Statistics

Week 1: Probability theory recap

Etienne Wijler

Econometrics and Data Science
Econometrics and Operations Research
Bachelor Program

VU VRIJE UNIVERSITEIT AMSTERDAM
SCHOOL OF BUSINESS AND ECONOMICS

The team

Etienne Wijler: Coordinator of P4+P5. Lecturer of P4.
Address organizational questions to him.

Juan Cai: Lecturer of P5.

Jan Bauer: Tutor of P4+P5.

What is statistics?

Definition (Oxford dictionary)

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

Personal view: statistics is the art and science of extracting actionable information from data.

Fact: Statistics forms the foundation of econometrics and data science.

Why study statistics?

Statistics enables us to:

- validate scientific theories based on empirical evidence,
- make predictions and quantify their uncertainty,
- infer (causal) relationships from the data,
- enable data-driven decision making.

Moreover, recent increases in data availability and computational power make statistics now more relevant than ever before.

Motivating examples

- Goal: A statistician aims to extract meaningful information from the data.
- Challenge: True relationships in the data are often obscured in the observed data.
- Examples:
  - Sales decreased: poor marketing campaign?
  - Temperatures increased: global warming?
  - Profitable portfolio: stock returns predictable?
  - Less patients diseased: new treatment effective?
  - Inflation is increasing: interest increase needed?

Course objectives

Goals: In this course you will learn how to:

1. define a statistical model for a research question,
2. derive general estimators for the parameters in these models (e.g. LS, ML, MM),
3. analyse the behaviour of these general estimators (e.g. unbiasedness and limit distributions),
4. decide on which estimators are optimal for a given model,
5. conduct inference on the estimated parameters.

Course overview

P4: Estimation

Week 1 Probability Recap
Week 2 Statistical Models
Week 3 Data Reduction and MME
Week 4 MLE and Evaluation
Week 5 Estimator Optimality
Week 6 Consistency

P5: Inference

Week 7 Hypothesis testing
Week 8 Mean and Variance testing
Week 9 Finding test statistics
Week 10 Evaluating tests
Week 11 Interval estimation
Week 12 Asymptotic tests

Course overview

P4: Estimation

Week 1 Probability Recap
Week 2 Statistical Models
Week 3 Data Reduction and MME
Week 4 MLE and Evaluation
Week 5 Estimator Optimality
Week 6 Consistency

P5: Inference

Week 7 Hypothesis testing
Week 8 Mean and Variance testing
Week 9 Finding test statistics
Week 10 Evaluating tests
Week 11 Interval estimation
Week 12 Asymptotic tests

Course logistics

Literature

- Course book: Casella & Berger (2nd ed.) Statistical inference, Chapters 5-9
- Lecture slides: posted before each lecture and updated with potential typo corrections
- Lecture notes: convenient summary for exam preparations

Evaluation

- Exams: one intermediate exam after P4 and one final exam after P5.
- Resit: one final resit covering P4 and P5.
- Passing requirements: An average grade of 5.5

Additional learning support

Additional support is available upon your own request.

- **Discussion boards**: ask questions on theory and/or tutorial exercises online
- **Knowledge clips**: e-mail me if your and your peers find a specific topic difficult
- **Practice exercises**: additional exercises to practice at home provided per tutorial
- **Practice exams**: see Canvas for a large collection of old exams

Let's get started

P4: Estimation

Week 1 Probability Recap
Week 2 Statistical Models
Week 3 Data Reduction and MME
Week 4 MLE and Evaluation
Week 5 Estimator Optimality
Week 6 Consistency

P5: Inference

Week 7 Hypothesis testing
Week 8 Mean and Variance testing
Week 9 Finding test statistics
Week 10 Evaluating tests
Week 11 Interval estimation
Week 12 Asymptotic tests

Probability Theory

Statistics is a mathematical and abstract subject, especially when encountered for the first time. Its foundation is build on probability theory, which provides a means of modelling the world around us.

Probability theory simplifies our life (yes, really!) by:

- summarizing complex sources of uncertainty in simple distributions
- providing the tools to compute probabilities of uncertain events happening

Random variables

Definition (Sample space)

The set S of possible outcomes of an experiment is called the sample space for the experiment.

Definition (Events)

An event A is a collection of possible outcomes of the experiment, that is $A \subseteq S$.

Definition (Random variable)

A random variable is a function from a sample space S to the real numbers $\mathbb{R}$.

Random variables: examples

For each of these experiments, define the sample space S, the events and the random variable.

- **Betting**: Tossing a coin two times and betting on the number of heads.
- **Marketing analysis**: Asking three people their preference between a sugar-free and sugary soft-drink to measure consumer preference.
- **Medical study**: Provide a new painkiller to 4 migraine patients and measure the reduction in headaches from 1 (no reduction) to 10 (headaches are gone).

Probability structures

Note: A probability structure is obtained once we define a probability measure $\mathbb{P}$ on "the subsets" of S.

Example (Coin tossing)

The sample space is given by $S = \{\{h, h\}, \{h, t\}, \{t, h\}, \{t, t\}\}$. Then, the probability measure of a fair coin is defined by $\mathbb{P}(A) = \frac{\#A}{\#S}$ for all $A \subseteq S$, where $\#X$ denote the number of elements in the set X.

Probability measure induced by random variables

Note: The sample space S can sometimes be very abstract and complex.

However: a probability measure defined on S can be easily extended to $\mathbb{R}$.

Example

Given a sample space S, probability measure $\mathbb{P}$ and random variable X, we can compute the probability of each event on the real line $A \subseteq \mathbb{R}$, as

$$\mathbb{P}(X \in A) = \mathbb{P}_X(A) = \mathbb{P}(\{s \in S \mid X(s) \in A\}).$$

In practice, we directly define the probabilities on $\mathbb{R}$. For example, $X \sim N(0, 1)$ means that the distribution of outcomes on the real line follows a normal distribution.

Distribution functions

Definition (cdf)

The cumulative distribution function (cdf) of a random variable X is given by $F(x) = \mathbb{P}(X \leq x)$ for all $x \in \mathbb{R}$.

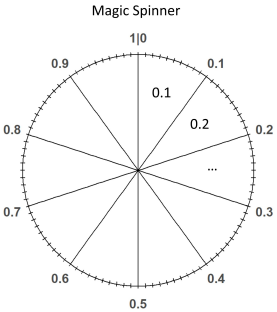
Definition (pmf)

The probability mass function (pmf) of a discrete random variable X is given by $f(x) = \mathbb{P}(X = x)$.

Definition (pdf)

The probability density function (pdf) of a continuous random variable X is given by $f(x) = \frac{d}{dx}F(x)$.

Distribution functions: review



Let X denote the outcome of one spin on the left spinner.

- What is $f(0.1)$? How about $F_X(0.25)$?
- Imagine that this spinner becomes infinitely precise (the pie slices become infinitely thin). What is $f(0.1)$? How about $F(0.25)$?

Transformations

Transformations: Let X be a random variable taking values in $\mathcal{X}$ and $Y = g(X)$ for a "measurable" function g, then $Y = g(X)$ is also a random variable.

Monotonicity: If g is a monotonic function, then

$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \begin{cases} \mathbb{P}(X \leq g^{-1}(y)) & \text{if} \quad \frac{d}{dx} g(x) \geq 0, \\ \mathbb{P}(X \geq g^{-1}(y)) & \text{if} \quad \frac{d}{dx} g(x) \leq 0. \end{cases}$$

Example

Suppose $X \sim \text{Uniform}(0, 1)$ and let $Y = -\log(X)$. Blackboard.

Order statistics

Definition (Order statistics)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)'$ be an iid vector and order $X_1, \ldots, X_n$ from small to big. Then, the k-th element in the new ordering is called the k-th order statistic, denoted $X_{(k)}$. Special order statistics are $X_{(1)} = \min(X_1, \ldots, X_n)$ and $X_{(n)} = \max(X_1, \ldots, X_n)$.

Example

Suppose $X_1, \ldots, X_n \sim \text{Exponential}(\lambda)$ with pdf $f(x) = \lambda e^{-\lambda x}$. We derive the distribution of $X_{(1)}$. Blackboard.

Note: A financial application of this example would relate to modelling the *shortest time until default* on extended loans.

Moments

Definition (Expected value)

The expected value or mean of a random variable $g(X)$ is given by

$$\mathbb{E}(g(X)) = \begin{cases} \sum_{x \in \mathcal{X}} g(x)\mathbb{P}(X = x) & \text{if } X \text{ is discrete,} \\ \int_{x \in \mathcal{X}} g(x) f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

If $\mathbb{E}\,|g(X)| = \infty$, we say that $\mathbb{E}(g(X))$ does not exist.

Example (Normal distribution)

Suppose $X \sim \text{Normal}(\mu, \sigma^2)$. Show that $\mathbb{E}X = \mu$. Blackboard.

Example (Cauchy distribution)

Let $Y \sim \text{Cauchy}(0, 1)$. Show that $\mathbb{E}\,|X| = \infty$. Blackboard.

Other important expectations

Definition (2.3.2)

Let X be a random variable. The variance is given by

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

Interpretation: The variance is the "expected squared deviation from the mean" of a random variable. Simply put: it measures the spread of the distribution.

Definition (2.3.1)

Let $k \in \mathbb{N}$. We define the k-th moment of X as $\mathbb{E}X^k$.

Note: the more moments exist, the faster the tails of the distribution decay to zero, i.e. "less chance of extreme values".

Important properties of expectation and variance

Lemma

Let X and Y be random variables and let $a, b \in \mathbb{R}$ be real numbers. Then,

- $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.
- $\mathbb{V}ar(aX + bY) = a^2\mathbb{V}ar(X) + b^2\mathbb{V}ar(Y) + 2ab\mathbb{C}ov(X, Y)$.

Expectation: is a linear operator.

Variance: is almost a linear operator when the random variables are uncorrelated.

Moment Generating Function (MGF)

Definition (2.3.6, Moment Generating Function (MGF))

The MGF of X, denoted by $M_X(t)$, is defined as

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Distributional equality: The MGF is often used to establish that (the limits of) two distribution are equal.

Moments: We will use the MFG as a quick method to compute moments!

Theorem (2.3.7)

If X has MGF $M_X(t)$, then

$$\mathbb{E}X^n = \frac{d^n}{dt^n} M_X(t)|_{t=0}.$$

MGF Example: Normal distribution

Example (Normal Distribution)

Let $X \sim \text{Normal}(\mu, \sigma^2)$. The MGF of X is given by

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}.$$

Then,

$$\mathbb{E}X = \frac{d}{dt} M_X(t)|_{t=0} = (\mu + \sigma^2 t) e^{\mu t + \sigma^2 t^2/2}|_{t=0} = \mu.$$

Note: this derivation of $\mathbb{E}(X)$ is substantially shorter than integrating the pdf!

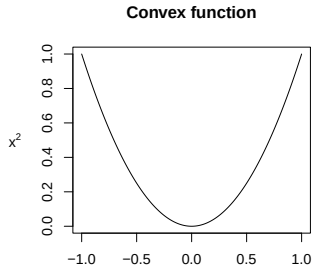
Jensen's inequality

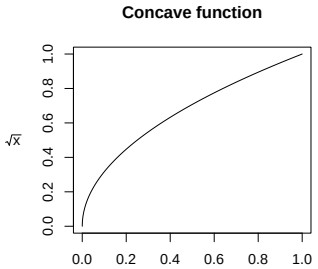
Definition (Convex and concave)

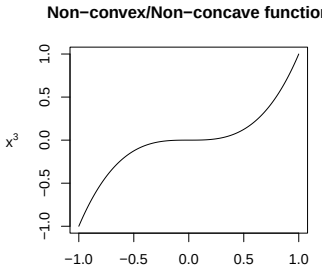
A function $g(x)$ is convex if it holds that

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$$

for all x, y and $0 < \lambda < 1$. The function $g(x)$ is concave if $-g(x)$ is convex.







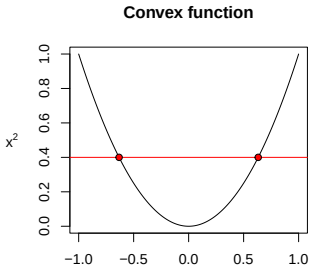
Convexity and concavity

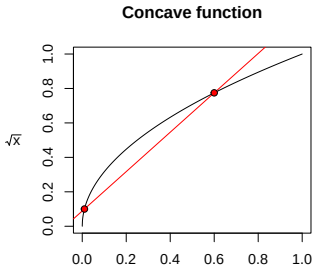
Definition (Convex and concave)

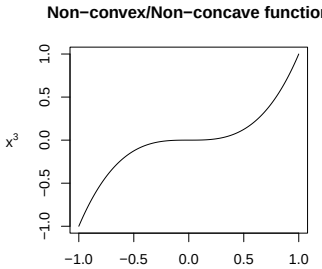
A function $g(x)$ is convex if it holds that

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$$

for all x, y and $0 < \lambda < 1$. The function $g(x)$ is concave if $-g(x)$ is convex.







Jensen's inequality

Theorem (4.7.7, Jensen's inequality)

For any random variable X, it holds that

- $\mathbb{E}(g(X)) \leq g(\mathbb{E}(X))$ *if g is convex*
- $\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$ *if g is concave*

Reminder: $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \geq 0 \Rightarrow \mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

Applications: Jensen's inequality will become super useful later to show that some estimators are biased, but for now have a look at this:

Lemma (Existence of moments)

If $\mathbb{E}\,|X|^k < \infty$, then $\mathbb{E}\,|X|^m < \infty$ for all $k, m \in \mathbb{N}$ with $m < k$. Blackboard proof.

Multiple random variables

Definition (4.1.1)

An n-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a vector where each element is a random variable.

Note: If X and Y are both random variables, the random vector (X, Y) has a joint pdf (pmf) denoted by $f(x, y)$ ($f(x, y) = \mathbb{P}(X = x, Y = y)$).

Theorem (4.1.6)

Let X and Y be two random variables and let $f(x, y)$ denote their joint density. Then the univariate densities can be obtained via

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

Conditional probabilities

Definition (4.2.3, conditional pdf)

Let X and Y be two random variables with joint density $f(x, y)$. The conditional pdf of Y given that $X = x$ is the function of y defined by

$$f(y \mid x) = f(x, y)/f_X(x).$$

Conditional expectation: $\mathbb{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f(y \mid x) dy$

Example

Suppose X and Y have joint density

$$f(x, y) = e^{-y} \qquad \text{if } 0 < x < y < \infty.$$

Derive the density for X and $Y \mid X$, and $\mathbb{E}(Y \mid X = x)$.

The Law of Iterated Expectations

Theorem (4.4.3 & 4.4.7)

Let X and Y be random variables. Then

- $\mathbb{E}X = \mathbb{E}(\mathbb{E}(X \mid Y))$.
- $\mathbb{V}ar(X) = \mathbb{E}(\mathbb{V}ar(X \mid Y)) + \mathbb{V}ar(\mathbb{E}(X \mid Y))$.

Useful when dealing with random variables of the form $Z = g(X)h(Y)$. Why?

Example (Light bulb lifetime)

Let the lifetime of a light bulb from your local DIY store be given by $X \sim \text{Exponential}(1 + Y)$, where $Y \sim \text{Binomial}(2, p)$. How long do you expect this light bulb to work? Blackboard.

Independence

In this course, we will exclusively be working with independent random variables.

Definition (4.2.5)

Let X and Y be random variables with joint pdf $f(x, y)$. Then X and Y are called independent if $f(x, y) = f_X(x)f_Y(y)$.

Extension: For a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ containing independent random variables with marginal distributions g_i, we have

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} g_i(x_i).$$

Implications of independence

Lemma (1.28)

Let X and Y be two independent random variables. Then

- $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.
- $\mathbb{V}ar(X+Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y)$.

Note: This even holds for any "measurable" functions of X and Y, i.e. $\mathbb{E}(g(X)h(y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$ if $X \perp Y$.

Example

Let $C \sim \text{Gamma}(\alpha, \beta)$ denote your monthly energy consumption and $P \sim \text{Normal}(\mu, \sigma^2)$ the monthly energy price. Your monthly energy bill is then $B = CP$. What is your expected energy bill? Blackboard.

Almost sure convergence

Intuitively, the more data we use to estimate some characteristic of a population, the better our estimate.

However, formally proving this requires a notion of convergence for sequences of random variables (typically estimators based on increasing sample sizes).

Definition (5.5.6)

An infinite sequence of random variables $X_1, X_2, \ldots$ is said to converge almost surely (a.s.) to $\tilde{X}$ if for all $\epsilon > 0$ we have

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - \tilde{X}| < \epsilon\right) = 1.$$

Convergence in probability

Note: A slightly weaker form of convergence, which is often easier to establish, is the following:

Definition (5.5.1)

An infinite sequence of random variables $X_1, X_2, \ldots$ is said to converge in probability to $\tilde{X}$ if for all $\epsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}\left(|X_n - \tilde{X}| < \epsilon \right) = 1.$$

Hierarchy: If X_n converges a.s. to $\tilde{X}$ as $n \to \infty$, then it also converges in probability.

The Law of Large Numbers

The sample average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_n$ plays an essential role in statistics, as it:

- measures the center of a distribution, and
- reduces noise in measurements.

The following fundamental theorem proves that the sample average is a good estimator of the expected outcome in very general settings!

Theorem (5.5.9, Law of Large Numbers)

Suppose $X_1, X_2, \ldots$ is a sequence of iid random variables with $\mathbb{E}|X_1| < \infty$. Then, almost surely,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}(X_1).$$

A little bit of history

The LLN is a remarkable piece of mathematics with a rich history:

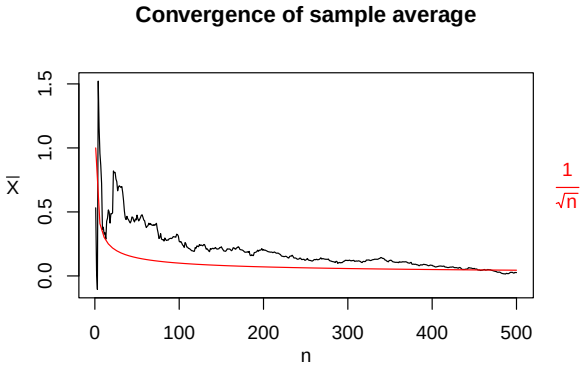
- Cardano (1500s), famous mathematician and gambler, stated the LLN without proof.
- Bernoulli (1713) provided a first proof for a special case of the LLN (for binary variables) which he proudly names The Golden Theorem. This took him 20 years to derive!
- Poisson (1873) generalized the theorem (slightly) and called it "la loi des grand nombre"
- Chebyshev (1874) derived a simpler proof using an unproven inequality.
- Markov (1884) completed the proof.

How close are we?

Note: The LLN tells us that $\bar{X}$ will get closer and closer to $\mathbb{E}(X_1)$ when n increases.

Question: For our sample (of size n), how close is $\bar{x}$ to $\mathbb{E}(X_1)$?

The LLN does not give us the answer! But look at this:



- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} t$ with 3 df.
- Clearly $\bar{X} \xrightarrow{p} \mathbb{E}(X_1) = 0$ as $n \to \infty$.
- However, $1/\sqrt{n} \to 0$ as $n \to \infty$ equally fast.
- Idea: $\sqrt{n}\bar{X}$ should not converge or diverge.

Central Limit Theorem

Solution: The answer to our question is provided by the Central Limit Theorem.

Theorem (5.5.16, Central Limit Theorem)

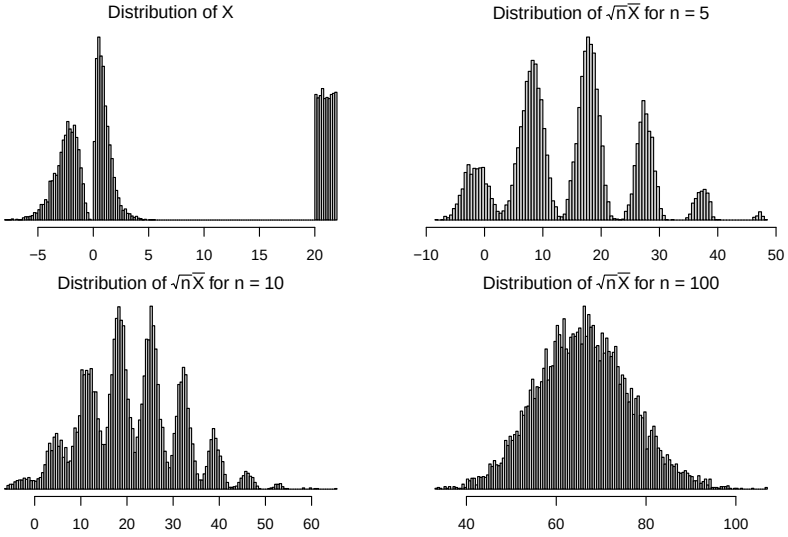
Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathbb{E}(X_1) = \mu < \infty$ and $\mathbb{V}ar(X_1) = \sigma^2 < \infty$. Then

$$\sqrt{n}(\overline{X}_n - \mu)/\sigma \xrightarrow{d} Normal(0, 1).$$

Note: This theorem enables us to quantify the uncertainty of our estimates in very general settings!

Important: The CLT holds for any distribution, as long as the first two moments exist!

Visualization of CLT



Hyping up the CLT

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along."

- Sir Francis Galton (1889)

Continuous mapping theorem (CMT)

Problem: Often, we are interested in sequences of transformations of random variables.

Solution: The following theorem shows that we don't need to do anything extra if the transformation is continuous!

Theorem (5.5.4, continuous mapping theorem)

Let $X_1, X_2, \ldots$ be a sequence of random variables and let h be a continuous real function. Then

- $X_n \xrightarrow{as} \tilde{X}$ *implies that* $h(X_n) \xrightarrow{as} h(\tilde{X})$.
- $X_n \xrightarrow{p} \tilde{X}$ *implies that* $h(X_n) \xrightarrow{p} h(\tilde{X})$.
- $X_n \xrightarrow{d} \tilde{X}$ *implies that* $h(X_n) \xrightarrow{d} h(\tilde{X})$.

Application of the CMT

Note: The CMT has tons of applications later in the course, as many estimators are functions of sample averages.

Example

Let $X_1, \ldots, X_N \sim \text{Normal}(\mu, \sigma^2)$. Consider estimating the variance σ^2 by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. Then, we can show that

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \\ &\xrightarrow{p} \mathbb{E}(X_1^2) - (\mathbb{E}X_1)^2 = \sigma^2,\end{aligned}$$

where the convergence follows from the LLN and CMT (why?).