)}$
- The standard normal corresponds to a t-distribution with $\nu = \infty$.

Multivariate Randomness

Definition 2.27 (Random Vector)

Let $(\Omega, \mathcal{A})$ be a measurable space. A random variable is a measurable mapping that assigns a real *d-dimensional vector* $\mathbb{X}(\omega)$ to each outcome ω. It means that

$$\mathbb{X} : \Omega \to \mathbb{R}^d \quad \& \quad \{\omega \in \Omega : \mathbb{X}(\omega) \leq \mathbb{x}\} \in \mathcal{A}, \forall \mathbb{x} \in \mathbb{R}^d.$$

Definition 2.28 (Multivariate CDF)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. The *multivariate cumulative distribution function* (mCDF) of $\mathbb{X}$ is the function $F_{\mathbb{X}} : \mathbb{R}^d \to [0, 1]$ defined by

$$F_{\mathbb{X}}(\mathbb{x}) = \mathbb{P}(\mathbb{X} \leq \mathbb{x}).$$

Multivariate CDF

Theorem 2.29 (Properties of mCDF)

If F is a mCDF of a random vector $\mathbb{X}$, then

(i) *F is element-wise non-decreasing and right-continuous;*

(ii)

$$\lim_{x_\ell \downarrow -\infty} F(\mathbb{x}) = 0 \quad \text{for any } \ell = 1, \ldots, d$$

and

$$\lim_{x_\ell \uparrow +\infty \, \forall \ell} F(\mathbb{x}) = 1;$$

Bivariate Distributions – Discrete

Definition 2.30 (Joint Mass Function)

Given a pair of discrete random variables X and Y,

$$f(x, y) = f_{(X,Y)}(x, y) = \mathbb{P}[X = x, Y = y]$$

is called the *joint probability mass function* of (X, Y).

Example 2.31

A bivariate distribution for two random variables X and Y each taking values 0 or 1:

	$Y = 0$	$Y = 1$	
$X = 0$	1/9	2/9	1/3
$X = 1$	2/9	4/9	2/3
	1/3	2/3	1

Thus, $f(1, 1) = \mathbb{P}(X = 1, Y = 1) = 4/9$.

Bivariate Distributions – Continuous

Definition 2.32 (Joint Probability Density Function)

Given a pair of continuous random variables X and Y, we call a function $f(x, y)$ a *joint probability density function* for (X, Y) if

(i) $f(x, y) \geq 0 \ \forall (x, y)$,

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \mathrm{d}x \mathrm{d}y = 1$, and

(iii) for any Borel set $A \subseteq \mathbb{R} \times \mathbb{R}$, $\mathbb{P}[(X, Y) \in A] = \int_A f(x, y) \mathrm{d}x \mathrm{d}y$.

Example 2.33

Let (X, Y) be uniform on the unit square. Then, $f(x, y) = \mathbb{1}\{(x, y) \in [0, 1]^2\}$. Find $\mathbb{P}[X < 1/2, Y < 1/2]$. The event $\{X < 1/2, Y < 1/2\}$ corresponds to a subset of the unit square. Integrating f over this subset corresponds, in this case, to computing the area of the set $\{x < 1/2, y < 1/2\}$ which is $1/4$. So, $\mathbb{P}[X < 1/2, Y < 1/2] = 1/4$.

Marginal Distribution – Discrete & Continuous

Theorem 2.34 (Marginal Distributions)

If (X, Y) have the joint mass function $f_{(X,Y)}$, then the marginal mass function for X is

$$f_X(x) = \mathbb{P}[X = x] = \sum_y \mathbb{P}[X = x, Y = y] = \sum_y f_{(X,Y)}(x, y).$$

If (X, Y) have joint probability density function $f_{(X,Y)}$, then the marginal probability density function for X is

$$f_X(x) = \int f_{(X,Y)}(x, y)dy.$$

Independence II

Note that

$$\lim_{x_\ell \uparrow +\infty \forall j \in \{1,\dots,k\}\setminus\{\ell\}} F_{\mathbb{X}}(\mathbb{x}) = \lim_{x_j \uparrow +\infty \forall j \in \{1,\dots,k\}\setminus\{\ell\}} \mathbb{P}[X_1 \le x_1, \dots, X_k \le x_k]$$
$$= \mathbb{P}[X_\ell \le x_\ell] = F_{X_\ell}(x_\ell)$$

Definition 2.35 (Independent Random Variables)

Random variables $X_1, \dots, X_k$ are independent if

$$F_{\mathbb{X}}(\mathbb{x}) = \prod_{\ell=1}^{k} F_{X_\ell}(x_\ell) \quad \text{for every } \mathbb{x} = [x_1, \dots, x_k]^\intercal \in \mathbb{R}^k,$$

where $\mathbb{X} = [X_1, \dots, X_k]^\intercal$.

If X and Y are independent random variables, we write $X \perp\!\!\!\perp Y$. If they are not independent (are dependent), we write $X \not\perp\!\!\!\perp Y$.

Discrete and Continuous Independence

- Support of a discrete random variable X ... $\mathcal{S}(X) = \{x \in \mathbb{R} : \mathbb{P}[X = x] > 0\}$
- Support of a continuous random variable X ... $\mathcal{S}(X) = \{x \in \mathbb{R} : f_X(x) > 0\}$

Theorem 2.36 (Equivalent Characterization of Independence)

Let the joint PDF of $X_1, \ldots, X_k$ be $f_{\mathbb{X}}(\mathbb{x})$. $\perp\!\!\!\perp \{X_1, \ldots, X_k\}$ iff

$$f_{\mathbb{X}}(\mathbb{x}) = \prod_{\ell=1}^{k} f_{X_\ell}(x_\ell) \quad \text{for almost all}^{a}\ \mathbb{x} \in \times_{\ell=1}^{k} \mathcal{S}(X_\ell).$$

Let the joint PMF of $X_1, \ldots, X_k$ be $\mathbb{P}[\mathbb{X} = \mathbb{x}]$. $\perp\!\!\!\perp \{X_1, \ldots, X_k\}$ iff

$$\mathbb{P}[\mathbb{X} = \mathbb{x}] = \prod_{\ell=1}^{k} \mathbb{P}[X_\ell = x_\ell] \quad \text{for all } \mathbb{x} \in \times_{\ell=1}^{k} \mathcal{S}(X_\ell).$$

[a] I.e., for all $\mathbb{x} \in \times_{\ell=1}^{k} \mathcal{S}(X_\ell) \setminus N$, where N is a Borel set having measure zero.

Cartesian Support and Independence

- Support of a discrete random vector $\mathbf{X}$ ... $\mathcal{S}(\mathbf{X}) = \{\mathbf{x} : \mathbb{P}[\mathbf{X} = \mathbf{x}] > 0\}$
- Support of a continuous random vector $\mathbf{X}$ ... $\mathcal{S}(\mathbf{X}) = \{\mathbf{x} : f_{\mathbf{X}}(\mathbf{x}) > 0\}$

Theorem 2.37 (Cartesian Product of Supports and Independence)

Suppose that $\mathcal{S}(X, Y) = \mathcal{S}(X) \times \mathcal{S}(Y)$. *If* $f_{(X,Y)}(x, y) = g(x)h(y)$ *or* $\mathbb{P}[X = x, Y = y] = g(x)h(y)$ *for some functions* g *and* h *(not necessarily PDFs or PMFs) for almost all* $[x, y]^\intercal \in \mathcal{S}(X, Y)$, *then* $X \perp\!\!\!\perp Y$.

Conditional Distributions

Definition 2.38 (Conditional PDF and PMF)

The *conditional probability mass function* of X given $Y = y$ is

$$f_{X|Y}(x|y) \equiv \mathbb{P}[X = x | Y = y] := \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} \equiv \frac{f_{(X,Y)}(x, y)}{f_Y(y)}, \quad \text{if } \mathbb{P}[Y = y] > 0.$$

The *conditional probability density function* of X given $Y = y$ is

$$f_{X|Y}(x|y) := \frac{f_{(X,Y)}(x, y)}{f_Y(y)}, \quad \text{if } f_Y(y) > 0.$$

- Conditional PMF an PDF are *functions* of argument x with parameter y
- Otherwise defined *arbitrarily*
- Discrete ... $\mathbb{P}[X \in A | Y = y] = \sum_{x \in A} \mathbb{P}[X = x | Y = y]$
- Continuous ... $\mathbb{P}[X \in A | Y = y] = \int_A f_{X|Y}(x|y) \mathrm{d}x$

Multivariate Normal Distribution

Definition 2.39 (Multivariate Normal)

The d-dimensional random vector $\mathbb{X} = [X_1, \ldots, X_d]^\intercal$ has a *d-variate normal distribution* with parameters $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$, if it has the PDF

$$f_{\mathbb{X}}(\mathbb{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(\mathbb{x} - \boldsymbol{\mu})^\intercal \Sigma^{-1} (\mathbb{x} - \boldsymbol{\mu})\right\}, \quad \mathbb{x} \in \mathbb{R}^d,$$

where Σ is a positive definite matrix.

- notation: $\mathbb{X} \sim \mathsf{N}_d(\boldsymbol{\mu}, \Sigma)$
- special case: standard d-variate normal distribution $\boldsymbol{\mu} = \mathbb{0}$ and $\Sigma = I_d$, i.e., $\mathbb{X} \sim \mathsf{N}_d(\mathbb{0}, I_d)$

Multivariate Standard Normal Distribution

Square root of a positive definite matrix Σ ... denoted by $\Sigma^{1/2}$

(i) $\Sigma^{1/2}$ is symmetric

(ii) $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$

(iii) $\Sigma^{1/2}\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2} = I_d$, where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$

Theorem 2.40 (Standardization)

If $\mathbb{Z} \sim \mathsf{N}_d(\mathbb{0}, I_d)$ *and* $\mathbb{X} = \boldsymbol{\mu} + \Sigma^{1/2}\mathbb{Z}$, *then* $\mathbb{X} \sim \mathsf{N}_d(\boldsymbol{\mu}, \Sigma)$. *Conversely, if* $\mathbb{X} \sim \mathsf{N}_d(\boldsymbol{\mu}, \Sigma)$, *then* $\Sigma^{-1/2}(\mathbb{X} - \boldsymbol{\mu}) \sim \mathsf{N}_d(\mathbb{0}, I_d)$.

Normal Margins

Suppose we *partition* a random normal vector $\mathbb{X}$ as $\mathbb{X} = [\mathbb{X}_a^\intercal, \mathbb{X}_b^\intercal]^\intercal$. We can similarly partition $\boldsymbol{\mu} = [\boldsymbol{\mu}_a^\intercal, \boldsymbol{\mu}_b^\intercal]^\intercal$, such that $\boldsymbol{\mu}_a \in \mathbb{R}^s$, $\boldsymbol{\mu}_b \in \mathbb{R}^{d-s}$, and

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}.$$

Theorem 2.41 (Properties of MND)

Let $\mathbb{X} \sim \mathsf{N}_d(\boldsymbol{\mu}, \Sigma)$. *Then,*

(i) *the marginal distribution of* $\mathbb{X}_a$ *is* $\mathbb{X}_a \sim \mathsf{N}_s(\boldsymbol{\mu}_a, \Sigma_{aa})$;

(ii) *the conditional distribution of* $\mathbb{X}_b$ *given* $\mathbb{X}_a = \mathbb{x}_a$ *is*

$$\mathbb{X}_b | \mathbb{X}_a = \mathbb{x}_a \sim \mathsf{N}_{d-s}(\boldsymbol{\mu}_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\mathbb{x}_a - \boldsymbol{\mu}_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab});$$

(iii) $\mathbb{a}^\intercal \mathbb{X} \sim \mathsf{N}(\mathbb{a}^\intercal \boldsymbol{\mu}, \mathbb{a}^\intercal \Sigma \mathbb{a})$ *for* $\mathbb{a} \in \mathbb{R}^d$;

(iv) $(\mathbb{X} - \boldsymbol{\mu})^\intercal \Sigma^{-1} (\mathbb{X} - \boldsymbol{\mu}) \sim \chi^2_d$.

Transformations of Discrete Random Variables

- A transformation $Y = t(X)$ (not necessarily monotonic)
- Having a *known discrete* distribution of X, i.e., $\mathbb{P}[X = x]$
- The goal is $\mathbb{P}[Y = y]$:

$$\mathbb{P}[Y = y] = \mathbb{P}[t(X) = y] = \mathbb{P}[\omega : t(X(\omega)) = y] = \mathbb{P}[X \in t^{-1}(y)] = \sum_{x}^{t(x)=y} \mathbb{P}[X = x],$$

where t^{-1} gives all the *preimages*

Transformations of Continuous Random Variables

- A transformation $Y = t(X)$ (not necessarily monotonic)
- Having a *known continuous* distribution of X, i.e., $f_X(x)$
- The aim is $f_Y(y)$:

1. For each y, find the set $\mathcal{T}(y) = \{x : t(x) \leq y\}$
2. The CDF is

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[t(X) \leq y] = \mathbb{P}[\omega : t(X(\omega)) \leq y] = \int_{\mathcal{T}(y)} f_X(x)\mathrm{d}x$$

3. If F_Y is absolutely continuous, then the PDF is $f_Y(y) = F'_Y(y)$

Transformations of Discrete Random Vectors

- A transformation $Z = t(X, Y)$, where $t : \mathbb{R}^2 \to \mathbb{R}$
- Having a *known discrete* distribution of $[X, Y]$, i.e., $\mathbb{P}[X = x, Y = y]$
- The target is $\mathbb{P}[Z = z]$:

$$\mathbb{P}[Z = z] = \mathbb{P}[t(X, Y) = z] = \mathbb{P}[\omega : t([X, Y](\omega)) = z] = \mathbb{P}[[X, Y] \in t^{-1}(z)]$$
$$= \sum_{[x,y]}^{t(x,y)=z} \mathbb{P}[X = x, Y = y],$$

where t^{-1} gives all the *preimages*

Transformations of Continuous Random Vectors

- A transformation $Z = t(X, Y)$, where $t : \mathbb{R}^2 \to \mathbb{R}$
- Having a *known continuous* distribution of $[X, Y]$, i.e., $f_{(X,Y)}(x, y)$
- The target is $f_Z(z)$:

1. For each z, find the set $\mathcal{T}(z) = \{[x, y] : t(x, y) \leq z\}$
2. The CDF is

$$F_Z(z) = \mathbb{P}[Z \leq z] = \mathbb{P}[t(X, Y) \leq y] = \mathbb{P}[\omega : t([X, Y](\omega)) \leq y]$$
$$= \iint_{\mathcal{T}(z)} f_{(X,Y)}(x, y) \mathrm{d}x\mathrm{d}y$$

3. If F_Z is absolutely continuous, then the PDF is $f_Z(z) = F'_Z(z)$

Agenda

3. Expectations

3.1 Definition
3.2 Discrete and Continuous Case
3.3 Expectation of Transformation
3.4 Moments
3.5 Properties
3.6 Variance
3.7 Covariance and Correlation
3.8 Variance-covariance Matrix
3.9 Conditional Expectation
3.10 Conditional Variance
3.11 Moment Generating Function
3.12 Characteristic Function

Expectation

Pre-definition:

$$\mathbb{E}[X] \equiv \mathbb{E}X := \int X \mathrm{d}\mathbb{P} \equiv \int X(\omega)\mathrm{d}\mathbb{P}(\omega)$$

Using the *pushforward measure* $\mathbb{P}_X(\cdot) := \mathbb{P}[X \in \cdot] \equiv \mathbb{P}[X^{-1}(\cdot)]$

$$\mathbb{E}X = \int x \mathrm{d}\mathbb{P}_X(x)$$

Here, $\mathbb{P}_X$ is the distribution or law of X.

Or, the *Lebesgue–Stieltjes measure* associated with the CDF F_X:

Definition 3.1 (Expectation or Expected Value or Mean Value)

The *expected value* of X is

$$\mathbb{E}X = \int_{\mathbb{R}} x \mathrm{d}F_X(x),$$

if the r.h.s. exists.

Discrete and Continuous Expectation

Using the *Radon–Nikodym derivative*:

Theorem 3.2 (Discrete and Continuous Mean)

The expected value of X is

$$\mathbb{E}X = \begin{cases} \int_{-\infty}^{+\infty} x f_X(x)\,dx & \text{if } X \text{ is continuous;} \\ \sum_{x \in \mathcal{S}(X)} x \mathbb{P}[X = x] & \text{if } X \text{ is discrete,} \end{cases}$$

given that the r.h.s. is well-defined.

Caution with Expectation

Example 3.3 (Cauchy Distribution)

If $X \sim$ Cauchy, then $\mathbb{E}X$ does not exist. Using per partes,

$$\int_0^\infty \frac{x}{\pi(1+x^2)} \mathrm{d}x = [x\arctan(x)]_0^\infty - \int_0^\infty \arctan(x)\mathrm{d}x = \infty.$$

Thus, for $\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)} \mathrm{d}x$, $\infty - \infty$ is not defined.

Expectation of Transformation

A transformation $Y = t(X)$.

Theorem 3.4 (The Rule of the Lazy Statistician)

Let $Y = t(X)$. Then,

$$\mathbb{E}\,Y = \int t(x)\,dF_X(x),$$

if the r.h.s. exists.

- A.k.a. *The Law of The Unconscious Statistician*
- Discrete ... $\mathbb{E}\,Y = \sum_{x \in \mathcal{S}(X)} t(x)\mathbb{P}[X = x]$
- Continuous ... $\mathbb{E}\,Y = \int_{-\infty}^{+\infty} t(x) f_X(x)\mathrm{d}x$

Moments

Definition 3.5 (Moment)

The kth *moment* of X is defined to be $\mathbb{E}[X^k]$ assuming that $\mathbb{E}[|X|^k] < \infty$.

Theorem 3.6 (Higher Moments)

If the kth moment exists and, then the ℓth moment exists for any $\ell \leq k$.

Example 3.7 (Student's t-distribution as "Counter" Example)

t-distribution with $\nu = 3$ degrees of freedom: $\mathbb{E}X = 0$, $\mathbb{E}X^2 = \infty$, $\mathbb{E}X^3$ does not exist.

Definition 3.8 (Absolute Moment)

The kth *absolute moment* of X is defined to be $\mathbb{E}[|X|^k]$ assuming that it exists.

Properties of Expectations

Definition 3.9 ($\mathcal{L}^p$-spaces of Random Variables)

$X \in \mathcal{L}^p$, if $\mathbb{E}[|X|^p] < \infty$.

Theorem 3.10 (Linearity)

If $X_1, \ldots, X_k \in \mathcal{L}^1$ and $a_1, \ldots, a_k$ are constants, then $\mathbb{E}\left(\sum_{\ell=1}^k a_\ell X_\ell\right) = \sum_{\ell=1}^k a_\ell \mathbb{E}X_\ell$.

Theorem 3.11 (Multiplication Under Independence)

If $X_1, \ldots, X_k \in \mathcal{L}^1$ are independent random variables, then $\mathbb{E}\left(\prod_{\ell=1}^k X_\ell\right) = \prod_{\ell=1}^k \mathbb{E}X_\ell$.

Variance

The variance measures the "spread" of a distribution.

Definition 3.12 (Variance and Standard Deviation)

The *variance* of X is defined by

$$\mathbb{V}\mathrm{ar}\,X = \mathbb{E}(X - \mathbb{E}X)^2.$$

The *standard deviation* of X is $\mathrm{sd}(X) = \sqrt{\mathbb{V}\mathrm{ar}\,X}$.

We can't use $\mathbb{E}(X - \mathbb{E}X)$ as a measure of spread since $\mathbb{E}(X - \mathbb{E}X) = 0$. We can and sometimes do use $\mathbb{E}|X - \mathbb{E}X|$ as a measure of spread, but more often we use the variance.

Properties of Variance

Theorem 3.13

Assuming that the considered second moments are finite, it has the following properties:

- $\mathbb{V}\mathrm{ar}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2 \geq 0$;
- *if* $a, b \in \mathbb{R}$, *then* $\mathbb{V}\mathrm{ar}(aX + b) = a^2\mathbb{V}\mathrm{ar}X$;
- *if* $X_i, \ldots, X_k$ *are independent and* $a_1, \ldots, a_k$ *are real constants, then* $\mathbb{V}\mathrm{ar}\left(\sum_{\ell=1}^{k} a_\ell X_\ell\right) = \sum_{\ell=1}^{k} a_\ell^2 \mathbb{V}\mathrm{ar}X_\ell$.

Covariance

The covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition 3.14 (Covariance and Correlation)

The *covariance* between X and Y is defined by

$$\mathbb{C}\text{ov}(X, Y) = \mathbb{E}\{(X - \mathbb{E}X)(Y - \mathbb{E}Y)\}.$$

If $\mathbb{V}\text{ar}(X)\mathbb{V}\text{ar}(Y) > 0$, then the *correlation* between X and Y is defined by

$$\rho_{X,Y} \equiv \mathbb{C}\text{orr}(X, Y) = \frac{\mathbb{C}\text{ov}(X, Y)}{\sqrt{\mathbb{V}\text{ar}(X)\mathbb{V}\text{ar}(Y)}}.$$

Note that $\mathbb{E}[t(X, Y)] = \int_{\mathbb{R}^2} t(x, y)\mathrm{d}F_{(X,Y)(x,y)}$ and, thus,

$$\mathbb{E}XY = \begin{cases} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_{(X,Y)}(x, y)\mathrm{d}x\mathrm{d}y & \text{if } X \text{ and } Y \text{ are continuous,} \\ \sum_{x\in\mathcal{S}(X), y\in\mathcal{S}(Y)} xy\mathbb{P}[X = x, Y = y] & \text{if } X \text{ and } Y \text{ are discrete} \end{cases}$$

Properties of Covariance and Correlation

Theorem 3.15

- $\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$.
- $-1 \leq \mathbb{C}\mathrm{orr}(X, Y) \leq 1$.
- $|\mathbb{C}\mathrm{orr}(X, Y)| = 1 \iff \exists a \neq 0 \,\&\, b \in \mathbb{R} :\ Y = aX + b$ *with probability 1.*
- *If* X *and* Y *are independent, then* $\mathbb{C}\mathrm{ov}(X, Y) = 0$.

In general, $\mathbb{C}\mathrm{ov}(X, Y) = 0 \nRightarrow X \perp\!\!\!\perp Y$.

Corollary 3.16

If $X_1, \ldots, X_k \in \mathcal{L}^2$ *and* $a_1, \ldots, a_k$ *are real constants, then*

$$\mathbb{V}\mathrm{ar}\left(\sum_{\ell=1}^{k} a_\ell X_\ell\right) = \sum_{\ell=1}^{k} a_\ell^2 \mathbb{V}\mathrm{ar}\, X_\ell + 2 \sum_{1 \leq j < \ell \leq k} \sum a_j a_\ell \mathbb{C}\mathrm{ov}(X_j, X_\ell).$$

Variance-covariance Matrix

Definition 3.17 (Multivariate Expectation)

The *expected value* of a random vector $\mathbb{X} = [X_1, \ldots, X_k]^\intercal$ is defined by

$$\mathbb{E}\mathbb{X} = [\mathbb{E}X_1, \ldots, \mathbb{E}X_k]^\intercal.$$

Note that $\mathbb{C}\text{ov}(X, Y) = \mathbb{C}\text{ov}(Y, X)$ and $\mathbb{C}\text{ov}(X, X) = \mathbb{V}\text{ar}X$.

Definition 3.18 (Variance-covariance Matrix)

The *variance-covariance matrix* of a random vector $\mathbb{X}$ is defined by

$$\mathbb{V}\text{ar}\mathbb{X} = \begin{bmatrix} \mathbb{V}\text{ar}X_1 & \mathbb{C}\text{ov}(X_1, X_2) & \ldots & \mathbb{C}\text{ov}(X_1, X_k) \\ \mathbb{C}\text{ov}(X_2, X_1) & \mathbb{V}\text{ar}X_2 & \ldots & \mathbb{C}\text{ov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}(X_k, X_1) & \mathbb{C}\text{ov}(X_k, X_2) & \ldots & \mathbb{V}\text{ar}X_k \end{bmatrix}.$$

Properties of Variance-covariance Matrix

Theorem 3.19

- *If* $\mathbb{X}$ *is a random vector and* $\mathbb{a}, \mathbb{b}$ *are real vectors, then*

$$\mathbb{E}(\mathbb{a}^\intercal \mathbb{X} + \mathbb{b}) = \mathbb{a}^\intercal \mathbb{E}\mathbb{X} + \mathbb{b} \quad \textit{and} \quad \mathbb{V}\mathrm{ar}(\mathbb{a}^\intercal \mathbb{X} + \mathbb{b}) = \mathbb{a}^\intercal (\mathbb{V}\mathrm{ar}\mathbb{X})\mathbb{a}.$$

- *If* $\mathbb{X}$ *is a random vector and* $\mathbb{A}, \mathbb{B}$ *are real matrices, then*

$$\mathbb{E}(\mathbb{A}\mathbb{X} + \mathbb{B}) = \mathbb{A}\mathbb{E}\mathbb{X} + \mathbb{B} \quad \textit{and} \quad \mathbb{V}\mathrm{ar}(\mathbb{A}\mathbb{X} + \mathbb{B}) = \mathbb{A}(\mathbb{V}\mathrm{ar}\mathbb{X})\mathbb{A}^\intercal.$$

Conditional Expectation

What is the mean of X among those times when $Y = y$?

Definition 3.20 (Deterministic and Random Conditioning)

The *conditional expectation* of $t(X, Y)$ given $Y = y$ is

$$g_{t(X,Y)}(y) \equiv \mathbb{E}[t(X, Y)|Y = y] = \begin{cases} \sum_{x \in \mathcal{S}(X)} t(x, y) f_{X|Y}(x|y), & \text{discrete case;} \\ \int t(x, y) f_{X|Y}(x|y) \mathrm{d}x, & \text{continuous case.} \end{cases}$$

The *conditional expectation* of $t(X, Y)$ given Y is $g_{t(X,Y)}(Y)$.

Theorem 3.21 (The Rule of Iterated Expectations)

$$\mathbb{E}\{\mathbb{E}[X|Y]\} = \mathbb{E}X, \quad \mathbb{E}\{\mathbb{E}[Y|X]\} = \mathbb{E}Y, \quad \textit{and} \quad \mathbb{E}\{\mathbb{E}[t(X, Y)|Y]\} = \mathbb{E}[t(X, Y)].$$

Properties of Conditional Expectations

Definition 3.22 ($\mathbb{P}$-almost surely)

We say that something *holds $\mathbb{P}$-almost surely* if it holds for all $\omega \in \Omega \setminus N$, where $N \in \mathcal{A}$ is a set of zero probability, i.e., $\mathbb{P}(N) = 0$. If it is clear what probability measure $\mathbb{P}$ we are using, we just say: it *holds almost surely*. We abbreviate $\mathbb{P}$-a.s., or only a.s.

Theorem 3.23

- $\mathbb{E}[a|X] = a$ *almost surely.*
- $\mathbb{E}[aX + bY|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z]$ *almost surely.*
- $\mathbb{E}[h(X)Y|X] = h(X)\mathbb{E}[Y|X]$ *almost surely.*

Conditional Variance

Definition 3.24

The *conditional variance* of X given $Y = y$ is

$$v_X(y) \equiv \mathbb{V}\mathrm{ar}[X \mid Y = y] = \begin{cases} \sum_{x \in \mathcal{S}(X)} \{x - g_X(y)\}^2 f_{X \mid Y}(x \mid y), & \text{discrete case;} \\ \int \{x - g_X(y)\}^2 f_{X \mid Y}(x \mid y) \mathrm{d}x, & \text{continuous case.} \end{cases}$$

The *conditional variance* of X given Y is $v_X(Y)$.

Theorem 3.25 (The Law of Total Variance)

$$\mathbb{V}\mathrm{ar}X = \mathbb{E}\mathbb{V}\mathrm{ar}[X \mid Y] + \mathbb{V}\mathrm{ar}\mathbb{E}[X \mid Y].$$

Moment Generating Function

Definition 3.26 (MGF)

The *moment generating function* (MGF), or *Laplace transform*, of X is defined by

$$\psi_X(t) = \mathbb{E}[\exp\{tX\}] = \int_{\mathbb{R}} \exp\{tx\} \mathrm{d}F_X(x), \quad t \in \mathbb{R},$$

if the r.h.s. exists.

A *characteristic function* $\varphi_X(t) = \mathbb{E}[\exp\{itX\}]$ is, however, well defined for all $t \in \mathbb{R}$.

Theorem 3.27 (Properties of the MGF)

- $\psi_X^{(m)}(0) = \mathbb{E}X^m$, $m \in \mathbb{N}_0$.
- *If* $Y = aX + b$, *then* $\psi_Y(t) = \exp\{bt\}\psi_X(at)$.
- *If* $X_i, \ldots, X_k$ *are independent and* $Y = \sum_{\ell=1}^k X_\ell$, *then* $\psi_Y(t) = \prod_{\ell=1}^k \psi_{X_\ell}(t)$.

If $\psi_X(t) = \psi_Y(t)$ for all t in an open interval around 0, then $X \stackrel{d}{=} Y$.

Characteristic Function

Definition 3.28 (CF)

The *characteristic function* (CF) of X is defined by

$$\varphi_X(t) = \mathbb{E}[\exp\{itX\}] = \int_{\mathbb{R}} \exp\{itx\} \mathrm{d}F_X(x), \quad t \in \mathbb{R}.$$

- *Fourier transform*: $\mathbb{E}[\exp\{-itX\}] = \varphi_X(-t)$
- $\varphi_X(t) = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)]$

Properties of the Characteristic Function

Theorem 3.29 (Properties of the CF)

(i) φ_X *exists for any distribution of* X

(ii) $\varphi_X(0) = 1$

(iii) $|\varphi_X(t)| \leq 1 \quad \forall t \in \mathbb{R}$

(iv) φ_X *is uniformly continuous* – $\forall \varepsilon > 0 \ \exists \delta > 0$*:* $|\varphi_X(t) - \varphi_X(s)| \leq \varepsilon$ *whenever* $|t - s| \leq \delta$

(v) $\varphi_{a+bX}(t) = e^{iat}\varphi_X(bt) \quad \forall t \in \mathbb{R} \quad \forall a, b \in \mathbb{R}$

(vi) $\varphi_{-X}(t) = \bar{\varphi}_X(t) \quad \forall t \in \mathbb{R}$ *(complex conjugate)*

(vii) $\varphi_X(t) \in \mathbb{R} \quad \forall t \in \mathbb{R} \iff \mathbb{P}[X > x] = \mathbb{P}[X < -x] \ \forall x \geq 0$ *(distribution symmetric about zero)*

(viii) $X \perp\!\!\!\perp Y \Rightarrow \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) \quad \forall t \in \mathbb{R}$

Unique Characterization of Distribution

Theorem 3.30 (Inversion Formula)

For any $a < b$,

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t)\,dt = \mathbb{P}[a < X < b] + \frac{\mathbb{P}[X = a] + \mathbb{P}[X = b]}{2}.$$

CF uniquely determines the distribution.

Corollary 3.31

$$\varphi_X = \varphi_Y \iff X \overset{d}{=} Y$$

Multivariate Characteristic Function

Definition 3.32 (CF for a Random Vector)

The *CF* of the random vector $\mathbb{X}$ is defined by

$$\varphi_{\mathbb{X}}(\mathbb{t}) = \mathbb{E}[\exp\{i\mathbb{t}^\top \mathbb{X}\}] = \int_{\mathbb{R}^k} \exp\{i\mathbb{t}^\top \mathbb{x}\} \mathrm{d}F_{\mathbb{X}}(\mathbb{x}), \quad \mathbb{t} \in \mathbb{R}^k.$$

- Remark that the univariate properties of the CF can be extended into the multivariate setting.
- Multivariate corollary 3.31 can be used to prove the properties of MND, cf Theorem 2.41.

Agenda

4. Stochastic Inequalities

4.1 Markov-type Inequalities

4.2 Inequalities for Expectations

Markov-type Inequalities I

Theorem 4.1 (Markov's Inequality)

Let X be a non-negative random variable and suppose that $\mathbb{E}[X]$ exists. For any $\varepsilon > 0$,

$$\mathbb{P}[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}.$$

Corollary 4.2

Let X be a non-negative random variable and suppose that $\mathbb{E}[X^r]$ exists for some $r > 0$. For any $\varepsilon > 0$,

$$\mathbb{P}[X \geq \varepsilon] \leq \frac{\mathbb{E}[X^r]}{\varepsilon^r}.$$

Markov-type Inequalities II

Theorem 4.3 (Chebyshev's Inequality)

Let X be a random variable and suppose that $\mathbb{E}[X]$ is finite. For any $\varepsilon > 0$,

$$\mathbb{P}[|X - \mathbb{E}X| \geq \varepsilon] \leq \frac{\mathbb{V}\mathrm{ar}[X]}{\varepsilon^2}.$$

Inequalities for Expectations

Theorem 4.4 (Cauchy-Schwarz Inequality)

If X and Y have finite variances, then

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

and

$$|\mathbb{C}\mathrm{ov}(X, Y)| \leq \sqrt{\mathbb{V}\mathrm{ar}X\,\mathbb{V}\mathrm{ar}Y}.$$

Inequalities for Expectations (cont.)

Theorem 4.5 (Jensen's Inequality)

If g is convex, then

$$\mathbb{E}g(X) \geq g(\mathbb{E}X).$$

If g is concave, then

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

Agenda

5. Stochastic Convergence

5.1 Types of Stochastic Convergence
5.2 Relationships Between the Types of Convergence
5.3 Continuous Mapping Theorem
5.4 Slutsky's Theorem
5.5 Lévy's Continuity Theorem
5.6 Weak Law of Large Numbers
5.7 Central Limit Theorem
5.8 Delta Method

Modes of Stochastic Convergence I

Definition 5.1 (Converges in Probability and in Distribution)

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

(i) X_n *converges to* X *in probability*, written $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, if, for every $\varepsilon > 0$,

$$\mathbb{P}[|X_n - X| > \varepsilon] \to 0 \text{ as } n \to \infty.$$

(ii) X_n *converges to* X *in distribution*, written $X_n \xrightarrow[n\to\infty]{\mathbb{D}} X$, if

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ at all } x \text{ for which } F \text{ is continuous.}$$

Modes of Stochastic Convergence II

Definition 5.2 (Converges in Lebesgue Spaces and Almost Surely)

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable.

(i) X_n *converges to* X *in* $\mathbb{L}_p$ for $p \geq 1$, written $X_n \xrightarrow[n\to\infty]{\mathbb{L}_p} X$, if

$$\mathbb{E}|X_n - X|^p \to 0 \text{ as } n \to \infty.$$

(ii) X_n *converges to* X $\mathbb{P}$*-almost surely*, written $X_n \xrightarrow[n\to\infty]{\mathbb{P}\text{-a.s.}} X$, if

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

Convergence in $L_1 \equiv$ convergence in *expectation*
Convergence in $L_2 \equiv$ convergence in *quadratic mean*

Relationships Between the Convergences

Theorem 5.3 (Implications Between the Modes of Convergence)

(a) $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X$

(b) $p \geq 1$: $X_n \xrightarrow{\mathbb{L}_p} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X$

(c) $p \geq q \geq 1$: $X_n \xrightarrow{\mathbb{L}_p} X \Rightarrow X_n \xrightarrow{\mathbb{L}_q} X$

(d) $X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{\mathbb{D}} X$

(e) *If* $X_n \xrightarrow{\mathbb{D}} X$ *and* $\mathbb{P}[X = c] = 1$ *for some* $c \in \mathbb{R}$, *then* $X_n \xrightarrow{\mathbb{P}} X$.

Reverse Implications Does Not Hold I

Example 5.4 (Convergence in probability does not imply almost sure convergence.)

$\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}(\Omega)$, $\mathbb{P} = \lambda$. We can uniquely write any positive integer by $2^n + m$, $m = 0, 1, \ldots, 2^n - 1$ and define

$$X_{2^n+m}(\omega) = \mathbb{1}\{\omega \in (m2^{-n}, (m+1)2^{-n}]\}, \quad \omega \in [0,1].$$

For instance, since $33 = 2^5 + 1$, we obtain $X_{33}(\cdot) = \mathbb{1}\{\cdot \in (2^{-5}, 2^{-4}]\}$. Then, for any $\varepsilon \in (0,1)$, we get $\mathbb{P}[|X_{2^n+m}| > \varepsilon] = 2^{-n} \to 0$ as $n \to \infty$. Thus, $X_n \xrightarrow{\mathbb{P}} 0$.

However, for each $\omega \in (0,1]$, $X_j(\omega) = 1$ and $X_j(\omega) = 0$ for *infinitely many* j's and so the sequence *does not converge almost surely*, i.e., $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} \not\to$.

Reverse Implications Does Not Hold II

Example 5.5 (Convergence in probability does not imply $\mathbb{L}_p$ convergence.)

$\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}(\Omega)$, $\mathbb{P} = \lambda$. We can define

$$X_{2^n+m}(\omega) = 2^n \mathbb{1}\{\omega \in ((m-1)2^{-n}, m2^{-n}]\}, \quad \omega \in [0,1].$$

Then again, for any $\varepsilon \in (0,1)$, we get $\mathbb{P}[|X_{2^n+m}| > \varepsilon] = 2^{-n} \to 0$ as $n \to \infty$. Thus, $X_n \xrightarrow{\mathbb{P}} 0$.

However, $\mathbb{E}|X_{2^n+m} - 0| = 2^n \mathbb{P}[X_{2^n+m} = 2^n] = 2^n 2^{-n} = 1$ and so the sequence *does not converge in* $\mathbb{L}_1$, i.e., $X_n \xrightarrow{\mathbb{L}_1}$. Hence, $X_n \xrightarrow{\mathbb{L}_p}$, $p \geq 1$.

Reverse Implications Does Not Hold III

Example 5.6 ($\mathbb{L}_q$ convergence does not imply $\mathbb{L}_p$ convergence, when $p > q \geq 1$.)

$\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}(\Omega)$, $\mathbb{P} = \lambda$. We can define

$$X_{2^n+m}(\omega) = 2^{\lfloor n/2 \rfloor} \mathbb{1}\{\omega \in ((m-1)2^{-n}, m2^{-n}]\}, \quad \omega \in [0,1].$$

Then, $\mathbb{E}|X_{2^n+m} - 0| = 2^{\lfloor n/2 \rfloor}\mathbb{P}[X_{2^n+m} = 2^n] = 2^{\lfloor n/2 \rfloor}2^{-n} \to 0$ as $n \to \infty$. Thus, $X_n \xrightarrow{\mathbb{L}_1} 0$.

However, $\mathbb{E}|X_{2^n+m} - 0|^2 = 2^n\mathbb{P}[X_{2^n+m} = 2^n] = 2^{2\lfloor n/2 \rfloor}2^{-n} \to 1$ as $n \to \infty$ and so the sequence *does not converge in* $\mathbb{L}_2$, i.e., $X_n \xrightarrow{\mathbb{L}_2}$.

Reverse Implications Does Not Hold IV

Example 5.7 (Convergence in distribution does not imply convergence in probability.)

$X \sim \mathrm{N}(0,1)$ and $X_n := -X$, $n \in \mathbb{N}$. Hence, $X_n \sim \mathrm{N}(0,1)$ for every $n \in \mathbb{N}$. So, trivially, $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$. Therefore, $X_n \xrightarrow{\mathbb{D}} X$.

However, $\mathbb{P}[|X_n - X| > \varepsilon] = \mathbb{P}[|2X| > \varepsilon] = \mathbb{P}[|X| > \varepsilon/2] \neq 0$ (which does not depend on n) and so the sequence *does not converge in probability*, i.e., $X_n \xrightarrow{\mathbb{P}} \not\to$.

Continuous Mapping Theorem

Theorem 5.8 (Continuous Mapping Theorem (CMT))

Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ be k-dimensional random vectors and $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P[\mathbf{X} \in C] = 1$.

- $\mathbf{X}_n \xrightarrow{\mathbb{P}\text{-a.s.}} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{\mathbb{P}\text{-a.s.}} g(\mathbf{X})$
- $\mathbf{X}_n \xrightarrow{\mathbb{P}} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{\mathbb{P}} g(\mathbf{X})$
- $\mathbf{X}_n \xrightarrow{\mathbb{D}} \mathbf{X} \Rightarrow g(\mathbf{X}_n) \xrightarrow{\mathbb{D}} g(\mathbf{X})$

! In general: $\mathbf{X}_n \xrightarrow{\mathbb{L}_p} \mathbf{X} \not\Rightarrow g(\mathbf{X}_n) \xrightarrow{\mathbb{L}_p} g(\mathbf{X})$

Cramér–Slutsky Theorem

Theorem 5.9 (Slutsky's Theorem)

If $X_n \xrightarrow{\mathbb{D}} X$ *and* $Y_n \xrightarrow{\mathbb{P}} c \in \mathbb{R}$, *then*

- $X_n + Y_n \xrightarrow{\mathbb{D}} X + c$;
- $X_n Y_n \xrightarrow{\mathbb{D}} cX$.

Corollary 5.10

If $X_n \xrightarrow{\mathbb{P}} a \in \mathbb{R}$ *and* $Y_n \xrightarrow{\mathbb{P}} b \in \mathbb{R}$, *then*

- $X_n + Y_n \xrightarrow{\mathbb{P}} a + b$;
- $X_n Y_n \xrightarrow{\mathbb{P}} ab$.

Convergence in Distribution and Pointwise Convergence of Characteristic Functions

Theorem 5.11 (Lévy's Continuity Theorem)

$\mathbb{X}_n \xrightarrow{\mathbb{D}} \mathbb{X} \Leftrightarrow \varphi_{\mathbb{X}_n}(\mathbb{t}) \to \varphi_{\mathbb{X}}(\mathbb{t}),\ \forall \mathbb{t} \in \mathbb{R}^k$

Example 5.12 (CF of a Normal Distribution)

$X \sim \mathrm{N}(\mu, \sigma^2) \Rightarrow \varphi_X(t) = \exp\{i\mu t - \sigma^2 t^2/2\},\ t \in \mathbb{R}$

Definition 5.13 (Sequence of Independent Random Variables)

$\{X_n\}_{n\in\mathbb{N}}$ is a *sequence of independent* random variables if

$$F_{\{X_j\}_{j\in J}}(\{x_j\}_{j\in J}) = \prod_{j\in J} F_{X_j}(x_j) \quad \forall \{x_j\}_{j\in J} \in \mathbb{R}^{|J|},\ \forall J \subseteq \mathbb{N},\ |J| < \infty.$$

Weak Law of Large Numbers

Definition 5.14 (IID Sequence; IID = Independent and Identically Distributed)

$\{X_n\}_{n\in\mathbb{N}}$ is a *sequence of IID* random variables if it is a sequence of independent random variables having the same CDF.

The above can be defined for *random vectors* as well.

Theorem 5.15 (Weak Law of Large Numbers (WLLN))

If $\{X_n\}_{n\in\mathbb{N}}$ *is an IID sequence of random variables with* $\mathbb{E}|X_1| < \infty$*, then* $\overline{X}_n := n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}X_1$ *as* $n \to \infty$.

- Strong Law of Large Numbers (SLLN): replacing $\xrightarrow{\mathbb{P}}$ by $\xrightarrow{\mathbb{P}\text{-a.s.}}$
- Finite variance is *not* required. Although, the underlying proof would be simpler, cf. Chebyshev's inequality.
- Independence can be relaxed.
- Identical distribution can be relaxed.

Central Limit Theorem for IID

Theorem 5.16 (Central Limit Theorem (CLT))

If $\{X_n\}_{n\in\mathbb{N}}$ is an IID sequence of random variables with $\mathbb{E}X_1^2 < \infty$ and $\mathbb{V}\mathrm{ar}X_1 > 0$, then

$$Z_n := \sqrt{n}\frac{\overline{X}_n - \mathbb{E}X_1}{\sqrt{\mathbb{V}\mathrm{ar}X_1}} \xrightarrow{\mathbb{D}} Z \sim \mathsf{N}(0,1), \quad n \to \infty.$$

In other words,

$$\lim_{n\to\infty}\mathbb{P}[Z_n \le x] = \Phi(x) \equiv \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\exp\{-t^2/2\}dt, \quad \forall x \in \mathbb{R}.$$

In short,

$$Z_n \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}(0,1).$$

Multivariate Central Limit Theorem

Corollary 5.17 (Cramér–Wold Device)

$\mathbb{X}_n \xrightarrow{\mathbb{D}} \mathbb{X} \Leftrightarrow \mathbb{t}^\intercal \mathbb{X}_n \xrightarrow{\mathbb{D}} \mathbb{t}^\intercal \mathbb{X}$, $\forall \mathbb{t} \in \mathbb{R}^k$

The Cramér–Wold Theorem is a trivial consequence of the Lévy's Continuity Theorem.

Theorem 5.18 (Multivariate CLT)

If $\{\mathbb{X}_n\}_{n\in\mathbb{N}}$ is an IID sequence of k-dimensional random vectors with the positive definite variance-covariance matrix $\mathbb{V}\mathrm{ar}\mathbb{X}_1$, then

$$\sqrt{n}\left(\overline{\mathbb{X}}_n - \mathbb{E}\mathbb{X}_1\right) \xrightarrow{\mathbb{D}} \mathsf{N}_k(\mathbb{0}, \mathbb{V}\mathrm{ar}\mathbb{X}_1), \quad n \to \infty.$$

Delta Method

Theorem 5.19 (Delta Method)

If $\sqrt{n}(Y_n - \mu) \xrightarrow{\mathbb{D}} \mathsf{N}(0, \sigma^2)$ *and* g *is continuously differentiable on the neighborhood of* μ *such that* $g'(\mu) \neq 0$*, then*

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{\mathbb{D}} \mathsf{N}\Big(0, \big(g'(\mu)\big)^2 \sigma^2\Big), \quad n \to \infty.$$

Agenda

6. Statistical Learning

6.1 Random Sample
6.2 Statistical Experiment
6.3 Stochastic Models
6.4 Parametric Models
6.5 Non-parametric Models
6.6 Fundamental Concepts in Inference
6.7 Estimation
6.8 Standard Error
6.9 Mean Squared Error
6.10 Confidence Sets

Statistical Learning

- *Statistical Learning* a.k.a. Statistical Inference

= The process of using *data to infer the distribution that generated the data*

? Given a random sample $X_1, \ldots, X_n \overset{IID}{\sim} F$, how do we infer F ?

Random Sample

Definition 6.1 (Random Sample and Sample Size)

If $X_1, \ldots, X_n$ are *independent* and each has the *same marginal distribution* with CDF F, we say that $X_1, \ldots, X_n$ are IID (independent and identically distributed) and we write

$$X_1, \ldots, X_n \overset{IID}{\sim} F.$$

We also call a *random sample* of size n from F.

Data From Experiment

- We consider / think of / assume:
  - *measurable mappings* $X_i : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $i = 1, \ldots, n$
- We observe / measure / obtain:
  - *real-valued data* $X_i(\omega) \in \mathbb{R}$, $i = 1, \ldots, n$ for one particular $\omega \in \Omega$

Stochastic Model

- *Stochastic model* $\approx$ Probabilistic model, Statistical model, ...

- Parametric model
  - A set $\mathcal{F}$ that can be parameterized by a finite number of parameters

- Non-parametric model
  - A set $\mathcal{F}$ that cannot be parameterized by a *finite* number of parameters

Parametric Models

Example 6.2 (Normal Model)

$$\mathcal{F} = \left\{ f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\},\ \mu \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

- Data come from a Normal distribution with *two* parameters μ and σ^2
- In general,

$$\mathcal{F} = \left\{ f(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d \right\}$$

- Some notation: $\mathbb{P}_\theta[X \in A] = \int_A f(x;\theta)\mathrm{d}x$, $E_\theta[g(X)] = \int_{\mathbb{R}} g(x) f(x;\theta)\mathrm{d}x$

Non-parametric Models

Example 6.3 (Sobolev Space Model)

$$\mathcal{F} = \left\{ f : \int_{\mathbb{R}} \{f''(x)\}^2 \mathrm{d}x < \infty \right\}.$$

- Data come from a distribution having a *density*, which is not too "wiggly"
- The distinction between parametric and non-parametric is more subtle than this but we don't need a rigorous definition for our purposes
- For instance, the whole PDF can be considered as *infinite dimensional* parameter
- Semi-parametric models

Point Estimation

Definition 6.4 (Estimator)

A point estimator $\widehat{\theta}_n$ of a parameter θ is a measurable function t of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = t(X_1, \ldots, X_n).$$

- Here, it is not generally required that $X_1, \ldots, X_n$ are IID
- Important:
  - *Parameter* θ is a fixed real number (vector), but unknown
  - *Estimator* $\widehat{\theta}_n$ is a random variable (a measurable function of random variables), but known (i.e., obtainable from the data)

Consistent Estimator

- $\widehat{\theta}_n$ is unbiased if $\mathbb{E}\left[\widehat{\theta}_n\right] = \theta$ for every $n \in \mathbb{N}$
- Unbiasedness used to receive much attention but these days is less important
- Many of the estimators we will use are biased
- The bias of an estimator is defined by $\text{bias}(\widehat{\theta}_n) = \mathbb{E}\left[\widehat{\theta}_n\right] - \theta$
- A reasonable requirement for an estimator is that it should "converge" to the true parameter value *as we collect more and more data*

Definition 6.5 (Consistent Estimator)

A point estimator $\widehat{\theta}_n$ of a parameter θ is *consistent* if $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$.

- *(weak) consistency* $\Leftrightarrow$ in probability $\leftrightarrow$ *strong consistency* $\Leftrightarrow$ almost surely
- *Qualitative* property

Standard Error

- The distribution of $\widehat{\theta}_n$ is called the *sampling distribution*
- The *standard deviation* of $\widehat{\theta}_n$ is called the standard error: $\mathrm{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}\mathrm{ar}\widehat{\theta}_n}$
- Often, the standard error depends on the *unknown* F
- In those cases, se is an unknown quantity (*parameter*), but we usually can estimate it
- The *estimated* standard error is denoted by $\widehat{\mathrm{se}}$

Example 6.6 (Standard Error in Alternative Model – Flipping a Coin)

Bernoulli random sample $X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{Be}(p)$ and parameter $p \in (0, 1)$
$\ldots$ $\widehat{p}_n = n^{-1} \sum_{i=1}^{n} X_i$ relative frequency as an estimator $\ldots$ $\widehat{\mathrm{se}}(\widehat{p}_n) = \sqrt{\widehat{p}_n(1 - \widehat{p}_n)/n}$.

Mean Squared Error

- The quality of a point estimate is sometimes assessed by the mean squared error

$$\mathrm{MSE}(\widehat{\theta}_n) = \mathbb{E}_\theta\left[\widehat{\theta}_n - \theta\right]^2$$

- Keep in mind that $\mathbb{E}_\theta$ refers in case of IID X's to expectation with respect to the distribution $f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$ that generated the data
- *Quantitative* property

Theorem 6.7 (Variance-Bias Decomposition of MSE)

$$\mathrm{MSE}(\widehat{\theta}_n) = \mathrm{bias}^2(\widehat{\theta}_n) + \mathbb{V}\mathrm{ar}(\widehat{\theta}_n)$$

Weak Consistency

Theorem 6.8 (Sufficient Condition for Consistency)

$\text{bias}(\widehat{\theta}_n) \to 0$ & $\mathbb{V}\text{ar}(\widehat{\theta}_n) \to 0 \Rightarrow \widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ *(a consistent estimator)*

Example 6.9 (Consistency – Flipping a Coin)

Example 6.6: $\mathbb{E}(\widehat{p}_n) = p$ & $\mathbb{V}\text{ar}(\widehat{p}_n) = p(1-p)/n \to 0 \Rightarrow \widehat{p}_n \xrightarrow{\mathbb{P}} p$

Definition 6.10 (Asymptotically Standard Normal Estimator)

An estimator $\widehat{\theta}_n$ of a parameter θ is *asymptotically standard normal* if

$$\frac{\widehat{\theta}_n - \theta}{\text{se}(\widehat{\theta}_n)} \xrightarrow{\mathbb{D}} \text{N}(0,1), \quad n \to \infty.$$

Confidence Sets

Definition 6.11 (Confidence Interval)

A $(1-\alpha)$-confidence interval for a parameter θ is an interval $C_n = (a, b)$, where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are measurable functions of the data such that

$$\mathbb{P}_\theta[\theta \in C_n] = 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

An approximate $(1-\alpha)$-confidence interval for a parameter θ is an interval C_n such that

$$\lim_{n\to\infty} \mathbb{P}_\theta[\theta \in C_n] = 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

- In words, (a, b) traps θ with probability (approximately) $1-\alpha$
- We call $1-\alpha$ the *coverage* of the confidence interval (CI)
- Warning: C_n is *random* and θ is *fixed*
- $\theta \in \mathbb{R}^d$, $d > 1$: a *confidence set* (such as a sphere/ellipse) instead of an interval
- A CI is not a probability statement about θ since θ is fixed, not a random variable

CI Normal Model

Example 6.12 (CI – Normality)

Example 6.2: $X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma^2)$ and parameter $\mu \in \mathbb{R}$ is *unknown* and to be be estimated (point estimator and confidence interval), $\sigma^2 > 0$ is supposed to be *known*

- Point estimator of μ: $\widehat{\mu}_n = n^{-1} \sum_{i=1}^{n} X_i \equiv \overline{X}_n$
- Linearity of the normal distributions: $\sqrt{n}(\widehat{\mu}_n - \mu)/\sigma \sim \mathrm{N}(0,1)$, $\forall n \in \mathbb{N}$
- Qunatiles of the standard normal distribution:

$$\mathbb{P}\left[-u_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2}} \leq +u_{1-\alpha/2}\right] = 1 - \alpha, \quad \forall n \in \mathbb{N} \text{ and } \forall \mu \in \mathbb{R}$$

- Confidence interval for μ: $\left(\overline{X}_n - u_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \overline{X}_n + u_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$

Length of the Confidence Interval

Length of the CI in the normal model for μ with $\sigma^2 > 0$ known (provided):

$$\text{CI: } \overline{X}_n \pm u_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \Rightarrow \quad \text{Length: } 2u_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

- Higher coverage $1-\alpha \Rightarrow$ wider CI
- Larger variability $\sigma^2 \Rightarrow$ wider CI
- Smaller sample size $n \Rightarrow$ wider CI

How many observations do I need for a CI narrower than d? ... $n \geq \left\lfloor \frac{4u_{1-\alpha/2}^2\sigma^2}{d^2} \right\rfloor + 1$

Normal-based Confidence Interval

Theorem 6.13 (Normal-based CI)

Suppose that $\widehat{\theta}_n$ *is an asymptotically standard normal estimator of the parameter* θ *and* $\widehat{\mathrm{se}}(\widehat{\theta}_n)$ *is a consistent estimator of* $\mathrm{se}(\widehat{\theta}_n)$*, i.e.,* $\widehat{\mathrm{se}}(\widehat{\theta}_n) - \mathrm{se}(\widehat{\theta}_n) \xrightarrow{\mathbb{P}} 0$*. Let* $u_{1-\alpha/2}$ *be the* $(1-\alpha/2)$*-quantile of the standard normal distribution and*

$$C_n = \left(\widehat{\theta}_n - u_{1-\alpha/2}\widehat{\mathrm{se}}(\widehat{\theta}_n), \widehat{\theta}_n + u_{1-\alpha/2}\widehat{\mathrm{se}}(\widehat{\theta}_n)\right).$$

Then,

$$\mathbb{P}_\theta\left[\theta \in C_n\right] \to 1-\alpha, \quad n \to \infty.$$

- Informally: $\widehat{\theta}_n \approx \mathrm{N}\left(\theta, \widehat{\mathrm{se}}(\widehat{\theta}_n)\right)$
- Approximately: For 95%-confidence intervals, $\alpha = 0.05$ and $u_{.975} \doteq 1.96 \approx 2$ leading to the approximate 95%-confidence interval $\widehat{\theta}_n \pm 2\widehat{\mathrm{se}}(\widehat{\theta}_n)$

CLT Confidence Interval

Example 6.14 (CI – Flipping a Coin)

Example 6.9:

- $\widehat{p}_n \xrightarrow{\mathbb{P}} p$ & $\mathsf{se}(\widehat{p}_n) = \sqrt{p(1-p)/n}$ & $\widehat{\mathsf{se}}(\widehat{p}_n) := \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$ & Slutsky's $\Rightarrow$ $\widehat{\mathsf{se}}(\widehat{p}_n) - \mathsf{se}(\widehat{p}_n) \xrightarrow{\mathbb{P}} 0$
- CLT $\Rightarrow \frac{\widehat{p}_n - p}{\mathsf{se}(\widehat{p}_n)} \xrightarrow{\mathbb{D}} \mathsf{N}(0,1)$

Then, by Slutsky's Theorem once again, $\Rightarrow \frac{\widehat{p}_n - p}{\widehat{\mathsf{se}}(\widehat{p}_n)} \xrightarrow{\mathbb{D}} \mathsf{N}(0,1)$. Thus,

$$\widehat{p}_n \pm u_{1-\alpha/2}\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$$

is an approximate $(1-\alpha)$-confidence interval for p.

Agenda

7. Statistical Functionals

7.1 Empirical Distribution Function
7.2 Properties of ECDF
7.3 Statistical Functional
7.4 Plug-in Estimator
7.5 Linear Statistical Functional
7.6 Plug-in Estimator For Linear Statistical Functional

Empirical Distribution Function

- $X_1, \ldots, X_n \overset{IID}{\sim} F$ is a *random sample* from F with *sample size* n
- To estimate F with its empirical *counterpart*

Definition 7.1 (ECDF)

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}, \quad x \in \mathbb{R}.$$

- The ECDF puts mass $1/n$ at each data point X_i
- The *relative frequency* of X's being smaller or equal to a fixed x, i.e., $\#\{X_i \leq x\}/n$

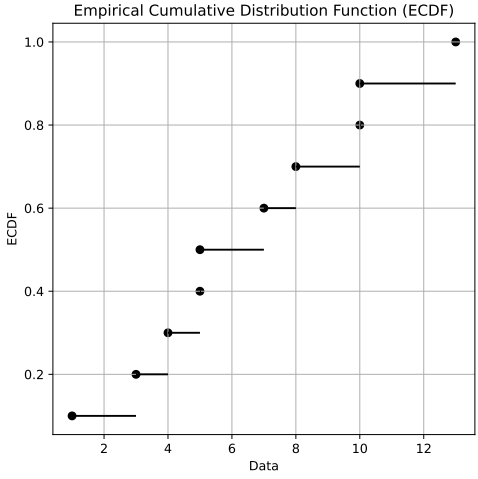
ECDF in Python

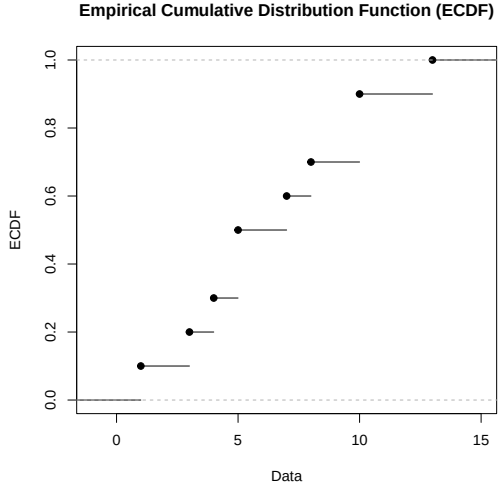
```python
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
# Example data
data = [1, 3, 4, 5, 5, 7, 8, 10, 10, 13]
# Compute ECDF
ecdf = sm.distributions.empirical_distribution.ECDF(data)
# Plot ECDF
plt.figure(figsize=(6, 6))
plt.hlines(ecdf.y[:-1],ecdf.x[:-1],ecdf.x[1:], color='black')
plt.scatter(ecdf.x, ecdf.y, marker='o', color='black')
plt.xlabel('Data')
plt.ylabel('ECDF')
plt.title('Empirical Cumulative Distribution Function (ECDF)')
plt.grid(True)
plt.show()
```

ECDF in R

```r
data <- c(1, 3, 4, 5, 5, 7, 8, 10, 10, 13)
plot(ecdf(data), xlab='Data', ylab='ECDF', main='Empirical Cumulative
    Distribution Function (ECDF)')
```

ECDF in Python & R





Properties of ECDF

Theorem 7.2 (Pointwise Properties of ECDF)

At any fixed $x \in \mathbb{R}$,

- $\mathbb{E}\left[\widehat{F}_n(x)\right] = F(x)$;
- $\mathbb{V}\text{ar}\left[\widehat{F}_n(x)\right] = \frac{F(x)\{1-F(x)\}}{n}$;
- $\text{MSE}\left(\widehat{F}_n(x)\right) = \frac{F(x)\{1-F(x)\}}{n} \to 0$ *as* $n \to \infty$;
- $\widehat{F}_n(x) \xrightarrow{\mathbb{P}} F(x)$ *as* $n \to \infty$.

Statistical Functional

A functional is just a function of a function.

Definition 7.3 (Functional)

A functional is a mapping $T : \mathcal{F} \to \mathbb{R}$, where $\mathcal{F}$ is a set of functions.

Definition 7.4 (Statistical Functional)

A statistical functional is a map T that maps a distribution function F to a real number.

A vector functional can be defined as well (just to replace the "output" by $\mathbb{R}^d$).

Example 7.5 (Mean, Variance, Median)

- $\mu \equiv \mathbb{E}X = \int x \mathrm{d}F(x)$ ... mean
- $\sigma^2 \equiv \mathbb{V}\mathrm{ar}X = \int (x-\mu)^2 \mathrm{d}F(x)$ ... variance
- $\mathrm{median}(X) = F^{-1}(1/2)$ ... median

Plug-in (Empirical) Estimator

Definition 7.6 (Plug-in Estimator)

The plug-in estimator of $\theta = T(F)$ is defined by $\widehat{\theta}_n = T(\widehat{F}_n)$.

Just plug in the *empirical* $\widehat{F}_n$ for the *unknown* F ... a.k.a. Empirical estimator

Linear Statistical Functional

Definition 7.7 (Linear Statistical Functional)

If $T(F) = \int r(x)\mathrm{d}F(x)$ for some measurable function r, then T is called a linear statistical functional.

The reason $T(F) = \int r(x)\mathrm{d}F(x)$ is called a linear functional is because T satisfies

$$T(aF + bG + c) = aT(F) + bT(G) + c, \quad a, b, c \in \mathbb{R};$$

if both sides exist. Hence, T is *linear* in its arguments.

Plug-in Estimator For Linear Statistical Functional

Theorem 7.8 (Empirical Estimator For Linear Statistical Functional)

The plug-in estimator for the linear statistical functional $T(F) = \int r(x)dF(x)$ *is*

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

Example 7.9 (The Mean)

$\mu = T(F) = \int x \mathrm{d}F(x) \Rightarrow \widehat{\mu}_n = \int x \mathrm{d}\widehat{F}_n(x) = \overline{X}_n$

Empirical Variance

Example 7.10 (The Variance)

$$\mathbb{V}\mathrm{ar}X = \sigma^2 = T(F) = \int (x-\mu)^2 \mathrm{d}F(x) = \int x^2 \mathrm{d}F(x) - \left(\int x \mathrm{d}F(x)\right)^2 \Rightarrow$$

$$\widehat{\sigma}_n^2 = \int x^2 \mathrm{d}\widehat{F}_n(x) - \left(\int x \mathrm{d}\widehat{F}_n(x)\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$$

Another reasonable estimator of σ^2 is the *sample variance*

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$$

Empirical Correlation

Example 7.11 (Correlation)

Let $Z = (X, Y)$ and let $\rho = T(F) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)/(\sigma_X \sigma_Y)$ denote the correlation between X and Y, where $F(x, y)$ is bivariate. We can write

$$T(F) = a\big(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F)\big),$$

where

$$T_1(F) = \int x \mathrm{d}F(x, y); \qquad T_2(F) = \int y \mathrm{d}F(x, y); \qquad T_3(F) = \int xy \mathrm{d}F(x, y);$$

$$T_4(F) = \int x^2 \mathrm{d}F(x, y); \qquad T_5(F) = \int y^2 \mathrm{d}F(x, y); \qquad \text{and}$$

$$a(t_1, t_2, t_3, t_4, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}.$$

Empirical Correlation (cont.)

Example 7.12 (Correlation (cont.))

Replace F with $\widehat{F}_n$ in $T_1(F)$, $T_2(F)$, $T_3(F)$, $T_4(F)$, $T_5(F)$, and take

$$\widehat{\rho} = T(\widehat{F}_n) = a\big(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n F), T_4(\widehat{F}_n), T_5(\widehat{F}_n)\big),$$

We get

$$\widehat{\rho} = \frac{\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_i (X_i - \overline{X}_n)^2 \sum_i (Y_i - \overline{Y}_n)^2}}$$

which is called the sample correlation.

Empirical Quantiles

Example 7.13 (Quantiles)

For $p \in (0,1)$, the *p-th quantile* is defined by

$$T(F) = F^{-1}(p) = \inf\Big\{x : \ F(x) > p\Big\}.$$

Now, we define

$$T(\widehat{F}_n) = \widehat{F}_n^{-1}(p) = \inf\Big\{x : \ \widehat{F}_n(x) > p\Big\}$$

and we call it the *p*-th sample quantile. Thus, the sample median is $\widehat{F}_n^{-1}(1/2)$. Moreover, the *interquartile range (IQR)* $\widetilde{T}(F) = F^{-1}(3/4) - F^{-1}(1/4)$ can be estimated through sample interquartile range $\widetilde{T}(\widehat{F}_n) = \widehat{F}_n^{-1}(3/4) - \widehat{F}_n^{-1}(1/4)$.

Agenda

8. Bootstrap

Bootstrapping the Statistics

Bootstrap is a class of methods for

- estimating standard errors;
- computing confidence intervals;
- testing hypotheses;
- calculation prediction intervals; ...

Definition 8.1 (Statistic)

A statistic $T_n \equiv T_n(X_1, \ldots, T_n)$ is *any* measurable function T_n of data $X_1, \ldots, X_n$.

- The simplest case of data $X_1, \ldots, X_n$ is a *random sample*, i.e., *IID data*
- For instance, every estimator is a statistic
- However, an estimator is related to some quantity of the distribution

Variance of Statistic

Suppose we want to know $\mathbb{V}\mathrm{ar}_F\, T_n$, i.e., the variance of T_n

- We have written $\mathbb{V}\mathrm{ar}_F$ to emphasize that the variance usually *depends* on the unknown distribution function F
- For example, if $T_n = \overline{X}_n$ for IID X_i's, then $\mathbb{V}\mathrm{ar}_F\, T_n = \sigma^2/n$, where $\sigma^2 = \int (x-\mu)^2 \mathrm{d}F(x)$ and $\mu = \int x \mathrm{d}F(x)$
- Thus, the variance of T_n is a function of F

The bootstrap idea has two steps:

(1) *Estimate* $\mathbb{V}\mathrm{ar}_F\, T_n$ with $\mathbb{V}\mathrm{ar}_{\widehat{F}_n}\, T_n$

(2) *Approximate* $\mathbb{V}\mathrm{ar}_{\widehat{F}_n}\, T_n$ using simulations

Suppose $T_n = \overline{X}_n$ and let $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$

(i) $\mathbb{V}\mathrm{ar}_{\widehat{F}_n}\, T_n = \widehat{\sigma}^2/n$ and Step (1) is enough

(ii) However in more complicated cases, we cannot write down a simple formula for $\mathbb{V}\mathrm{ar}_{\widehat{F}_n}\, T_n$, which is why we need Step (2)

Simulation

- Suppose we draw an IID sample $Y_1, \ldots, Y_B$ from a distribution G
- By the LLN,

$$\overline{Y}_B = \frac{1}{B} \sum_{b=1}^{B} Y_b \xrightarrow[B \to \infty]{\mathbb{P}} \int y \mathrm{d}G(y) = \mathbb{E}\, Y$$

- So if we draw a *large sample* from G, we can use the sample mean $\overline{Y}_B$ to approximate $\mathbb{E}\, Y$
- In a simulation, we can make B *as large as we like*, in which case, the difference between Y and $\mathbb{E}\, Y$ is *negligible*

Simulation (cont.)

- More generally, if h is any function with finite mean, then

$$\frac{1}{B}\sum_{b=1}^{B} h(Y_b) \xrightarrow[B\to\infty]{\mathbb{P}} \int h(y)\mathrm{d}G(y) = \mathbb{E}[h(Y)]$$

- In particular,

$$\frac{1}{B}\sum_{b=1}^{B}\left(Y_b - \overline{Y}_B\right)^2 = \frac{1}{B}\sum_{b=1}^{B} Y_b^2 - \left(\frac{1}{B}\sum_{b=1}^{B} Y_b\right)^2$$

$$\xrightarrow[B\to\infty]{\mathbb{P}} \int y^2\mathrm{d}G(y) - \left(\int y\mathrm{d}G(y)\right)^2 = \mathbb{V}\mathrm{ar}\, Y$$

- Hence, we can use the *sample variance of the simulated values* to approximate $\mathbb{V}\mathrm{ar}\, Y$

Bootstrap Variance Estimation

- We just learned, we can approximate $\mathbb{V}\text{ar}_{\widehat{F}_n} T_n$ by simulation
- Now, $\mathbb{V}\text{ar}_{\widehat{F}_n} T_n$ means "the variance of T_n if the distribution of the data is $\widehat{F}_n$"
- How can we simulate from the distribution of T_n, when the data are assumed to have distribution $\widehat{F}_n$?
- The answer is to simulate $X_1^\star, \ldots, X_n^\star$ from $\widehat{F}_n$ and, then, compute $T_n^\star = T_n(X_1^\star, \ldots, X_n^\star)$
- The idea is illustrated in the following diagram:

$$\begin{array}{rccccc} \text{Real world} & F & \Rightarrow & X_1, \ldots, X_n & \Rightarrow & T_n = T_n(X_1, \ldots, X_n) \\ \text{Bootstrap world} & \widehat{F}_n & \Rightarrow & X_1^\star, \ldots, X_n^\star & \Rightarrow & T_n^\star = T_n(X_1^\star, \ldots, X_n^\star) \end{array}$$

Bootstrap Variance Estimation (cont.)

- How do we simulate $X_1^\star, \ldots, X_n^\star$ from $\widehat{F}_n$?
- Notice that in puts mass $1/n$ at each data point $X_1, \ldots, X_n$
- Therefore, *drawing an observation from $\widehat{F}_n$ is equivalent to drawing one point at random from the original data set*

Example 8.2 (Summary of Bootstrap Variance Estimation)

(1) Simulate $X_1^\star, \ldots, X_n^\star \sim \widehat{F}_n \Leftrightarrow$ draw n observations with replacement from $X_1, \ldots, X_n$

(2) Compute $T_n^\star = T_n(X_1^\star, \ldots, X_n^\star)$

(3) Repeat steps (1) and (2), B times, to get ${}_{(1)}T_n^\star, \ldots, {}_{(B)}T_n^\star$

(4) Let

$$\widehat{\mathbb{V}\text{ar}}^\star T_n = \frac{1}{B}\sum_{b=1}^{B}\left({}_{(b)}T_n^\star - \frac{1}{B}\sum_{j=1}^{B}{}_{(j)}T_n^\star\right)^2$$

Bootstrap Variance of Sample Mean

$$\mathbb{V}\text{ar}_F\, T_n \approx \mathbb{V}\text{ar}_{\widehat{F}_n}\, T_n \equiv \mathbb{V}\text{ar}^\star\, T_n \approx \widehat{\mathbb{V}\text{ar}}^\star\, T_n$$

```
# scientifc computing
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# set seed for the random number generator
np.random.seed(2024)

# example data
dt = [1, 3, 4, 5, 5, 7, 8, 10, 10, 13]
```

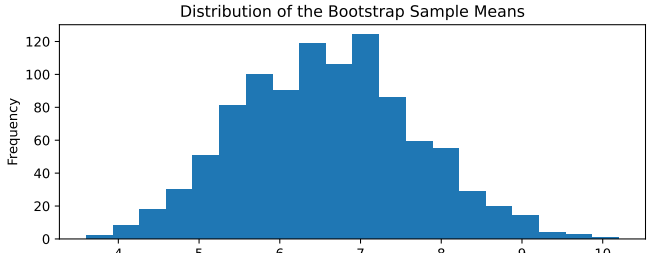
Bootstrap Variance of Sample Mean (cont.)

```
def create_bootstrap_samples(sample_size = len(dt), B_samples = 1000):

# create a list for sample means
sample_means = []

# loop n_samples times
for i in range(B_samples):
# create a bootstrap sample of sample_size with replacement
bootstrap_sample = np.random.choice(dt, size = sample_size, replace =
    True)
# calculate the bootstrap sample mean
sample_mean = bootstrap_sample.mean()
# add this sample mean to the sample means list
sample_means.append(sample_mean)

return pd.Series(sample_means)
```

Bootstrap Variance of Sample Mean (cont. II)

```
# create bootstrap samples
sample_means = create_bootstrap_samples()
# calculate bootstrap variance of the sample mean
sample_means.var()
# plot the distribution
sample_means.plot(kind = 'hist', bins = 20, title = 'Distribution of the
    Bootstrap Sample Means')
```

$$\widehat{\mathbb{V}\mathrm{ar}}^{\star} T_n \approx 1.1818$$



Bootstrap Confidence Intervals

- *Several ways* to construct bootstrap CIs
- Here, we discuss three nonparametric methods, although there are also *parametric* bootstrap methods

Method 1: Pivotal Intervals

- $\theta = T(F)$ and $\widehat{\theta}_n = T(\widehat{F}_n)$
- Define the pivot $R_n := \widehat{\theta}_n - \theta$
- Let ${}_{(1)}\widehat{\theta}^{\star}_n, \ldots, {}_{(B)}\widehat{\theta}^{\star}_n$ be the bootstrap replications of $\widehat{\theta}_n$
- Denote the CDF of the pivot

$$H(r) = \mathbb{P}[R_n \leq r]$$

- Define $C^{\star}_n = (a, b)$, where

$$a = \widehat{\theta}_n - H^{-1}(1 - \alpha/2) \quad \text{and} \quad b = \widehat{\theta}_n - H^{-1}(\alpha/2)$$

Pivotal Intervals

- Now, it follows (if H is continuous)

$$\begin{aligned}
&\mathbb{P}[a < \theta < b] = \mathbb{P}[a - \widehat{\theta}_n < \theta - \widehat{\theta}_n < b - \widehat{\theta}_n] = \mathbb{P}[\widehat{\theta}_n - b < \widehat{\theta}_n - \theta < \widehat{\theta}_n - a] \\
&= \mathbb{P}[\widehat{\theta}_n - b < R_n < \widehat{\theta}_n - a] = H(\widehat{\theta}_n - a) - H(\widehat{\theta}_n - b) \\
&= H\big(H^{-1}(1-\alpha/2)\big) - H\big(H^{-1}(\alpha/2)\big) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha
\end{aligned}$$

- Hence, $C_n^\star$ is an exact $1-\alpha$ confidence interval for θ
- Unfortunately, a and b depend on the *unknown* distribution H, but we can form a bootstrap estimate of H

$$\widehat{H}_n(r) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{{}_{(b)}R_n^\star \leq r\}, \quad \text{where } {}_{(b)}R_n^\star = {}_{(b)}\widehat{\theta}_n^\star - \widehat{\theta}_n$$

Pivotal Intervals (cont.)

- Let $r_\beta^\star$ denote the β sample quantile of $\left({}_{(1)}R_n^\star, \ldots, {}_{(B)}R_n^\star\right)$
- Let $\theta_\beta^\star$ denote the β sample quantile of $\left({}_{(1)}\widehat{\theta}_n^\star, \ldots, {}_{(B)}\widehat{\theta}_n^\star\right)$
- Since $r_\beta^\star = \theta_\beta^\star - \widehat{\theta}_n$, then $C_n = (\widehat{a}, \widehat{b})$ is an approximate $(1-\alpha)$-confidence interval, where

$$\widehat{a} = \widehat{\theta}_n - \widehat{H}^{-1}(1-\alpha/2) = \widehat{\theta}_n - r_{1-\alpha/2}^\star;$$
$$\widehat{b} = \widehat{\theta}_n - \widehat{H}^{-1}(\alpha/2) = \widehat{\theta}_n - r_{\alpha/2}^\star$$

- Then, the $(1-\alpha)$-bootstrap pivotal confidence interval is

$$C_n = \left(2\widehat{\theta}_n - \theta_{1-\alpha/2}^\star, 2\widehat{\theta}_n - \theta_{\alpha/2}^\star\right)$$

Normal Intervals

Method 2: Normal Intervals

- Suppose that $\widehat{\theta}_n$ is an *asymptotically standard normal* estimator of the parameter θ
- The simplest method gives the $(1-\alpha)$-bootstrap normal confidence interval

$$\widehat{\theta}_n \pm u_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}^{\star}\widehat{\theta}_n}$$

- Note that $\sqrt{\widehat{\mathbb{V}\mathrm{ar}}^{\star}\widehat{\theta}_n} =: \mathrm{se}^{\star}(\widehat{\theta}_n)$ is the *bootstrap estimate of the standard error*

Percentile Intervals

Method 3: Percentile Intervals

- The $(1-\alpha)$-bootstrap percentile confidence interval is

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

- The idea in-behind: Suppose there exists a monotone transformation $U = m(T)$ such that $U \sim \mathsf{N}(\phi, c^2)$ where $\phi = m(\theta)$
- We do not suppose we know the transformation, only that one *exists*
- Let $U^\star_t = m\left({}_{(b)}\widehat{\theta}^\star_n\right)$
- Let $u^\star_\beta$ be the β sample quantile of the $U^\star_b$'s
- Since a monotone transformation *preserves* quantiles, we have that $u^\star_\beta = m(\theta^\star_\beta)$

Percentile Intervals (cont.)

- Also, since $U \sim \mathsf{N}(\phi, c^2)$, the $\alpha/2$ quantile of U is $\phi + u_{\alpha/2}c$
- Hence $u^\star_{\alpha/2} = \phi + u_{\alpha/2}c$
- Similarly, $u^\star_{1-\alpha/2} = \phi + u_{1-\alpha/2}c$
- Therefore,

$$\begin{aligned}
&\mathbb{P}[\theta^\star_{\alpha/2} < \theta < \theta^\star_{1-\alpha/2}] = \mathbb{P}\left[m\left(\theta^\star_{\alpha/2}\right) < m(\theta) < m\left(\theta^\star_{1-\alpha/2}\right)\right] \\
&= \mathbb{P}[u^\star_{\alpha/2} < \phi < u^\star_{1-\alpha/2}] = \mathbb{P}[U + u_{\alpha/2}c < \phi < U + u_{1-\alpha/2}c] \\
&= \mathbb{P}\left[u_{\alpha/2} < \frac{U-\phi}{c} < u_{1-\alpha/2}\right] = 1 - \alpha
\end{aligned}$$

- An exact normalizing transformation will rarely exist, but there may exist *approximate normalizing transformations*

Computing Bootstrap CIs

Example 8.3 (Skewness as a Measure of Asymmetry)

Let $\theta = T(F) = \int (x-\mu)^3 \mathrm{d}F(x)/\sigma^3 =: \widetilde{\mu}_3$ be the skewness. A normal distribution, for example, has skewness 0. The plug-in estimator of the skewness is

$$\widehat{\theta}_n = T(\widehat{F}_n) = \frac{\int (x-\mu)^3 \mathrm{d}\widehat{F}_n(x)}{\widehat{\sigma}^3}.$$

Let us consider a standard lognormal distribution. This means, assume that $Y_1, \ldots, Y_n \overset{IID}{\sim} \mathrm{N}(0,1)$ and we call that $X_i = \exp\{Y_i\}$, $i = 1, \ldots, n$ are IID having the standard lognormal distribution. Then, $\mu \equiv \mathbb{E}X_1 = \mathbb{E}\exp\{Y_1\} = \sqrt{e}$, $\sigma^2 = \mathbb{V}\mathrm{ar}X_1 = e^2 - e$, and $\theta \equiv \widetilde{\mu}_3 = (e+2)\sqrt{e-1} \doteq 6.185$.

Skewness for Lognormal Data

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from scipy.stats import lognorm
np.random.seed(2024)
def create_data(n=1000):
  y = norm.rvs(size=n)
  return np.exp(y)
def skewness(x):
  n = len(x)
  mu = sum(x) / n
  var = sum((x - mu)**2) / n
  return sum((x - mu)**3) / (n * var**(3/2))

# Creating the data
x = create_data(n=1000)
```

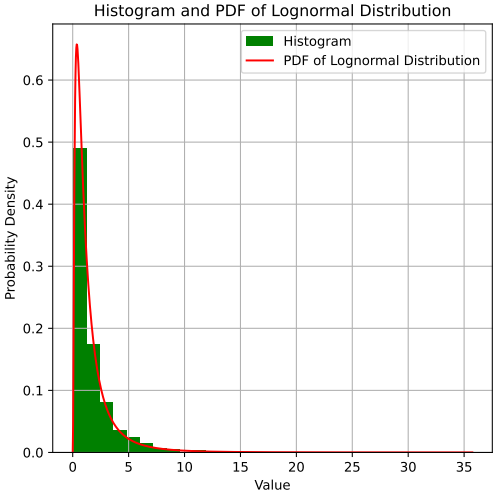
Histogram for Simulated Lognormal Data

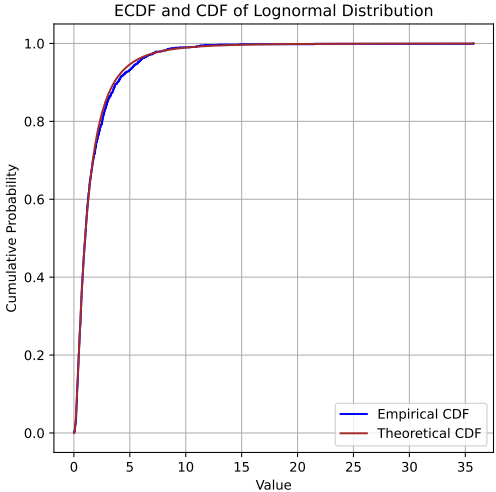
```
# Plot histogram
plt.figure(figsize=(6, 6))
plt.hist(x, bins=30, density=True, color='green', label='Histogram')

# Plot theoretical PDF of lognormal distribution
xx = np.linspace(0, max(x), 1000)
pdf = lognorm.pdf(xx, 1, scale=np.exp(0))

plt.plot(xx, pdf, 'red', label='PDF of Lognormal Distribution')
plt.title('Histogram and PDF of Lognormal Distribution')
plt.xlabel('Value')
plt.ylabel('Probability Density')
plt.legend()
plt.grid(True)
plt.show()
```

ECDF for Simulated Lognormal Data

```
# Sort the data for plotting ECDF
sorted_data = np.sort(x)
nn = len(sorted_data)
ecdf = np.arange(1, nn + 1) / nn

# Plot ECDF
plt.figure(figsize=(6, 6))
plt.step(sorted_data, ecdf, label='Empirical CDF', color='blue', where='
    post')

# Plot theoretical CDF of lognormal distribution
xx = np.linspace(0, max(sorted_data), 1000)
cdf = lognorm.cdf(xx, 1, scale=np.exp(0))
plt.plot(xx, cdf, color='brown', label='Theoretical CDF')
plt.title('ECDF and CDF of Lognormal Distribution')
plt.xlabel('Value'); plt.ylabel('Cumulative Probability')
plt.legend(); plt.grid(True); plt.show()
```

Lognormal Data in Python





Bootstrap CIs in Python

```
def bootstrap_values(x, B=10000):
  n = len(x)
  t_boot = np.empty(B)
  for i in range(B):
  xx = np.random.choice(x, n, replace=True)
  t_boot[i] = skewness(xx)
  return t_boot
def bootstrap_intervals(theta_hat, t_boot, alpha=0.05):
  se = t_boot.std()
  u = norm.ppf(1 - alpha/2)
  q_half_alpha = np.quantile(t_boot, alpha/2)
  q_c_half_alpha = np.quantile(t_boot, 1 - alpha/2)
  pivotal_conf = (2*theta_hat - q_c_half_alpha, 2*theta_hat -
    q_half_alpha)
  normal_conf = (theta_hat - u * se, theta_hat + u * se)
  percentile_conf = (q_half_alpha, q_c_half_alpha)
  return pivotal_conf, normal_conf, percentile_conf
```

Bootstrap CIs in Python (cont.)

```
# Nonparametric Bootstrap
theta_hat = skewness(x)
t_boot = bootstrap_values(x, B=10000)
pivotal_conf, normal_conf, percentile_conf = bootstrap_intervals(
    theta_hat, t_boot, alpha=0.05)

print('empirical skewness: \t %.3f' % theta_hat)
print('95%% confidence interval (pivotal): \t %.3f, %.3f' % pivotal_conf
    )
print('95%% confidence interval (Normal): \t %.3f, %.3f' % normal_conf)
print('95%% confidence interval (percentile): \t %.3f, %.3f' %
    percentile_conf)
```

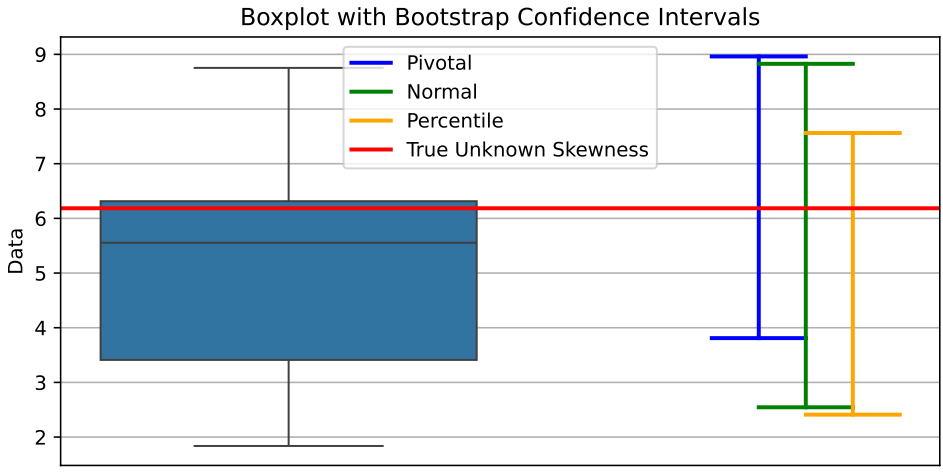
$\theta = 6.185, \quad \widehat{\theta}_n = 5.686,$

$C_n^{pivot} = (3.810, 8.962), \quad C_n^{norm} = (2.545, 8.827), \quad C_n^{perc} = (2.410, 7.562)$

Bootstrap Distribution with CIs in Python

```
intervals=[(pivotal_conf[0],pivotal_conf[1],'blue','Pivotal'),(
    normal_conf[0],normal_conf[1],'green','Normal'),(percentile_conf[0],
    percentile_conf[1],'orange','Percentile')] # Define three CIs
plt.figure(figsize=(7, 4))
sns.boxplot(data=t_boot, orient='v') # Create the box plot
for i, (start, end, color, name) in enumerate(intervals):
  plt.plot([0.9+i*0.1,1.1+i*0.1],[start,start],color=color,linestyle='-'
    ,linewidth=2,label=name)
  plt.plot([1+i*0.1,1+i*0.1],[start,end],color=color,linestyle='-',
    linewidth=2)
  plt.plot([0.9+i*0.1,1.1+i*0.1],[end,end],color=color,linestyle='-',
    linewidth=2) # Plot CIs
plt.axhline(y=6.185, color='r', linestyle='-', linewidth=2, label='True
    Unknown Skewness')
plt.title('Boxplot with Bootstrap Confidence Intervals')
plt.ylabel('Data'); plt.xticks([]); plt.grid(True)
plt.legend(loc='upper center'); plt.show()
```

Bootstrap Distribution of Skewness



Agenda

9. Parametric Inference

9.1 Parametric Family
9.2 Method of Moments
9.3 Multivariate Normal Distribution Revisited
9.4 Multivariate Delta Method
9.5 Asymptotic Properties of Method of Moments
9.6 Maximum Likelihood
9.7 Asymptotic Properties of Maximum Likelihood
9.8 Score Function and Fisher Information
9.9 Asymptotic Normality of MLE
9.10 Parametric Bootstrap
9.11 Computing Maximum Likelihood Estimates

Parametric Family

- Parametric family of models (or Family of parametric models):

$$\mathcal{F} = \left\{ f(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d \right\},$$

  where $\boldsymbol{\Theta}$ is the parameter space and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ is the parameter

? How would we ever know that the distribution that generated the data is in some parametric model?

! A reasonable *approximation* & *diagnostics* (Goodness-of-fit tests, ...)

Parameter of Interest

- We are only interested in some function $T(\boldsymbol{\theta})$
- For example, if $X_i \overset{IID}{\sim} \mathrm{N}(\mu, \sigma^2)$, then the parameter is $\boldsymbol{\theta} = (\mu, \sigma^2)$
- If our goal is to estimate μ, then $\mu = T(\boldsymbol{\theta})$ is called the parameter of interest and σ^2 is called a nuisance parameter
- The parameter of interest might be a *complicated* function of $\boldsymbol{\theta}$

Parameter of Interest (cont.)

Example 9.1 (Gamma Model)

Recall that X has a Gamma(α, β) distribution if

$$f_X(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\{-x/\beta\}, \quad x > 0;$$

where $\alpha, \beta > 0$ and

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} \exp\{-y\} \mathrm{d}y$$

is the Gamma function. The parameter is $\boldsymbol{\theta} = (\alpha, \beta)$. The Gamma distribution is sometimes used to model lifetimes of people, animals, and electronic equipment. Suppose we want to estimate the mean lifetime. Then, $T(\alpha, \beta) = \mathbb{E}_{\boldsymbol{\theta}} X = \alpha\beta$.

Parametric Estimators

Consider a *random sample* $X_1, \ldots, X_n \overset{IID}{\sim} F \in \mathcal{F}$

Example 9.2 (Just Finite Mean)

$\mathcal{F} = \left\{F(\mu) : \mathbb{E}_{F(\mu)} = \mu \ \& \ |\mu| < \infty\right\}$

- $\overline{X}_n$ is a *consistent* and *unbiased* estimator of μ
- X_1 is an *unbiased*, but *not a consistent* estimator of μ

Example 9.3 (Just Finite Variance)

$\mathcal{F} = \left\{F(\sigma^2) : \mathbb{V}\text{ar}_{F(\sigma^2)} = \sigma^2 \ \& \ \sigma^2 < \infty\right\}$

- $\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$ is a *consistent*, but *not an unbiased* estimator of σ^2
- $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$ is a *consistent* and *unbiased* estimator of σ^2

Parametric Estimators (cont.)

Example 9.4 (Poisson Model)

$\mathcal{F} = \{\text{Po}(\lambda), \lambda > 0\}$ and $\theta = \mathbb{P}[X_i = 0]$

- $\widehat{\theta}_n = n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i = 0\}$ is a *consistent* and *unbiased* estimator of λ
- $\widetilde{\theta}_n = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}$ is also a *consistent* and *unbiased* estimator of λ

Example 9.5 (Poisson Model – Awkward Case)

$\mathcal{F} = \{\text{Po}(\lambda), \lambda > 0\}$ and $\theta = \exp\{-2\lambda\}$

- The only *unbiased* estimator of θ is $(-1)^{X_1}$, however it never reaches an admissible value of $\exp\{-2\lambda\}$

Method of Moments

- A method for obtaining parametric estimators
- MoM stimators are not optimal, but they are often easy to compute
- They are are also useful as *starting values for other methods* that require iterative numerical routines
- Define the k-th (raw) moment, $1 \leq k \leq d$,

$$\mu'_k \equiv \mu'_k(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} X^k = \int x^k \mathrm{d}F_{\boldsymbol{\theta}}(x)$$

- Define the k-th sample moment

$$\widehat{\mu}'_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

MoM Estimators

Definition 9.6

The method of moments estimator $\widehat{\boldsymbol{\theta}}_n$ is defined to be the value of $\boldsymbol{\theta}$ such that

$$\mu_1'(\widehat{\boldsymbol{\theta}}_n) = \widehat{\mu}_1', \quad \mu_2'(\widehat{\boldsymbol{\theta}}_n) = \widehat{\mu}_2', \quad \ldots \quad \mu_d'(\widehat{\boldsymbol{\theta}}_n) = \widehat{\mu}_d'.$$

Alternatively, the *k-th centered moments* together with their empirical counterparts can be used instead.

Example 9.7 (Method of Moments Estimator for Alternative Distribution)

$X_1, \ldots, X_n \overset{IID}{\sim} \text{Be}(p)$. Then, $\mu_1' = \mathbb{E}_p X_1 = p$ and $\widehat{\mu}_1' = \overline{X}_n$. By equating these, we get the estimator

$$\widehat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and it is indeed the same *plug-in (empirical) estimator*.

MoM Estimators (cont.)

Example 9.8 (Method of Moments Estimator for Normal Distribution)

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma^2)$. Then, $\mu_1' = \mathbb{E}_{(\mu,\sigma^2)} X_1 = \mu$ and $\mu_2' = \mathbb{E}_{(\mu,\sigma^2)} X_1^2 = \mathbb{V}\mathrm{ar}_{(\mu,\sigma^2)} X_1 + (\mathbb{E}_{(\mu,\sigma^2)} X_1)^2 = \sigma^2 + \mu^2$. We need to solve the equations

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \widehat{\mu}_n^2 + \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

This is a system of 2 equations with 2 unknowns. The solution is

$$\widehat{\mu}_n = \overline{X}_n \quad \text{and} \quad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

MoM Estimators (cont. II)

Example 9.9 (MoM in Gamma Model – Alternative Parametrization)

$X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(a, p)$, where

$$f_X(x; a, p) = \frac{a^p}{\Gamma(p)} x^{p-1} \exp\{-ax\}, \quad x > 0;$$

(an alternative *parametrization*) where $a, p > 0$ and $\Gamma(p) = \int_0^\infty y^{p-1} \exp\{-y\} \mathrm{d}y$ is the Gamma function. Then, the MoM estimators

$$\widehat{a}_n = \frac{\overline{X}_n}{\widehat{\sigma}_n^2} \quad \text{and} \quad \widehat{p}_n = \frac{\overline{X}_n^2}{\widehat{\sigma}_n^2}$$

are *consistent* and *AN*.

MoM Estimators (cont. III)

Example 9.10 (MoM in Uniform Model – Both Sided)

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{U}(\theta_1, \theta_2)$. The MoM estimators

$$\widehat{\theta}_{1,n} = \overline{X}_n - \sqrt{3\widehat{\sigma}_n^2} \quad \text{and} \quad \widehat{\theta}_{2,n} = \overline{X}_n + \sqrt{3\widehat{\sigma}_n^2}$$

are *consistent* and *AN*.

Example 9.11 (MoM in Beta Model)

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{Beta}(\alpha, \beta)$. The MoM estimators

$$\widehat{\alpha}_n = \overline{X}_n \left\{ \frac{\overline{X}_n(1-\overline{X}_n)}{\widehat{\sigma}_n^2} - 1 \right\} \quad \text{and} \quad \widehat{\beta}_n = (1-\overline{X}_n) \left\{ \frac{\overline{X}_n(1-\overline{X}_n)}{\widehat{\sigma}_n^2} - 1 \right\}$$

are *consistent* and *AN*.

Multivariate Normal Distribution Revisited

Definition 9.12 (Multivariate Normal Via CF)

The d-dimensional random vector $\mathbb{X} = [X_1, \ldots, X_d]^\top$ has a *d-variate normal distribution* with parameters $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$, if it has the CF

$$\varphi_{\mathbb{X}}(\mathbb{t}) = \exp\{i\boldsymbol{\mu}^\top \mathbb{t} - \mathbb{t}^\top \Sigma \mathbb{t}/2\}, \quad \mathbb{t} \in \mathbb{R}^d.$$

- Notation $\mathbb{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$
- Now, MND is defined even for a singular variance-covariance matrix Σ
- The alternative Definition 9.12 is more *general* than Definition 2.39 using the density

Theorem 9.13 (Equivalence of MND Definitions Under Regular Σ)

If $\mathbb{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ and Σ is regular, then $\mathbb{X}$ has the density from Definition 2.39.

Multivariate Delta Method

Theorem 9.14 (Multivariate Delta Method)

If $\sqrt{n}\big(\mathbb{Y}_n - \boldsymbol{\mu}\big) \xrightarrow{\mathbb{D}} \mathsf{N}_d(\mathbb{0}_d, \Sigma)$, *function* $g : \mathbb{R}^d \to \mathbb{R}^k$ *is continuously differentiable on the neighborhood of* $\boldsymbol{\mu}$, *and* $\nabla g(\boldsymbol{\mu}) = \big[\partial g_j(\boldsymbol{\mu})/\partial \mu_\ell\big]_{j,\ell=1}^{k,d}$, *then*

$$\sqrt{n}\big(g(\mathbb{Y}_n) - g(\boldsymbol{\mu})\big) \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}_k\big(\mathbb{0}_k, \nabla g(\boldsymbol{\mu})\Sigma\nabla^\intercal g(\boldsymbol{\mu})\big).$$

Asymptotic Properties of MoM Estimators

Theorem 9.15 (Consistency and Asymptotic Normality for MoM)

Let $\widehat{\boldsymbol{\theta}}_n$ denote the method of moments estimator for parameter $\boldsymbol{\theta}$, which true value is $\boldsymbol{\theta}_0$. Suppose that $\mathbb{E}|X_1^d| < \infty$ and $t : \mathbb{R}^d \to \mathbb{R}^d : \boldsymbol{\theta} \mapsto \big(\mu'_1(\boldsymbol{\theta}), \ldots, \mu'_d(\boldsymbol{\theta})\big)^\intercal$ is invertible on $\mathcal{U}(\boldsymbol{\mu}'_0)$, which is some neighborhood of $\boldsymbol{\mu}'_0 = \big(\mu'_1(\boldsymbol{\theta}_0), \ldots, \mu'_d(\boldsymbol{\theta}_0)\big)^\intercal$.

(1) *The MoM estimator $\widehat{\boldsymbol{\theta}}_n$ exists with probability tending to 1 as $n \to \infty$.*

(2) *If t^{-1} is continuous on $\mathcal{U}(\boldsymbol{\mu}'_0)$, then the MoM estimator is consistent: $\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{\mathbb{P}} \boldsymbol{\theta}_0$.*

(3) *If t^{-1} is continuously differentiable on $\mathcal{U}(\boldsymbol{\mu}'_0)$ such that $\det\big[\nabla(t^{-1})(\boldsymbol{\mu}'_0)\big] \neq 0$, then the MoM estimator is asymptotically normal: $\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}_d(\mathbb{0}, \Sigma)$, where $\Sigma = \nabla(t^{-1})(\boldsymbol{\mu}'_0)\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{Y}\mathbb{Y}^\intercal]\nabla^\intercal(t^{-1})(\boldsymbol{\mu}'_0)$, $\nabla(t^{-1})(\boldsymbol{\mu}) = \big[\partial(\mu'_j)^{-1}(\boldsymbol{\mu})/\partial\mu_k\big]_{j,k=1}^{d,d}$, and $\mathbb{Y} = (X_1^1, \ldots, X_1^d)^\intercal$.*

- (3) $\Rightarrow$ *standard errors* & *confidence intervals* $\leftrightarrow$ an easier way: *parametric bootstrap*

Maximum Likelihood

- The most common method for estimating parameters in a parametric model is the maximum likelihood method

Definition 9.16

The likelihood function is defined by

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i; \boldsymbol{\theta}).$$

The log-likelihood function is defined by $\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta})$.

- The likelihood function is just the *joint density of the data*, except that we treat it is a function of the parameter $\boldsymbol{\theta}$
- Thus, $L_n : \boldsymbol{\Theta} \to [0, \infty)$
- The likelihood function is not a density function: in general, it is not true that $L_n(\boldsymbol{\theta})$ integrates to 1 (with respect to $\boldsymbol{\theta}$)

Maximum Likelihood Estimator

Definition 9.17 (MLE)

The maximum likelihood estimator MLE, denoted by $\widehat{\boldsymbol{\theta}}_n$, is the value of $\boldsymbol{\theta}$ that maximizes $L_n(\boldsymbol{\theta})$.

- The maximum of $L_n(\boldsymbol{\theta})$ occurs *at the same place* as the maximum of $\ell_n(\boldsymbol{\theta})$

Example 9.18 (Maximum Likelihood Estimator for Alternative Distribution)

$X_1, \ldots, X_n \overset{IID}{\sim} \text{Be}(p)$. The PF is $f(x; p) = p^x(1-p)^{1-x}$, $x \in \{0, 1\}$. Then,

$$L_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

Hence, $\ell_n(p) = (\log p)\sum_{i=1}^n X_i + \{\log(1-p)\}(n - \sum_{i=1}^n X_i)$. Set $\partial \ell_n(p)/\partial p \overset{!}{=} 0$. The MLE is $\widehat{p}_n = \overline{X}_n$.

ML Estimator

Example 9.19 (Maximum Likelihood Estimator for Normal Distribution)

$X_1, \ldots, X_n \overset{IID}{\sim} \mathsf{N}(\mu, \sigma^2)$. Then,

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

Set $\partial \ell_n(\mu, \sigma^2)/\partial(\mu, \sigma^2)^\intercal \overset{!}{=} \mathbb{0}_2$. The MLE is $\widehat{\mu}_n = \overline{X}_n$ and $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

ML Estimator (cont.)

Example 9.20 (Maximum Likelihood Estimator for Uniform Distribution – Hard)

$X_1, \ldots, X_n \overset{IID}{\sim} \mathsf{U}(0, \theta)$. Then,

$$L_n(\theta) = \theta^{-n} \mathbb{1}\{\theta \geq X_{(n)}\},$$

which is not differentiable w.r.t. θ. However, it can be maximized and, thus, the MLE is $\widehat{\theta}_n = X_{(n)} \equiv \max_{1 \leq i \leq n}\{X_i\}$.

ML Estimator (cont. II)

Example 9.21 (MLE in Exponential Model)

$X_1, \ldots, X_n \overset{IID}{\sim} \text{Exp}(\lambda)$. The MLE is $\widehat{\lambda}_n = 1/\overline{X}_n$.

Example 9.22 (MLE in Gamma Model – Reparametrized)

$X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(a, p)$. The MLE $\widehat{p}_n$ for the parameter p solves the *non-linear equation*

$$\log \widehat{p}_n - \frac{\Gamma'(\widehat{p}_n)}{\Gamma(\widehat{p}_n)} = \frac{\overline{X}_n}{\sqrt[n]{\prod_{i=1}^n X_i}}.$$

The MLE $\widehat{a}_n$ for the parameter a is $\widehat{a}_n = \widehat{p}_n / \overline{X}_n$.

Properties of Maximum Likelihood Estimators

MLE is:

- *consistent*
- *equivariant* ... $\boldsymbol{\theta} \rightsquigarrow \widehat{\boldsymbol{\theta}}_n \leftrightarrow g(\boldsymbol{\theta}) \rightsquigarrow g\big(\widehat{\boldsymbol{\theta}}_n\big)$
- *asymptotically normal*
- *asymptotically optimal* or *efficient* ... roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large samples
- approximately the *Bayes* estimator

Identifiability

Definition 9.23 (Identifiable Model)

We say that the model $\mathcal{F}$ is identifiable if $\boldsymbol{\theta} \neq \boldsymbol{\vartheta}$ implies that $\int f(x;\boldsymbol{\theta}) \log\left(\frac{f(x;\boldsymbol{\theta})}{f(x;\boldsymbol{\vartheta})}\right) \mathrm{d}x > 0$.

- This means that *different values of the parameter* correspond to *different distributions*
- We will assume *from now on* the the model is identifiable
- *Kullback–Leibler divergence* $D(f,g) := \int f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x$ ... not a distance, because it is not symmetric
- $D(\boldsymbol{\theta},\boldsymbol{\vartheta}) \equiv \int f(x;\boldsymbol{\theta}) \log\left(\frac{f(x;\boldsymbol{\theta})}{f(x;\boldsymbol{\vartheta})}\right) \mathrm{d}x$

Asymptotic Properties of ML Estimators

Theorem 9.24 (MLE Consistency)

Let $\boldsymbol{\theta}_0$ be the true unknown value of $\boldsymbol{\theta}$. Define

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i;\boldsymbol{\theta})}{f(X_i;\boldsymbol{\theta}_0)}$$

and $M(\boldsymbol{\theta}) = -D(\boldsymbol{\theta}_0, \boldsymbol{\theta})$. Suppose that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

and that, for every $\varepsilon > 0$,

$$\sup_{\boldsymbol{\theta}:\, \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \geq \varepsilon} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0).$$

Then, the MLE $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ is consistent, i.e., $\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{\mathbb{P}} \boldsymbol{\theta}_0$.

Uniform Law of Large Numbers

- By LLN,

$$M_n(\boldsymbol{\theta}) \xrightarrow[n\to\infty]{\mathbb{P}} -D(\boldsymbol{\theta}_0, \boldsymbol{\theta})$$

- However, we need *ULLN* (uniform law of large numbers)

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

- This can be achieved by assuming a *dominating integrable majorant* $d(x)$, i.e.,

$$\left|\log \frac{f(x;\boldsymbol{\theta})}{f(x;\boldsymbol{\theta}_0)}\right| \leq d(x) \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \quad \text{and} \quad \mathbb{E}[d(X)] < \infty$$

- Additionally, we require that $\boldsymbol{\Theta}$ is *compact* and $f(x;\boldsymbol{\theta})$ is *continuous* in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for almost all x's

Asymptotic Properties of ML Estimators (cont.)

Theorem 9.25 (MLE Equivariancy)

Let $\boldsymbol{\tau} = g(\boldsymbol{\theta})$ be a measurable function of $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\theta}}_n$ be the MLE of $\boldsymbol{\theta}$. Then, $\widehat{\boldsymbol{\tau}}_n := g(\widehat{\boldsymbol{\theta}}_n)$ is the MLE of $\boldsymbol{\tau}$.

Example 9.26

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\theta, 1)$. The MLE for θ is $\widehat{\theta}_n = \overline{X}_n$. Let $\tau = \exp\{\theta\}$, Then, the MLE for τ is $\widehat{\tau}_n = \exp\{\overline{X}_n\}$.

Score and Information

Definition 9.27 (Score Function and Fisher Information)

The score function is defined to be

$$S(X;\boldsymbol{\theta}) = \frac{\partial \log f(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

The Fisher information is defined to be

$$I_n(\boldsymbol{\theta}) = \mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{n} S(X_i;\boldsymbol{\theta})\right] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}[S(X_i;\boldsymbol{\theta})].$$

- For $n = 1$, we sometimes write $I(\boldsymbol{\theta})$ instead of $I_1(\boldsymbol{\theta})$
- $S(X;\boldsymbol{\theta})$ is a d-dimentional random vector
- $I_n(\boldsymbol{\theta})$ is a $(d \times d)$-variate deterministic matrix

Score and Information

Denote $[\cdot]^{\otimes 2} := [\cdot][\cdot]^\intercal$

Lemma 9.28 (Mean and Variance of Score)

$$\mathbb{E}_{\boldsymbol{\theta}}[S(X;\boldsymbol{\theta})] = \mathbb{0}_d \quad \textit{and} \quad \mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}[S(X;\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[S^{\otimes 2}(X;\boldsymbol{\theta})].$$

Lemma 9.29 (Alternative Definition of Information)

$$I_n(\boldsymbol{\theta}) = nI(\boldsymbol{\theta}) \quad \textit{and} \quad I(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\intercal \partial \boldsymbol{\theta}}\right].$$

Corollary 9.30 (Limiting Information)

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(X_i;\boldsymbol{\theta}) \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}_d\big(\mathbb{0}_d, I(\boldsymbol{\theta})\big).$$

Asymptotic Normality of MLE

Theorem 9.31 (AN of MLE)

Let $\boldsymbol{\theta}_0$ be the true unknown value of $\boldsymbol{\theta}$. Let the following regularity conditions hold:

(i) $\mathcal{S}(X)$ *does not depend on* $\boldsymbol{\theta}$;

(ii) $\boldsymbol{\theta} \in \operatorname{int} \boldsymbol{\Theta}$;

(iii) $I(\boldsymbol{\theta})$ *is positive definite on some neighborhood of* $\boldsymbol{\theta}_0$;

(iv) $f(x; \boldsymbol{\theta})$ *is twice continuously differentiable w.r.t.* $\boldsymbol{\theta}$;

(v) $\int \frac{\partial}{\partial \boldsymbol{\theta}} h(x; \boldsymbol{\theta}) d\nu(x) = \frac{\partial}{\partial \boldsymbol{\theta}} \int h(x; \boldsymbol{\theta}) d\nu(x)$ *for* $h(x; \boldsymbol{\theta}) = f(x; \boldsymbol{\theta})$ *and* $h(x; \boldsymbol{\theta}) = \partial f(x; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$.

Then, the MLE $\widehat{\boldsymbol{\theta}}_n$ *for* $\boldsymbol{\theta}$ *is asymptotically normal such that*

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}_d\big(\mathbb{0}_d, I^{-1}(\boldsymbol{\theta}_0)\big).$$

- Normal-based confidence intervals ($d = 1$) ... Theorem 6.13

MLE Examples

Example 9.32 (MLE in Special Case of Beta)

$X_1, \ldots, X_n \overset{IID}{\sim} f(x;\theta) = \theta(1-x)^{\theta-1}\mathbb{1}\{x \in (0,1)\}$ and θ_0 be the true unknown value of the parameter $\theta > 1$. Then,

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big) \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}(0, \theta_0^2).$$

Example 9.33 (MLE in Tied Normal)

$X_1, \ldots, X_n \overset{IID}{\sim} f(x;\theta) = \theta(1-x)^{\theta-1}\mathbb{1}\{x \in (0,1)\}$ and θ_0 be the true unknown value of the parameter $\theta > 1$. Then,

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big) \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}\big(0, 2\theta_0^2/(2\theta_0+1)\big).$$

MLE Examples (cont.)

Example 9.34 (MLE in Bernoulli Model – Flipping the Coin)

Consider Example 9.18. Consequently,

$$S(X;p) = \frac{X}{p} - \frac{1-X}{1-p} \quad \text{and} \quad S'(X;p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2}.$$

Thus,

$$I(p) = \frac{1}{p(1-p)} \quad \text{and} \quad \widehat{\mathrm{se}}(\widehat{p}_n) = \frac{1}{\sqrt{I_n(\widehat{p}_n)}} = \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}.$$

An approximate 95% percent confidence interval is $\overline{X}_n \pm u_{.975}\widehat{\mathrm{se}}(\widehat{p}_n)$.

MLE Examples (cont. II)

Example 9.35 (MLE in Normal Model)

Consider Example 9.19, where σ^2 is *known*. The MLE of μ is $\widehat{\mu}_n = \overline{X}_n$. Consequently,

$$S(X;\mu) = \frac{X-\mu}{\sigma^2} \quad \text{and} \quad S'(X;\mu) = -\frac{1}{\sigma^2}.$$

Thus,

$$I(\mu) = \frac{1}{\sigma^2} \quad \text{and} \quad \widehat{\text{se}}(\widehat{\mu}_n) = \frac{1}{\sqrt{I_n(\widehat{\mu}_n)}} = \sqrt{\frac{\sigma^2}{n}}.$$

Thus, $\widehat{\mu}_n \approx \text{N}(\mu, \sigma^2/n)$ (informaly; for large n). The normal approximation is actually *exact*, i.e., $\widehat{\mu}_n \sim \text{N}(\mu, \sigma^2/n)$.

MLE Examples (cont. III)

Example 9.36 (MLE in Poisson Model)

$X_1, \ldots, X_n \overset{IID}{\sim} \text{Po}(\lambda)$, where $\lambda > 0$. The MLE of λ is $\widehat{\lambda}_n = \overline{X}_n$. Consequently,

$$S(X;\lambda) = \frac{X}{\lambda} - 1 \quad \text{and} \quad S'(X;\lambda) = -\frac{X}{\lambda^2}.$$

Thus,

$$I(\lambda) = \frac{1}{\lambda} \quad \text{and} \quad \widehat{\text{se}}(\widehat{\lambda}_n) = \frac{1}{\sqrt{I_n(\widehat{\lambda}_n)}} = \sqrt{\frac{\overline{X}_n}{n}}.$$

An approximate 95% percent confidence interval is $\overline{X}_n \pm u_{.975}\sqrt{\overline{X}_n/n}$.

Delta Method and Approximate Normal-based CIs

Example 9.37 (Flipping a Coin and Logit Transform)

Let us continue with Example 9.18. Consider the *logit transform* $\psi = g(p) = \log\frac{p}{1-p}$. The MLE of p is $\widehat{p}_n = \overline{X}_n$. The Fisher information function is $I(p) = 1/(p(1-p))$. So, the estimated standard error of the MLE $\widehat{p}_n$ is $\widehat{\mathrm{se}}(\widehat{p}_n) = \frac{1}{\sqrt{I_n(\widehat{p}_n)}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$. The MLE of ψ is $\widehat{\psi}_n = \log\frac{\widehat{p}_n}{1-\widehat{p}_n}$. Since $g'(p) = 1/(p(1-p))$, according to the delta method,

$$\widehat{\mathrm{se}}(\widehat{\psi}_n) = |g'(\widehat{p}_n)|\widehat{\mathrm{se}}(\widehat{p}_n) = \frac{1}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}.$$

An approximate 95% percent confidence interval is $\widehat{\psi}_n \pm \frac{2}{\sqrt{n\overline{X}_n(1-\overline{X}_n)}}$, because $u_{.975} \doteq 2$.

Transformations Using Delta Method

Example 9.38 (Relative Standard Error)

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma^2)$. Then, a relative dispersion can be characterized by the se-to-mean ratio $\tau = g(\mu, \sigma) = \sigma/\mu$. Here, the Fisher information matrix is $I_n(\mu, \sigma) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}$. Hence, its inverse becomes $I_n^{-1}(\mu, \sigma) = \frac{1}{n}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2/2 \end{bmatrix}$. The gradient of g is $\nabla g(\mu, \sigma) = [-\sigma/\mu^2, 1/\mu]$. Thus,

$$\widehat{\mathrm{se}}(\widehat{\tau}_n) = \sqrt{\nabla g(\widehat{\mu}_n, \widehat{\sigma}_n)\widehat{I}_n^{-1}(\widehat{\mu}_n, \widehat{\sigma}_n)\nabla^\intercal g(\widehat{\mu}_n, \widehat{\sigma}_n)} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{\widehat{\mu}_n^4} + \frac{\widehat{\sigma}_n^2}{2\widehat{\mu}_n^2}}.$$

An approximate 95% percent confidence interval is $\widehat{\sigma}_n/\widehat{\mu}_n \pm u_{.975}\widehat{\mathrm{se}}(\widehat{\tau}_n)$.

Parametric Bootstrap

- For parametric models, SEs and CIs (and many more things) may also be estimated using the bootstrap
- There is only one *difference* to the nonparametric bootstrap
- In the nonparametric bootstrap, we sampled $X_1^\star, \ldots, X_n^\star$ from the empirical distribution $\widehat{F}_n$
- In the parametric bootstrap, we sample instead from $f(x; \widehat{\boldsymbol{\theta}}_n)$
- Here, $\widehat{\boldsymbol{\theta}}_n$ could be the MLE or the MoM estimator

Parametric Bootstrap Example

Example 9.39 (Relative Dispersion via the SE-to-Mean Ratio)

Consider Example 9.38. To get the bootstrap standard error, simulate $X_1^\star, \ldots, X_n^\star \sim \mathsf{N}(\widehat{\mu}_n, \widehat{\sigma}_n^2)$, compute $\widehat{\mu}_n^\star = n^{-1} \sum_{i=1}^n X_i^\star$ and $\widehat{\sigma}_n^{2\star} = n^{-1} \sum_{i=1}^n \left(X_i^\star - \widehat{\mu}_n^\star\right)^2$. Then, calculate $\widehat{\tau}_n^\star = g\left(\widehat{\mu}_n^\star, \widehat{\sigma}_n^{2\star}\right) = \widehat{\mu}_n^\star / \widehat{\sigma}_n^{2\star}$. Repeating this B times yields bootstrap replications

$$ {}_{(1)}\widehat{\tau}_n^\star, \ldots, {}_{(B)}\widehat{\tau}_n^\star. $$

The estimated standard error becomes

$$ \widehat{\mathsf{se}}^\star\left(\widehat{\tau}_n\right) = \sqrt{\frac{\sum_{b=1}^B \left({}_{(b)}\widehat{\tau}_n^\star - \widehat{\tau}_n\right)^2}{B}}. $$

- The bootstrap is much easier than the delta method
- On the other hand, the delta method has the advantage that it gives a closed form expression for the standard error

Parametric Bootstrap vs Delta Method

Example 9.40 (Comparing Two Treatments)

n_1 people are given Treatment 1 and n_2 people are given Treatment 2. Let X_1 be the number of people on Treatment 1 who respond favorably to the treatment and let X_2 be the number of people on Treatment 2 who respond favorably. Assume that $X_1 \sim \text{Bi}(n_1, p_1)$, $X_2 \sim \text{Bi}(n_2, p_2)$. Let $\psi = p_1 - p_2$.

(a) Find the MLE of ψ.

(b) Find the Fisher Information Matrix $I(p_1, p_2)$.

(c) Use the delta method to find the asymptotic standard error of $\widehat{\psi}_n$.

(d) Suppose that $n_1 = n_2 = 200$, $X_1 = 160$ and $X_2 = 148$. Find $\widehat{\psi}_n$. Find an approximate 90% confidence interval for ψ using (i) the delta method and (ii) the parametric bootstrap.

Parametric Bootstrap vs Delta Method (cont.)

Example 9.41 (Comparing Two Treatments – Solution)

(a) The MLE is equivariant, so $\widehat{\psi}_n = \widehat{p}_1 - \widehat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$, where $n = n_1 + n_2$.

(b) The probability mass function is

$$f([x_1, x_2]; \psi) = f_1(x_1; p_1) f_2(x_2; p_2) = \binom{n_1}{x_1} p_1^{x_1} (1 - p_1)^{n_1 - x_1} \binom{n_2}{x_2} p_2^{x_2} (1 - p_2)^{n_2 - x_2}.$$

The log-likelihood is

$$\ell_n = \log\{f([x_1, x_2]; \psi)\} = \sum_{i=1}^{2} \log\left\{ \binom{n_i}{x_i} + x_i \log p_i + (n_i - x_i) \log(1 - p_i) \right\}.$$

Parametric Bootstrap vs Delta Method (cont. II)

Example 9.42 (Comparing Two Treatments – Solution (cont.))

(b) Calculating the partial derivatives and their expectations

$$H_{11} = \frac{\partial^2 \ell_n}{\partial p_1^2} = \frac{\partial}{\partial p_1}\left(\frac{x_1}{p_1} - \frac{n_1 - x_1}{1 - p_1}\right) = -\frac{x_1}{p_1^2} - \frac{n_1 - x_1}{(1-p_1)^2},$$

$$\mathbb{E}H_{11} = -\frac{\mathbb{E}[x_1]}{p_1^2} - \frac{\mathbb{E}[n - x_1]}{(1-p_1)^2} = -\frac{n_1 p_1}{p_1^2} - \frac{n_1(1-p_1)}{(1-p_1)^2} = -\frac{n_1}{p_1(1-p_1)},$$

$$H_{22} = -\frac{x_2}{p_2^2} - \frac{n_2 - x_2}{(1-p_2)^2}, \quad \mathbb{E}H_{22} = -\frac{n_2}{p_2(1-p_2)}, \quad H_{12} = \frac{\partial^2 \ell_n}{\partial p_1 \partial p_2} = 0, \quad H_{21} = 0.$$

So, the Fisher information matrix is $I_n(p_1, p_2) = \begin{bmatrix} \frac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1-p_2)} \end{bmatrix}$.

Parametric Bootstrap vs Delta Method (cont. III)

Example 9.43 (Comparing Two Treatments – Solution (cont. II))

(c) Using the multivariate delta method for $g(\psi) = p_1 - p_2$, we obtain $\nabla g = [\partial g/\partial p_1, \partial g/\partial p_2] = [1, -1]$. The inverse of the Fisher information matrix is

$$I_n^{-1}(p_1, p_2) = \begin{bmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{bmatrix}.$$

Then, the estimated asymptotic standard error of $\widehat{\psi}_n$ becomes

$$\widehat{\mathrm{se}}(\widehat{\psi}_n) = \sqrt{\nabla g(\widehat{p}_1, \widehat{p}_2)\widehat{I}_n^{-1}(\widehat{p}_1, \widehat{p}_2)\nabla^\intercal g(\widehat{p}_1, \widehat{p}_2)} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

(d) Python code

Parametric Bootstrap vs Delta Method in Python

```python
import numpy as np
from scipy.stats import norm, binom

np.random.seed(2024)

n = 200
X1 = 160
X2 = 148

p1_hat = X1 / n
p2_hat = X2 / n
psi_hat = p1_hat - p2_hat

print("Estimated psi: \t %.3f" % psi_hat)
```

Parametric Bootstrap vs Delta Method in Python (cont.)

```
# Confidence using delta method

z = norm.ppf(.95)

se_delta = np.sqrt(p1_hat * (1 - p1_hat)/n + p2_hat * (1 - p2_hat) / n)
confidence_delta = (psi_hat - z * se_delta, psi_hat + z * se_delta)

print("90%% confidence interval (delta method): \t %.3f, %.3f" %
    confidence_delta)
```

Parametric Bootstrap vs Delta Method in Python (cont. II)

```
# Confidence using parametric bootstrap

B = 1000
xx1 = binom.rvs(n, p1_hat, size=B)
xx2 = binom.rvs(n, p2_hat, size=B)
t_boot = xx1 / n - xx2 / n

se_bootstrap = t_boot.std()
confidence_delta = (psi_hat - z * se_bootstrap, psi_hat + z *
    se_bootstrap)

print("90%% confidence interval (parametric bootstrap): \t %.3f, %.3f" %
     confidence_delta)
```

$\widehat{\psi}_n \doteq 0.060, \quad C_n^{delta}(90\%) = (-0.009, 0.129), \quad C_n^{parboot}(90\%) = (-0.008, 0.128)$

Computing Maximum Likelihood Estimates

- In some cases, we can find the MLE $\widehat{\boldsymbol{\theta}}_n$ *analytically*
- More often, we need to find the MLE by numerical methods, cf. Example 9.22
- Two *iterative methods* – NR & EM – that produce a sequence of values $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, which, under ideal conditions, converge to the MLE $\widehat{\boldsymbol{\theta}}_n$
- In each case, it is helpful to use a good *starting value* $\boldsymbol{\theta}^{(0)}$
- Often, the *MoM estimator* is a good starting value

Newton-Raphson Algorithm

- Expand the derivative of the log-likelihood around $\theta^{(j)}$:

$$0 = \ell_n'(\widehat{\theta}_n) \approx \ell_n'(\theta^{(j)}) + (\widehat{\theta}_n - \theta^{(j)})\ell_n''(\theta^{(j)})$$

- Solving for $\widehat{\theta}_n$ gives

$$\widehat{\theta}_n \approx \theta^{(j)} - \frac{\ell_n'(\theta^{(j)})}{\ell_n''(\theta^{(j)})}$$

- This suggests the following iterative scheme

$$\theta^{(j+1)} := \theta^{(j)} - \frac{\ell_n'(\theta^{(j)})}{\ell_n''(\theta^{(j)})}$$

- In the multiparameter case,

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - H^{-1}\nabla\ell_n(\boldsymbol{\theta}^{(j)}),$$

where H is the matrix of second derivatives (Hessian matrix) of the log-likelihood

Expectation-Maximization Algorithm

- Suppose we have data Y whose density $f(y;\boldsymbol{\theta})$ leads to a log-likelihood that is *hard to maximize*
- But suppose we can find another random variable Z such that $f(y;\boldsymbol{\theta}) = \int f(y,z;\boldsymbol{\theta})\mathrm{d}z$ and such that the likelihood based on $f(y,z;\boldsymbol{\theta})$ is *easy to maximize*
- In other words, the model of interest is the marginal of a model with a simpler likelihood
- In this case, we call Y the observed data and Z the hidden (or latent or missing) data
- If we could just "fill in" the missing data, we would have an easy problem
- Conceptually, the EM algorithm works by filling in the missing data, maximizing the log-likelihood, and iterating

EM Algorithm

The idea is to iterate between taking an expectation then maximizing

(0) Pick a starting value $\widehat{\boldsymbol{\theta}}_n^{(0)}$ and repeat:

(1) [The E-step]: Calculate

$$J(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) = \mathbb{E}_{\boldsymbol{\theta}^{(j)}}\left[\log \frac{f(\mathbb{Y},\mathbb{Z};\boldsymbol{\theta})}{f(\mathbb{Y},\mathbb{Z};\boldsymbol{\theta}^{(j)})}\Big|\mathbb{Y}=\mathbb{y}\right]$$

The expectation is over the missing data $\mathbb{Z}$ treating $\boldsymbol{\theta}^{(j)}$ and the observed data $\mathbb{Y}$ as fixed

(2) [The M-step]: Find $\boldsymbol{\theta}^{(j+1)}$) to maximize $J(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)})$

The EM algorithm always increases the likelihood, that is, $L_n(\boldsymbol{\theta}^{(j+1)}) \geq L_n(\boldsymbol{\theta}^{(j)})$

Agenda

10. Hypothesis Testing

Testing – Motivation Example

Example 10.1 (Comparing Two Prediction Algorithms)

Performance two prediction algorithms is tested. We test Algorithm 1 on the first test set and Algorithm 2 on the second test set.

The Null Hypothesis: The probability of correct prediction is the same for both algorithms.

The Alternative Hypothesis: The probability of correct prediction is NOT the same.

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \neq p_2.$$

Formal Testing Framework

- We stay with *parametric models*
- We partition the parameter space $\boldsymbol{\Theta}$ into two disjoint sets $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$, i.e., $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_1 = \emptyset$
- We wish to test

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$$

- We call H_0 the *null hypothesis* and H_1 the *alternative hypothesis*

Decisions by Hypothesis Testing

- Let $\mathbb{X}$ be a random vector (e.g., a random sample) and let $\mathcal{X}$ be the range of $\mathbb{X}$
- We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the rejection region
- Decision made:
  - $\mathbb{X} \in R \Rightarrow$ reject H_0
  - $\mathbb{X} \notin R \Rightarrow$ do not reject H_0 against H_1 (retain H_0)
- The rejection region R can usually be rewritten in the form

$$R = \{\mathbb{x} : \ T(\mathbb{x}) \in C\},$$

where T is a test statistic and C is a critical region

Definition 10.2 (Statistical Test)

A statistical test (T, C) is a couple consisting of a test statistic T and a critical region C.

Outcomes of Hypothesis Testing

- There are two types of errors we can make:
  - Rejecting H_0, when H_0 is true is called a type I error
  - Retaining H_0, when H_1 is true is called a type II error
  - To reject or not ?

	Do not reject the null H_0	Reject the null H_0 against the alternative H_1
H_0 true	✓	type I error
H_1 true	type II error	✓

- We retain H_0 unless there is strong evidence to reject H_0

Level – Size – Power

Definition 10.3 (Power Function, Size of the Test, Significance Level)

The *power function* of a test with rejection region R is defined by

$$\beta(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[\mathbb{X} \in R].$$

The *size of a test* is defined to be

$$\alpha = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \beta(\boldsymbol{\theta}).$$

A test is said to have *level* α if its size is less than or equal to α.

- Alternatively, $\beta(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[T(\mathbb{X}) \in C]$, where (T, C) is a statistical test

Types of Hypotheses and Tests

- A hypothesis of the form $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, i.e., $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0\}$ is called a simple hypothesis
- A hypothesis of the form $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, where $\boldsymbol{\Theta}_0$ contains more than one element, is called a composite hypothesis
- A test of the form $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ is called a two-sided test
- A test of the form $H_0 : \boldsymbol{\theta} \leq \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} > \boldsymbol{\theta}_0$ or $H_0 : \boldsymbol{\theta} \geq \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} < \boldsymbol{\theta}_0$ is called a one-sided test
- The most common tests are *two-sided*

Power Function for Mean Parameter I

Example 10.4 (Power Function for Mean in Normal Model)

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Hence, $\boldsymbol{\Theta}_0 = (-\infty, 0]$ and $\boldsymbol{\Theta}_1 = (0, +\infty)$. Consider the test:

$$\text{reject } H_0 \text{ if } T(\mathbb{X}) > c,$$

where $T(\mathbb{X}) = \overline{X}_n$. The rejection region is

$$R = \{(x_1, \ldots, x_n) : \ T(x_1, \ldots, x_n) > c\}.$$

Since $\sqrt{n}(\overline{X}_n - \mu)/\sigma \sim \mathrm{N}(0, 1)$, the *power function* becomes

$$\beta(\mu) = \mathbb{P}_\mu \left[\overline{X}_n > c \right] = \mathbb{P}_\mu \left[\sqrt{n} \frac{\overline{X}_n - \mu}{\sigma} > \sqrt{n} \frac{c - \mu}{\sigma} \right] = 1 - \Phi\left(\sqrt{n} \frac{c - \mu}{\sigma} \right).$$

Power Function for Mean Parameter II

Example 10.5 (Power Function for Mean in Normal Model (cont.))

The power function $\beta(\mu)$ is increasing in μ (cf. the next slide). Hence, the *size* equals

$$\sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\sqrt{n}\frac{c}{\sigma}\right).$$

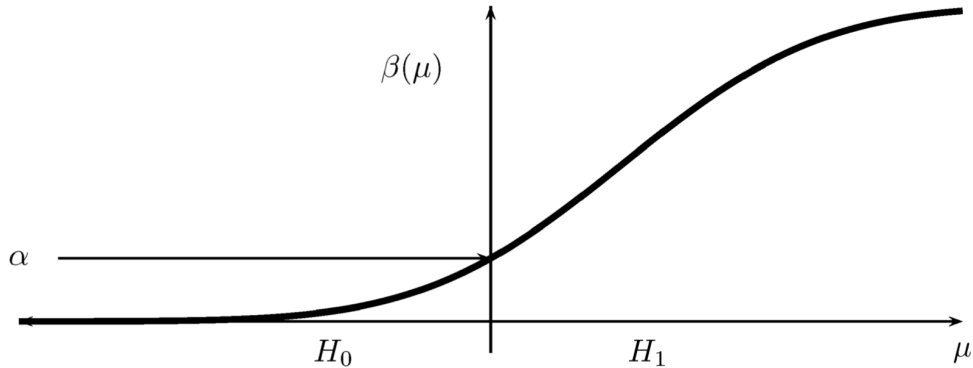
For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

We reject, when $\overline{X}_n > \sigma\Phi^{-1}(1-\alpha)/\sqrt{n}$. Note that $\mu_0 = 0$. Equivalently, we reject when

$$\sqrt{n}\frac{\overline{X}_n - 0}{\sigma} > u_{1-\alpha} \equiv \Phi^{-1}(1-\alpha).$$

Plot of Power Function for Mean Parameter



- The size of the test is the largest probability of rejecting H_0, when H_0 is true
- This occurs at $\mu = 0$, hence the size is $\beta(0)$
- We choose the critical value c so that $\beta(0) = \alpha$

The Wald Test

- Let θ a *scalar* parameter, let $\widehat{\theta}$ an estimate of θ and let $\widehat{\mathsf{se}}(\widehat{\theta})$ be the estimated standard error of $\widehat{\theta}$

Definition 10.6 (Wald Test)

Consider *testing* $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Assume that

$$W := \frac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}(\widehat{\theta})} \xrightarrow[n\to\infty]{\mathbb{D}} \mathsf{N}(0,1).$$

The size α Wald test is: *reject* H_0, when $|W| > u_{1-\alpha/2}$.

Theorem 10.7 (Asymptotic Size of the Wald Test)

$$\mathbb{P}_{\theta_0}[|W| > u_{1-\alpha/2}] \to \alpha, \quad n \to \infty.$$

Power of the Wald Test

- An alternative version of the Wald test statistic is $W = (\widehat{\theta} - \theta_0)/\text{se}_0(\widehat{\theta})$ where $\text{se}_0(\widehat{\theta})$ is the standard error computed at $\theta = \theta_0$
- Both versions of the test are valid

Theorem 10.8 (Power of the Wald Test Under the Alternative)

Suppose the true value of θ is $\theta_1 \neq \theta_0$. The power $\beta(\theta_1)$ – the probability of correctly rejecting the null hypothesis – is given approximately by

$$1 - \Phi\left(\frac{\theta_0 - \theta_1}{\widehat{\text{se}}(\widehat{\theta})} + u_{1-\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_1}{\widehat{\text{se}}(\widehat{\theta})} - u_{1-\alpha/2}\right).$$

- Recall that $\widehat{\text{se}}(\widehat{\theta})$ tends to 0 as the sample size increases
- The power is large if θ_1 is far from θ_0
- The power is large, if the sample size is large

Testing Two Probabilities – Two Sample Comparison

Example 10.9 (Comparing Two Prediction Algorithms (cont.))

Continue with the motivation Example 10.1. We test a prediction Algorithm 1 on a test set of size m and we test a prediction Algorithm 2 on a second test set of size n. Let X be the number of correct predictions for Algorithm 1 and let Y be the number of correct predictions for Algorithm 2. Then, $X \sim \text{Bi}(m, p_1)$ and $Y \sim \text{Bi}(n, p_2)$. To test the null hypothesis that $p_1 = p_2$, write $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, where $\delta = p_1 - p_2$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\widehat{\text{se}}(\widehat{\delta}) = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}.$$

The size α Wald test is to reject H_0 when $|W| > u_{1-\alpha/2}$, where $W = (\widehat{\delta} - 0)/\widehat{\text{se}}(\widehat{\delta})$. The power of the test will be largest when p_1 is far from p_2 and when the sample sizes are large.

Testing Two Probabilities – Paired Comparison

Example 10.10 (Comparing Two Prediction Algorithms (revised))

Continue with the previous Example 10.9. What if we used the *same test* set to test both algorithms? The two samples are *no longer independent*. Instead we use the following strategy. Let $X_i = 1$ if Algorithm 1 is correct on test case i and $X_i = 0$ otherwise. Let $Y_i = 1$ if Algorithm 2 is correct on test case i, and Yi = 0 otherwise. Define $D_i = X_i - Y_i$. Let

$$\delta = \mathbb{E}D_i = \mathbb{E}X_i - \mathbb{E}Y_i = \mathbb{P}[X_i = 1] - \mathbb{P}[Y_i = 1] = p_1 - p_2.$$

The nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{D}_n = n^{-1}\sum_{i=1}^n D_i$ and $\widehat{\mathrm{se}}(\widehat{\delta}) = S/\sqrt{n}$, where $S = n^{-1}\sum_{i=1}^n (D_i - \overline{D}_n)^2$. To test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, we use $W = \widehat{\delta}/\widehat{\mathrm{se}}(\widehat{\delta})$ and reject H_0 if $|W| > u_{1-\alpha/2}$.

Testing Two Expectations – Two Sample Problem

Example 10.11 (Comparing Two Means)

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from populations with some finite positive variances and having means μ_1 and μ_2, respectively. Let's test the null hypothesis that $\mu_1 = \mu_2$. Write this as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, where $\delta = \mu_1 - \mu_2$. Recall that the nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{X}_m - \overline{X}_n$ with estimated standard error

$$\widehat{\mathrm{se}}(\widehat{\delta}) = \sqrt{\frac{\widehat{\sigma}_1^2}{m} + \frac{\widehat{\sigma}_2^2}{n}},$$

where $\widehat{\sigma}_1^2 = m^{-1} \sum_{i=1}^m (X_i - \overline{X}_m)^2$ and $\widehat{\sigma}_2^2 = n^{-1} \sum_{i=1}^n (Y_i - \overline{Y}_n)^2$ are the sample variances. The size α Wald test is to reject H_0 when $|W| > u_{1-\alpha/2}$, where $W = (\widehat{\delta} - 0)/\widehat{\mathrm{se}}(\widehat{\delta})$.

Duality in Hypothesis Testing and Confidence Intervals

- There is a relationship between the size α Wald test and the $(1-\alpha)$ asymptotic CI $\widehat{\theta} \pm u_{1-\alpha/2}\widehat{\mathsf{se}}(\widehat{\theta})$
- Testing the hypothesis is equivalent to checking whether the null value is in the confidence interval

Theorem 10.12 (Equivalence in Hypothesis Testing and Confidence Interval)

The size α Wald test is rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ iff $\theta_0 \notin C$, where

$$C = \left(\widehat{\theta} - u_{1-\alpha/2}\widehat{\mathsf{se}}(\widehat{\theta}), \widehat{\theta} + u_{1-\alpha/2}\widehat{\mathsf{se}}(\widehat{\theta})\right).$$

p-values

- Reporting "reject H_0" or "retain H_0" is not very informative
- Instead, we could ask, for every α, whether the test rejects at that level
- Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$
- Hence, there is a smallest α at which the test rejects and we call this number the p-value

Definition 10.13 (p-value)

For a particular $\omega \in \Omega$, let the observed data be $\mathbb{X}(\omega) = \mathbb{x}$. Suppose that for every $\alpha \in (0, 1)$, we have a size α test with rejection region R_α. Then,

$$p\text{-value} = \inf \{\alpha : \mathbb{x} \in R_\alpha\}.$$

That is, the p-value is the *smallest level at which we can reject* H_0.

p-values (cont.)

- Informally, the p-value is a measure of the evidence against H_0: the smaller the p-value, the stronger the evidence against H_0
- Typically, researchers use the following evidence scale:
  - $< .01$ ... very strong evidence against H_0
  - $.01 - .05$ ... strong evidence against H_0
  - $.05 - .10$ ... weak evidence against H_0
  - $> .1$ ... little or no evidence against H_0

! A large p-value is NOT strong evidence in favor of H_0

- A large p-value can occur for two reasons:
  (i) H_0 is true or
  (ii) H_0 is false, but the test has low power
- Do not confuse the p-value with $\mathbb{P}[H_0|Data]$

! The p-value is NOT the probability that the null hypothesis is true

How to Compute the p-value

Theorem 10.14 (Calculate the p-value)

Suppose that the size α test is of the form

$$\text{reject } H_0 \text{ iff } T(\mathbb{X}) \geq c_\alpha.$$

Then,

$$p\text{-value} = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbb{P}_{\boldsymbol{\theta}} \left[T(\mathbb{X}) \geq T(\mathbb{x}) \right],$$

where $\mathbb{x}$ is the observed value of $\mathbb{X}$. If $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0\}$, then

$$p\text{-value} = \mathbb{P}_{\boldsymbol{\theta}_0} \left[T(\mathbb{X}) \geq T(\mathbb{x}) \right].$$

- The p-value is the probability (under H_0) of observing a value of the test statistic the same as or more extreme than what was actually observed

p-value for the Wald Test

Theorem 10.15 (p-value for the Wald Test)

Let $w = \frac{\widehat{\theta}(\mathbf{x}) - \theta_0}{\widehat{\mathrm{se}}(\widehat{\theta})}$ *denote the observed value of the Wald statistic* W*. The p-value becomes*

$$p\text{-value} = \mathbb{P}_{\theta_0}\left[|W| > |w|\right] \longrightarrow 2\Phi(-|w|), \quad n \longrightarrow \infty.$$

Permutation Test

- A nonparametric method for testing whether two distributions are the same
- An exact test meaning that it is not based on large sample theory approximations
- $X_1, \ldots, X_m \overset{IID}{\sim} F_X \perp\!\!\!\perp Y_1, \ldots, Y_n \overset{IID}{\sim} F_Y$
- $H_0 : F_X = F_Y$ versus $H_1 : F_X \neq F_Y$
- Let T be some test statistic, for example, $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = |\overline{X}_m - \overline{Y}_n|$
- Let $N = m + n$ and consider forming all $N!$ permutations of the data $X_1, \ldots, X_m, Y_1, \ldots, Y_n$
- For each permutation, compute the test statistic T
- Denote these values by $T_1, \ldots, T_{N!}$
- Under the null hypothesis, each of these values is equally likely
- More precisely, under H_0, given the ordered data values, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is uniformly distributed over the $N!$ permutations of the data

Permutation Test (cont.)

- The distribution $\mathbb{P}_0$ that puts mass $1/N!$ on each T_j is called the permutation distribution of T
- Let t_{obs} be the observed value of the test statistic
- Assuming we reject when T is large,

$$p\text{-value} = \mathbb{P}_0[T > t_{obs}] = \frac{1}{N!}\sum_{j=1}^{N!} \mathbb{1}\{T_j > t_{obs}\}$$

Algorithm for Permutation Test

- Usually, it is not practical to evaluate all $N!$ permutations
- We can *approximate* the p-value by *sampling randomly from the set of permutations*
- The fraction of times $T_j > t_{obs}$ among these samples approximates the p-value

1. Compute the observed value of the test statistic

$$t_{obs} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$$

2. Randomly permute the data and compute the statistic again using the permuted data
3. Repeat the previous step B times and let $T_1, \ldots, T_B$ denote the resulting values
4. The approximate p-value is

$$p\text{-value} = \frac{1}{B} \sum_{j=1}^{B} \mathbb{1}\{T_j > t_{obs}\}$$

Wald vs Permutation Tests – by Mark Twain

Example 10.16 (Mark Twain)

In 1861, 10 essays appeared in the New Orleans Daily Crescent. They were signed 'Quintus Curtuis Snodgrass' and some people suspected they were actually written by Mark Twain. To investigate this, we will *investigate the proportion of three letter words* found in an author's work.

From eight Twain essays, we have:

$$0.225 \quad 0.262 \quad 0.217 \quad 0.240 \quad 0.230 \quad 0.229 \quad 0.235 \quad 0.217$$

From 10 Snodgrass essays, we have:

$$0.209 \quad 0.205 \quad 0.196 \quad 0.210 \quad 0.202 \quad 0.207 \quad 0.224 \quad 0.223 \quad 0.220 \quad 0.201$$

Recall Example 10.11.

Mark Twain by Abraham Wald in Python

```
X = [0.225,0.262,0.217,0.240,0.230,0.229,0.235,0.217]
Y = [0.209,0.205,0.196,0.210,0.202,0.207,0.224,0.223,0.220,0.201]

import numpy as np
from scipy.stats import norm

np.random.seed(2024)

X = np.array(X)
Y = np.array(Y)

x_hat = X.mean()
y_hat = Y.mean()

diff_hat = x_hat - y_hat
se_hat = np.sqrt(X.var(ddof=1)/len(X) + Y.var(ddof=1)/len(Y))
```

Mark Twain by Abraham Wald in Python (cont.)

```
u = norm.ppf(0.975)
confidence_interval = (diff_hat - u * se_hat, diff_hat + u * se_hat)

w = diff_hat / se_hat
p_value = 2 * (1 - norm.cdf(abs(w)))

print('Estimated difference of means:\t %.3f' % diff_hat)
print('Estimated SE: \t\t\t %.3f' % se_hat)
print('95%% confidence interval:\t (%.3f, %.3f)' % confidence_interval)
print('Wald statistic: \t\t %.3f' % w)
print('Wald test p-value: \t\t %.4f' % p_value)
```

$$\widehat{\mu}_T - \widehat{\mu}_S = 0.022, \quad \widehat{\mathsf{se}}(\widehat{\mu}_T - \widehat{\mu}_S) = 0.006, \quad C_n^{AN}(95\%) = (0.010, 0.034)$$

$$W = 3.704, \quad p\text{-value} = 0.0002$$

Mark Twain via Permutation Test in Python

```
# Permutation test using random shuffling
B = 1000000
full_series = np.concatenate([X, Y])
nx = len(X)
diff_boot_count = 0
for i in range(B):
np.random.shuffle(full_series)
xx, yy = full_series[:nx], full_series[nx:]
diff_boot = xx.mean() - yy.mean()
if diff_boot > diff_hat:
diff_boot_count += 1

p_value_perm = diff_boot_count / B
print('Permutation test p-value: \t\t %.4f' % p_value_perm)
```

p-value = 0.0005

Likelihood Ratio Test

Definition 10.17 (Likelihood Ratio Test Statistic)

Consider *testing* $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $H_1 : \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$. The *likelihood ratio statistic* is

$$\lambda = 2\log\frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} L_n(\boldsymbol{\theta})} = 2\log\frac{L_n(\widehat{\boldsymbol{\theta}})}{L_n(\widehat{\boldsymbol{\theta}}_0)},$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE and $\widehat{\boldsymbol{\theta}}_0$ is the MLE when $\boldsymbol{\theta}$ is restricted to lie in $\boldsymbol{\Theta}_0$.

Theorem 10.18 (Likelihood Ratio Test)

Under the null $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^d : [\theta_{q+1}, \ldots, \theta_d]^\intercal = [\theta_{q+1,0}, \ldots, \theta_{d,0}]^\intercal\}$,

$$\lambda \xrightarrow[n\to\infty]{\mathbb{D}} \chi^2_{d-q},$$

where $d - q$ *is the dimension of* $\boldsymbol{\Theta}_0$ *minus the dimension of* $\boldsymbol{\Theta}_0$.

Agenda

11. References

References


Casella, G. and Berger, R.L. (2001)
Statistical Inference, 2nd Edition
Pacific Grove, CA: Duxbury


Chung, K.L. (2001)
A Course in Probability Theory, 3rd Edition
San Diego, CA: Academic Press


Dupač, V. and Hušková, M. (2013)
Pravděpodobnost a matematická statistika
Praha, CZ: Karolinum


Resnick, S.I. (2013)
A Probability Path, 2014th Edition
Basel, CH: Birkhäuser


Rosenthal, J.S. (2006)
A First Look at Rigorous Probability Theory, 2nd Edition
Singapore, SG: World Scientific


Ross, S.M. (2020)
A First Course in Probability, 10th Edition
London, UK: Pearson


Wasserman, L. (2013)
All of Statistics: A Concise Course in Statistical Inference
New York, NY: Springer

The End

michal.pesta@mff.cuni.cz