SM-4331 Advanced Statistics
Chapter 1 (Estimation Theory)

Dr Haziq Jamil
FOS M1.09

Universiti Brunei Darussalam

Semester II 2019/20

Outline

Inequalities

Inequalities are useful tools in establishing various properties of statistical inference methods. They may also provide estimates for probabilities with little assumption on probability distributions.

There are four main inequalities that we will learn:

- Markov's inequality
- Chebyshev's inequality
- Cauchy-Schwarz inequality
- Jensen's inequality

Markov's inequality

In probability theory, Markov's inequality gives an upper bound for the probability that a non-negative random variable (r.v.) exceeds some positive constant.

Theorem 1 (Markov's inequality)

Let X be a non-negative r.v. and $\mathrm{E}(X) < \infty$. Then, for any $t > 0$,

$$P(X \geq t) \leq \frac{\mathrm{E}(X)}{t}.$$

Markov's inequality relate probabilities to expectations, and provides bounds for the cumulative distribution function of a r.v..

Proof of Markov's inequality

Proof.

Let $f(x)$ be the pdf of X. Since $X > 0$,

$$\begin{aligned}
\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x &= \int_{0}^{\infty} x\, f(x)\, \mathrm{d}x \\
&= \int_{0}^{t} x\, f(x)\, \mathrm{d}x + \int_{t}^{\infty} x\, f(x)\, \mathrm{d}x \\
&\geq \int_{t}^{\infty} x\, f(x)\, \mathrm{d}x \\
&\geq t \int_{t}^{\infty} f(x)\, \mathrm{d}x \\
&= tP(X \geq t)
\end{aligned}$$

Corollary to Markov's inequality

Corollary 2

For any r.v. X and any constant $t > 0$,

$$P(|X| \geq t) \leq \frac{E|X|}{t} \quad \text{provided } E|X| < \infty$$

$$P(|X| \geq t) \leq \frac{E\left(|X|^k\right)}{t^k} \quad \text{provided } E\left(|X|^k\right) < \infty$$

The tail probability $P(|X| \geq t)$ is a useful measure in insurance and risk management in finance. The more moments X has, the smaller the tail probabilities are.

Chebyshev's inequality

In probability theory, Chebyshev's inequality guarantees that no more than a certain fraction of values can be more than a certain distance from the mean.

Theorem 3 (Chebyshev's inequality)

Suppose a r.v. X has mean μ and variance σ^2. Then, for any $t > 0$,

$$P(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}.$$

Because it can be applied to completely arbitrary distributions (provided they have a known finite mean and variance), the inequality generally gives a poor bound, compared to what might be deduced if more aspects are known about the distribution involved.

Chebyshev's inequality (cont.)

Example 4

Suppose X has mean 0 and variance 1. By Chebyshev's inequality,

$$P(|X| \geq 1) \leq 1.00$$
$$P(|X| \geq 2) \leq 0.25$$
$$P(|X| \geq 3) \leq 0.11$$

In contrast, suppose that we know that X is normally distributed. Then

$$P(|X| \geq 1) \leq 0.318$$
$$P(|X| \geq 2) \leq 0.046$$
$$P(|X| \geq 3) \leq 0.002$$

Proof of Chebyshev's inequality

Prove this as an exercise. *Hint: Use Markov's inequality.*

Cauchy-Schwartz inequality

This is a very useful inequality that crops up in many different areas of mathematics, such as linear algebra, analysis, probability theory, vector algebra, etc.

Theorem 5 (Cauchy-Schwartz inequality)

Let $E(X^2) < \infty$ and $E(Y^2) < \infty$. Then

$$|E(XY)|^2 \leq E(X^2)E(Y^2).$$

Cauchy-Schwartz inequality (cont.)

The covariance inequality can be proved using Cauchy-Schwartz.

Lemma 6 (Covariance inequality)

Let X and Y be random variables. Then

$$\operatorname{Var}(Y) \geq \frac{\operatorname{Cov}(Y,X)\operatorname{Cov}(Y,X)}{\operatorname{Var}(X)}$$

Proof.

Let $\mu := \mathrm{E}\, X$ and $\nu := \mathrm{E}\, Y$. Then

$$\begin{aligned} |\operatorname{Cov}(X,Y)|^2 &= \left|\mathrm{E}\left[(X-\mu)(Y-\nu)\right]\right|^2 \\ &\leq \mathrm{E}\left[(X-\mu)^2\right]\mathrm{E}\left[(Y-\nu)^2\right] \\ &= \operatorname{Var}(X)\operatorname{Var}(Y) \end{aligned}$$

Convex functions

- A function g is *convex* if for any x, y and any $\alpha \in [0, 1]$,

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y).$$

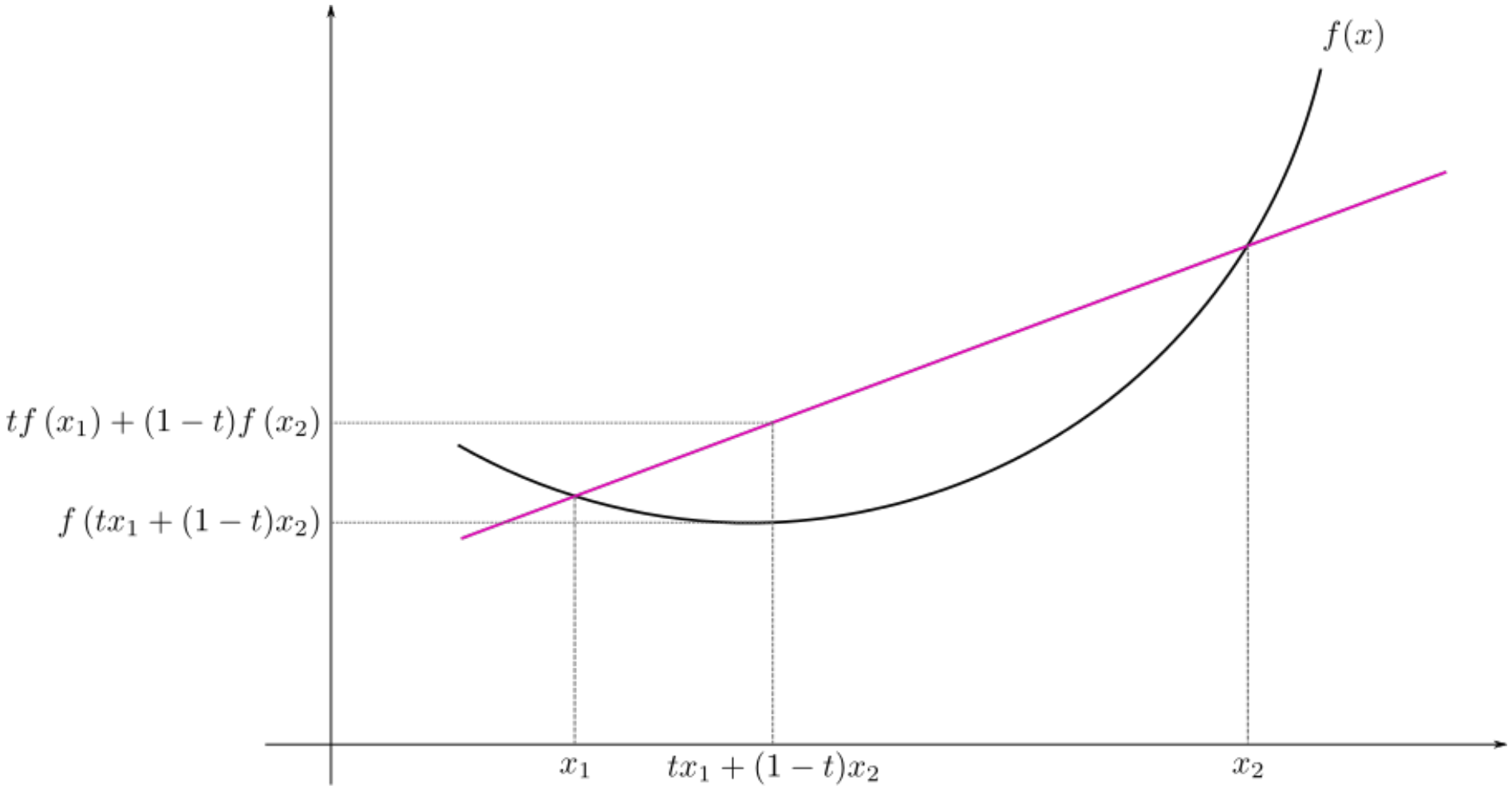
- If $g''(x) > 0$ for all x, then g is convex.
- A function g is *concave* if $-g$ is convex.

Example 7

Examples of convex functions: $g_1(x) = x^2$ and $g_2(x) = e^x$, since $g_1''(x) = 2 > 0$ and $g_2(x) = e^x > 0$ for all x.

Examples of concave functions: $g_3(x) = -x^2$ and $g_4(x) = log(x)$.

Convex functions



Jensen's inequality

In the context of probability theory,

Theorem 8 (Jensen's inequality)

If g is convex,

$$E\left[g(X)\right] \geq g\left(\mathrm{E}\, X\right)$$

Example 9

It follows directly from Jensen's inequality, the following:

$$\mathrm{E}(X^2) \geq \{\mathrm{E}(X)\}^2$$
$$\mathrm{E}(1/X) \geq 1/\mathrm{E}\, X$$
$$\mathrm{E}(\log X) \geq \log(\mathrm{E}\, X)$$

Limits

Recall the limits of sequences of real numbers $x_1, x_2, \ldots$:

Definition 10 (Limits of real sequences)

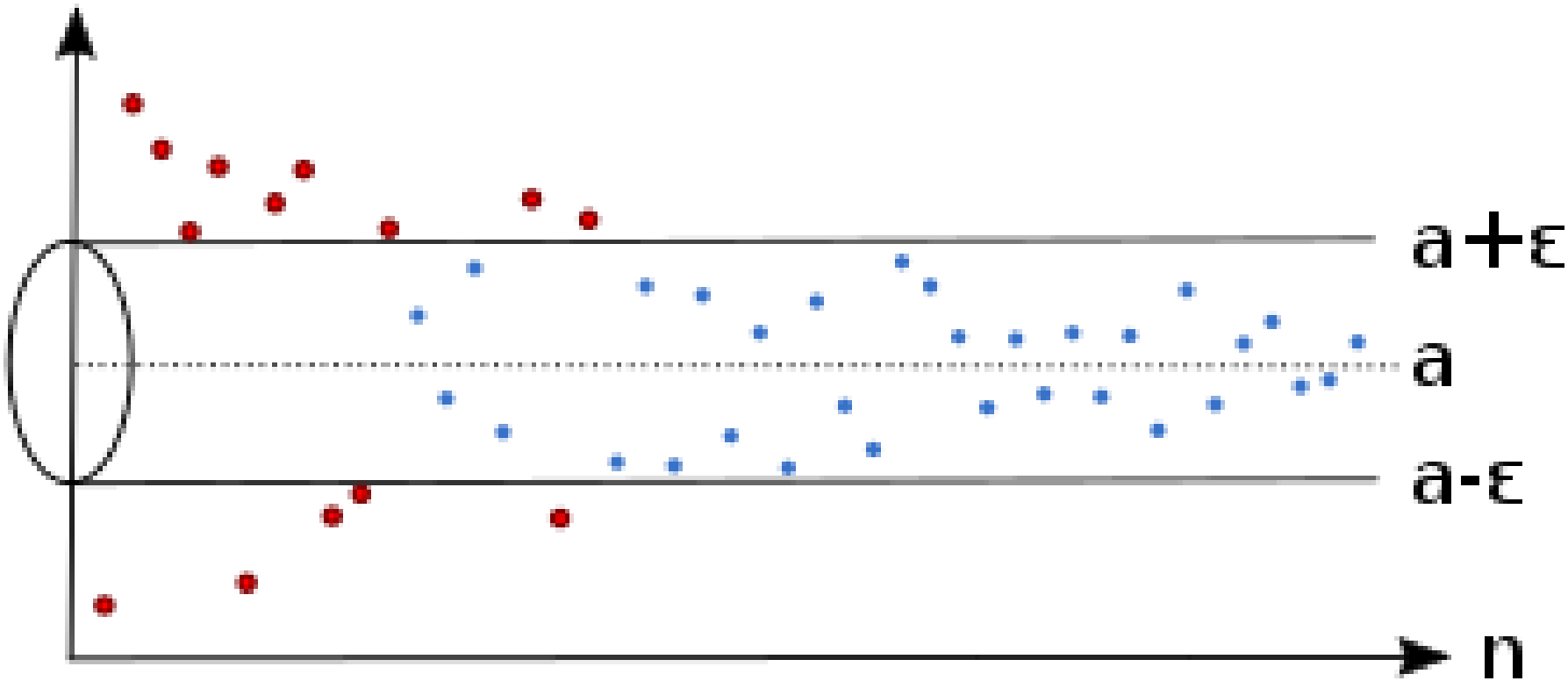
We call x the limit of the real sequence (x_n) if for each real number $\epsilon > 0$, there exists a natural number N such that, for every natural number $n \geq N$, we have $|x_n - x| < \epsilon$.

We write $\lim_{n\to\infty} x_n = x$, or simply $x_n \to x$. This also means that $|x_n - x| \to 0$ as $n \to \infty$. For every measure of closeness ϵ, the sequence's terms are eventually that close to the limit.

Example 11

- If $x_n = c$ for some constant $c \in \mathbb{R}$, then $x_n \to c$.
- If $x_n = 1/n$, then $x_n \to 0$.
- $\lim_{n\to\infty}(1 + 1/n)^n = e$.

Limits (cont.)



Convergence of random variables

We can say similar things about sequences of **random variables**, e.g. X is the limit of a sequence (X_n) if $|X_n - X| \to 0$ as $n \to \infty$. There are some subtle issues here:

1. $|X_n - X|$ itself is a r.v., i.e. it takes difference values in the sample space Ω. Therefore, $|X_n - X| \to 0$ should hold (almost) entirely on the sample space. This calls for a probability statement.

2. Since r.v. have distributions, we may also consider convergence of their distributions $F_{X_n}(x) \to F_X(x)$ for all x.

Types of convergence

Let $X_1, X_2, \ldots$ be a sequence of r.v., and X be another r.v.. The two main types of convergence for r.v. are defined as follows.

Definition 12 (Convergence in probability)

X_n converges to X in probability if for any constant $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$. We write $X_n \xrightarrow{P} X$, or $\text{plim}_{n\to\infty} X_n = X$.

Definition 13 (Convergence in distribution)

X_n converges to X in distribution if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$. We write $X_n \xrightarrow{D} X$.

Remarks:

1. X may be a constant, since a constant is a r.v. with probability mass concentrated on a single point.
2. If $X_n \xrightarrow{P} X$, it also holds that $X_n \xrightarrow{D} X$, but **not** vice versa.

Types of convergence (cont.)

Example 14

Let $X \sim \mathrm{N}(0,1)$ and $X_n = -X$ for all $n \geq 1$. Then, clearly $F_{X_n} \equiv F_X$ (by linearity of normal distributions). Hence, $X_n \xrightarrow{\mathrm{D}} X$.

However, $X_n \not\xrightarrow{\mathrm{P}} X$, as for any $\epsilon > 0$,

$$\begin{aligned} \mathrm{P}(|X_n - X| > \epsilon) &= \mathrm{P}(2|X| > \epsilon) \\ &= \mathrm{P}(|X| > \epsilon/2) > 0. \end{aligned}$$

So we cannot have that $\mathrm{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

Types of convergence (cont.)

The convergence of r.v. is especially important when considering how good our estimator is. For example, suppose we have collected data $X_1, \ldots, X_n$ for estimation purposes, which can be treated to be realisations of a sequence of r.v.. Let $\hat{\theta}_n = h(X_1, \ldots, X_n)$ be an estimator for θ.

- Naturally, we require $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.
- But $\hat{\theta}_n$ is a r.v.; it takes different values with different samples. To consider how good it is an estimator of θ, we hope that the distribution of $(\hat{\theta}_n - \theta)$ becomes more concentrated around zero when n increases.

Mean square convergence

It is sometimes more convenient to consider the mean square convergence:

Definition 15 (Mean square convergence)

X_n converges in mean square to X if $\mathrm{E}\left[(X_n - X)^2\right] \to 0$ as $n \to \infty$. We write $X_n \xrightarrow{\text{m.s.}} X$.

It follows that from Markov's inequality,

$$\begin{aligned} \mathrm{P}(|X_n - X| \geq \epsilon) &= \mathrm{P}(|X_n - X|^2 \geq \epsilon^2) \\ &\leq \frac{\mathrm{E}\left[(X_n - X)^2\right]}{\epsilon^2} \end{aligned}$$

Therefore, if $X_n \xrightarrow{\text{m.s.}} X$, it also holds that $X_n \xrightarrow{\mathrm{P}} X$.

Mean square convergence (cont.)

Example 16

Let

$$U \sim \text{Unif}(0,1) \text{ and } X_n = \begin{cases} n & \text{if } U < 1/n \\ 0 & \text{otherwise} \end{cases}$$

Then, for some $\epsilon > 0$, $\text{P}(|X_n| > \epsilon) \leq \text{P}(U < 1/n) = 1/n \to 0$ as $n \to \infty$.
Hence, $X_n \xrightarrow{\text{P}} X$.

However,

$$\text{E}(X_n^2) = n^2 \, \text{P}(U < 1/n) = n \to \infty$$

hence $X_n \xrightarrow{\text{m.s.}} X$.

Mean square convergence (cont.)

Example 17

Let $X_n = n$ w.p. $1/n$, and 0 w.p. $1 - 1/n$. Then $X_n \xrightarrow{\mathrm{P}} 0$, since

$$\mathrm{P}(|X_n| \leq \epsilon) \leq 1/n \to 0$$

as $n \to \infty$. However, $\mathrm{E}(X_n) = n \times 1/n = 1 \not\to 0$.

Caution

$X_n \xrightarrow{\mathrm{P}} X$ does not imply $\mathrm{E}(X_n) \to \mathrm{E}(X)$.

Relationship between convergences

In general,

$$\begin{array}{ccccc} \text{Convergence in} & \Rightarrow & \text{Convergence} & \Rightarrow & \text{Convergence} \\ \text{mean square} & & \text{in probability} & & \text{in distribution} \end{array}$$

Caution

When $X_n \xrightarrow{\text{D}} X$, we also write $X_n \xrightarrow{\text{D}} F_X$, where F_X is the cdf of X.

However, the notation $X_n \xrightarrow{\text{P}} F_X$ does not make sense!

Slutzky's Theorem

Theorem 18 (Slutzky's Theorem)

Let X_n, Y_n, X, and Y be r.v., g a continuous function, and c a real constant. Then,

1. If $X_n \xrightarrow{\text{P}} X$ and $Y_n \xrightarrow{\text{P}} Y$, then
   - $X_n + Y_n \xrightarrow{\text{P}} X + Y$;
   - $X_n Y_n \xrightarrow{\text{P}} XY$; and
   - $g(X_n) \xrightarrow{\text{P}} g(X)$.
2. If $X_n \xrightarrow{\text{D}} X$ and $Y_n \xrightarrow{\text{D}} c$, then
   - $X_n + Y_n \xrightarrow{\text{D}} X + c$;
   - $X_n Y_n \xrightarrow{\text{D}} cX$; and
   - $g(X_n) \xrightarrow{\text{D}} g(X)$.

Caution

Note that $X_n \xrightarrow{\text{D}} X$ and $Y_n \xrightarrow{\text{D}} Y$ does **not** in general imply $X_n + Y_n \xrightarrow{\text{D}} X + Y$.

The (weak) Law of Large Numbers (LLN)

Let $X_1, X_2, \ldots$ be iid with mean μ and variance σ^2. Let $\bar{X}_n$ denote the sample mean, i.e.

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

Recall two simple facts:

$$\mathrm{E}(\bar{X}_n) = \mu \qquad \text{and} \qquad \mathrm{Var}(\bar{X}_n) = \sigma^2/n.$$

Exercise: Prove this.

The (weak) Law of Large Numbers (LLN) (cont.)

Definition 19 (The weak Law of Large Numbers)

As $n \to \infty$, $\bar{X}_n \xrightarrow{\mathrm{P}} \mu$.

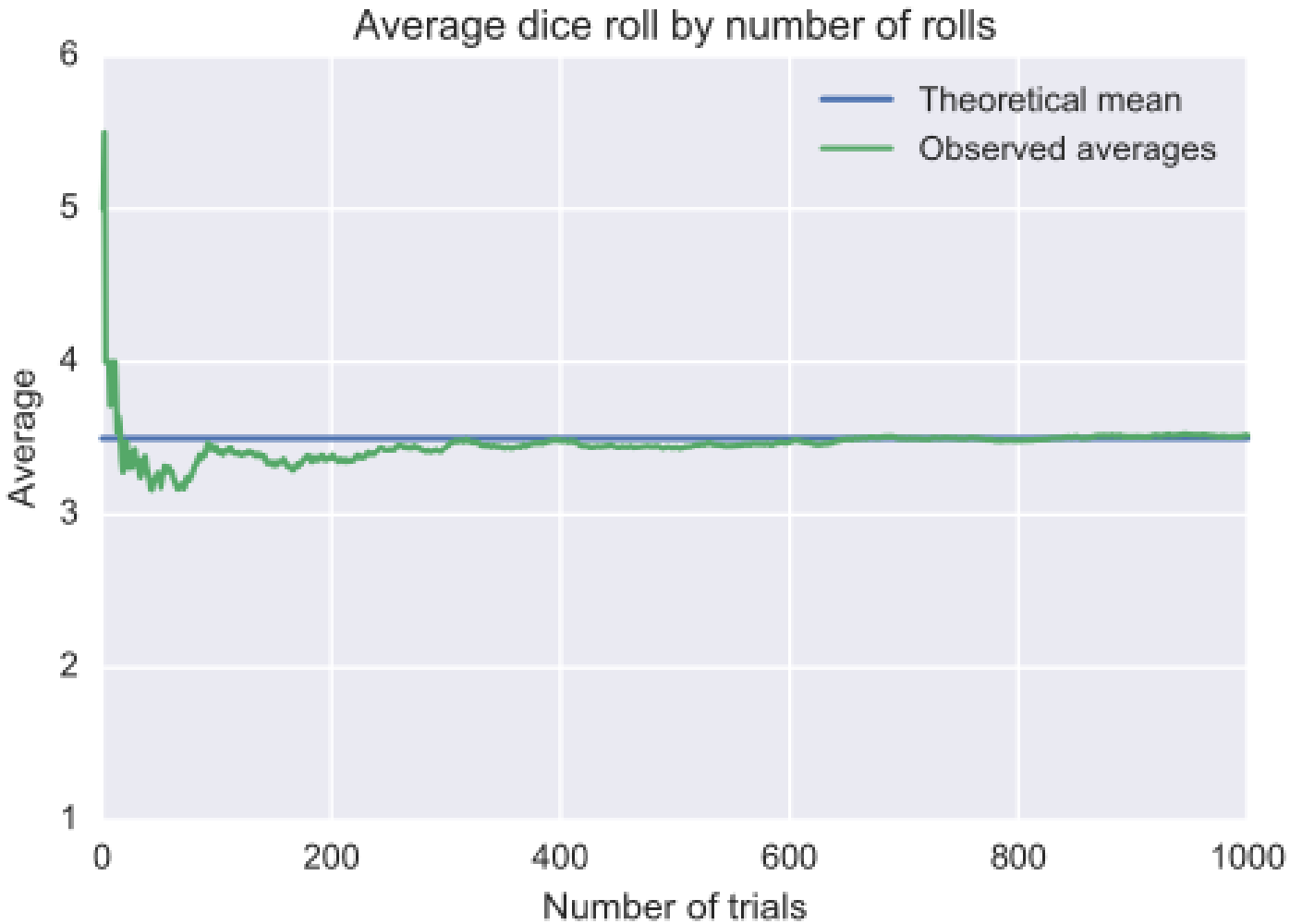
The LLN is very natural: When the sample size increases, the sample mean becomes more and more close to the population mean. Furthermore, the distribution of $\bar{X}_n$ degenerates to a single point distribution at μ.

Proof.

Prove this as an exercise. *Hint: Use Chebyshev's inequality.*

The (weak) Law of Large Numbers (LLN) (cont.)

Let $X_1, X_2, \ldots$ be the score of randomly thrown dice.



The Central Limit Theorem (CLT)

Definition 20 (The Central Limit Theorem)

As $n \to \infty$,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$$

The standardised sample mean $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is approximately standard normal when the sample size is large. Hence, we can make statements such as

- $\sqrt{n}(\bar{X}_n - \mu)/\sigma \approx \mathrm{N}(0,1)$
- $\sqrt{n}(\bar{X}_n - \mu) \approx \mathrm{N}(0, \sigma^2/n)$
- $\bar{X}_n - \mu \approx \mathrm{N}(0, \sigma^2/n)$
- $X_n \approx \mathrm{N}(\mu, \sigma^2/n)$

The CLT is one of the reasons why the normal distribution is the most useful and important distribution in statistics.

The Central Limit Theorem (CLT) (cont.)

Example 21

If we take a sample $X_1, \ldots, X_n$ from Unif$(0, 1)$, the standardised histogram will resemble the density function $f(x) = \mathbb{1}_{(0,1)}(x)$ (i.e. horizontal line at 1 from $x = 0$ to $x = 1$). Now, the sample mean when calculated will be close to $\mu = \mathrm{E}(X_i) = 0.5$, provided n is sufficiently large (LLN).

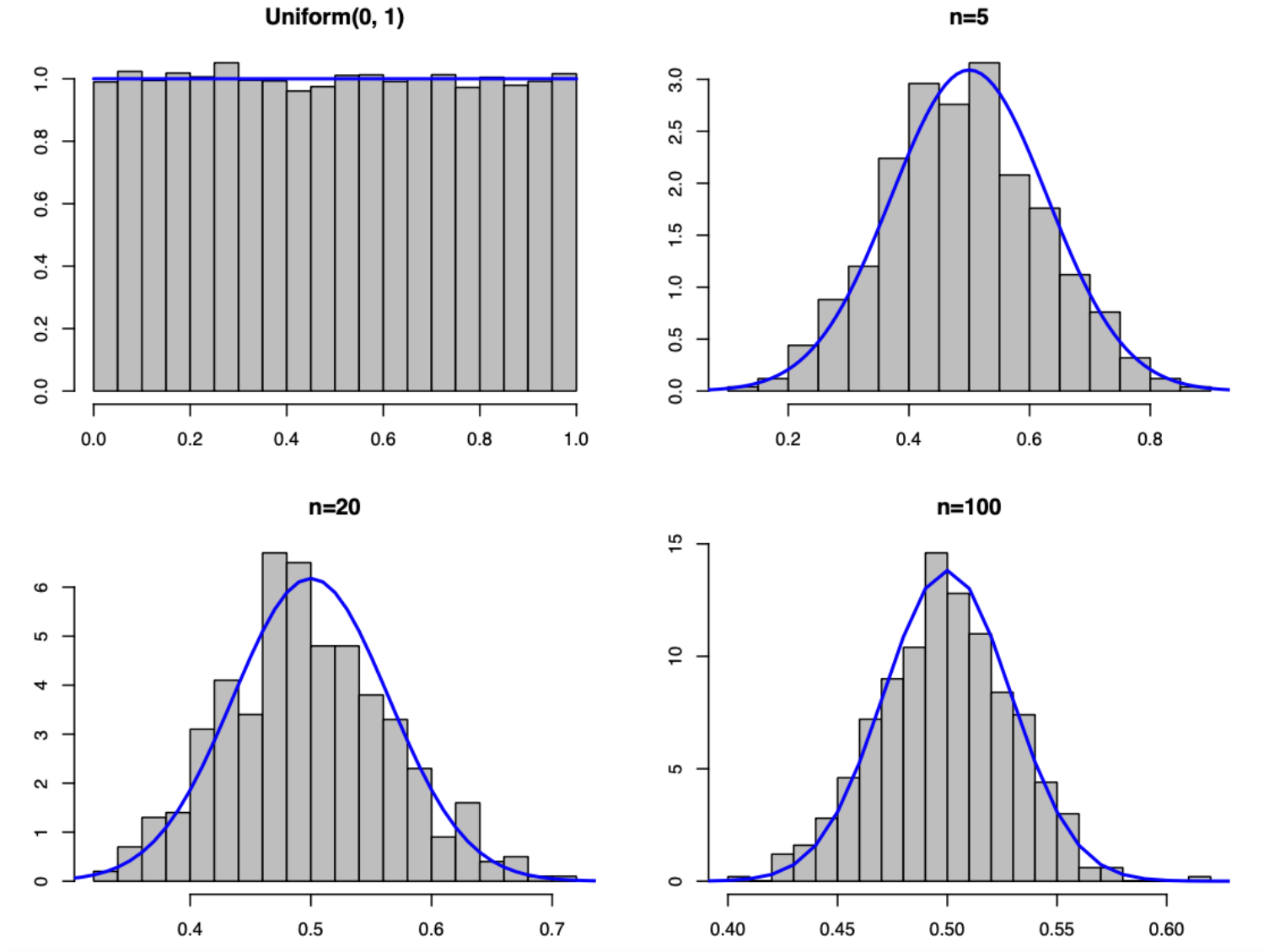
However, the CLT implies that

$$\bar{X}_n \approx \mathrm{N}(0.5, (12n)^{-1})$$

since $\mathrm{Var}(X_i) = 1/12$.

If we take many samples of size n and compute the sample mean for each sample, we then obtain many sample means. The standardised histogram of those samples resembles the pdf of $\mathrm{N}(0.5, (12n)^{-1})$.

The Central Limit Theorem (CLT) (cont.)



The Central Limit Theorem (CLT) (cont.)

Example 22

Suppose $X_1, \ldots, X_n$ is an iid sample. A natural estimator for the population mean $\mu = \mathrm{E}(X_i)$ is the sample mean $\bar{X}_n$. By the CLT, we can easily gauge the error of this estimation as follows:

$$\begin{aligned}
\mathrm{P}(|\bar{X}_n - \mu| > \epsilon) &= \mathrm{P}(|\sqrt{n}(\bar{X}_n - \mu)/\sigma| > \sqrt{n}\epsilon/\sigma) \\
&\approx \mathrm{P}(|Z| > \sqrt{n}\epsilon/\sigma), \text{where } Z \sim \mathrm{N}(0,1) \\
&= 2\,\mathrm{P}(Z > \sqrt{n}\epsilon/\sigma) \\
&= 2\big(1 - \Phi(\sqrt{n}\epsilon/\sigma)\big)
\end{aligned}$$

So with ϵ,n given, we can find the value $\Phi(\sqrt{n}\epsilon/\sigma)$ from the table for the standard normal distribution, *if we know* σ.

The Central Limit Theorem (CLT) (cont.)

Remarks.

- Let $\epsilon := 2\sigma/\sqrt{n} = 2\sqrt{\text{Var}(\bar{X}_n)}$. Then

$$\text{P}(|\bar{X}_n - \mu| < \epsilon) \approx 2\Phi(2) - 1 = 0.954$$

  Hence, if one estimates μ by $\bar{X}_n$, and repeats it a large number of times, about 95% of times, μ is within $2 \times \text{s.d.}(\bar{X}_n)$ distance away from $\bar{X}_n$. Recall the "68-95-99.7" rule.

- Typically, $\sigma^2 = \text{Var}(X_i)$ is unknown in practice. We estimate is using the (unbiased) sample variance estimator

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

The Central Limit Theorem (CLT) (cont.)

- Note that the estimate of $\sqrt{\text{Var}(\bar{X}_n)} = \sigma/\sqrt{n}$, given by $s_n/\sqrt{n}$, is called the **standard error** of the sample mean. In full,

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

- In fact, it still holds that as $n \to \infty$,

$$\frac{\bar{X}_n - \mu}{s_n \sqrt{n}} \xrightarrow{\text{D}} \text{N}(0, 1)$$

  which implies that replacing σ with s_n in the above example yields the same results. Hence, if one estimates μ by $\bar{X}_n$, and repeats it a large number of times, about 95% of times, μ is within $2 \times \text{s.d.}(\bar{X}_n)$ distance away from $\bar{X}_n$.

Fundamental concepts in statistical inference

Let $X_1, \ldots, X_n$, be a sample from a population that follows some distribution with pdf $f(x|\theta)$. The form of the pdf f is known, but the parameter θ of the pdf is unknown. Often, we may specify $\theta \in \Theta$, where Θ is the parameter space. Note that θ may be a vector $\theta = (\theta_1, \ldots, \theta_p)^\top$.

Example 23

- For $\mathrm{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)^\top$, so $p = 2$ and $\Theta = \mathbb{R} \times \mathbb{R}_{\geq 0}$.
- For $\mathrm{Poi}(\lambda)$, $\theta = \lambda$, so $p = 1$ and $\Theta = \mathbb{R}_{\geq 0}$.

Most inference problems can be identified as one of three types:

1. Point estimation
2. Confidence sets
3. Hypothesis testing

for the parameter θ.

Point estimation

GOAL: Provide a single "best guess" of θ, based on observations $X_1, \ldots, X_n$. Formally, we may write

$$\hat{\theta} = \hat{\theta}_n = g(X_1, \ldots, X_n)$$

as a point estimator for θ, where $g(X_1, \ldots, X_n)$ is a statistic.

Example 24

A natural point estimator for the mean $\mu = \mathrm{E}(X_1)$ is the sample mean $\hat{\mu} = \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

Remark: Parameters to be estimated are **unknown constants**. Their estimators are viewed as r.v., although in practice, $\hat{\theta}$ admit some concrete values.

Point estimation (cont.)

Estimator vs Estimate

We use the term "estimator" to denote the function that gives the estimate. On the other hand, an "estimate" is the realised value of the estimator function.

Estimator/Estimate vs true value

The standard convention is to denote estimators/estimates of parameters with hats on the respective symbols (e.g. $\hat{\theta}$), whereas true values do not have hats (c.f. θ).

A good estimator should make $|\hat{\theta} - \theta|$ as small as possible. However,

1. θ is unknown; and
2. the value of $\hat{\theta}$ changes with the sample observed.

Hence, we seek for an estimator $\hat{\theta}$ which makes the **mean squared error** (MSE) as small as possible for all possible values of θ.

MSE, bias and standard error

Definition 25 (Bias)

The bias of an estimator $\hat{\theta}$ is defined to be

$$\text{Bias}_\theta(\hat{\theta}) = \text{E}_\theta(\hat{\theta}) - \theta.$$

When $\text{E}_\theta(\hat{\theta}) = \theta$, $\text{Bias}_\theta(\hat{\theta}) = 0$ for all possible values of θ, and $\hat{\theta}$ is called an **unbiased estimator** for θ.

Definition 26 (Mean squared error)

The MSE of the estimator $\hat{\theta}$ is defined as

$$\text{MSE}_\theta(\hat{\theta}) = \text{E}_\theta\left[(\hat{\theta} - \theta)^2\right] = \{\text{Bias}_\theta(\hat{\theta})\}^2 + \text{Var}_\theta(\hat{\theta}).$$

Note: E_θ means that expectations are taken with respect to the distribution which uses the **true value** θ.

MSE, bias and standard error (cont.)

Definition 27 (Standard error)

The standard error of the estimator $\hat{\theta}$ is defined as

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}_{\hat{\theta}}(\hat{\theta})}$$

Note that in the definition of the standard error of an estimator, the expectations (variance) are taken using the distribution with *estimated parameters* $\hat{\theta}$.

MSE, bias and standard error (cont.)

Example 28

Let $Y_1, \ldots, Y_n$ be a sample from Bern(p), with p unknown. Let $\hat{p} := \bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$. Then,

$$\mathrm{E}(\hat{p}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(Y_i) = p; \text{ and}$$

$$\mathrm{Var}(\hat{p}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{p(1-p)}{n}.$$

Therefore, $\bar{Y}_n$ is an unbiased estimator for p, with standard deviation $\sqrt{p(1-p)/n}$, and standard error $\mathrm{SE}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$.

For example, if $n = 10$ and $\hat{Y}_n = 0.3$, we have $\hat{p} = 0.3$ and $\mathrm{SE}(\hat{p}) = 0.1449$, while the standard deviation of $\hat{p}$ is unknown.

Consistency

Definition 29 (Consistency)

$\hat{\theta}_n$ is a consistent estimator for θ if $\hat{\theta}_n \to \theta$ as $n \to \infty$.

Consistency is a natural condition for a reasonable estimator as $\hat{\theta}_n$ should converge to θ if we have a (theoretically) infinite amount of information. Therefore, a **non-consistent estimator should not be used in practice!**

Remark

If $\text{MSE}(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n \xrightarrow{\text{m.s.}} \theta$ (by definition). Therefore, $\hat{\theta}_n \xrightarrow{\text{P}} \theta$ too, so $\hat{\theta}$ is a consistent estimator for θ.

Example 30

(Cont. e.g. 28) Since $\text{Bias}(\hat{p}) = 0$, $\text{MSE}(\hat{p}) = \text{Var}(\hat{p}) = p(1-p)/n \to 0$ as $n \to \infty$. Hence, $\hat{p}$ is consistent.

Consistency vs Unbiasedness

Consistency and unbiasedness are two different concepts:

- Unbiasedness ($\mathrm{E}(\hat{\theta}) = \theta$) is a statement about the expected value of the sampling distribution of the estimator.
- Consistency ($\mathrm{plim}_n \hat{\theta}_n = \theta$) is a statement about "where the sampling distribution of the estimator is going" as the sample size increases.

Both are desirable properties of estimators, though it might be possible for one to be satisfied but not the other (see Example 32).

Consistency vs Unbiasedness (Cont.)

In both examples, let $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(\mu, \sigma^2)$.

Example 31

Define $\hat{\mu} = X_1$. Then $\hat{\mu}$ is unbiased since $\mathrm{E}(X_1) = \mu$, but it is not consistent since the distribution of $\hat{\mu}$ is always $\mathrm{N}(\mu, \sigma^2)$ and will never concentrate around μ even with infinite sample size.

Example 32

Define $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$. It is a fact that $\mathrm{E}(\sigma^2) = \frac{n-1}{n}\sigma^2$, which shows that $\hat{\sigma}^2$ is biased in finite samples. On the other hand, we can show

$$\mathrm{Var}(\hat{\sigma}^2) = \frac{2\sigma^4(n-1)}{n^2} \to 0$$

as $n \to \infty$. Therefore, $\mathrm{MSE}(\hat{\sigma}^2) \to 0$, and $\hat{\sigma}^2$ is therefore consistent.

Asymptotic normality

Definition 33 (Asymptotic normality)

An estimator $\hat{\theta}_n$ is asymptotically normal if

$$\frac{\hat{\theta}_n - \theta}{\mathrm{SE}(\hat{\theta}_n)} \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1)$$

Alternatively, it can be written as $\hat{\theta}_n \xrightarrow{\mathrm{D}} \mathrm{N}(\theta, \mathrm{SE}(\hat{\theta}_n)^2)$.

Remark:

- Many good estimators such as maximum likelihood estimator (MLE), least squares estimator (LSE) and method of moments estimator (MME) are asymptotically normal under some mild conditions.
- The desire for asymptotic normality is simply for convenience (e.g. hypothesis testing of parameters).

Estimation methods

Thus far, we have only discussed desirable properties of estimators, but not any specific way of actually obtaining estimates (apart from the sample mean for μ). Many methods exist, and to name a few:

- **Method of moments**: Express population moments as functions of parameters, then substituting in the sample moments.
- **Maximum likelihood estimation**: Parameter estimates are those which maximise the likelihood function.
- **Least squares estimation**: Parameter estimates are those which minimise some error function.

And many others... (generalised MoM, minimum mean squared error, Bayesian least squared error, maximum a posteriori, particle filter, Markov chain Monte Carlo methods, etc.)

In this course, we will be focusing on ML estimation.

Likelihood

The likelihood is one of the most fundamental concepts in all types of statistical inference.

Definition 34 (Likelihood)

Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ has pdf or pmf $f(\mathbf{x}|\boldsymbol{\theta})$, and we have observed $\mathbf{X} = \mathbf{x}^*$. Then, the likelihood function with observation $\mathbf{x}^*$ is defined as

$$\mathrm{L}(\boldsymbol{\theta}) \equiv \mathrm{L}(\boldsymbol{\theta}|\mathbf{x}^*) = f(\mathbf{x}^*|\boldsymbol{\theta})$$

Pdf/Pmf vs likelihood

Pdf/Pmf is a function of $\mathbf{x}$, specifying the distribution of the random variable $\mathbf{X}$.

Likelihood is a function of $\boldsymbol{\theta}$, reflecting information on θ contained in observation $\mathbf{x}$.

Likelihood (cont.)

HJ (2018, Ch. 3)

It was Fisher in 1922 who introduced the method of maximum likelihood as an objective way of conducting statistical inference. This method of inference is distinguished from the Bayesian school of thought in that only the data may inform deductive reasoning, but not any sort of prior probabilities. Towards the later stages of his career, his work reflected the view that the likelihood is to be more than simply a device to obtain parameter estimates; it is also a vessel that carries uncertainty about estimation. In this light and in the absence of the possibility of making probabilistic statements, one should look to the likelihood in order to make ***rational conclusions about an inference problem****. Specifically, we may ask two things of the likelihood function: where is the maxima and what does the graph around the maxima look like? The first of these two problems is maximum likelihood estimation, while the second concerns the Fisher information.*

Likelihood (cont.)

Example 35

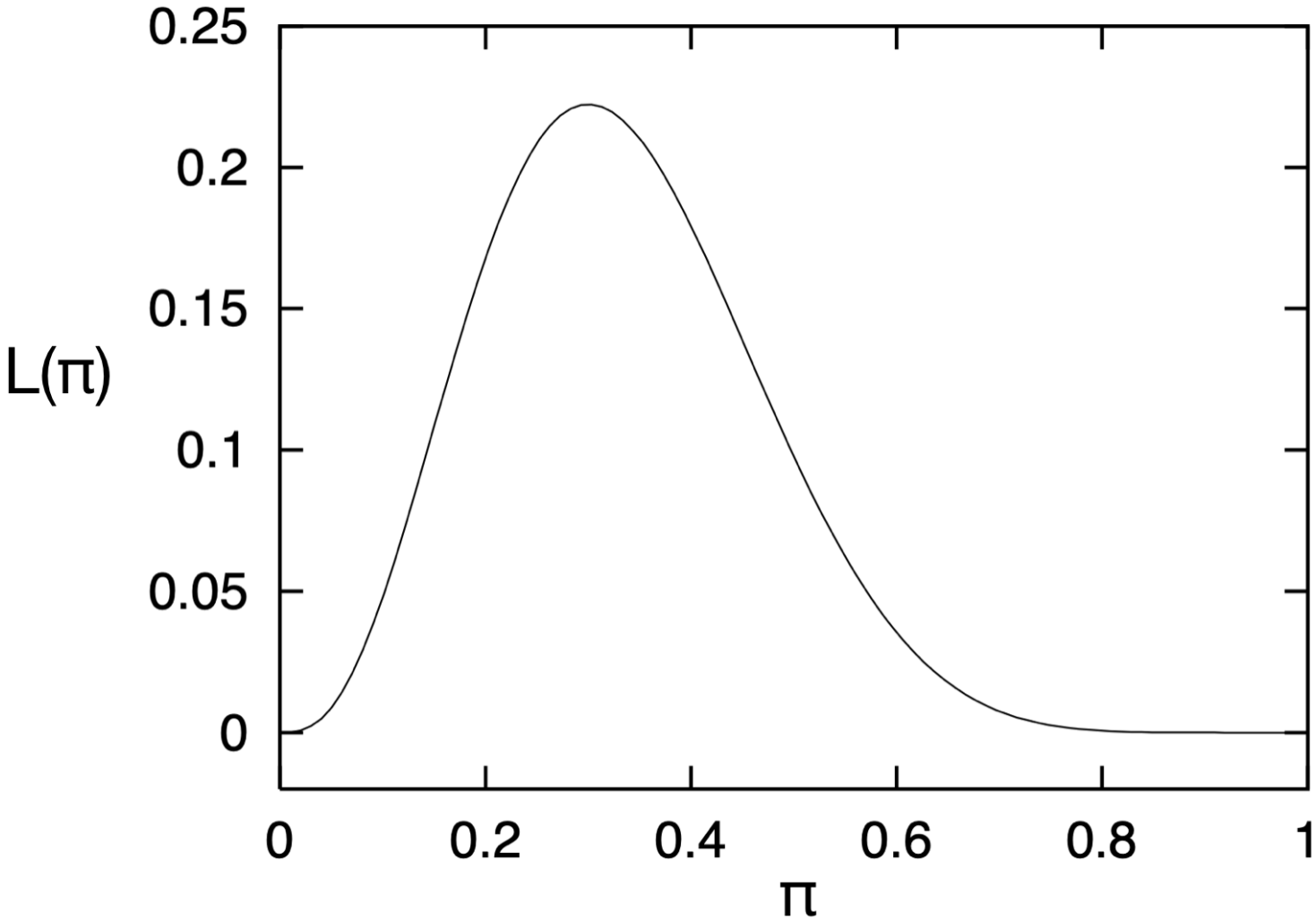
Suppose that x is the number of success from a known number n of independent trials with unknown probability of success π. The probability function, and so the likelihood function, is

$$L(\pi) = f(x|\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$$

Remember, x is known, while π is unknown.

The likelihood function can be graphed as a function of π. It changes shape for different values of x. As an example, a likelihood function for $x = 3$ when $n = 10$ is shown in the figure in the next slide.

Likelihood (cont.)



Likelihood (cont.)

Notice that the likelihood function shown is **not** a density function. It does not have an area of one below it.

We use the likelihood function to compare the plausibility of different possible parameter values.

- For instance, the likelihood is much larger for $\pi = 0.3$ than for $\pi = 0.8$.
- That is, the data $x = 3$ have a greater probability of being observed if $\pi = 0.3$ than if $\pi = 0.8$.

Note

In the above argument, we do not need to calculate exact probabilities under different values of π. Only the order (magnitude) of those quantities matter.

Log-likelihood

Let $X_1, \ldots, X_n$ be iid with pdf $f(x|\boldsymbol{\theta})$. Write $\mathbf{X} = (X_1, \ldots, X_n)^\top$. Then, the likelihood function is a **product** of n terms:

$$\mathrm{L}(\boldsymbol{\theta}) \equiv \mathrm{L}(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}).$$

Definition 36 (Log-likelihood)

The log-likelihood function is defined to be

$$l(\boldsymbol{\theta}) \equiv l(\boldsymbol{\theta}|\mathbf{X}) = \log f(\mathbf{X}|\boldsymbol{\theta}).$$

Thus, for iid observations, the log-likelihood becomes a **sum** of n terms:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i|\boldsymbol{\theta}).$$

This explains why log-likelihood functions are often used with independent observations.

Maximum likelihood estimator

Definition 37 (Maximum likelihood estimator)

A maximum likelihood estimator (MLE), $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X}) \in \Theta$ of parameter $\boldsymbol{\theta}$ is defined to be

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathrm{L}(\boldsymbol{\theta}|\mathbf{X}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \{\log \mathrm{L}(\boldsymbol{\theta}|\mathbf{X})\}.$$

By definition, the MLE satisfies $L(\hat{\boldsymbol{\theta}}|\mathbf{X}) \geq L(\boldsymbol{\theta}|\mathbf{X})$ for all $\boldsymbol{\theta} \in \Theta$. Obviously, an MLE is the most plausible value for $\boldsymbol{\theta}$ as judged by the likelihood function.

Maximum likelihood estimator (cont.)

Definition 38 (Score function)

The score function is defined to be

$$S(\boldsymbol{\theta}|\mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathrm{L}(\boldsymbol{\theta}|\mathbf{X}).$$

In many cases where Θ is continuous and the maximum does not occur at a boundary of Θ, $\hat{\boldsymbol{\theta}}$ is often the solution of the equation

$$S(\boldsymbol{\theta}|\mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathrm{L}(\boldsymbol{\theta}|\mathbf{X}) = 0.$$

This requires solving a system of p (one for each θ_k in $\boldsymbol{\theta}$) simultaneous equations.

Maximum likelihood estimator (cont.)

Example 39

Suppose that $Y_1, \ldots, Y_n$ is an iid random sample from $\mathrm{N}(\mu, \sigma^2)$, where neither μ nor σ^2 is known. Then, the log-likelihood function is

$$\begin{aligned} l(\mu, \sigma^2) &= \log\left\{(\sqrt{2\pi\sigma^2})^{-n} e^{-\sum_{i=1}^{n}(Y_i-\mu)^2/(2\sigma^2)}\right\} \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2 \end{aligned}$$

Taking derivative with respect to μ gives us the first component of the score function:

$$S_1(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)$$

Maximum likelihood estimator (cont.)

Example 39

Equating this to zero gives the MLE for μ

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu) = 0$$

$$\sum_{i=1}^{n} Y_i - n\mu = 0$$

$$\Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} Y_i =: \bar{Y}_n$$

Thus, $\hat{\mu} = \bar{Y}_n$.

Maximum likelihood estimator (cont.)

Example 39

The **profile log-likelihood** remaining is

$$l(\hat{\mu}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{\mu})^2$$

$$= \text{const.} - \frac{n}{2}\log\sigma^2 - \frac{n\hat{\sigma}^2}{2\sigma^2}$$

where we defined $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu})^2$. Taking derivatives with respect to σ^2, we get

$$S_2(\boldsymbol{\theta}) = -\frac{n}{2\sigma^2} + \frac{n\hat{\sigma}^2}{2\sigma^4}$$

Maximum likelihood estimator (cont.)

Example 39

Equating S_2 to zero we get

$$-\frac{n}{2\sigma^2} + \frac{n\hat{\sigma}^2}{2\sigma^4} = 0$$
$$\Rightarrow \sigma^2 = \hat{\sigma}^2$$

Thus the MLE for σ^2 is $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$.

Remark

The MLE for σ^2 is biased since

$$\mathrm{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Invariance property of MLE

Definition 40 (One-to-one transformation)

Let $g : \mathcal{X} \to \mathcal{Y}$ be a function. Then g is said to be **one-to-one** if and only if for every $y \in \mathcal{Y}$ there is at most **one** $x \in \mathcal{X}$ such that $g(x) = y$. Equivalently, g is one-to-one if and only if for $x_1, x_2 \in \mathcal{X}$, $g(x_1) = g(x_2)$ implies $x_1 = x_2$.

Definition 41 (Invariance property of MLE)

Suppose $\mathbf{X} \sim f(\mathbf{x}|\boldsymbol{\theta})$, and $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ is 1 one-to-one transformation. Let $\hat{\boldsymbol{\theta}}$ be the MLE for $\boldsymbol{\theta}$, i.e.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathrm{L}(\boldsymbol{\theta}).$$

Then, the MLE for $\boldsymbol{\psi}$ is

$$\hat{\boldsymbol{\psi}} = \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}).$$

Invariance property of MLE (cont.)

Example 42

Let $\hat{\pi}$ be the MLE for π after observing data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(\pi)$. The log-odds of an event happening is given by $\nu = \log\left(\pi / \log(1-\pi)\right)$, which is a one-to-one transformation of π. Therefore, the MLE for ν is given by

$$\hat{\nu} = \log \frac{\hat{\pi}}{1-\hat{\pi}}.$$

Numerical computation of MLEs

In modern statistical applications, it is typically difficult to find explicit analytical forms for the MLE. These estimators are found more often by iterative procedures built into computer software.

- An iterative scheme starts with some guess at the MLE and then steadily improves it with each iteration.
- The estimator is considered to be found when it has become numerically stable.
- Sometimes, the iterative procedures become trapped at a local maximum which is not a global maximum.
- There may be a very large number of parameters in a model, which makes such local entrapment more common.
- Some iterative schemes include: Newton-Raphson scheme, Fisher scoring algorithm, quasi-Newton methods, gradient descent, conjugate gradients, etc.–we won't go into details in this course.

Fisher information

In simple terms, the Fisher information measures the amount of information that an observable random variable X carries about an unknown parameter θ of the statistical model that models X.

Consider the unidimensional case for now: let $X \sim f(x|\theta)$, where $\theta \in \mathbb{R}$.

Definition 43 (Fisher information)

The Fisher information is defined to be the expectation of the second moment of the score function, i.e.

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta)\right)^2\right] \in \mathbb{R}$$

Fisher information (cont.)

Lemma 44 (Expectation of the score is zero)

Under certain regularity conditions, $\mathrm{E}[S(\theta)] = 0$.

Proof.

$$\begin{aligned}
\mathrm{E}[S(\theta)] &= \int \frac{\mathrm{d}}{\mathrm{d}\theta} \log\{f(x|\theta)\} f(x|\theta)\, \mathrm{d}x \\
&= \int \frac{\mathrm{d}}{\mathrm{d}\theta} f(x|\theta)\, \mathrm{d}x \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta} \underbrace{\int f(x|\theta)\, \mathrm{d}x}_{1} = 0
\end{aligned}$$

Fisher information (cont.)

Corollary 45 (The Fisher information is the variance of the score)

$$\mathcal{I}(\theta) = \mathsf{Var}\left(S(\theta)\right)$$

Proof.

Prove this as an exercise.

Fisher information (cont.)

Lemma 46 (Another definition for the Fisher information)

Under certain regularity conditions, the Fisher information can also be defined to be negative expected value of the second derivative of the score

$$\mathcal{I}(\theta) = -\,\mathsf{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X|\theta)\right] \in \mathbb{R}$$

Fisher information (cont.)

Proof.

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X|\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta)\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}f(X|\theta)}{f(X|\theta)}\right) \\
&= \left(f(X|\theta)\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(X|\theta) - \frac{\mathrm{d}}{\mathrm{d}\theta}f(X|\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}f(X|\theta)\right) \Big/ f(X|\theta)^2 \\
&= \frac{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(X|\theta)}{f(X|\theta)} - \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta)\right)^2
\end{aligned}$$

Fisher information (cont.)

Proof.

Note that

$$\mathrm{E}\left[\frac{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f(X|\theta)}{f(X|\theta)}\right] = \int \frac{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f(X|\theta)}{f(X|\theta)} f(X|\theta)\,\mathrm{d}x = \int \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f(X|\theta)\,\mathrm{d}x$$

$$= \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \underbrace{\int f(X|\theta)\,\mathrm{d}x}_{1} = 0$$

Therefore,

$$-\mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X|\theta)\right] = \mathrm{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X|\theta)\right)^2\right] = \mathcal{I}(\theta).$$

□

Fisher information (cont.)

The above discussion was concerning a single random variable X. If we observe instead $\mathbf{X} = (X_1, \ldots, X_n)^\top$, where each X_i are iid, then

$$\mathcal{I}(\theta) = \mathcal{I}_{\mathbf{X}}(\theta) = \sum_{i=1}^{n} \mathcal{I}_{X_i}(\theta) = n\mathcal{I}_{X_1}.$$

That is, the information is additive: the total Fisher information from n iid random variables $X_1, \ldots, X_n$ is simply the sum of the n unit Fisher information.

Fisher information matrix

If instead $\boldsymbol{\theta}$ is p-dimensional, then we also have similar results for the Fisher information, which is now a matrix:

- $\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X|\boldsymbol{\theta}) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(X|\boldsymbol{\theta})\right)^{\top}\right] \in \mathbb{R}^{p \times p}$.
- $\mathcal{I}(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\top}} \log f(X|\boldsymbol{\theta})\right] \in \mathbb{R}^{p \times p}$.
- $\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{I}_{X_i}(\boldsymbol{\theta}) = n\mathcal{I}_{X_1}$.

Cramér-Rao inequality

Theorem 47 (Cramér-Rao inequality)

Let $\mathbf{X} \sim f(\mathbf{x}|\theta)$ satisfying some regularity conditions. Let $T = T(\mathbf{X})$ be a statistic with $g(\theta) = \mathrm{E}_\theta(T)$. Then, for any $\theta \in \Theta$,

$$\mathrm{Var}_\theta(T) \geq \frac{\{g'(\theta)\}^2}{\mathcal{I}(\theta)}.$$

The Cramér-Rao inequality specifies a lower bound for any **unbiased estimator** for the paramter $g(\theta)$. When the equality holds T is called the **minimum variance unbiased estimator (MVUE)** of $g(\theta)$.

Important case

For any unbiased estimator $\hat{\theta} = T(\mathbf{X})$, i.e. $g(\theta) = \theta$ (the identity function),

$$\mathrm{Var}(\hat{\theta}) \geq 1/\mathcal{I}(\theta).$$

Cramér-Rao inequality (cont.)

We can state a similar theorem for the multivariate case:

Theorem 48 (Multivariate Cramér-Rao inequality)

Let $\mathbf{X} \sim f(\mathbf{x}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$ satisfying some regularity conditions. Let $\mathcal{I}(\boldsymbol{\theta}) \in \mathbb{R}^{p \times p}$ be the Fisher information matrix, and let $\mathbf{T} = \big(T_1(\mathbf{X}), \ldots, T_p(\mathbf{X})\big)$ be a statistic with $g(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}(\mathbf{T})$. Then, for any $\theta \in \Theta$,

$$\mathrm{Var}_\theta(\mathbf{T}) \geq \mathcal{J}(\boldsymbol{\theta})\mathcal{I}(\boldsymbol{\theta})^{-1}\mathcal{J}(\boldsymbol{\theta})^\top.$$

where $\mathcal{J}(\boldsymbol{\theta})$ is called the Jacobian matrix whose (i, j)-th element is given by $\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j}$. Note that the matrix inequality $\mathbf{A} \geq \mathbf{B}$ is understood to mean that the matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite.

Cramér-Rao inequality (cont.)

Important case

For any unbiased estimator $\hat{\boldsymbol{\theta}} = \mathbf{T}(\mathbf{X})$,

$$\text{Var}(\mathbf{T}) \geq \mathcal{I}(\theta)^{-1}.$$

It is convenient to compute the inverse of the Fisher information matrix, then one can simply take the reciprocal of the corresponding diagonal element to find a (possibly loose) lower bound.

$$\begin{aligned} \text{Var}(T_k) &= \text{Var}(\mathbf{T})_{kk} \quad \text{the } k\text{th diagonal entry} \\ &\geq \left[\mathcal{I}(\boldsymbol{\theta})^{-1}\right]_{kk} \\ &\geq \left[\mathcal{I}(\boldsymbol{\theta})_{kk}\right]^{-1} \end{aligned}$$

Cramér-Rao inequality (cont.)

Example 49

Let $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(\mu, \sigma^2)$. We consider estimators for μ, treating σ^2 as known. The score function for a single observation is

$$\begin{aligned} S(\mu) &= \frac{\partial}{\partial \mu} \log \left\{ \text{const.} \times e^{-\frac{1}{2\sigma^2}(X_1-\mu)^2} \right\} \\ &= \frac{\partial}{\partial \mu} \left\{ -\frac{1}{2\sigma^2}(X_1-\mu)^2 \right\} \\ &= \frac{X_1-\mu}{\sigma^2} \end{aligned}$$

Also note that $I''(\mu) = S'(\mu) = \sigma^{-2}$, hence $\mathcal{I}_{X_1}(\mu) = \sigma^{-2}$. Therefore,

$$\mathcal{I}(\mu) = \sum_{i=1}^{n} \mathcal{I}_{X_i}(\mu) = n/\sigma^2.$$

Cramér-Rao inequality (cont.)

Example 49

For any unbiased estimator $\hat{\mu}$ for μ, the Cramér-Rao theorem says that

$$\text{Var}(\hat{\mu}) \geq \mathcal{I}(\mu)^{-1} = \sigma^2/n.$$

Previous example has shown that

$$\text{Var}(\bar{X}_n) = \sigma^2/n$$

so therefore $\bar{X}_n$ is the MVUE for μ.

Asymptotic properties of MLE

Let $X_1, \ldots, X_n$ be iid with pdf $f(x|\boldsymbol{\theta})$. Write

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}).$$

Let $\hat{\boldsymbol{\theta}}$ be the MLE which maximises $l(\boldsymbol{\theta})$, and suppose that f fulfils certain regularity conditions.

The MLE satisfies (usually) the following two properties:

1. Consistency; and
2. Asymptotic normality.

Consistency

The MLE is consistent in the sense that as $n \to \infty$,

$$\mathrm{P}\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| > \epsilon\right] \to 0$$

for any $\epsilon > 0$. The operator $\|\cdot\|$ is some norm (absolute value norm, 2-norm, d-norm, et.c) for vectors, analogous to taking absolute values in the unidimensional case.

Consistency requires that an estimator converges to the parameter to be estimated. It is a very mild and modest condition that any reasonable estimator should fulfil. The consistency condition is often used to *rule out bad estimators*.

Asymptotic normality

As $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} \mathrm{N}_p(\mathbf{0}, \mathcal{I}_{X_1}(\boldsymbol{\theta})^{-1})$$

In other words, for large n, it approximately holds that

$$\hat{\boldsymbol{\theta}} \sim \mathrm{N}_p(\boldsymbol{\theta}, \mathcal{I}_{X_1}(\boldsymbol{\theta})^{-1}/n).$$

Therefore, asymptotically the MLE is unbiased and attains the Cramér-Rao bound. Any estimator fulfilling this condition is called **efficient**.

Approximate standard error

An approximate standard error of the jth component of $\hat{\boldsymbol{\theta}}$ (i.e. $\hat{\theta}_j$) is the square root of the (j,j)th element of $\mathcal{I}_{X_1}(\boldsymbol{\theta})^{-1}$ divided by $\sqrt{n}$.

Asymptotic properties of MLE (cont.)

Example 50

The family of Bernoulli distributions has pmf $f(x|p) = p^x(1-p)^{1-x}$, $x \in \{0, 1\}$. Taking logarithms,

$$\log f(x|p) = x \log p + (1-x) \log(1-p),$$

and the first and second derivatives are

$$S(p) = \frac{x}{p} - \frac{1-x}{1-p}$$

$$S'(p) = -\frac{x}{p^2} + \frac{1-x}{(1-p)^2}$$

Asymptotic properties of MLE (cont.)

Example 50

The (unit) Fisher information can then be computed as

$$\begin{aligned}\mathcal{I}(p) = -\,\mathrm{E}\left[S'(p)\right] &= \frac{\mathrm{E}\,X}{p^2} - \frac{1-\mathrm{E}\,X}{(1-p)^2} \\ &= \frac{p}{p^2} - \frac{1-p}{(1-p)^2} \\ &= \frac{1}{p} - \frac{1}{1-p} = \frac{1}{p(1-p)}.\end{aligned}$$

As we know (or can be shown), the MLE for p is $\hat{p} = \bar{X}_n$ when a sample $X_1, \ldots, X_n$ is observed, and the asymptotic normality result states that

$$\sqrt{n}(\hat{p} - p) \xrightarrow{\mathrm{D}} \mathrm{N}(0, p(1-p)).$$

Asymptotic properties of MLE (cont.)

Example 50

Of course, since the MLE is $\hat{p} = \bar{X}_n$ (the sample mean), we can also appeal to the CLT which gives the exact same result:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathrm{D}} \mathrm{N}(0, \sigma^2)$$

as $n \to \infty$, where $\mu = \mathrm{E}\, X = p$ and $\sigma^2 = \mathrm{Var}\, X = p(1-p)$ for the Bernoulli distribution.

Remark

Recall the "normal approximation to the binomial": For large n, $X \sim \mathrm{Bin}(n, p)$ can be approximated by $X \sim \mathrm{N}(np, np(1-p))$. This is exactly the reason why this approximation is used. Note that the Bernoulli and Binomial distributions are related.

Confidence sets

A point estimator is simple to construct and use, but it is not very informative.

- If a different sample is used, the value of the estimate changes.
- The point estimate does not reflect the uncertainty in the estimation.

The **confidence interval** is the most commonly used confidence set. It is more informative than a point estimator.

Confidence sets (cont.)

Example 51

A random sample $X_1, \ldots, X_n$ is drawn from $\mathrm{N}(\mu, 1)$. Then we know that $\sqrt{n}(\bar{X}_n - \mu) \sim \mathrm{N}(0, 1)$ approximately or asymptotically. From this, we can deduce that

$$\mathrm{P}\left(-1.96 \leq \sqrt{n}(\bar{X}_n - \mu) \leq 1.96\right) = 0.95$$

or

$$\mathrm{P}\left(\bar{X}_n - 1.96/\sqrt{n} \leq \mu \leq \bar{X}_n + 1.96/\sqrt{n}\right) = 0.95,$$

so a 95% confidence interval for μ is

$$\left(\bar{X}_n - 1.96/\sqrt{n}, \bar{X}_n + 1.96/\sqrt{n}\right).$$

Suppose $n = 4$, $\bar{X}_n = 2.25$, then a 95% CI is

$$(2.25 - 1.96/2, 2.25 + 1.96/2) = (1.27, 3.23)$$

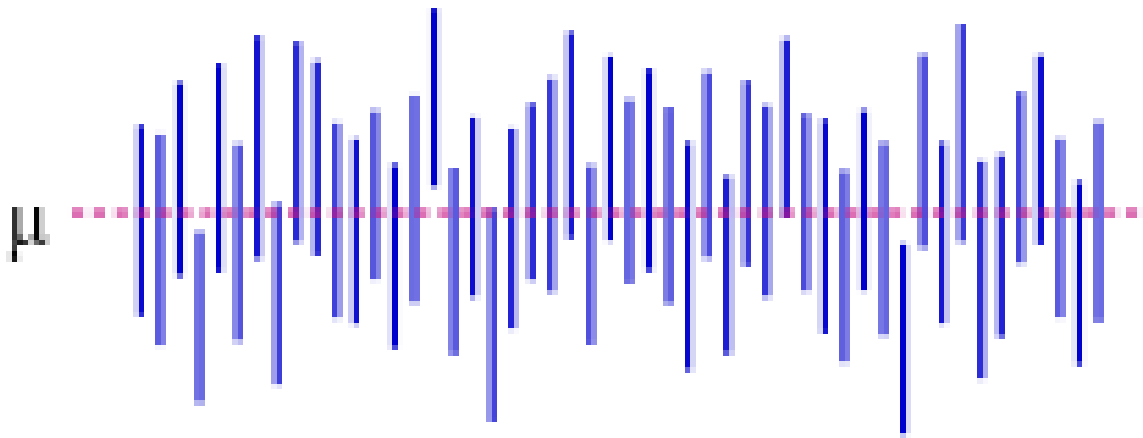
Confidence sets (cont.)

QUESTION: What is P($1.27 < \mu < 3.23$)? Note that μ is an unknown constant!

ANSWER: $(1.27, 3.23)$ is one instance of the **random interval** which covers μ with probability 0.95.

If one draws 10,000 samples with size n each in order to construct 10,000 intervals of the form $\left(\bar{X}_n - 1.96/\sqrt{n}, \bar{X}_n + 1.96/\sqrt{n}\right)$, then about 95% of the intervals, i.e. 9,500 intervals, will cover the true value of μ.

Confidence sets (cont.)



The blue lines represent CI constructed with a replicate of the sample of size n. In this case, hypothetically the true value μ is known.

Confidence interval

Definition 52 (Confidence interval)

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a sample from a pdf with parameter θ. If $L(\mathbf{X})$ and $U(\mathbf{X})$ are two statistics for which

$$P\big(L(\mathbf{X}) < \theta < U(\mathbf{X})\big) = 1 - \alpha,$$

then $\big(L(\mathbf{X}), U(\mathbf{X})\big)$ is called a $100(1-\alpha)\%$ confidence interval for θ.

Remark

$1 - \alpha$ is called the confidence level, which is usually set at 0.90, 0.95 or 0.99. Naturally for given α, we shall search for the interval with the shortest length $U - L$, which gives the most accurate estimation.

Approximate CI based on asymptotic normality

Definition 53 (Approximate CI based on asymptotic normality)

If $(\hat{\theta} - \theta)/\text{SE}(\hat{\theta}) \xrightarrow{\text{D}} \text{N}(0, 1)$, then

$$\left(\hat{\theta} - Z_{\alpha/2}\text{SE}(\hat{\theta}), \hat{\theta} + Z_{\alpha/2}\text{SE}(\hat{\theta})\right)$$

is an approximate $1 - \alpha$ confidence interval for θ, where Z_α is the top-α point of $\text{N}(0, 1)$, i.e. $\text{P}(Z > Z_\alpha) = \alpha$, where $Z \sim \text{N}(0, 1)$.

Sometimes this approximate interval is known as the Wald interval.

Here are some values for $Z_{\alpha/2}$ for various $\alpha = 0.1$, $Z_{\alpha/2} = 1.64$; α values: $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$ (most commonly used); $\alpha = 0.01$, $Z_{\alpha/2} = 2.58$.

Confidence interval (cont.)

Example 54

We saw earlier that for $X_1, \ldots, X_n \sim \text{Bern}(p)$,

$$\sqrt{n}(\hat{p} - p) \xrightarrow{\text{D}} \text{N}(0, p(1-p))$$

as $n \to \infty$, where $\hat{p} = \bar{X}_n$. Now an approximate 95% confidence interval for p is

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}.$$

For example, if $n = 10$ and $\hat{p} = 0.3$, an approximate 95% confidence interval for p is

$$0.3 \pm 1.96\sqrt{0.3(1-0.3)/10} = (0.155, 0.445).$$

Whereas if $n = 100$ and $\hat{p} = 0.3$, an approximate 95% confidence interval for p is $(0.254, 0.346)$.

Confidence interval (cont.)

Remark

The point estimator $\hat{p}$ in the above example did not change with $n = 10$ or $n = 100$ (of course this is just a made-up example, in reality it might change, but for the argument's sake, let's say it did not change). However, the confidence interval is much shorter when $n = 100$, giving much more accurate estimation.

References I

Casella, G. and R. L. Berger (2002). *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury. ISBN: 978-0-534-24312-8.

Pawitan, Y. (2001). *In All Likelihood*. Statistical Modelling and Inference Using Likelihood. Oxford University Press. ISBN: 978-0-19-850765-9.

Jamil, H. (Oct. 2018). "Regression modelling using priors depending on Fisher information covariance kernels (I-priors)". PhD thesis. London School of Economics and Political Science.

Wassermann, L. (2006). *All of Nonparametric Statistics*. New York: Springer-Verlag. ISBN: 978-0-387-25145-5. DOI: 10.1007/0-387-30623-4.