Mathematical statistics

Winter 2024/2025

Thomas Nagler

Version: January 8, 2025

Contents

1 Introduction and overview

1.1 Mathematical statistics

This course is about a deep understanding of statistical methods. You have already encountered many such methods in your studies: data summaries (like the mean, variance, or quantiles), maximum-likelihood estimators, hypothesis tests, histograms, regression estimators, etc. But do they even achieve what they intend, and why? Are they biased? How accurate/certain are the results?

Let's make this more concrete. You might have already heard that the median of a sample $X_1, \dots, X_n$ can be estimated by

$$\widehat{m} = \arg\min \sum_{i=1}^{n} |X_i - m|.$$

But why is this a good estimator? Would it at least approach the true median with infinite data? If so, how certain can we be about the estimate? What if observations aren't independent or contaminated by outliers? To a certain extent, these questions can be answered empirically. We may simulate some data, compute the median as above, and see how it behaves. However, this only answers the question for the specific model from which we generated the data and how we estimated the median. What if we change the model or the estimator? We can generate more data and repeat the experiment, but this is not feasible for every possible model and estimator. In particular, we have not gained any understanding of why the estimator behaves the way it does.

Mathematical statistics is all about understanding the properties of statistical methods and deriving sound answers to the questions above in the form of mathematical guarantees.

1.2 Aims and scope of this course

This course provides an introduction to modern mathematical statistics, emphasizing the principles and tools essential for a profound understanding of statistical methods. The course is intended for students who enjoy mathematics and prepares for research-level work in methodological statistics. On a high level, this course will teach you:

- Fundamental principles of what makes a statistical method work or fail.
- Important results and concepts in mathematical statistics.
- The mathematical tools to derive them.

Specific statistical methods are discussed to illustrate the concepts and tools; most of them should already be known from previous courses. While mathematical theory can be fun and interesting in itself, this course has a very practical intent.

> *"There is nothing more practical than a good theory." — Kurt Lewin*

Understanding the fundamental mechanisms governing the behavior of statistical methods is an invaluable asset in practice. It helps to choose the right method for a given problem, understand its limitations, and develop new methods when needed. The tools and concepts you learn in this course provide a general framework for analyzing statistical methods. This is extremely helpful for developing intuition and navigating new challenges in statistical practice.

For those considering an academic career, I feel the urge to speak about another (possibly unpleasant) truth. If you browse through papers in the most prestigious statistics journals[1], you will find a whole lot of mathematics, theorems, and proofs. A thorough mathematical analysis of any statistical procedure you propose is almost mandatory for publication in these venues. That does absolutely not mean that one cannot be a great statistician or have a lasting impact on the field without it. In fact, quite a few researchers are annoyed by the strong emphasis on mathematics in methodological statistics. But this emphasis grew for a reason. Essentially all methods and recommendations we teach today are based on deep mathematical understanding and results. This theory gives reassurance to a level that empirical studies cannot provide. Convincing others of the validity of a new method or recommendation is much easier if you can back it up with a solid mathematical argument.

1.3 What's expected from you

The course is designed under the assumption that students

- already had a first course in probability theory and statistical inference,
- have a solid grasp of basic concepts in mathematical analysis (e.g., convergence, continuity, differentiability).

It is OK if things are a bit rusty, you can catch up on the way. However, the course will be quite challenging if you are not comfortable with these prerequisites. The mathematical arguments in the lectures are quite advanced at times. In general, students are not expected to develop or reproduce such advanced arguments in the exam or exercise. Here, focus will be put on applying the main results to specific statistical procedures and simpler derivations using the key concepts.

1.4 These notes

These lecture notes are meant to support the oral lectures during class. The notes are more detailed than what is done in class. I will sometimes omit or shorten mathematical

[1] See, for example, the Google Scholar ranking.

arguments and proofs if additional detail doesn't add much to our understanding. Sections marked with a star are considered optional and probably not covered in class. Reading the additional details in the notes is an optional offer for interested students. The general rule is: what we don't discuss in class isn't necessary for succeeding in this course.

1.5 Outlook

1.5.1 Statistical methods

Generally speaking, a statistical method takes some observations $X_1, \dots, X_n$ from a distribution P and computes some output $T_n(X_1, \dots, X_n)$. The output can be a number, a vector, a function, etc. The goal is to understand the properties of the output, depending on the properties of the data. Statistical procedures can have many different goals and solve many different problems. Here are some simple examples:

Example 1.5.1 (Sample mean)**.** *The sample mean* $T_n(X_1, \dots, X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *is a common estimator for the population mean* $\mu = \mathbb{E}[X]$*. The mean itself is a measure of* location*. An alternative measure of location is the median mentioned above.*

Example 1.5.2 (Sample quantiles)**.** *The sample quantile* $T_n(X_1, \dots, X_n) = \widehat{Q}(\alpha)$*, defined as the* $\lceil n\alpha \rceil$*-smallest observation from the sample* $X_1, \dots, X_n$*. Quantiles for large or small* α *are often used to quantify* risks*, i.e., unlikely events with negative consequences.*

Example 1.5.3 (Statistical tests)**.** *A statistical test computes a test statistic* $T_n = T_n(X_1, \dots, X_n)$ *to make a yes-or-no decision about a hypothesis. For example, the t-test tests the hypothesis that the mean of a sample is equal to a given value.*

Example 1.5.4 (Maximum-likelihood estimator)**.** *The maximum-likelihood estimator (MLE) is a method that fits a parametric model to data. A parametric model is a collection of densities* $\{f_\theta \colon \theta \in \Theta\}$ *indexed by a parameter* θ*. The maximum-likelihood estimator gives the parameter under which the observations are most likely:*

$$T_n(X_1, \dots, X_n) = \widehat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \sum_{i=1}^n \ln f_\theta(X_i).$$

After finding the MLE, the fitted model can make predictions or other inferences about the world.

Example 1.5.5 (Empirical distribution function)**.** *The empirical cumulative distribution function (ECDF)* F_n *is a statistical approximation of the distribution* F_X *of a random*

variable X. It is defined as

$$T_n(X_1, \dots, X_n) = \widehat{F}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X \leq \cdot\}.$$

Note that the ECDF is a function, not a number or vector.

Example 1.5.6 (Histogram)**.** *The histogram is a statistical approximation of the density f_X of a random variable. It is defined as*

$$T_n(X_1, \dots, X_n) = \widehat{f}_X(\cdot) = \sum_{b=1}^{B} \mathbb{1}_{[x_{b-1}, x_b)}(\cdot) \frac{\sum_{i=1}^{n} \mathbb{1}_{[x_{b-1}, x_b)}(X_i)}{n},$$

where $x_0, \dots, x_B$ are fixed boundaries of the bins.

Example 1.5.7 (Bootstrap)**.** *The bootstrap is a procedure gauging the uncertainty of an estimator $\widehat{\theta}(X_1, \dots, X_n)$. It works by resampling the data many times and computing the estimator on each resample. For each $b = 1, \dots, B$, let $X_1^{(b)}, \dots, X_n^b$ be a sample from the empirical distribution of $X_1, \dots, X_n$. The bootstrap estimator of the sampling variance is*

$$T_n(X_1, \dots, X_n; \xi) = \widehat{\sigma}_n^2 = \frac{1}{B} \sum_{b=1}^{B} (\widehat{\theta}(X_1^{(b)}, \dots, X_n^{(b)}) - \widehat{\theta}(X_1, \dots, X_n))^2,$$

where ξ is some external source of randomness generating the bootstrap samples. This estimator is often used to quantify uncertainty through confidence intervals constructed from a normal approximation of $\widehat{\theta}$. Other bootstrap confidence intervals based on quantiles are also possible.

Despite their apparent differences, the methods above can be understood and analyzed from a core set of principles. The methods are so common that we all know they are reasonable for the problems they try to solve. But do you know why? What if I give you a new method you haven't seen before? How would you know if it is a good method? And for the methods you know, are they still reasonable when the data is heavy-tailed or non-*iid*? All these questions can (and have been) answered theoretically with mathematical statistics.

1.5.2 The role of asymptotics

The *statistics* $T_n(X_1, \dots, X_n)$ are random quantities because they are functions of the random sample $X_1, \dots, X_n$. The probabilistic behavior of $T_n(X_1, \dots, X_n)$ usually depends on the distribution of the data in a complicated way. It is often infeasible to derive exact results about the distribution of $T_n(X_1, \dots, X_n)$, especially for complex statistics or distributions.

A powerful technique to cope with this is *asymptotic analysis*. It refers to the study

of the behavior of statistical methods as the sample size n grows to infinity. The idea is that many statistical methods become more predictable in the limit of large n. The influence of the data distribution becomes simpler, and the method's behavior can be understood in an insightful way.

Consistency

For example, the sample mean $\overline{X}_n$ converges to the population mean μ as $n \to \infty$ under mild conditions. Here, the limit does not depend at all on the data distribution, so many of the complications of a finite-sample analysis disappear. Convergence of a statistical method T_n to the quantity $T = T(P)$ it tries to estimate is known as *consistency*. It is, in essence, a minimal requirement for a good statistical method. If a method does not converge to the right quantity, it is probably not a good one. Consistency is a very rough property, though. It only tells us that the method is not entirely off. It does not tell us how fast the method converges or how certain we can be about the estimate.

Asymptotic normality

Asymptotic normality results are more nuanced. They take the form

$$\mathbb{P}\left(r_n(T_n - B_n - T)/\sigma \leq x\right) \overset{n\to\infty}{\to} \Phi(x),$$

where Φ is the standard normal distribution function, r_n is some convergence rate (often $r_n = \sqrt{n}$), B_n is the *asymptotic bias*, and σ is the *asymptotic variance*.

Asymptotic normality tells us how fast the method converges to the target and how much the estimates fluctuate around it. These more refined properties give us a more detailed understanding of the method. In particular, B_n and σ usually depend on the distribution of the data, but in a tractable and insightful way. For example, the bias of the histogram can be shown to be large when the bins are large or the density is erratic, and the variance is large when the bins are small or the true density is small (= few observations in a bin). This allows us to select an appropriate number of bins and assess how certain we can be about the estimated density values.

Limitations

Asymptotic results are powerful, but they have limitations. They only tell us about the method's behavior in the limit of large n. At face value, they do not reveal anything about the behavior of the method for finite n. In fact, some methods can be proven to be optimal asymptotically but produce complete garbage on small or moderate samples. Luckily, asymptotic approximations are often very good, even for moderate n. Further, the asymptotic bias and variance almost always reveal some fundamental relations between the method and the data-generating distribution.

Not all of mathematical statistics is about asymptotics. There are many results that hold for finite n, but they usually take the form of probabilistic bounds. Such results are often more complicated and loose than asymptotic results. Asymptotic analysis

has proven to be the most effective and insightful tool for understanding statistical methods, so it is the main focus of this course. Nevertheless, it is important to be aware of its limitations and, generally, a good idea to do basic checks on the asymptotic results via finite-sample simulations.

1.5.3 Topics

The following gives an overview of the topics we will likely cover in this course. We may skip some topics or add others depending on time and interest.

Stochastic convergence

Asymptotic results are usually based on the concept of convergence of sample averages. We will start with a basic study of this phenomenon, particularly the law of large numbers and central limit theorem. There is a joke about statisticians taking averages all day; in a sense, this is true. Many estimators can be expressed as averages; and even when they can't, they can often be approximated by a suitable average. This property is called *asymptotic linearity* and is key to asymptotic statistics.

M-estimators

Most statistical methods can be written as so-called *M-estimators*. The M in M-estimator stands for either maximum or minimum. Important examples are the maximum-likelihood estimator or empirical risk minimizers. We will study the general properties of M-estimators, such as consistency, asymptotic normality, and efficiency.

Robust statistics

Robust statistics deals with statistical methods that aren't affected too much by outliers. For example, the median is a robust estimator of location because it remains reliable even if a significant portion of the data is contaminated. We will discuss the basic principles of robustness, such as breakdown point and influence function, and introduce robust estimators that provide reliable results even in the presence of outliers or model misspecification.

Plug-in estimators

Plug-in estimators involve estimating a functional of the distribution by plugging in an estimator of the distribution itself. This approach is commonly used in non-parametric statistics and machine learning for tasks like density estimation and regression function estimation. For example, to estimate the entropy of a distribution, one could use the empirical distribution function to plug into the entropy formula, yielding a plug-in estimator of the entropy. More generally, almost all statistical procedures can be cast as *functionals* of the empirical distribution function. We develop some general concepts and theoretical results for analyzing estimators through this lens

U-statistics

U-statistics are generalizations of sample averages. Instead of averaging over (functions of) individual observations, U-statistics average over k-tuples of observations. For example, a second-order U-statistic averages over functions of pairs of observations:

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j),$$

for some function h. Simple examples are the sample variance, the Gini coefficient, or the Wilcoxon test statistics. Further, U-statistics often arise naturally when analyzing more complex statistical methods, where one estimate is substituted to construct another. Under mild conditions, U-statistics are asymptotically linear (i.e., they can be approximated by a sample average), for which the usual laws of statistics apply.

Dependent data

Many real-world data sets involve dependent observations, such as time series or spatial data. For example, today's air temperature is correlated with yesterday's temperature, and the temperature in Munich is correlated with the temperature in Augsburg. This conflicts with the assumption of *iid* data commonly made in limit theorems. We will see that meaningful asymptotics are still possible if the dependence fades out as two points in time or space are sufficiently far apart. In particular, we discuss the concept of mixing and its implications for statistical inference.

1.6 Further reading

The contents, results, and proofs in these notes are selected and presented in a way that reflects my own taste and understanding. The notes are supposed to be self-contained, and no further reading is required. They build on several other textbooks and lecture notes. One of the main references is the book "Asymptotic statistics" by Van der Vaart (2000), which covers many additional topics.

2 Stochastic convergence

We start the course with the basics. Asymptotic analysis is concerned with the convergence of statistical methods as the sample size grows. Because the data is random, the convergence is not deterministic but stochastic. We therefore need to develop adequate notions of stochastic convergence and some key results that we put in our toolbox for studying statistical methods. We'll also use this chapter to ease into the mathematical notation and style that we will use throughout the course. The mathematical level in the beginning is rather elementary and should feel comfortable. It builds up as we progress.

A bit later in the course, we will study the behavior of fairly advanced statistical methods. To motivate and build up the tools without too much distraction, we mostly focus on the simplest of all examples: sample averages. This may sound a bit dull, but it really isn't. Let's first get a sense for why sample averages are so important.

2.1 Sample averages

Let $Y_1, \dots, Y_n \sim P$ be a sequence of *iid* random variables. The sample average

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

is the most common and natural estimator of the mean $\mathbb{E}[Y_1]$. As already mentioned in the introduction, most estimators can at least be approximated by sample averages. Here, the actual data is a sequence of random variables $X_1, \dots, X_n \sim P$ and the estimator T_n is such that

$$T_n(X_1, \dots, X_n) \approx \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

for some function g and $Y_i = g(X_i)$. We will see this many times later in the course. But even without such approximations, sample means are already extremely powerful. They often come in the disguised form above (but with equality):

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} g(X_i). \tag{2.1}$$

Here are some common examples.

Example 2.1.1 (Moments)**.** *For $g(x) = x^k$, $k \in \mathbb{N}$, (2.1) is an estimator for the k-th moment of X. From this, we can, for example derive estimators for the variance.*

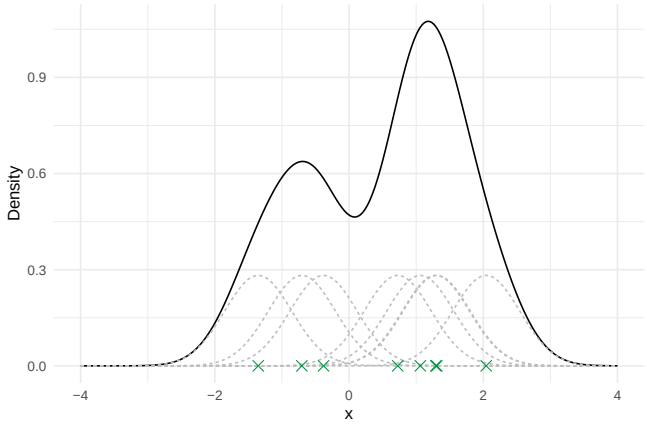

Figure 2.1: The kernel density estimator in action (Example 2.1.5). Eight data points are shown as cross on the x axis. The KDE places a bump $h^{-1}K((x - X_i)/h)/n$ on each data point (dashed lines). The final estimate is the sum of all bumps (solid line).

Example 2.1.2 (Tail probabilities)**.** *Now suppose X_i is an observed loss in an insurance portfolio. We want to estimate the probability of a large loss, e.g., one exceeding some risk budget r. This probability can be estimated by*

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > r\},$$

which corresponds to (2.1) *with $g(x) = \mathbb{1}\{x > r\}$.*

Example 2.1.3 (Empirical distribution function)**.** *The empirical distribution function is of the form* (2.1) *with $g(x) = \mathbb{1}\{X_i \leq x\}$.*

Example 2.1.4 (Histogram)**.** *The histogram estimator in Example 1.5.6 is of this form. To estimate the density at a fixed point $x \in B = [x_{b-1}, x_b)$, we may take $g(x) = \mathbb{1}\{x \in B\}$,*

Example 2.1.5 (Kernel density estimator, KDE)**.** *Kernel density estimators are a slightly more advanced technique to estimate densities. They are of the form* (2.1) *with $g(x) = h^{-1}K((x - X_i)/h)$, where K is a probability density function symmetric around zero, and h a bandwidth parameter. The kernel density estimator also counts points in a neighborhood of x, but weights them according to the kernel. Points far away from x count less, points close to x count more. The KDE is illustrated in Fig. 2.1.*

Example 2.1.6 (Monte-Carlo integration)**.** *The Monte-Carlo method is one of the most easy to use and powerful methods for numerically integrating some function $g\colon \mathbb{R}^d \to \mathbb{R}$.*

Let S be a bounded subset of $\mathbb{R}^d$ and $U_1, \dots, U_d$ iid *samples from* $\mathrm{Unif}(S)$*, the uniform distribution on S. Then*

$$\int_S g(x)\,\mathrm{d}x = \mathbb{E}[g(U_1)],$$

which can be estimated by the sample average

$$\frac{1}{n}\sum_{i=1}^{n} g(U_i).$$

The same idea also works for infinite integration domains. For example, let ϕ be the multivariate standard normal density and $X_1, \dots, X_n$ be an iid *sample from it. Then*

$$\int_{\mathbb{R}^d} g(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \frac{g(x)}{\phi(x)}\phi(x)\,\mathrm{d}x = \mathbb{E}\left[\frac{g(X_1)}{\phi(X_1)}\right],$$

can be estimated by the sample average

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{\phi(X_i)}.$$

Other densities can be used in place of ϕ to give more weight to regions where g is large to make the estimator more efficient.

Example 2.1.7 (Simulation studies)**.** *Simulation studies are a useful tool for developing and assessing statistical methods. In a simulation study, we generate many sets of* iid *samples $\{X_1^{(m)} \dots, X_n^{(m)}\}$, $m = 1, \dots, M$ from a known distribution P and apply the method we want to study to each sample. This gives us a sense of how the method behaves in practice. For example, to study the bias $\mathbb{E}[\widehat{\theta}] - \theta$ of an estimator $\widehat{\theta} = T(X_1, \dots, X_n)$, we can compute*

$$\widehat{\theta}^{(m)} = T(X_1^{(m)}, \dots, X_n^{(m)})$$

for each sample $m = 1, \dots, M$, and compare their average

$$\frac{1}{M}\sum_{m=1}^{M} \widehat{\theta}^{(m)}$$

to the true value θ. The estimator's mean-squared error $\mathrm{MSE} = \mathbb{E}[(\widehat{\theta} - \theta)^2]$ *can be estimated by*

$$\frac{1}{M}\sum_{m=1}^{M} (\widehat{\theta}^{(m)} - \theta)^2.$$

I hope you're convinced by now that studying sample averages more deeply is worthwhile.

2.2 Convergence in probability and consistency

2.2.1 Convergence in probability

In many of the above examples, it was taken for granted that the sample average $\overline{Y}_n$ is a good estimator of the mean $\mathbb{E}[Y_1]$. But what do we mean by that? Well, we know that the sample average $\overline{Y}_n$ approaches $\mathbb{E}[Y_1]$ somehow as we gather enough data. There are several ways to express this more precisely. For statistical applications, the most relevant one is convergence in probability.

Definition 2.2.1 (Convergence in probability)**.** *Let $Y, Y_1, Y_2, \ldots \in \mathbb{R}^d$ be random vectors. We say that Y_n converges to Y in probability or $Y_n \to_p Y$, if for every $\epsilon > 0$,*

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

In plain words, $Y_n \to_p Y$ means: as $n \to \infty$, the probability that Y_n is ϵ away from Y goes to 0. You could also write the definition the other way around: for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\|Y_n - Y\| \leq \epsilon) = 1.$$

The variables Y_n and Y become arbitrarily close to each other with probability going to 1. The general definition above involves a random variable Y as the limit. In most cases of interest, the limit Y is actually a constant (i.e., a random vector with zero variance).

2.2.2 The law of large numbers

Now we're all set to state what I like to call the *fundamental theorem of statistics*.

Theorem 2.2.2 (The law of large numbers, LNN)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be* iid *random vectors with $\max_j \mathbb{E}[X_{1,j}^2] < \infty$ and define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then*

$$\overline{X}_n \to_p \mathbb{E}[X_1].$$

While we can't know $\mathbb{E}[X_1]$, $\overline{X}_n$ is something we observe. The LNN implies that the sample mean $\overline{X}_n$ is a reasonable approximation of μ. Hence, $\overline{X}_n$ gives us a "feeling" what the actual mean μ might be. The LNN makes this intuition mathematically precise. It allows us to learn about the expected value of an unknown random mechanism just from seeing the data.

Example 2.2.3. *Let's illustrate the LLN with a small experiment: We simulate $X_1, \ldots, X_n \sim$ Bernoulli(0.5) and compute $\overline{X}_n$ for each n. We repeat this experiment five times. By the law of large numbers, we expect the five resulting sequences to converge to the expected value $\mathbb{E}[X_1] = 0.5$. The results are shown in Figure 2.2. Each line (color) corresponds to a sequence $\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots$, one line for each repetition of the experiment. We see that for small n, $\overline{X}_n$ can be quite far away from the mean. As*

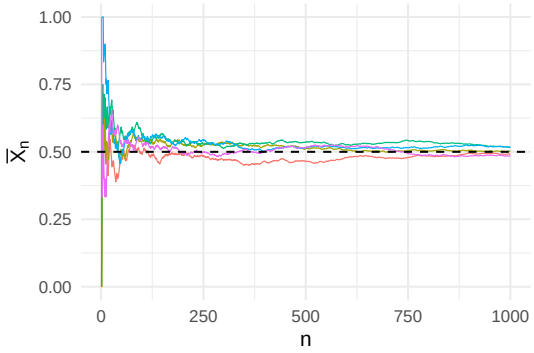

Figure 2.2: The law of large numbers in action (Example 2.2.3). Each line corresponds to a sequence $\overline{X}_n$ after simulating from n *iid* Bernoulli(0.5) random variables.

we increase the amount of data, all three lines seem to stabilize around 0.5. However, the three lines are different, reflecting the randomness of the samples. The green line lies mainly above 0.5, the others mainly below. The LLN states that, despite this randomness, it becomes less and less likely that one of the lines ends up away from 0.5.

Remark 2.2.4. *The result in Theorem 2.2.2 is sometimes called* weak *law of large numbers, because there is a* strong *version that employs a different notion of convergence. More on that later.*

2.2.3 Proof using Markov's inequality

The assumption $\max_j \mathbb{E}[X_{1,j}^2] < \infty$ in Theorem 2.2.2 is slightly stronger than necessary. While $\mathbb{E}[|X_{1,j}|] < \infty$ is sufficient, our stronger assumption is OK in 99.9% of applications and greatly simplifies the proof. The (optional) proof under the weaker condition is given at the end of this chapter. The simpler proof only requires a simple, but fundamental probabilistic inequality.

Theorem 2.2.5 (Markov's inequality)**.** *For any real-valued random variable $Y \geq 0$ and $t > 0$,*

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}[Y]}{t}.$$

Proof. We calculate

$$\mathbb{E}[Y] = \int_0^\infty y dP(y) \geq \int_t^\infty y dP(y) \geq t \int_t^\infty dP(y) = t\mathbb{P}(Y \geq t). \qquad \square$$

The theorem is more powerful than it may look at first sight. Other than being non-negative the random variable Y is arbitrary.

Proof of Theorem 2.2.2. Let $Y_n = \overline{X}_n - \mathbb{E}[X_1]$ and note that $\mathbb{E}[Y_n] = 0$. It holds

$$\begin{aligned}
&\mathbb{P}(\|\overline{X}_n - \mathbb{E}[X_1]\| > \varepsilon) \\
&= \mathbb{P}(\|\overline{X}_n - \mathbb{E}[X_1]\|^2 > \varepsilon^2) && [g(x) = x^2 \text{ strictly increasing for } x \geq 0] \\
&\leq \frac{\mathbb{E}[\|\overline{X}_n - \mathbb{E}[X_1]\|^2]}{\varepsilon^2} && [\text{Markov's inequality}] \\
&= \frac{\sum_{k=1}^d \mathbb{E}[(\overline{X}_{n,k} - \mathbb{E}[X_{1,k}])^2]}{\varepsilon^2} \\
&= \frac{\sum_{k=1}^d \mathbb{V}\mathrm{ar}[\overline{X}_{n,k}]}{\varepsilon^2}.
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\overline{X}_{n,k}] = \mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^n X_{i,k}\right] &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{C}\mathrm{ov}[X_{i,k}, X_{j,k}] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}\mathrm{ar}[X_{i,k}] && [\text{independence of } X_{i,k}\text{'s}] \\
&= \frac{1}{n}\mathbb{V}\mathrm{ar}[X_{1,k}]. && [\text{identical distribution}]
\end{aligned}$$

Because $\max_k \mathbb{V}\mathrm{ar}[X_{1,k}] \leq \max_k \mathbb{E}[X_{1,k}^2] < \infty$, we have

$$\mathbb{P}(\|\overline{X}_n - \mathbb{E}[X_1]\| > \varepsilon) \leq \frac{\sum_{k=1}^d \mathbb{V}\mathrm{ar}[\overline{X}_{n,k}]}{\varepsilon^2} = \frac{1}{n}\frac{\sum_{k=1}^d \mathbb{V}\mathrm{ar}[X_{1,k}]}{\varepsilon^2} \overset{n\to\infty}{\to} 0,$$

for every $\varepsilon > 0$. This proves $\overline{X}_n \to_p \mathbb{E}[X_1]$. □

2.2.4 Estimators and consistency

The statement "$\overline{X}_n$ is a good approximation of $\mathbb{E}[X_1]$" is made mathematically precise by $X_n \to_p \mathbb{E}[X_1]$. In that case, we say that $\overline{X}_n$ is a *consistent estimator* for $\mathbb{E}[X_1]$. Let us put this in a slightly more abstract setting.

Definition 2.2.6 (Estimator). *If $X_1, \ldots, X_n$ is our data, any quantity that can be expressed as $T_n(X_1, \ldots, X_n)$ for some function T_n is called an **estimator**.*

Less formally, an estimator is any quantity that you compute from data.

Definition 2.2.7 (Consistency). *Let θ be an unknown quantity that we are interested in. An estimator $\widehat{\theta}_n$ is called **consistent** for θ if*

$$\widehat{\theta}_n \to_p \theta.$$

In particular, $\widehat{\theta}_n = \overline{X}_n$ is a consistent estimator for $\theta = \mathbb{E}[X]$. This now applies in a straightforward way to the examples provided in Section 2.1. In case of the empirical distribution function, the histogram, and the kernel density estimator, we

are actually estimating functions instead of numbers. For example, the empirical distribution function $\widehat{F}_n$ from Example 1.5.5 is an estimator of a function. The LLN in Theorem 2.2.2 implies pointwise consistency

$$\widehat{F}_n(x) \to_p \mathbb{E}[\mathbb{1}\{X \leq x\}] = F(x).$$

It does not imply consistency of the entire function with respect to the sup-metric $\sup_x |\widehat{F}_n(x) - F(x)|$. There are, in fact, methods that are pointwise consistent but not uniformly consistent (i.e., consistent with respect to the sup-metric). The ECDF is also uniformly consistent, but this is a more advanced result that we will cover in a few weeks. Nevertheless, this illustrates that consistency is a concept intimately linked to the choice of metric.

2.2.5 Some useful facts and tools

We close this section with some useful tools. The first is about the convergence of tuples of random vectors.

Lemma 2.2.8 (Convergence in product spaces)**.** *Let $Y, Y_1, Y_n, \dots \in \mathbb{R}^d$ and $X, X_1, \dots, X_n \in \mathbb{R}^p$ be sequences of random vectors. If $Y_n \to_p Y$ and $X_n \to_p X$, then $(Y_n, X_n) \to_p (Y, X)$.*

Proof. Exercise. □

Another useful tool is the continuous mapping theorem. It allows us to deduce convergence of functions of random variables from the convergence of the random variables themselves. Recall that a function $g\colon \mathbb{R}^d \to \mathbb{R}^q$ is continuous if $\|y_1 - y_2\| \to 0$ implies $\|g(y_1) - g(y_2)\| \to 0$.

Theorem 2.2.9 (Continuous mapping theorem)**.** *Let $Y_n \to_p Y$ and $g\colon \mathbb{R}^d \to \mathbb{R}^q$ be continuous everywhere on a set S with $\mathbb{P}(Y \in S) = 1$. Then $g(Y_n) \to_p g(Y)$.*

Proof. Fix $\varepsilon > 0$. We want to control the probability of the event $\{\|g(Y_n) - g(Y)\| > \varepsilon\}$. We split this event into two parts: one where Y_n is far away from Y (this is unlikely because $Y_n \to_p Y$), and one where Y_n is close to Y but $g(Y_n)$ is far away from $g(Y)$ (this is unlikely because g is continuous). Define the set

$$S_{\delta,\varepsilon} = \{y\colon S\colon \exists y'\colon \|y - y'\| < \delta, \|g(y) - g(y')\| > \varepsilon\}.$$

These are points y whose neighborhood contains other points that g maps to points far away from $g(y)$. In such regions, g is highly erratic. Because g is continuous, these regions become small as $\delta \to 0$: $S_{\delta,\varepsilon} \downarrow \emptyset$. It holds

$$\begin{aligned}\mathbb{P}\{\|g(Y_n) - g(Y)\| > \varepsilon\} &= \mathbb{P}\{Y \in S_{\delta,\varepsilon}\} + \mathbb{P}\{\|g(Y_n) - g(Y)\| > \varepsilon, Y \notin S_{\delta,\varepsilon}\} \\ &\leq \mathbb{P}\{Y \in S_{\delta,\varepsilon}\} + \mathbb{P}\{\|Y_n - Y\| \geq \delta\}.\end{aligned}$$

The first term goes to 0 as $\delta \to 0$ because $S_{\delta,\varepsilon} \to \emptyset$. The second term goes to 0 as $n \to \infty$ because $Y_n \to_p Y$. This proves $g(Y_n) \to_p g(Y)$. □

Example 2.2.10 (Empirical variance). *Consider the estimator $S_n^2 = \overline{X_n^2} - (\overline{X}_n)^2$ for $\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. The LLN for the bivariate observations (X_i, X_i^2) implies $(\overline{X_n}, \overline{X_n^2}) \to_p (\mathbb{E}[X], \mathbb{E}[X^2])$. The continuous mapping theorem then implies that $S_n^2 \to_p \mathbb{V}\mathrm{ar}[X]$, because the function $g(x, y) = y - x^2$ is continuous.*

Remark 2.2.11. *The continuous mapping theorem and Lemma 2.2.8 imply that if $X_n \to_p X$ and $Y_n \to_p Y$, then also $X_n + Y_n \to_p X + Y$ and $X_n Y_n \to_p XY$.*

2.3 Convergence in distribution and asymptotic normality

If the estimator is consistent, we know that it converges for infinitely many observations. But on finite samples, there is some *uncertainty* how close we are to the truth. The estimation error $\widehat{\theta}_n - \theta$ is a random variable, so it has a distribution. The main question is therefore what this distribution is. In special cases, the distribution can be derived exactly. But more commonly, we need to rely on asymptotic approximations. Consistency tells us that the distribution converges to a point mass in the limit. But that's not helpful to quantify uncertainty. We first need a definition for convergence of distributions.

2.3.1 Convergence in distribution and the CLT

Definition 2.3.1 (Convergence in distribution). *Let $Y_n \in \mathbb{R}^d$ be a sequence of random vectors and $Y \in \mathbb{R}^d$ be another random vector. Denote the CDF of Y by F. Then we say that Y_n **converges in distribution** to Y or*

$$Y_n \to_d Y,$$

if for all $y \in \mathbb{R}^d$ where F is continuous,

$$\mathbb{P}(Y_n \leq y) \to F(y), \qquad \text{as } n \to \infty.$$

The restriction to continuity points is necessary to allow for non-continuous distributions.

For uncertainty quantification, we are really interested in situations where the limit is genuinely random. The CLT provides just that.

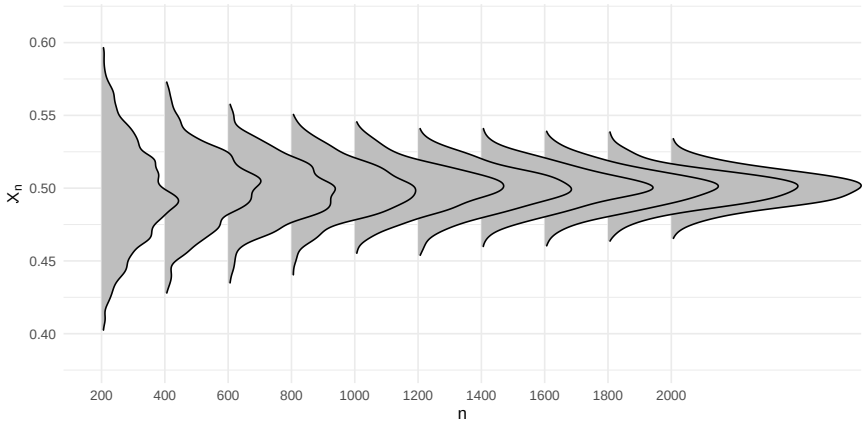

Figure 2.3: Illustration of the central limit theorem. The plots shows kernel density estimates of $\overline{X}_n$ after simulating many *iid* data sets from $X_i \sim$ Bernoulli(0.5).

Theorem 2.3.2 (Central limit theorem, CLT). *Let* $Y_1, \ldots, Y_n \in \mathbb{R}^d$ *be* iid *with mean* $\mu = \mathbb{E}[Y_1]$ *and covariance matrix* $\Sigma = \mathbb{V}\text{ar}[Y_1]$. *The sample mean* $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ *satisfies*

$$\sqrt{n}(\overline{Y}_n - \mu) \to_d \mathcal{N}(0, \Sigma),$$

and we say that the sequence $\overline{Y}_n$ *is* ***asymptotically normal****.*

Remark 2.3.3. *The statement of the theorem uses the common short notation* $\sqrt{n}(\overline{Y}_n - \mu) \to_d \mathcal{N}(0, \Sigma)$. *The long form is "there is a random variable* $Y \sim \mathcal{N}(0, \Sigma)$ *such that* $\sqrt{n}(Y_n - \mu) \to_d Y$*."*

A proof of the one-dimensional version can be found in Section 2.8.

Let's consider the one-dimensional case for simplicity. The interpretation of the CLT is that, for large enough n, the sample average $\overline{Y}_n \in \mathbb{R}$ behaves approximately[1] like a $\mathcal{N}(\mu, \sigma^2/n)$ random variable. This is illustrated in Fig. 2.3. As $n \to \infty$, the variance $\mathbb{V}\text{ar}[\overline{Y}_n] = \sigma^2/n$ vanishes. Hence, in a probabilistic sense, the difference $\overline{Y}_n - \mu$ gets closer to 0 (that's the law of large numbers). The scaling with $\sqrt{n}$ allows us to obtain a non-trivial limit. You can think of it this way: multiplying a random variable by $\sqrt{n}$ blows up its variance. The rate $\sqrt{n}$ strikes just the right balance: $\mathbb{V}\text{ar}[\sqrt{n}\overline{Y}_n] = (\sqrt{n})^2\mathbb{V}\text{ar}[\overline{Y}_n] = \sigma^2 \in (0, \infty)$.

The central limit theorem is quite remarkable. The only assumptions are that the sequence is *iid* with finite variance. It is called *central* because it plays such a central role in probability and statistics. The name was first used by George Pólya[2] in 1920 (in German, "Zentraler Grenzwertsatz"), but the idea is older and many other famous

[1] "Approximately behaves like" refers to probability statements: probability statements concerning $\overline{Y}_n$ are approximated by probability statements concerning $\mathcal{N}(\mu, \sigma^2/n)$.

[2] You might have been tortured by his 'urn' in high school.

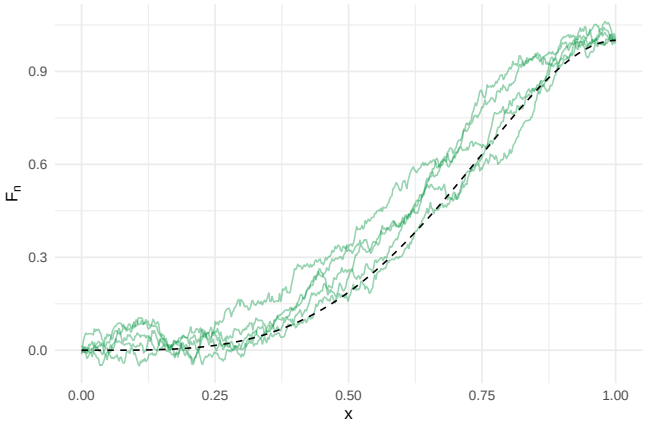

Figure 2.4: The F-Brownian bridge. The plot shows simulated curves from the asymptotic distribution of the empirical distribution function (solid lines) and the true distribution function (dashed line).

mathematicians contributed, including Laplace, Cauchy, Bessel, and Poisson.

2.3.2 Asymptotic normality of estimators

So how is this useful for uncertainty quantification? If $\widehat{\theta} - \theta^* \approx \mathcal{N}(0, \sigma^2/n)$, we can compute an (approximate) probability that $\widehat{\theta}$ is within some distance of θ^*. In particular, for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta} - \theta| < \epsilon) &= \mathbb{P}\left(\left|\frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma}\right| < \frac{\sqrt{n}\epsilon}{\sigma}\right) \\
&= \mathbb{P}\left(-\frac{\sqrt{n}\epsilon}{\sigma} < \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma} < \frac{\sqrt{n}\epsilon}{\sigma}\right) \\
\text{(CLT)} \quad &\approx \Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right) - \Phi\left(-\frac{\sqrt{n}\epsilon}{\sigma}\right).
\end{aligned}$$

If the variance σ^2 is known, we can actually compute this number. It is usually unknown, but can be estimated.

As $n \to \infty$, the probability above approaches 1: the more data we have, the more certain we are that $\widehat{\theta}$ is close to θ^*. Note that the standard deviation of $\widehat{\theta}$ is approximately $\sigma/\sqrt{n}$. This term is called *standard error* and often used as a measure of uncertainty. As $n \to \infty$, the standard error goes to zero, which reflects our increase in certainty.

The CLT applies directly to the sample average $\widehat{\theta} = \overline{X}_n$. This is an estimator for the parameter $\theta^* = \mathbb{E}[X]$. Let's revisit some of the other examples from the beginning. As always, we assume that the data are *iid* random variables $X_1, \ldots, X_n \sim F$.

Example 2.3.4 (Empirical CDF)**.** *The empirical distribution function is defined as*

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

Convince yourself that

$$\mathbb{E}[\widehat{F}_n(x)] = F(x), \qquad \mathbb{V}\mathrm{ar}[\widehat{F}_n(x)] = \frac{F(x)(1-F(x))}{n}.$$

(Hint: what's the distribution of $\mathbb{1}(X_i \leq x)$*?) By the CLT,*

$$\frac{\sqrt{n}(\widehat{F}_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \to_d \mathcal{N}(0,1).$$

The variance $\sigma^2 = F(x)(1-F(x))$ *is not known, because it involves the unknown distribution* F*. However, we can estimate it by* $\widehat{\sigma}^2 = \widehat{F}_n(x)(1-\widehat{F}_n(x))$*. Even more, we can say something about the behavior of* $\widehat{F}_n$ *at several points* $x_1, \ldots, x_m$ *simultaneously. The multivariate CLT implies that*

$$\sqrt{n}(\widehat{F}_n(x_1) - F(x_1), \ldots, \widehat{F}_n(x_m) - F(x_m)) \to_d \mathcal{N}(0, \Sigma),$$

with

$$\Sigma_{i,j} = F(\min\{x_i, x_j\}) - F(x_i)F(x_j).$$

This is the covariance function of a stochastic process called F-Brownian bridge*. Simulated curves from this process are shown in Fig. 2.4.*

Example 2.3.5 (Kernel density estimator, KDE)**.** *Suppose we want to estimate a continuous density* $f\colon \mathbb{R} \to \mathbb{R}$*. The kernel estimator (with uniform kernel* $K(x) = \mathbb{1}\{|x| \leq 1\}/2$*) is defined as*

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h}\mathbb{1}\{|X_i - x| \leq h\}.$$

We get

$$\mathbb{E}[\widehat{f}(x)] = \frac{\mathbb{E}[\mathbb{1}\{|X_i - x| \leq h\}]}{2h} = \frac{\mathbb{P}(X \leq x+h) - \mathbb{P}(X \leq x-h)}{2h} := f_h(x),$$
$$\mathbb{V}\mathrm{ar}[\widehat{f}(x)] = \frac{1}{n}\mathbb{V}\mathrm{ar}\left[\frac{1}{2h}\mathbb{1}\{|X_i - x| \leq h\}\right] = \frac{f_h(x)(1-2hf_h(x))}{2nh},$$

and the central limit theorem yields

$$\widehat{f}(x) \overset{d}{\approx} \mathcal{N}\left(f_h(x), \frac{f_h(x)(1-2hf_h(x))}{2nh}\right).$$

Note that $\mathbb{E}[\widehat{f}(x)] = f_h(x)$ is a smoothed version of f; in particular, $f_h(x) \neq f(x)$, so the KDE is biased. Because $f_h(x) \to f(x)$ as $h \to 0$, we prefer small bandwidths h to make the estimator less biased. However, the variance increases with smaller h, so there is a trade-off between bias and variance. When n is large, the variance is small anyway, so we may also afford smaller h.

2.3.3 Useful facts and tools

Besides the CLT, a few other tools are useful for analyzing the distributional limits of statistical methods. The first is the continuous mapping theorem for convergence in distribution.

Theorem 2.3.6 (Continuous mapping theorem, cnt'd)**.** *Let $Y_n \to_p Y$ and $g\colon \mathbb{R}^d \to \mathbb{R}^q$ be continuous everywhere on a set S with $\mathbb{P}(Y \in S) = 1$. Then $g(Y_n) \to_d g(Y)$.*

The proof is rather technical and omitted here; it can be found in, e.g., Van der Vaart (2000, Theorem 2.3).

Example 2.3.7. *Let $Y_n \to_d \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{p\times p}$ invertible. Because the function $g(y) = y^\top \Sigma^{-1} y$ is continuous at every $y \in \mathbb{R}^p$, the continuous mapping theorem implies*

$$Y_n^T \Sigma^{-1} Y_n \to_d Y^T \Sigma^{-1} Y,$$

Since $Z = \Sigma^{-1/2} Y \sim \mathcal{N}(0, I)$ and $Z^\top Z \sim \chi^2(p)$, this is equivalent to

$$Y_n^T \Sigma^{-1} Y_n \to_d \chi^2(p).$$

This technique is the basis for many tests in statistics, known as χ^2-tests.

Let us now establish some connections between the different modes of convergence.

Lemma 2.3.8. *Let Y_n, Y, Z_n be random vectors. Then:*

(i) If $Y_n \to_p Y$, then $Y_n \to_d Y$.

(ii) $Y_n \to_p c$ for some constant c if and only if $Y_n \to_d c$.

(iii) If $Y_n \to_d Y$ and $Z_n - Y_n \to_p 0$, then $Z_n \to_d Y$.

(iv) If $Y_n \to_d Y$ and $Z_n \to_p c$ for some constant c, then $(Y_n, Z_n) \to_d (Y, c)$.

The proof is left as an exercise. A combination of the last assertion of the lemma and the continuous mapping theorem gives the following useful result.

Lemma 2.3.9 (Slutsky's lemma)**.** *Let $Y_n \to_d Y$ and $Z_n \to_p c$ for some constant c. Then*

(i) $Y_n + Z_n \to_d Y + c$,

(ii) $Y_n Z_n \to_d cY$.

(iii) $Y_n / Z_n \to_d Y/c$ *if* $c \neq 0$.

Example 2.3.10 (t-test)**.** *Let $X_1, \dots, X_n$ be an* iid *sample with $\mathbb{V}\mathrm{ar}[X_1] > 0$ and $\mathbb{E}[X_1^4] < \infty$. We want to test the null hypothesis $H_0 \colon \mathbb{E}[X_1] = \mu_0$ against the alternative $H_1 \colon \mathbb{E}[X_1] \neq \mu_0$. We consider the t-statistic*

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n},$$

where $S_n^2 = \overline{X}_n^2 - (\overline{X}_n)^2$ is the sample variance. Under the null hypothesis, $\sqrt{n}(\overline{X}_n - \mu_0) \to_d \mathcal{N}(0, \mathbb{V}\mathrm{ar}[X_1])$ by the CLT and, further, $S_n^2 \to_p \mathbb{V}\mathrm{ar}[X_1]$ by Example 2.2.10. Now Slutsky's lemma implies that $T_n \to_d \mathcal{N}(0,1)$.

Another useful tool for functions of estimators is the delta method. It allows us to approximate the distribution of a function of an estimator by the distribution of the estimator itself.

Theorem 2.3.11 (Delta method)**.** *Let $g \colon \mathbb{R}^d \to \mathbb{R}^q$ be differentiable at some point θ. If $r_n(T_n - \theta) \to_d T$ for some sequence $r_n \to \infty$, then*

$$r_n(g(T_n) - g(\theta)) \to_d \nabla g(\theta) T,$$

where $\nabla g(\theta) = (\partial g_i(\theta)/\partial \theta_j)_{i,j} \in \mathbb{R}^{q \times d}$ is the Jacobian of g at θ.

Proof (optional). Note that because $1/r_n \to 0$ and $r_n(T_n - \theta) \to_d T$, Slutsky's lemma implies

$$T_n - \theta = \frac{1}{r_n} r_n(T_n - \theta) \to_d 0 \cdot T = 0.$$

Hence, $T_n - \theta \to_p 0$. Next define the function $\psi \colon \mathbb{R}^d \to \mathbb{R}^q$ with

$$\psi(h) = \frac{g(\theta + h) - g(\theta) - \nabla g(\theta) h}{\|h\|} \quad \text{for } h \neq 0, \quad \text{and } \psi(0) = 0.$$

Because g is differentiable at θ, ψ is continuous at 0. Now the continuous mapping theorem implies $\psi(T_n - \theta) \to_p \psi(0) = 0$. Now

$$r_n \left(g(T_n) - g(\theta) - \nabla g(\theta)(T_n - \theta)\right) = r_n \|T_n - \theta\| \psi(T_n - \theta) \to_d \|T\| \cdot 0 = 0,$$

by the continuous mapping theorem ($r_n \|T_n - \theta\| \to_d \|T\|$) and Slutsky's lemma. Now $Z_n = r_n \nabla g(\theta)(T_n - \theta) \to_d \nabla g(\theta) T$ by the continuous mapping theorem, and Lemma 2.3.8 (iii) implies that also $Y_n = r_n \left(g(T_n) - g(\theta)\right) \to_d \nabla g(\theta) T$. □

The delta method has many applications. A classical one is the distribution of the sample variance.

Example 2.3.12 (Sample variance). *Let $X, X_1, \ldots, X_n \overset{iid}{\sim} F$ be a sequence of random variables with finite variance $\sigma^2 = \mathbb{V}\mathrm{ar}[X]$. The sample variance is defined as*

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

Note that the sample variance does not change if we replace X_i by $X_i - \mathbb{E}[X_i]$, so we may assume $\mathbb{E}[X_i] = 0$ without loss of generality. The central limit theorem implies that

$$\sqrt{n}\begin{pmatrix}\overline{X}_n - 0 \\ \overline{X^2}_n - \sigma^2\end{pmatrix} \to_d Z \sim \mathcal{N}\left(0, \begin{pmatrix}\mathbb{E}[X^2] & \mathbb{E}[X^3] \\ \mathbb{E}[X^3] & \mathbb{E}[X^4] - \mathbb{E}[X^2]^2\end{pmatrix}\right).$$

The function $g(x, y) = y - x^2$ is differentiable at $\theta = (\mathbb{E}[X], \sigma^2)$ with Jacobian $\nabla g(\theta) = (0, 1)$. The delta method implies

$$\sqrt{n}(S_n^2 - \sigma^2) = \sqrt{n}\left(\overline{X^2}_n - (\overline{X}_n)^2 - \sigma^2\right) \to_d \nabla g(\theta) Z = Z_2 \overset{d}{=} \mathcal{N}(0, \mathbb{E}[X^4] - \mathbb{E}[X^2]^2).$$

Using Slutsky's lemma, we can also show that the same limit holds for the unbiased version $nS_n^2/(n-1)$, because $nS_n^2/(n-1) - S_n^2 \to_p 0$.

2.4 Stochastic O-notation

We now introduce some notation that allows us to describe the behavior of random variables in a concise way. The stochastic O-notation is a probabilistic version of the Landau O-notation from analysis. It allows us to describe the rate of convergence of random variables in a compact way. Recall that for two sequences a_n, b_n, we write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ for all n large enough, and $a_n = o(b_n)$ if $a_n/b_n \to 0$. In particular, $a_n = O(1)$ means that a_n is bounded, and $a_n = o(1)$ means that a_n converges to zero. This is now generalized to sequences of random variables.

Definition 2.4.1 (Bounded in probability). *A sequence of random vectors Y_n is called* bounded in probability *if for every ε, there is M such that $\mathbb{P}(\|Y_n\| > M) < \varepsilon$ for n large enough. We write $Y_n = O_p(1)$.*

Similarly, we write $Y_n = o_p(1)$ if $\|Y_n\| \to_p 0$. For example, the law of large numbers reads $\overline{X}_n - \mathbb{E}[X_1] = o_p(1)$. This can be generalized as follows.

Definition 2.4.2 (Stochastic O-symbols). *Let Y_n, Z_n, R_n be sequences of random variables. We write*

- *$Y_n = O_p(R_n)$ if $Y_n = Z_n R_n$ for some $Z_n = O_p(1)$.*
- *$Y_n = o_p(R_n)$ if $Y_n = Z_n R_n$ for some $Z_n \to_p 0$.*

Intuitively, $Y_n = O_p(R_n)$ means that, for n large enough and with high probability, Y_n is at most as large as some constant times R_n. The statement $Y_n = o_p(R_n)$ means that Y_n is much smaller than R_n. The notation is most often used with deterministic sequences $r_n \searrow 0$, describing a convergence rate. Then $Y_n = O_p(r_n)$ is equivalent to $Y_n/r_n = O_p(1)$, and $Y_n = o_p(r_n)$ is equivalent to $Y_n/r_n \to_p 0$. Here the interpretation is as follows: $Y_n = O_p(r_n)$ means that Y_n behaves like r_n times a bounded random variable; $Y_n = o_p(r_n)$ means that Y_n goes to zero in probability faster than r_n.

Example 2.4.3. *The sample mean $\overline{X}_n$ satisfies $\overline{X}_n - \mathbb{E}[X_1] = O_p(1/\sqrt{n})$, since*

$$\mathbb{P}\left(\frac{|\overline{X}_n - \mathbb{E}[X_1]|}{1/\sqrt{n}} > M\right) = \mathbb{P}(\sqrt{n}|\overline{X}_n - \mathbb{E}[X_1]| > M) \leq \frac{n\mathbb{V}\mathrm{ar}\overline{X}_n}{M^2} = \frac{\mathbb{V}\mathrm{ar}[X_1]}{M^2}.$$

The right hand side can be made arbitrarily small by choosing M large enough. Hence, $|\overline{X}_n - \mathbb{E}[X_1]|/(1/\sqrt{n}) = O_p(1)$. Note that the statement $X_n - \mathbb{E}[X_1] = O_p(1/\sqrt{n})$ is more accurate than just saying $X_n \to_p \mathbb{E}[X_1]$, because it additionally quantifies the rate of convergence.

The above is a simple example of a more general principle. Markov's equality gives $Y = O_p(\mathbb{E}[\|Y\|^k]^{1/k})$ for any $k \geq 1$.

Stochastic O-symbols are so convenient because there are many simple rules to calculate with them. Here are some examples:

Lemma 2.4.4 (Stochastic O-calculus)**.** *It holds*

(i) $O_p(A_n) + O_p(B_n) = O_p(\max\{A_n, B_n\})$,

(ii) $o_p(A_n) + o_p(B_n) = o_p(\max\{A_n, B_n\})$,

(iii) $O_p(A_n) + o_p(A_n) = O_p(A_n)$,

(iv) $O_p(A_n)O_p(B_n) = O_p(A_nB_n)$,

(v) $o_p(A_n)O_p(B_n) = o_p(A_nB_n)$,

(vi) $(1 - o_p(1))^{-1} = O_p(1)$.

Proof. The proofs are rather straightforward, working with the explicit definitions of the statements. We shall only prove (iii) for illustration. The term $O_p(A_n)$ is an alias for a random variable Z_nA_n, with $Z_n = O_p(1)$. Similarly, $o_p(A_n)$ is an alias for a random variable W_nA_n with $W_n = o_p(1)$. Hence, $O_p(A_n) + o_p(A_n) = (Z_n + W_n)A_n = O_p(A_n)$, because $Z_n + W_n = O_p(1)$. To see the latter statement, note that

$$\mathbb{P}(|Z_n + W_n| > M) \leq \mathbb{P}(|Z_n| > M/2) + \mathbb{P}(|W_n| > M/2).$$

The second term converges to zero as $n \to \infty$, because $W_n = o_p(1)$. The first term can be made arbitrarily small since $Z_n = O_p(1)$. Hence, $Z_n + W_n = O_p(1)$. □

Note that while these statements look like equalities, they must only be read from left

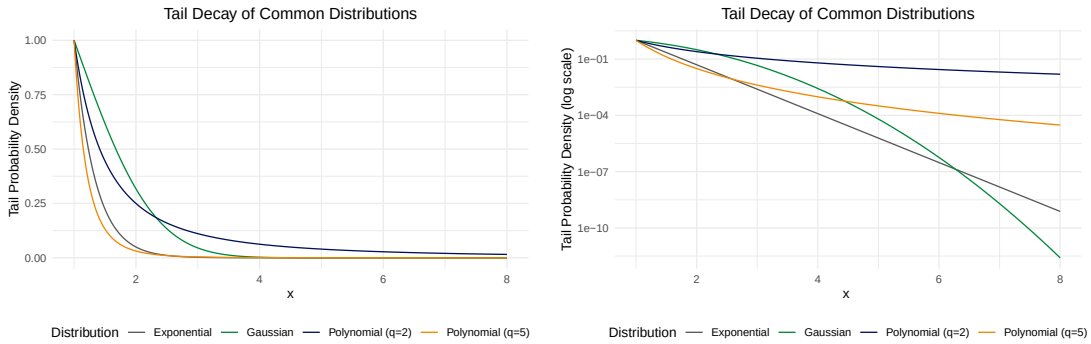

Figure 2.5: Common orders for the decay of tail probabilities $\mathbb{P}(|X| > s)$. The left plot shows shows the probabilities on log-scale to make the far tail better visible.

to right. For example, the statement $o_p(1) = O_p(1)$ ('every sequence converging to zero in probability is bounded in probability') is true in general; but the statement $O_p(1) = o_p(1)$ ('every bounded sequence converges to zero in probability') is false.

2.5 Excursion: moment conditions

The statement and proof of both LLN and CLT involve *moment conditions* of the form $\mathbb{E}[|X|^q] < \infty$. Moment conditions are essential in mathematical statistics. Whether a statistical law holds or fails often depends on the existence of certain moments, both in theory and in practice. It is worthwhile to spend some time understanding what they really mean.

Moment conditions are assumptions about the *tail* of a distribution. Recall that for any positive random variable Y,

$$\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y > t)\, dt.$$

This implies

$$\mathbb{E}[|X|^q] = \int_0^\infty \mathbb{P}(|X|^q > t)\, dt = \int_0^\infty \mathbb{P}(|X| > t^{1/q})\, dt.$$

The *tail probabilities* $\mathbb{P}(|X| > s)$ determine how unlikely it is to observe very large values of X. The moment condition $\mathbb{E}[|X|^q] < \infty$ requires that the tail of X is not too heavy. Whether or not the integral is finite is only determined by the far tail of the distribution, i.e., $\mathbb{P}(|X| > s)$ with s large. In particular, $\mathbb{E}[|X|^q] < \infty$ holds if

$$\mathbb{P}(|X| > s) = Cs^{-q'} \qquad \text{for some } q' > q, C < \infty \text{ and all } s \geq 1,$$

and it fails if

$$\mathbb{P}(|X| > s) \geq cs^{-q} \qquad \text{for some } c > 0 \text{ and all } s \geq 1.$$

The larger q, the faster the tail probabilities must decay for the moment condition to

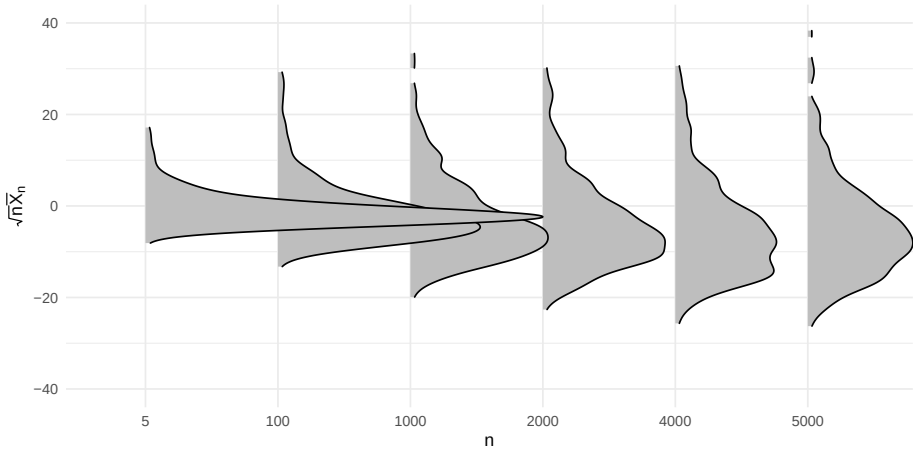

Figure 2.6: The CLT fails for the centered Pareto distribution with parameter $\alpha = 1.5$, because $\mathbb{E}[X^2] = \infty$.

hold. Hence, the larger q, the less likely we see extreme values of X.

Many standard distributions have finite moments of all orders $q \geq 1$, including the Gaussian, the χ^2, the exponential, and Poisson distributions. Their tail probabilities decay exponentially fast in s. The Student t distribution with ν degrees of freedom, for example, only has finite moments up to order $q < \nu$. Its tail decays only polynomially in s. Also the Pareto distribution has polynomial tail decay. Such distributions are called *heavy tailed* and appear frequently when analyzing extreme events in climate, finance, or insurance. Some common orders of tail decay are illustrated in Fig. 2.5

If the tail doesn't decay fast enough, the CLT will fail — also in practice. Fig. 2.6 shows an example of a centered Pareto distribution with tail $s^{-1.5}$, which has infinite variance and fails to converge to a Gaussian. In fact, the spread of $\sqrt{n}(\overline{X}_n - \mu)$ *increases* with n and also the asymmetry doesn't appear to go away. Taking this a step further, Fig. 2.7 shows paths of the sample average $\overline{X}_n$ of a Student t distribution with $\nu = 1$ degree of freedom (i.e., a Cauchy distribution). The sample average does not converge to the mean, but jumps around chaotically. This is because $\mathbb{E}[|X|] = \infty$, which makes the LLN fail.

Generally speaking, conditions on the first or second moment are considered very mild. In financial statistics, polynomial moment conditions with single-digit q are often considered appropriate. In high-dimensional statistics, bounded exponential moments, $\mathbb{E}[\exp(|X|)] < \infty$, (which implies boundedness of all polynomial moments) are often required. In finite-dimensional settings with independent data, bounded first or second moments are usually sufficient. So that's what we will encounter most frequently in this course.

2.6 Almost sure convergence and strong consistency*

There is a third mode of convergence that sometimes appears in the literature: almost sure convergence. It is the strongest form of convergence, implying convergence with

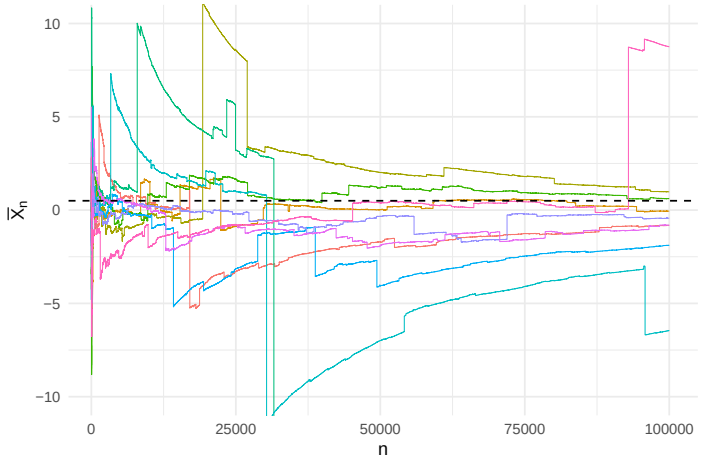

Figure 2.7: The LLN fails for the Cauchy distribution because $\mathbb{E}[|X|] = \infty$.

probability one.

Definition 2.6.1 (Almost sure convergence)**.** *Let $Y, Y_1, Y_2, \ldots \in \mathbb{R}^d$ be random vectors. We say that Y_n converges to Y almost surely (a.s.), or $Y_n \to_{a.s.} Y$, if*

$$\mathbb{P}\left\{\omega\colon \lim_{n\to\infty} \|Y_n(\omega) - Y(\omega)\| = 0\right\} = 1.$$

In the notation, we use the fundamental definition of a random variable being a map from the sample space Ω to $\mathbb{R}^d$. Upon fixing $\omega \in \Omega$, the sequence $\|Y_n(\omega) - Y(\omega)\|$ is deterministic and may converge or not. Almost sure convergence asserts that the collection of all states ω where the sequence converges to zero has probability 1. Almost sure convergence is a very strong form of convergence. It implies convergence in probability and distribution.

The almost sure version of consistency is called *strong consistency*.

Definition 2.6.2 (Strong consistency)**.** *Let θ be an unknown quantity that we are interested in. An estimator $\widehat{\theta}_n$ is called* ***strongly consistent*** *for θ if*

$$\widehat{\theta}_n \to_{a.s.} \theta.$$

Almost sure convergence appears to be of minor interest in statistics and stronger than needed in most applications. It is often difficult to establish and does not provide much additional information compared to convergence in probability. However, it is sometimes useful in theoretical analyses. Proofs of strong consistency rely on the strong law of large numbers, $\overline{X}_n \to_{a.s.} \mathbb{E}[X_1]$, or a direct argument based on the Borel-Cantelli lemma. The strong law holds under the condition $\mathbb{E}[|X_1|] < \infty$, but we shall state a version using a stronger requirement to illustrate the proof technique.

Theorem 2.6.3. *Suppose $X_1, X_2, \ldots$ is an* iid *sequence of random variables with $\mathbb{E}[X_1^4] < \infty$. Then $\overline{X}_n \to_{a.s.} \mathbb{E}[X_1]$.*

Proof. The statement $\overline{X}_n \to_{a.s.} \mathbb{E}[X_1]$ is the same as $\overline{X}_n - \mathbb{E}[X_1] \to_{a.s.} 0$, so we may assume without loss of generality that $\mathbb{E}[X_1] = 0$. The statement $\overline{X}_n(\omega) \to 0$ is equivalent to: for every $\varepsilon > 0$, $|\overline{X}_n(\omega)| < \varepsilon$ for all but finitely many n. Define the events $E_n = \{\omega \colon |\overline{X}_n(\omega)| \geq \varepsilon\}$. The Borel-Cantelli lemma states that if $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$, then $\mathbb{P}(E_n \text{ infinitely often}) = 0$. Markov's inequality gives

$$\mathbb{P}(E_n) = \mathbb{P}\left(|\overline{X}_n| \geq \varepsilon\right) \leq \frac{\mathbb{E}[\overline{X}_n^4]}{\varepsilon^4}.$$

Now

$$\mathbb{E}[\overline{X}_n^4] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^4\right] = \frac{1}{n^4} \sum_{1 \leq i,j,k,l \leq n} \mathbb{E}[X_i X_j X_k X_l].$$

Because the X_i are independent with mean zero, all summands where an index shows up only once vanish. This leaves us with only terms of the form $\mathbb{E}[X_i^4]$ and $\mathbb{E}[X_i^2 X_j^2]$, from which there are only $O(n^2)$ many, and all of them are finite. Hence, $\mathbb{E}[\overline{X}_n^4] = O(1/n^2)$, and $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$. □

2.6.1 Triangular arrays*

The KDE example illustrates the need for more general CLT results. To get good density estimates, we need to adjust the bandwidth $h = h_n$ according to the sample size n. But if we do, we can no longer apply the CLT from Theorem 2.3.2. The issue is somewhat subtle. For any fixed, h, we may define the random variables $Y_{i,h} = \frac{1}{h}\mathbb{1}\{|X_i - x| \leq h\}$ and write $\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_{i,h}$. This is a sample average over the *iid* sequence $Y_{1,h}, \ldots, Y_{n,h} \sim P_h$. However, if we change the bandwidth to h', we generate another *iid* sequence $Y_{1,h'}, \ldots, Y_{n,h'}$ coming from a different distribution $P_{h'}$. In particular, changing h with n, we generate a *triangular array* of *iid* sequences

$$\begin{aligned}
&Y_1^{(1)},\\
&Y_1^{(2)}, Y_2^{(2)},\\
&Y_1^{(3)}, Y_2^{(3)}, Y_3^{(3)},\\
&\qquad\vdots\\
&Y_1^{(n)}, Y_2^{(n)}, Y_3^{(n)}, \ldots, Y_n^{(n)}.
\end{aligned}$$

Every row of this array is an *iid* sequence, but the distribution of the variables changes from row to row. The Lindeberg-Feller CLT is one of the most general results that covers this situation.

Theorem 2.6.4 (Lindeberg-Feller CLT). *For each n, let $Y_1^{(n)}, \dots, Y_{k_n}^{(n)}$ be a sequence of independent random vectors such that*

$$\sum_{i=1}^{k_n} \mathbb{V}\mathrm{ar}[Y_i^{(n)}] \to \Sigma, \qquad \text{[converging covariance]}$$

$$\sum_{i=1}^{k_n} \mathbb{E}[\|Y_i^{(n)}\|^2 \mathbb{1}\{\|Y_i^{(n)}\| > \varepsilon\}] \to 0 \quad \textit{for every } \varepsilon > 0, \qquad \text{[Lindeberg condition]}$$

Then

$$\sum_{i=1}^{k_n} (Y_i^{(n)} - \mathbb{E}[Y_i^{(n)}]) \to_d \mathcal{N}(0, \Sigma).$$

Some comments:

- The theorem is most often applied with $k_n = n$, but the more general form is sometimes useful.
- The result does not even require the rows of the triangular array to be *iid*. It is sufficient that the variables in each row are independent.
- Note that the $\sqrt{n}$ factor is missing from the convergence statement. In the triangular array setup, this factor is included in the random variables $Y_i^{(n)}$. These are assumed to be standardized such that sum of their variances converges to a constant. This implies that most individual variances go to zero as $n \to 0$. We recover the usual CLT for an *iid* sequence $X_1, \dots, X_n$ by setting $Y_i^{(n)} = X_i/\sqrt{n}$. However, including the scaling in the variables additionally allows for convergence rates different from $\sqrt{n}$, which is often useful (for example, in the context of the KDE).
- Lindeberg's condition is a technical condition that gives additional control over the deviations form the mean. This additional control helps us deal with the fact that the distribution of the random variables changes with n. It excludes some pathological cases where the variances converge, but the distributions change in very unfortunate ways that prevent convergence to a Gaussian limit.

While establishing convergence of the variances is often straightforward, the Lindeberg condition can be tricky to verify. A common approach is to use the stricter Lyapunov condition, which is easier to check but implies the Lindeberg condition.

Lemma 2.6.5 (Lyapunov's condition). *If for some $\delta > 0$,*

$$\sum_{i=1}^{k_n} \mathbb{E}[\|Y_i^{(n)}\|^{2+\delta}] \to 0,$$

then the Lindeberg condition holds.

Proof. Observe

$$\sum_{i=1}^{k_n} \mathbb{E}\left[\|Y_i^{(n)}\|^2 \mathbb{1}\{\|Y_i^{(n)}\| > \varepsilon\}\right] \leq \frac{1}{\varepsilon^\delta} \sum_{i=1}^{k_n} \mathbb{E}\left[\|Y_i^{(n)}\|^{2+\delta} \mathbb{1}\{\|Y_i^{(n)}\| > \varepsilon\}\right].$$

If the far right-hand side converges to zero, also the left-hand side does. □

It is usually most convenient to apply Lyapnuov's condition with $\delta = 1$ or $\delta = 2$. Let us illustrate this with the KDE.

Example 2.6.6 (Kernel density estimator, cont'd)**.** *Consider again the kernel density estimator from Example 2.3.5. Suppose that f is smooth and bounded. For some bandwidth sequence $h_n \to 0$, define*

$$Y_i^{(n)} = \frac{1}{\sqrt{nh_n}}(\mathbb{1}\{|X_i - x| \leq h_n\},$$

so that

$$\sqrt{nh_n}(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]) = \sum_{i=1}^{n}(Y_i^{(n)} - \mathbb{E}[Y_i^{(n)}]).$$

We now check the conditions of Lindeberg-Feller CLT. It holds

$$\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[Y_i^{(n)}] = n\mathbb{V}\mathrm{ar}[Y_i^{(n)}] = f_{h_n}(x)(1 - h_n f_{h_n}(x)) \to f(x),$$

and

$$\sum_{i=1}^{n} \mathbb{E}\left[|Y_i^{(n)}|^3\right] = n\frac{1}{n^{3/2}h_n^{3/2}}\mathbb{E}[\mathbb{1}\{|X - x| \leq h_n\}^3] = \frac{1}{n^{1/2}h_n^{1/2}} f_{h_n}(x) \to 0,$$

provided $nh_n \to \infty$. Hence, if h_n does not vanish too fast, the KDE is asymptotically normal:

$$\sqrt{nh_n}(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]) \to_d \mathcal{N}(0, f(x)).$$

2.7 Proof of the weak LLN under minimal conditions*

We now prove the weak law of large numbers under the condition $\max_j \mathbb{E}[|X_{i,j}|] < \infty$. To simplify, we only consider the univariate case $X_i \in \mathbb{R}$. The proof is largely similar to the one given before, but we cannot use Markov's inequality for the second moment of $\overline{X}_n$ directly, because it might be infinite. Instead, we use a *truncation argument*, a common technique in probability theory. We truncate the random variables to a bounded interval where we can apply Markov's inequality, and then let the truncation threshold go to infinity.

Formally, define the truncated random variables $X_i' = X_i \mathbb{1}(|X_i| \leq \sqrt{n})$ and the truncated average $\overline{X}_n' = \frac{1}{n}\sum_{i=1}^n X_i'$. It holds

$$\overline{X}_n = \overline{X}_n' + \frac{1}{n}\sum_{i=1}^n X_i \mathbb{1}(|X_i| > \sqrt{n}).$$

For the second term, we have

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i \mathbb{1}(|X_i| > \sqrt{n})\right|\right] \leq \mathbb{E}[|X_i| \mathbb{1}(|X_i| > \sqrt{n})].$$

The right hand side is decreasing in n, and by the monotone convergence theorem (every bounded decreasing sequence converges), we have

$$\mathbb{E}[X_i \mathbb{1}(|X_i| > \sqrt{n})] \to 0 \quad \text{as } n \to \infty.$$

This further implies $\mathbb{E}[\overline{X}_n'] \to \mathbb{E}[\overline{X}_n] = \mathbb{E}[X_i]$ and, by Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^n X_i \mathbb{1}(|X_i| > \sqrt{n}) \to_p 0.$$

It remains to show that

$$\overline{X}_n' - \mathbb{E}[\overline{X}_n'] \to_p 0.$$

Because the X_i' are bounded, their second moment exists, and we can apply Markov's inequality as in the simpler proof: For any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}(|\overline{X}_n' - \mathbb{E}[\overline{X}_n']| > \varepsilon) &\leq \frac{\mathbb{V}\mathrm{ar}[\overline{X}_n']}{\varepsilon^2} && \text{[Markov]} \\
&\leq \frac{\mathbb{V}\mathrm{ar}[X_i']}{n\varepsilon^2} && [X_i' \text{ } iid] \\
&\leq \frac{\mathbb{E}[|X_i'|^2]}{n\varepsilon^2} && [\mathbb{V}\mathrm{ar}[Y] \leq \mathbb{E}[Y^2]] \\
&\leq \frac{\sqrt{n}\mathbb{E}[|X_i'|]}{n\varepsilon^2}. && [|X_i'| \leq \sqrt{n}]
\end{aligned}$$

This converges to zero as $n \to \infty$, which concludes the proof.

The strong law can be proven under the same condition with the same technique, but requires several additional tricks. In case you're interested, see https://terrytao.wordpress.com/2008/06/18/the-strong-law-of-large-numbers.

2.8 Proof of the CLT*

We give a short proof of the one-dimensional CLT.

Theorem 2.8.1 (Central Limit Theorem)**.** *Let $X, X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite variance σ^2. Then*

$$\sqrt{n}(\overline{X}_n - \mathbb{E}[X]) \xrightarrow{d} N(0, \sigma^2).$$

Proof. Let $\phi_X(t)$ be the characteristic function of X_1. We will prove that the characteristic function of $Z_n = \sqrt{n}(\overline{X}_n - \mathbb{E}[X_1])/\sigma$ converges to $e^{-t^2/2}$, which is the characteristic function of $N(0,1)$. By Lévy's continuity theorem, this proves that $Z_n \xrightarrow{d} N(0,1)$. Since $\overline{X}_n - \mathbb{E}[X]$ has mean zero, we may assume w.l.o.g. that $\mathbb{E}[X] = 0$.

The characteristic function of Z_n is:

$$\begin{aligned}
\phi_{Z_n}(t) &= \mathbb{E}[e^{itZ_n}] \\
&= \mathbb{E}\left[\exp\left(\sum_{j=1}^{n} \frac{it}{\sigma\sqrt{n}} X_j\right)\right] && \text{[defnition of } Z_n\text{]} \\
&= \mathbb{E}\left[\prod_{j=1}^{n} \exp\left(\frac{it}{\sigma\sqrt{n}} X_j\right)\right] && [e^{\sum_i a_i} = \textstyle\prod_i e^{a_i}] \\
&= \prod_{j=1}^{n} \mathbb{E}\left[\exp\left(\frac{it}{\sigma\sqrt{n}} X_j\right)\right] && [\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \text{ for independent } X, Y] \\
&= \left[\phi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n && \text{[identical distribution of } X_1, \ldots, X_n\text{]}
\end{aligned}$$

By Taylor's theorem, for small t:

$$\phi_X(t) = 1 - \frac{t^2}{2}\sigma^2 + o(t^2),$$

so

$$\phi_X\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right).$$

Using the fact that $(1 + x_n)^n \to e^x$ when $nx_n \to x$, we get

$$\phi_{Z_n}(t) = \left[\phi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \to \exp\left(-\frac{t^2}{2}\right),$$

as claimed. □

3 M- and Z-estimators

In the previous chapter we have learned how to deal with statistical methods that can be written as a (function of) a sample average. This is useful, but some of the most important statistical methods, such as the maximum likelihood estimator or sample median, are not of this form. Even more commonly, estimators are defined as solutions to optimization problems. Such estimators are called *M-estimators*, where the M stands for either minimum or maximum. A second important class is the class of *Z-estimators*, which are defined as the zero of an *estimating equation*.

3.1 M-estimators

Definition 3.1.1 (M-estimator). *Let Θ be a parameter space, $X_1, \dots, X_n \in \mathcal{X}$* iid *observations, and $m_\theta \colon \mathcal{X} \to \mathbb{R}$ a loss function. An* M-estimator *is defined as the solution to the optimization problem*

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i).$$

Existence of the minimizer can be guaranteed under mild conditions; for example, if Θ is compact and $m_\theta(x)$ is continuous in θ for all x. We will take existence as a given in what follows. Instead of minimizing the loss function, we could also maximize it, because the latter is equivalent to minimizing the negative of the loss function. Let's see some examples.

Example 3.1.2 (Least squares estimator in linear model). *In the context of linear regression, the least squares estimator is an M-estimator where the loss function is the squared error. Consider the linear model*

$$Y_i = \beta^\top X_i + \epsilon_i, \quad i = 1, \dots, n,$$

where ϵ_i are iid *random errors with $\mathbb{E}[\varepsilon_i \mid X_i] = 0$. The parameter vector β is estimated by minimizing the sum of squared residuals:*

$$\widehat{\beta} = \arg\min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \beta^\top X_i \right)^2.$$

Example 3.1.3 (Maximum Likelihood Estimator). *The maximum likelihood estimator (MLE) is an M-estimator where the loss function is the negative log-likelihood. Suppose*

that the observations $X_1, \dots, X_n$ follow a probability distribution with parameter $\theta \in \Theta$ and probability density function f_θ. The MLE is defined as

$$\widehat{\theta}_{MLE} = \arg\min_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^{n} - \log f_\theta(X_i) \right).$$

Example 3.1.4 (Sample quantiles)**.** *The α-quantile of a sample minimizes the expected value of an asymmetric loss function. For given $\alpha \in (0,1)$, the* pinball loss *is defined as*

$$\rho_\alpha(x) = \begin{cases} \alpha|x| & \text{if } x > \theta, \\ (1-\alpha)|x| & \text{if } x \leq \theta. \end{cases}$$

The α-sample quantile $\widehat{\theta}_\alpha$ is then given by

$$\widehat{\theta}_\alpha = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(x - \theta).$$

The sample median is a special case of the sample quantile with $\alpha = 1/2$.

Example 3.1.5 (Quantile Regression)**.** *Quantile regression extends the concept of quantiles to the estimation of conditional quantiles of the response variable. For example, the linear quantile regression estimator $\widehat{\beta}_\alpha$ is then given by*

$$\widehat{\beta}_\alpha = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(Y_i - \beta^\top X_i),$$

with ρ_α given as in Example 3.1.4.

3.2 Z-estimators

To solve an M-estimation problem, we need to minimize a loss function. If the loss function is differentiable in θ, we may take the derivative of the criterion function and set it to zero. Then $\widehat{\theta}$ is defined as the solution of the *estimating equation*

$$\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta m_\theta(X_i) = 0.$$

More generally, we can define estimators directly as solutions to estimating equations, without the need to minimize a loss function. By convention, the right-hand side of the estimating equation is set to zero, which is why these estimators are called *Z-estimators* (Z for zero).

Definition 3.2.1 (Z-estimator). *Let Θ be a parameter space, $X_1, \dots, X_n \in \mathcal{X}$* iid *observations, and $\psi_\theta \colon \mathcal{X} \to \mathbb{R}^p$ an estimating function. A* Z-estimator *is defined as the solution to the estimating equation*

$$\frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = 0.$$

Any of the examples from the previous section can be converted into a Z-estimator by taking the derivative. We shall restrict our attention to two interesting examples.

Example 3.2.2 (Score equations of MLE). *The Maximum Likelihood Estimator is a Z-estimator with $\psi_\theta(x) = \nabla_\theta \log f_\theta(x)$, the score function of the likelihood. The MLE is then defined as the solution to the* score equation

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f_\theta(X_i) = 0.$$

Example 3.2.3 (Sample quantile as Z-estimator). *The sample quantile can be written as a Z-estimator with $\psi_\theta(x) = \alpha \mathbb{1}(x > \theta) - (1-\alpha)\mathbb{1}(x \le \theta)$. Note that the function ρ_α from Example 3.1.4 is not differentiable at 0, but it suffices that it is differentiable almost everywhere. The sample quantile is then defined as the solution to the estimating equation*

$$\frac{1}{n}\sum_{i=1}^{n} [\alpha \mathbb{1}(X_i > \theta) - (1-\alpha)\mathbb{1}(X_i \le \theta)] = 0.$$

This equation can be rewritten as

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} [\alpha \mathbb{1}(X_i > \theta) - (1-\alpha)\mathbb{1}(X_i \le \theta)] = 0 \\
\Leftrightarrow \quad & \alpha \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i > \theta) = (1-\alpha)\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le \theta) \\
\Leftrightarrow \quad & \alpha \frac{1}{n}\sum_{i=1}^{n} [1 - \mathbb{1}(X_i \le \theta)] = (1-\alpha)\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le \theta) \\
\Leftrightarrow \quad & \alpha = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le \theta).
\end{aligned}$$

So indeed, the sample quantile is the value θ such that a fraction α of the observations are at most θ.

Z-estimators are sometimes easier to analyze, especially when considering asymptotic normality results. The choice between formulations as M- and Z-estimators is often a matter of convenience. There are some examples of Z-estimators that cannot be written as minimizing an average loss, however. One example comes from one of my

primary fields of research.

Example 3.2.4 (2-step estimator in copula model)**.** *A copula model decomposes the joint density f of a random vector (X, Y) into the product of the marginal densities f_X and f_Y and a* copula density *c:*

$$f(x, y; \theta) = f_X(x; \theta_X) f_Y(y; \theta_Y) c(F_X(x; \theta_x), F_Y(y; \theta_Y); \theta_C).$$

The marginal densities capture the individual behvior of X and Y, while the copula induces dependence between X and Y. When $c \equiv 1$, the variables are independent. To estimate the parameters $\theta = (\theta_X, \theta_Y, \theta_C)$, we typically use a 2-step procedure. In the first step, we estimate the marginal parameters θ_X and θ_Y by maximum likelihood:

$$\widehat{\theta}_X = \arg\max_{\theta_X} \frac{1}{n} \sum_{i=1}^{n} \log f_X(X_i; \theta_X),$$
$$\widehat{\theta}_Y = \arg\max_{\theta_Y} \frac{1}{n} \sum_{i=1}^{n} \log f_Y(Y_i; \theta_Y).$$

In the second step, we fix these estimates and estimate θ_C by maximizing the copula likelihood:

$$\widehat{\theta}_C = \arg\max_{\theta_C} \frac{1}{n} \sum_{i=1}^{n} \log c(F_X(X_i; \widehat{\theta}_X), F_Y(Y_i; \widehat{\theta}_Y); \theta_C).$$

The full parameter estimate $\widehat{\theta} = (\widehat{\theta}_x, \widehat{\theta}_Y, \widehat{\theta}_C)$ solves a sequence of maximization problems, which cannot be written as a single optimization problem over a sample average. However, by converting each step into a Z-estimation problem, we can write $\widehat{\theta}$ as the solution to the estimating equation

$$\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \nabla_{\theta_X} \log f_X(X_i; \theta_X) \\ \nabla_{\theta_Y} \log f_Y(Y_i; \theta_Y) \\ \nabla_{\theta_C} \log c(F_X(X_i; \theta_X), F_Y(Y_i; \theta_Y); \theta_C) \end{pmatrix} = 0.$$

Note that since $(\widehat{\theta}_X, \widehat{\theta}_Y)$ must solve the first two equations, the estimates are indeed equivalent to the (marginal) MLEs defined above. With these estimates being fixed, the third equation is equivalent to the second step of the 2-step procedure.

3.3 Consistency

3.3.1 Setup

In the above examples, it was tacitly understood that the M- and Z-estimators are sensible estimators of the quantities they're designed to estimate. But how do we know that's really the case? As we have discussed in the last section, a minimal condition for an estimator $\widehat{\theta}$ of some parameter θ_0 to be considered sensible, is consistency:

$\widehat{\theta} \to_p \theta_0$ as $n \to \infty$. Gladly, there is a very general theory of consistency for M- and Z-estimators.

To set things up, we assume that the true parameter θ_0 is the maximizer of some population criterion function $M\colon \Theta \to \mathbb{R}$, i.e.

$$\theta_0 = \arg\max_{\theta \in \Theta} M(\theta).$$

To construct an estimator, we replace the population criterion $M(\theta)$ by a sample criterion $M_n(\theta)$, and define

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} M_n(\theta).$$

Remark 3.3.1. *For M-estimators as defined above, the population criterion is $M(\theta) = \mathbb{E}[m_\theta(X)]$. For Z-estimators, $M(\theta) = \|\mathbb{E}[\psi_\theta(X)]\|$ is a valid choice. The population criterion involves an expectation over an unknown probability measure for X. The estimators are constructed by replacing the expectation by a sample average.*

3.3.2 Why pointwise convergence is not enough

The sample criterion function $M_n(\theta)$ is a random function, because it depends on the random sample $X_1, \dots, X_n$. To analyze the consistency of the estimator $\widehat{\theta}$, we need to understand how the sample criterion function behaves as $n \to \infty$. We usually have $M_n(\theta) \to_p M(\theta)$ for all $\theta \in \Theta$ by the law of large numbers. However, this is not sufficient to guarantee that the estimator $\widehat{\theta}$ is consistent. Informally, the reason is that we are comparing infinitely many possible values of $M_n(\theta)$ at the same time. While, in isolation, each $M_n(\theta)$ may be close to $M(\theta)$ with high probability, they are not necessarily close to each other simultaneously.

To illustrate this point, consider the following example. Let $\Theta = \{\theta_1, \dots, \theta_K\}$ be a finite set with K elements. Suppose that $M_n(\theta_1), \dots, M_n(\theta_K)$ are independent random variables with distribution $\mathcal{N}(M(\theta_k), 1/\sqrt{n})$ for all $k = 1, \dots, K$. A simple application of Markov's inequality shows $M_n(\theta_k) - M(\theta_k) \to_p 0$ for every $k = 1, \dots, K$. Now consider the uniform distance $\max_{1 \le k \le K} |M_n(\theta_k) - M(\theta_k)|$. It holds

$$\begin{aligned}
&\mathbb{P}\left(\max_{1 \le k \le K} |M_n(\theta_k) - M(\theta_k)| \le \varepsilon\right) \\
&= \mathbb{P}(|M_n(\theta_1) - M(\theta_1)| \le \varepsilon, \dots, |M_n(\theta_K) - M(\theta_K)| \le \varepsilon) \\
&= \prod_{k=1}^{K} \mathbb{P}(|M_n(\theta_k) - M(\theta_k)| \le \varepsilon) && \text{[independence of } M_n(\theta_k)\text{]} \\
&= [\Phi(\varepsilon/\sqrt{n}) - \Phi(-\varepsilon/\sqrt{n})]^K,
\end{aligned}$$

where Φ is the standard normal CDF. For any fixed K, the probability on the right-hand side converges to 1 as $n \to \infty$. However, this is no longer the case when we take

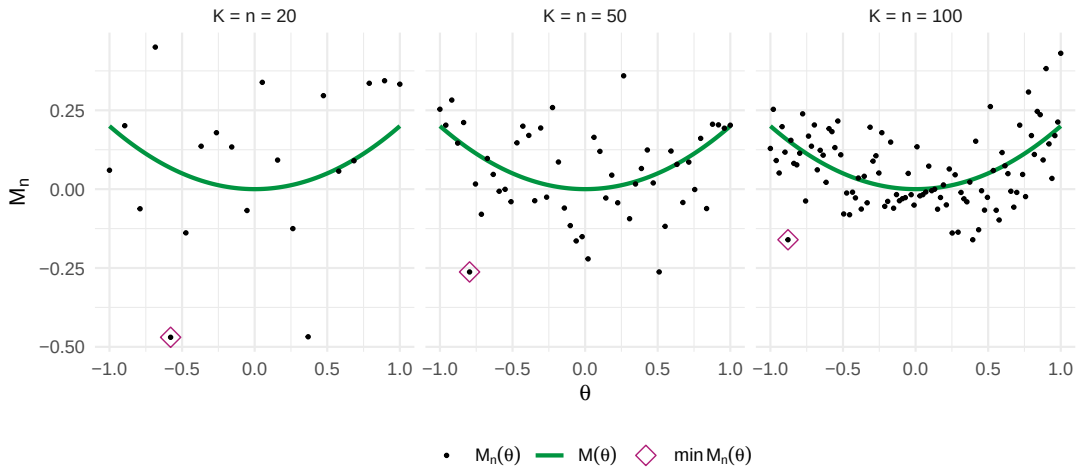

Figure 3.1: Illustration of the failure of pointwise convergence. The more values of $M_n(\theta)$ we compare, the more likely it is that at least one of them is far away from $M(\theta)$; the corresponding θ may be the minimizer.

$K \to \infty$:

$$\lim_{K\to\infty} \mathbb{P}\left(\max_{1\le k\le K} |M_n(\theta_k) - M(\theta_k)| \le \varepsilon\right) = 0.$$

Simply put, because we compare so many values $\{M_n(\theta) \colon \theta \in \Theta\}$, at the same time, there is a high probability that at least one of them is far away from $M(\theta)$. We can then no longer treat M_n as a sufficiently good approximation of M when optimizing over the entire loss surface. To illustrate how this can go wrong, a simulated example of the above toy model with $M(\theta) = \theta^2/5$ and $n = K$ is shown in Fig. 3.1. On the left ($n = 20$), the variance is still very high, and the minimizer is far away from $\theta_0 = 0$. As we move to the right, the variance decreases. But because we compare so many values of $M_n(\theta)$, there's always one damn θ_k far away from θ_0 that minimizes $M_n(\theta)$.

Not all hope is lost. The example above has an unrealistic feature. The functions $M_n(\theta)$ are usually continuous, so it is unlikely that the values $M_n(\theta)$ and $M_n(\theta')$ are far away from each other if θ and θ' are close. If that's the case, the two values $M_n(\theta)$ and $M_n(\theta')$ should be strongly dependent, not independent! Fig. 3.2 shows the same setting, but with $\operatorname{Corr}(M_n(\theta), M_n(\theta')) = 1 - |\theta - \theta'|/10$ for all $\theta \neq \theta'$. Although most $M_n(\theta)$ may be somewhat far away from $M(\theta)$, their distance from $M(\theta)$ is always similar. This is a consequence of the dependence between close-by values of M_n. As a result, their relative magnitude is mostly preserved, and the sample maximizer $\widehat{\theta}$ is also close to the true value $\theta_0 = 0$. Further, it doesn't really matter how many θ_k we compare, because they all lie on a smooth curve.

What we observe here is *uniform convergence* (in probability):

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \to_p 0.$$

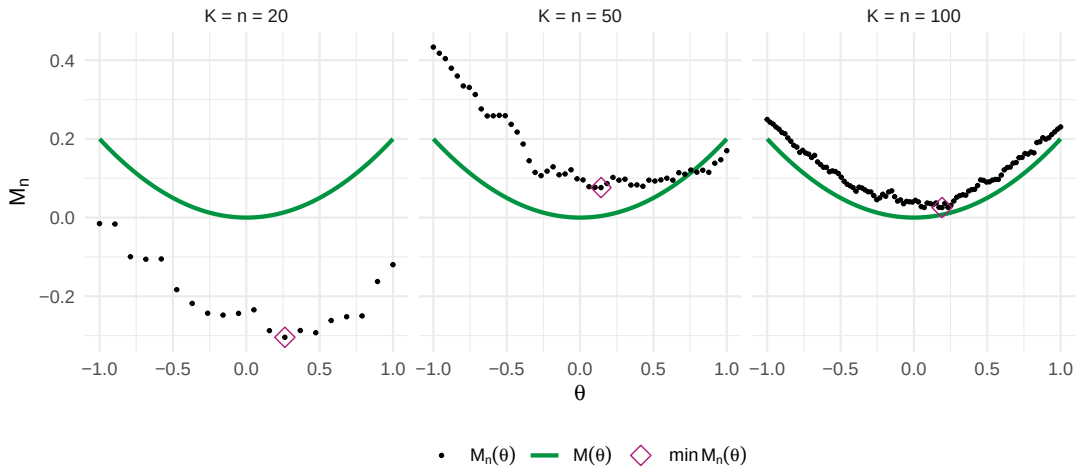

Figure 3.2: Illustration of the uniform convergence. No matter how many values of $M_n(\theta)$ we compare, their distance from $M(\theta)$ is uniformly small.

Uniform convergence is a much stronger property than pointwise convergence. All the distances $|M_n(\theta) - M(\theta)|$ must be small at the same time. As we shall see, this is exactly what we need to guarantee the consistency of the estimator $\widehat{\theta} = \arg\max M_n(\theta)$.

3.3.3 Main results

Theorem 3.3.2 (Consistency of M-estimators)**.** *Let*

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} M_n(\theta), \quad \theta_0 = \arg\min_{\theta\in\Theta} M(\theta).$$

Suppose that:

(i) The minimum is well-separated: for every $\varepsilon > 0$,

$$\inf_{\|\theta-\theta_0\|\geq\varepsilon} M(\theta) > M(\theta_0);$$

(ii) The sample criterion M_n converges uniformly to the population criterion M:

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \to_p 0.$$

Then $\widehat{\theta} \to_p \theta_0$.

Proof. Let $\varepsilon > 0$ be arbitrary. We want to show

$$\mathbb{P}\left\{\|\widehat{\theta} - \theta_0\| > \varepsilon\right\} \to 0 \quad \text{for all } \varepsilon > 0.$$

Because the minimum is well separated by (i), we can find $\eta > 0$ such that $M(\theta) \geq$

$M(\theta_0) + \eta$. We now have the following implications:

$$\begin{aligned}
& \|\widehat{\theta} - \theta_0\| > \varepsilon \\
\Rightarrow \quad & \inf_{\|\theta-\theta_0\|>\varepsilon} M_n(\theta) \leq \inf_{\|\theta-\theta_0\|\leq\varepsilon} M_n(\theta) \\
\Rightarrow \quad & \inf_{\|\theta-\theta_0\|>\varepsilon} M_n(\theta) \leq M_n(\theta_0) && [\inf_{\|\theta-\theta_0\|\leq\varepsilon} M(\theta) \leq M(\theta_0)] \\
\Rightarrow \quad & \inf_{\|\theta-\theta_0\|>\varepsilon} M_n(\theta) - M(\theta_0) \leq M_n(\theta_0) - M(\theta_0) \\
\Rightarrow \quad & \inf_{\|\theta-\theta_0\|>\varepsilon} M_n(\theta) - M(\theta) \leq M_n(\theta_0) - M(\theta_0) - \eta && [M(\theta) - \eta \geq M(\theta_0)] \\
\Rightarrow \quad & \inf_{\|\theta-\theta_0\|>\varepsilon} [M_n(\theta) - M(\theta)] - [M_n(\theta_0) - M(\theta_0)] \leq -\eta \\
\Rightarrow \quad & 2 \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \geq \eta
\end{aligned}$$

Thus,

$$\mathbb{P}\left\{\|\widehat{\theta} - \theta_0\| > \varepsilon\right\} \leq \mathbb{P}\left\{2 \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \geq \eta\right\} \to 0. \qquad \square$$

The well-separatedness condition on the population criterion $M(\theta)$ is mild. For example, it is satisfied if the population criterion M is continuous and the maximum is unique. A similar consistency result for Z-estimators follows as an easy corollary.

Theorem 3.3.3 (Consistency of Z-estimators). *Let $\Psi_n \colon \Theta \to \mathbb{R}^p$ be the sample criterion function of a Z-estimator, $\Psi \colon \Theta \to \mathbb{R}^p$ the population criterion, and $\widehat{\theta}$ and θ_0 be the solutions of the equations*

$$\Psi_n(\widehat{\theta}) = 0, \quad \Psi(\theta_0) = 0.$$

Suppose that:

(i) The zero is well-separated: for every $\varepsilon > 0$,

$$\inf_{\|\theta-\theta_0\|\geq\varepsilon} \|\Psi(\theta)\| > 0;$$

(ii) The sample criterion Ψ_n converges uniformly to the population criterion Ψ:

$$\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \to_p 0.$$

Then $\widehat{\theta} \to_p \theta_0$.

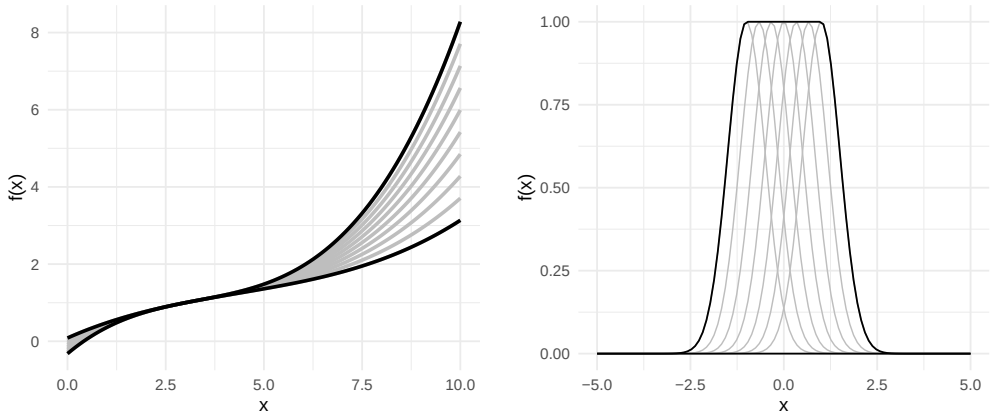

Figure 3.3: Illustration of brackets (black) for some sets of functions (grey).

Proof. Take $M_n(\theta) = \|\Psi_n(\theta)\|$ and $M(\theta) = \|\Psi(\theta)\|$. It holds

$$\begin{aligned}\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| &= \sup_{\theta\in\Theta} |\|\Psi_n(\theta)\| - \|\Psi(\theta)\|| \\ &\leq \sup_{\theta\in\Theta} |\|\Psi_n(\theta) - \Psi(\theta)\|| \qquad \text{[reverse triangle inequality]} \\ &\to_p 0.\end{aligned}$$

The result now follows from the previous theorem. □

3.3.4 Uniform laws of large numbers

The uniform convergence condition in the previous theorems is more demanding. Setting $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$, we can rewrite the uniform convergence condition as

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n m_\theta(X_i) - \mathbb{E}[m_\theta(X)] \right| \to_p 0.$$

This is a *uniform law of large numbers*, and classes of functions $\{m_\theta \colon \theta \in \Theta\}$ that satisfy such a law are sometimes called *Glivenko-Cantelli classes*. The uniform law is a much stricter statement than the usual law of large numbers applied to any point separately. The illustrative examples from Fig. 3.1 shows a situation where pointwise convergence holds, but uniform convergence fails. Establishing uniform convergence is often the most challenging part of proving consistency. It is a core topic in *empirical process theory*, an advanced subject that deserves its own course. Without going into too much detail, we can get quite reasonable results with rather elementary tools.

Definition 3.3.4 (Bracketing numbers). *Let $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ be a set of functions.*

- *A bracket $[\underline{f}, \overline{f}]$ is the set of all $f \in \mathcal{F}$ such that $\underline{f}(x) \leq f(x) \leq \overline{f}(x)$ $\forall x \in \mathcal{X}$.*
- *We call $[\underline{f}, \overline{f}]$ an ε-bracket (with respect to a norm $\|\cdot\|$ on $\mathcal{F}$) if $\|\overline{f} - \underline{f}\| \leq \varepsilon$.*
- *The minimal number N of ε-brackets needed to cover $\mathcal{F}$, i.e, $\mathcal{F} \subseteq \bigcup_{k=1}^{N} [\underline{f}_k, \overline{f}_k]$, is called the bracketing number of $\mathcal{F}$ and is denoted by $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$.*

Examples of brackets are illustrated in Fig. 3.3. Bracketing numbers measure the 'size' or 'complexity' of a class of functions. If $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is small, then $\mathcal{F}$ is 'simple'. Up to ε-error in the norm $\|\cdot\|$, all functions in $\mathcal{F}$ can be represented by only a few functions. If $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is large, then $\mathcal{F}$ is 'rich'. The behavior of the functions in $\mathcal{F}$ is so diverse that we need many brackets to represent them well. The size of the bracket is determined by the norm. The most relevant norms for us are the $L_q(P)$-norms, defined as

$$\|f - g\|_{L_q(P)} = \mathbb{E}_{X \sim P}[|f(X) - g(X)|^q]^{1/q} = \left(\int |f(x) - g(x)|^q \, dP(x) \right)^{1/q}.$$

We have the following result.

Theorem 3.3.5. *$\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ be a set of functions. Suppose that:*

(i) $\mathbb{E}[\sup_{f \in \mathcal{F}} |f(X)|] < \infty$;

(ii) $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ *for every* $\varepsilon > 0$.

Then

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \to_p 0.$$

Proof. Define

$$\mathbb{P}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \quad P(f) = \mathbb{E}[f(X)].$$

Fix some $\varepsilon > 0$ and choose finitely many ε-brackets $\{[\underline{f}_k, \overline{f}_k]\}_{k=1}^{N(\varepsilon)}$ as in assumption (ii). It then holds

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n(f) - P(f)| \leq \max_{1 \leq k \leq N(\varepsilon)} \sup_{\underline{f}_k \leq f \leq \overline{f}_k} |\mathbb{P}_n(f) - P(f)|.$$

Now for any $\underline{f}_k \leq f \leq \overline{f}_k$,

$$\begin{aligned} \mathbb{P}_n(f) - P(f) \leq \mathbb{P}_n(\overline{f}_k) - P(\underline{f}_k) &= \mathbb{P}_n(\overline{f}_k) - P(\overline{f}_k) + P(\overline{f}_k - \underline{f}_k) \\ &\leq \mathbb{P}_n(\overline{f}_k) - P(\overline{f}_k) + \varepsilon. \end{aligned}$$

Using the same argument for a lower bound shows that

$$\mathbb{P}_n(f) - P(f) \geq -|\mathbb{P}_n(\underline{f}_k) - P(\underline{f}_k)| - \varepsilon,$$

and, thus,

$$\sup_{\underline{f}_k \leq f \leq \overline{f}_k} |\mathbb{P}_n(f) - P(f)| \leq |\mathbb{P}_n(\overline{f}_k) - P(\overline{f}_k)| + |\mathbb{P}_n(\underline{f}_k) - P(\underline{f}_k)| + \varepsilon.$$

Because $\underline{f}_k, \overline{f}_k$ are fixed, the usual law of large numbers gives

$$\sup_{\underline{f}_k \leq f \leq \overline{f}_k} |\mathbb{P}_n(f) - P(f)| = \varepsilon + o_p(1).$$

Furthermore, because $N(\varepsilon)$ is finite, the continuous mapping theorem implies that

$$\max_{1 \leq k \leq N(\varepsilon)} \sup_{\underline{f}_k \leq f \leq \overline{f}_k} |\mathbb{P}_n(f) - P(f)| \leq \varepsilon + o_p(1).$$

The claim follows upon choosing ε arbitrarily small. □

We see that we only need to be able to cover the class of functions $\mathcal{F}$ with finitely many brackets (at every scale ε) to establish the uniform law of large numbers. This fails, for example, if $\mathcal{F}$ is the set of all measurable functions $f \colon \mathcal{X} \to \mathbb{R}$. This set of functions is simply too large. However, condition (ii) is satisfied for many sets of functions. One of the most important examples are parametrized families of functions.

Lemma 3.3.6 (Functions Lipschitz in a parameter)**.** *Let $\Theta \subseteq \{\|\theta\| \leq K\} \subset \mathbb{R}^p$ and $\mathcal{F} = \{f_\theta \colon \theta \in \Theta\}$ be a parametrized family of functions. Suppose that there is a function Λ and $q \in \mathbb{N}$ such that*

$$|f_\theta(x) - f_{\theta'}(x)| \leq \Lambda(x)\|\theta - \theta'\|, \quad \mathbb{E}[\Lambda^q(X)] < \infty.$$

Then for any $\varepsilon > 0$, it holds

$$N_{[]}(\varepsilon, \mathcal{F}, L_q(P)) \leq \left(\frac{6K\|\Lambda\|_{L_q(P)}}{\varepsilon} \right)^p.$$

Proof. Set $\eta = \varepsilon / 2\|\Lambda\|_{L_q(P)}$. Because $\Theta \subseteq \{\|\theta\| \leq K\}$, we can cover it by finitely many η-balls: $\Theta \subseteq \bigcup_{k=1}^{N(\eta)} B_\eta(\theta_k)$. In fact, it is known that we need at most $N(\eta) = (3K/\eta)^p$ many. Define $\underline{f}_k(x) = f_{\theta_k}(x) - \Lambda(x)\eta$ and $\overline{f}_k(x) = f_{\theta_k}(x) + \Lambda(x)\eta$. Then for any $\theta \in B_\eta(\theta_k)$, we have

$$\underline{f}_k(x) \leq f_\theta(x) \leq \overline{f}_k(x)$$

and

$$|\overline{f}_k(x) - \underline{f}_k(x)| \leq 2\Lambda(x)\eta.$$

Thus,

$$\|\overline{f}_k - \underline{f}_k\|_{L_q(P)} \leq 2\|\Lambda\|_{L_q(P)}\eta = \varepsilon.$$

We have shown that $N(\eta) = (3K/\eta)^p = (6K\|\Lambda\|_{L_q(P)}/\varepsilon)^p$ brackets are sufficient to cover $\mathcal{F}$. □

The bracketing number of the class is finite for every ε. Hence condition (ii) of Theorem 3.3.5 is satisfied for such classes. This is quite useful. For example, it allows us to establish the consistency of MLEs in parametric models.

Example 3.3.7. *Let $\mathcal{F} = \{f_\theta \colon \theta \in \Theta \subset \mathbb{R}^p\}$, be density functions in a parametric model. A first-order Taylor expansion gives*

$$|\log f_\theta(x) - \log f_{\theta'}(x)| = |\nabla \log f_{\theta^*}(x)(\theta - \theta)| \leq \|\nabla \log f_{\theta^*}(x)\|\|\theta - \theta\|,$$

for some θ^* on the line segment from θ to θ', i.e., $\theta^* = \theta + t(\theta' - \theta)$ for some $t \in (0, 1)$. We may therefore take $\Lambda(x) = \sup_{\theta\in\Theta} \|\nabla \log f_{\theta^*}(x)\|$.

A second common example are monotone functions.

Lemma 3.3.8. *Let $\mathcal{F} = \{f \colon \mathbb{R} \to [0, 1]\}$ be a set of monotone functions. There is a constant $K_q < \infty$ such that*

$$N_{[]}(\varepsilon, \mathcal{F}, L_q(P)) \leq K_q \exp(1/\varepsilon).$$

Proof. The proof is quite technical; see van der Vaart and Wellner (2023, Theorem 2.7.9). □

The result can also be applied to functions mapping to any other closed interval, by centering and scaling the functions. The set of monotone functions seems to be much larger than the previous example: the bracketing number grows exponentially with $1/\varepsilon$. Nevertheless, it is finite so that the uniform law of large numbers applies. A host of results on bracketing numbers of different function classes can be found in van der Vaart and Wellner (2023, Section 2.7).

3.3.5 Examples

Proposition 3.3.9. *Let $f_\theta, \theta \in \Theta \subset \mathbb{R}^p$, be density functions in a parametric model. Suppose Θ is bounded, the true parameter θ_0 is a well-separated maximizer of the population criterion $M(\theta) = \mathbb{E}[\log f_\theta(X)]$, and*

$$\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla \log f_\theta(X)\|\right] < \infty,$$

Then the MLE

$$\widehat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i)$$

is consistent: $\widehat{\theta}_{MLE} \to_p \theta_0$.

Proof. By Example 3.3.7, Lemma 3.3.6, and Theorem 3.3.5, it holds

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i) - \mathbb{E}[\log f_\theta(X_i)] \right| \to_p 0.$$

Now the result follows from Theorem 3.3.2. □

A sufficient condition for well-separatedness is that the expected Hessian is negative definite near the maximum:

$$\lambda_{\max}(\mathbb{E}[\nabla^2 \log f_\theta(X)]) < 0, \quad \text{for all } \|\theta - \theta_0\| < \delta.$$

As a simple corollary, we get the consistency of the ordinary least-squares estimator.

Corollary 3.3.10 (Consistency of OLS estimator)**.** *Let $Y_i = \beta^\top X_i + \varepsilon_i$ be a linear regression model with $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ and $\mathbb{E}[\varepsilon_i^2 \mid X_i] < C < \infty$, and $\beta_0 = \arg\min_{\beta\in\mathbb{R}^p} \mathbb{E}[(Y - X\beta)^2]$. If the matrix $\Sigma = \mathbb{E}[XX^\top] \in \mathbb{R}^{p\times p}$ is positive definite with $\|\Sigma\| < \infty$ and the parameter space is bounded, the OLS estimator*

$$\widehat{\beta}_{OLS} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i\beta)^2$$

is consistent: $\widehat{\beta}_{OLS} \to_p \beta_0$.

Proof. Note that the OLS estimator is the MLE in a linear regression model with Gaussian errors. It holds $\nabla^2 \log f_\theta(X) = -XX^\top$ and $\mathbb{E}[XX^\top] = \Sigma$. Because the latter is positive definite, the maximum is well separated. Further,

$$\|\nabla \log f_\theta(X)\| \le |Y - X^\top\beta| \|X\| \le Y\|X\| + \sup_{\beta\in\mathcal{B}} \|\beta\| \|X\|^2.$$

The expectation of this is bounded by assumption. Now the result follows from Proposition 3.3.9. □

As our last example for this section, we establish the consistency of sample quantiles.

Proposition 3.3.11 (Consisteny of sample quantiles)**.** *Let*

$$\rho_\alpha(x) = \begin{cases} \alpha|x| & \text{if } x \geq \theta, \\ (1-\alpha)|x| & \text{if } x < \theta, \end{cases} \qquad \text{and} \qquad m_\theta(x) = \rho_\alpha(x-\theta).$$

The α-quantile of X can be identified as $\theta_\alpha = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}[m_\theta(X)]$. If the minimum is well-separated, the α-sample quantile

$$\widehat{\theta}_\alpha = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \rho_\alpha(x-\theta),$$

is consistent: $\widehat{\theta}_\alpha \to_p \theta_\alpha$.

Proof. Let $m_\theta(x) = \rho_\alpha(x-\theta)$. Note that for $\theta \leq \theta' \leq x$, we have

$$|\rho_\alpha(x-\theta) - \rho_\alpha(x-\theta')| \leq \alpha||x-\theta| - |x-\theta'|| \leq \alpha|\theta-\theta'| \leq |\theta-\theta'|$$

by the reverse triangle inequality and $\alpha \in [0,1]$. A similar argument for the cases $x \leq \theta \leq \theta'$ and $\theta \leq x \leq \theta'$ shows that the conditions of Lemma 3.3.6 are satisfied with $\Lambda(x) \equiv 1$. Now Example 3.3.7, Lemma 3.3.6, and Theorem 3.3.5 yield

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \rho_\alpha(X_i-\theta) - \mathbb{E}[\rho_\alpha(X-\theta)]\right| \to_p 0,$$

and the result follows from Theorem 3.3.2. □

3.3.6 Remarks

The consistency results can be extended in several ways. Compactness of the parameter space Θ can be relaxed with some additional work. In a first step, we have to prove that the estimator $\widehat{\theta}_n$ is *uniformly tight*: for every $\varepsilon > 0$, there is a constant K such that $\sup_n \mathbb{P}(\widehat{\theta}_n > K) \leq \varepsilon$. The uniform convergence condition

$$\sup_{\theta\in S}|M_n(\theta) - M(\theta)| \to_p 0,$$

is only needed to hold for all compact subsets $S \subset \Theta$. Further, the population criterion, the loss functions, and the parameter space may change with n. This is commonly the case for spline models, where the number of basis functions increases with the sample size. This makes the results more technical but does not fundamentally change the argument.

We have used bracketing numbers to establish a uniform law of large numbers. There are other ways to achieve this. Two other prominent methods based on *uniform covering numbers* and *Rademacher complexity* are discussed extensively in my course on *Statistical Learning Theory*.

3.4 Asymptotic normality

As mentioned earlier, consistency should be considered a minimal requirement for any reasonable estimator. However, many estimators are consistent, and this alone does not tell us much about the estimators' quality. To assess accuracy or uncertainty we have to look at its distribution. The statement $\widehat{\theta} \to_p \theta_0$ only tells us that the limiting distribution is a point-mass, which isn't very helpful. To get a useful limit, we have to rescale the estimator in a way that the limiting distribution is non-degenerate. The central limit theorem suggests $\sqrt{n}$ as a natural scaling factor. This is typically the right scaling when a finite number of parameters are estimated. In the case of estimating a function, the scaling factor might be different. The limiting distribution is most often normal:

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathcal{N}(0, \Sigma).$$

To get an idea where this is coming from, we first give an informal argument in a simple case.

3.4.1 An informal argument

Asymptotic normality is most easily proved in the Z-estimator formulation: $\Psi_n(\widehat{\theta}) = 0$. Suppose $\Theta \subseteq \mathbb{R}$ and that ψ_θ is sufficiently differentiable. Recall the first-order Taylor expansion of a function $f\colon \mathbb{R} \to \mathbb{R}$ around a point x_0:

$$f(x) = f(x_0) + f'(\widetilde{x})(x - x_0),$$

for some $\widetilde{x} \in [x_0, x]$. It holds

$$0 = \Psi_n(\widehat{\theta}) = \Psi_n(\theta_0) + \Psi_n'(\widetilde{\theta})(\widehat{\theta} - \theta_0) \quad \Leftrightarrow \quad \sqrt{n}(\widehat{\theta} - \theta_0) = \frac{\sqrt{n}\Psi_n(\theta_0)}{\Psi_n'(\widetilde{\theta})}.$$

By the central limit theorem,

$$\sqrt{n}\Psi_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) \quad \to_d \quad \mathcal{N}(0, \mathbb{E}[\psi_{\theta_0}(X)^2]).$$

For the denominator we expect $\Psi_n'(\widetilde{\theta}) \to_p \Psi'(\theta_0)$, because Ψ_n' is a sample average, and $\widehat{\theta} \to_p \theta_0$. This is not immediate because $\widehat{\theta}$ is a random sequence that depends on the data. Taking that as given, Slutsky's lemma (Lemma 2.3.9) gives

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \frac{\sqrt{n}\Psi_n(\theta_0)}{\Psi_n'(\widetilde{\theta})} \quad \to_d \quad \mathcal{N}(0, \mathbb{E}[\psi_{\theta_0}(X)^2]/\Psi'(\theta_0)^2).$$

3.4.2 Main result

The informal argument took several shortcuts. First, we assumed that the parameter is one-dimensional. Second, we assumed differentiabilty of ψ_θ. This is violated, for

example, for the median: $\psi_\theta = \operatorname{sign}(x-\theta)$. Third, we did not formally prove $\Psi_n'(\widetilde{\theta}) \to_p \Psi'(\theta_0)$, which is the hardest part.

For a more general result, we first need a generalization of the Taylor expansion to multivariate functions. For any continuously differentiable $f\colon \mathbb{R}^p \to \mathbb{R}^q$, it holds

$$f(x) = f(x_0) + \nabla f(x_0 + t(x - x_0))(x - x_0), \quad \text{for some } t \in (0,1),$$

where $(\nabla f(x))_{kj} = \partial f_k(x)/\partial x_j$ is the Jacobian of f at x. This Taylor expansion (and high-order versions) is one of the most powerful tools in asymptotic statistics. Assuming sufficient regularity and $X \to_p X_0$, we may replace $f(X)$ by $f(X_0) + O_p(\|X - X_0\|)$. Additionally, we will need something similar to the uniform convergence condition for the consistency proof. But now we need something stronger, *stochastic equicontinuity* of the centered and scaled criterion at θ_0:

$$\sup_{\|\theta-\theta_0\|\le\delta_n} \sqrt{n}\,\|[\Psi_n(\theta) - \Psi(\theta)] - [\Psi_n(\theta_0) - \Psi(\theta_0)]\| \to_p 0, \quad \text{for all } \delta_n \to 0. \tag{3.1}$$

This is a much stronger condition than uniform convergence, not least because of the $\sqrt{n}$-blowup factor.

Theorem 3.4.1 (Asymptotic normality of Z-estimators)**.** *Let $\widehat{\theta}$ be the solution to the estimating equation*

$$\Psi_n(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n \psi_{\widehat{\theta}}(X_i) = 0.$$

(i) The estimator $\widehat{\theta}$ is consistent: $\widehat{\theta} \to_p \theta_0$.

(ii) The function $\theta \mapsto \Psi(\theta) = \mathbb{E}[\psi_\theta(X)]$ is continuously differentiable with invertible Jacobian $\dot{\Psi}(\theta_0) = \nabla_\theta \mathbb{E}[\psi_{\theta_0}(X)]$.

(iii) The stochastic equicontinuity condition (3.1) holds.

Then

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -\nabla\Psi(\theta_0)^{-1}\sqrt{n}\Psi_n(\theta_0) + o_p(1) \tag{3.2}$$

and

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathcal{N}(0, \Sigma), \quad \text{with} \quad \Sigma = \dot{\Psi}(\theta_0)^{-1}\mathbb{E}[\psi_{\theta_0}(X)\psi_{\theta_0}(X)^\top]\dot{\Psi}(\theta_0)^{-\top}.$$

Proof.

- Decompose

$$\begin{aligned} 0 = \Psi_n(\widehat{\theta}) &= [\Psi_n(\widehat{\theta}) - \Psi_n(\theta_0)] + \Psi_n(\theta_0) \\ &= [\Psi_n(\widehat{\theta}) - \Psi_n(\theta_0)] - [\Psi(\widehat{\theta}) + \Psi(\theta_0)] + [\Psi(\widehat{\theta}) - \Psi(\theta_0)] + \Psi_n(\theta_0). \end{aligned}$$

- We will first show that the difference of the first brackets is negligible. Because $\widehat{\theta} \to_p \theta_0$, there is a sequence $\delta_n \to 0$ such that $\|\widehat{\theta} - \theta_0\| \leq \delta_n$ with probability going to 1. In this event,

$$\begin{aligned}\left|[\Psi_n(\widehat{\theta}) - \Psi_n(\theta_0)] - [\Psi(\widehat{\theta}) + \Psi(\theta_0)]\right| &\leq \sup_{\|\theta-\theta_0\|\leq\delta_n} \|[\Psi_n(\theta) - \Psi(\theta)] - [\Psi_n(\theta_0) - \Psi(\theta_0)]\| \\ &= o_p(1/\sqrt{n}),\end{aligned}$$

  by the stochastic equicontinuity condition (3.1).

- For the third term, we have

$$\Psi(\widehat{\theta}) - \Psi(\theta_0) = \nabla\Psi(\widetilde{\theta})(\widehat{\theta} - \theta_0).$$

  Because $\widetilde{\theta} \to_p \theta_0$, and $\theta \mapsto \nabla\Psi(\theta) = \mathbb{E}[\nabla\psi_\theta(X)]$ is continuous, we have $\nabla\Psi(\widetilde{\theta}) - \nabla\Psi(\theta_0) \to_p 0$ by the continuous mapping theorem. Thus,

$$\Psi(\widehat{\theta}) - \Psi(\theta_0) = \nabla\Psi(\theta_0)(\widehat{\theta} - \theta_0) + o_p(\|\widehat{\theta} - \theta_0\|).$$

- We have shown that

$$\begin{aligned}0 &= [\Psi_n(\widehat{\theta}) - \Psi_n(\theta_0)] - [\Psi(\widehat{\theta}) + \Psi(\theta_0)] + [\Psi(\widehat{\theta}) - \Psi(\theta_0)] + \Psi_n(\theta_0) \\ &= \nabla\Psi(\theta_0)(\widehat{\theta} - \theta_0) + \Psi_n(\theta_0) + o_p(1/\sqrt{n} + \|\widehat{\theta} - \theta_0\|),\end{aligned}$$

  which is equivalent to

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -\nabla\Psi(\theta_0)^{-1}\sqrt{n}\Psi_n(\theta_0) + o_p(1 + \sqrt{n}\|\widehat{\theta} - \theta_0\|).$$

  The o_p terms are negligible, and by the central limit theorem and continuous mapping, the first term converges in distribution to a normal distribution with variance Σ. □

The second condition of the theorem still requires differentiability, but only for the expectation $\mathbb{E}[\psi_\theta(X)]$. This is fine for the median, because $\mathbb{E}[\text{sign}(X - \theta)] = \mathbb{P}(X < \theta) - \mathbb{P}(X > \theta)$ for all θ. The derivative is $2f_X(\theta)$ where p is the density of X. If this density is continuous and bounded away from zero, condition (ii) is satisfied.

The statement (3.2) says that $\widehat{\theta} - \theta_0$ is first-order equivalent to a (scaled) sample average over the functions ψ_{θ_0}. First-order equivalence means that the difference is negligible compared to the term dominating the behavior (here the sample average). When an estimator is first-order equivalent to a sample average, it is called *asymptotically linear*. This is a pleasant property many estimators enjoy. As soon as an estimator is asymptotically linear, asymptotic normality follows directly from the central limit theorem.

3.4.3 Conditions for stochastic equicontinuity

Stochastic equicontinuity can be guaranteed with similar tools as for uniform convergence. We start with a general result that bounds the supremum over differences between a sample average and expectation. Proving this is really hard, so we will not waste our time with this. See Chapter 19, specifically Corollary 19.35, of Van der Vaart (2000) for more details.

Lemma 3.4.2. *Let $\mathcal{F}$ by a class of functions with* envelope *F, i.e., $|f(x)| \leq F(x)$ for all $f \in \mathcal{F}$. Then if for some $C < \infty$, $\alpha \in (0, 2)$,*

$$\ln N_{[]}(\varepsilon \|F\|_{L_2(P)}, \mathcal{F}, L_2(P)) \leq C\varepsilon^{-\alpha},$$

it holds

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| = O_p\left(\frac{\|F\|_{L_2(P)}}{\sqrt{n}} \right).$$

We can now state the main result on stochastic equicontinuity of the criterion functions.

Theorem 3.4.3. *Let $\mathcal{F}_\delta = \{\psi_\theta - \psi_{\theta_0} \colon \|\theta - \theta_0\| \leq \delta\}$ and F_{δ_n} be an* envelope *for the class, i.e., $|f(x)| \leq F_{\delta_n}(x)$ for all $f \in \mathcal{F}_\delta$ and all x. Suppose*

(i) $\lim_{\delta \searrow 0} \|F_\delta\|_{L_2(P)} = 0$.

(ii) $\ln N_{[]}(\varepsilon \|F_\delta\|_{L_2(P)}, \mathcal{F}_\delta, L_2(P)) \leq C\varepsilon^{-\alpha}$ *for some $C < \infty, \alpha \in (0, 2)$.*

Then the stochastic equicontinuity condition (3.1) holds.

Proof. It holds

$$\begin{aligned} T_n &:= \sup_{\|\theta - \theta_0\| \leq \delta_n} \left| [\Psi_n(\widehat{\theta}) - \Psi_n(\theta_0)] - [\Psi(\widehat{\theta}) + \Psi(\theta_0)] \right| \\ &= \sup_{\|\theta - \theta_0\| \leq \delta_n} \left| \frac{1}{n} \sum_{i=1}^{n} [\psi_\theta(X_i) - \psi_{\theta_0}(X_i)] - \mathbb{E}[\psi_\theta(X) - \psi_{\theta_0}(X)] \right|. \end{aligned}$$

By our assumption on the bracketing numbers and Lemma 3.4.2,

$$T_n = O_p(\|F_{\delta_n}\|_{L_2(P)}/\sqrt{n}) = o_p(1/\sqrt{n}). \qquad \square$$

The stochastic equicontinuity condition is fairly easy to establish if the estimating functions ψ_θ are smooth in the parameter. This is typically the case. However, smoothness is not necessary. Let's see one example of each case.

Example 3.4.4 (Smooth models)**.** *Consider identifying functions $\psi_\theta(x) \in \mathbb{R}^p$ where each component $\psi_\theta(x)_k$ satisfies the conditions of Lemma 3.3.6 for some function $\Lambda_k(x)$.*

It suffices to show that the stochastic equicontinuity condition holds for every component of the estimating equation. Because

$$\sup_{\|\theta-\theta'\|\le\delta_n} |\psi_\theta(x)_k - \psi_{\theta'}(x)_k| \le \Lambda(x)\delta_n,$$

so we may take $F_{\delta_n}(x) = \Lambda(x)\delta_n$ *as the envelope. It holds* $\|F_{\delta_n}\|_{L_2(P)} = O(\delta_n) = o(1)$ *for every* $\delta_n = o(1)$*, so condition (i) of Theorem 3.4.3 is satisfied. Then*

$$\ln N_{[]}(\varepsilon\|F_{\delta_n}\|_{L_2(P)}, \mathcal{F}_\delta, L_2(P)) \le p\ln(6/\varepsilon) = O(\varepsilon^{-\alpha})$$

for any $\alpha > 0$*, and condition (ii) of Theorem 3.4.3 is satisfied.*

Example 3.4.5. *The median has* $\psi_\theta(x) = \operatorname{sign}(x-\theta)$*, which is not continuous. First, we construct an envelope. For simplicity let* $\Theta = [0,1]$*. Then*

$$\begin{aligned}\sup_{|\theta-\theta_0|\le\delta_n} |\operatorname{sign}(x-\theta) - \operatorname{sign}(x-\theta_0)| &\le \sup_{|\theta-\theta_0|\le\delta_n} 2\mathbb{1}\{x\in[\theta,\theta_0]\}\\ &\le 2\mathbb{1}\{x\in[\theta_0-\delta_n,\theta_0+\delta_n]\}\\ &=: F_{\delta_n}(x).\end{aligned}$$

If the density f_X *of* X *is continuous and bounded by* $C < \infty$ *near the median* θ_0*, it holds*

$$\|F_{\delta_n}\|_{L_2(P)} = 2\mathbb{E}[\mathbb{1}\{x\in[\theta_0-\delta_n,\theta_0+\delta_n]\}] = 2\mathbb{P}(\theta_0-\delta_n \le X \le \theta_0+\delta_n) \le 2C\delta_n = o(1),$$

so condition (i) of Theorem 3.4.3 is satisfied. Condition (ii) follows from Lemma 3.3.8.

In fact, the signs are particularly simple monotone functions, and it is easy to construct a much smaller bracketing explicitly. Let's do this for sake of illustration. Choose an equally spaced grid $\theta_j = j\varepsilon$ *for* $j = 0,\dots,\lceil 1/\varepsilon\rceil - 1$*. Define*

$$\underline{f}_k(x) = \operatorname{sign}(x-\theta_k), \qquad \overline{f}_k(x) = \operatorname{sign}(x-\theta_{k-1}).$$

Because $\theta_{k-1} \le \theta_k$ *and* $\operatorname{sign}(x-\theta)$ *is decreasing in* θ*, it holds* $\underline{f}_k \le \overline{f}_k$*. By similar arguments as before, we get*

$$\|\underline{f}_k - \underline{f}_k\|_{L_2(P)} \le 2\mathbb{P}(\theta_{k-1} \le X \le \theta_k) \le 2C\varepsilon.$$

Hence, $N_{[]}(2C\varepsilon, \mathcal{F}_\delta, L_2(P)) = O(1/\varepsilon)$*, which is exponentially smaller than the bracketing for all monotone functions.*

3.4.4 Examples

Let us apply these results to some common examples. We start with the MLE.

Proposition 3.4.6. *If the assumptions of Proposition 3.3.9 hold and $\theta \mapsto \mathbb{E}[\nabla f_\theta(X)]$ is continuously differentiable,*

$$\sqrt{n}(\widehat{\theta}_{MLE} - \theta_0) \to_d \mathcal{N}(0, \Sigma),$$

with

$$\Sigma = \mathbb{E}[\nabla^2 \log f_{\theta_0}(X)]^{-1} \mathbb{E}[\nabla \log f_{\theta_0}(X) (\nabla f_{\theta_0}(X))^\top] \mathbb{E}[\nabla^2 \log f_{\theta_0}(X)]^{-1}.$$

Proof. By the assumptions of Proposition 3.3.9, the MLE is consistent and, in view of Example 3.4.4, the stochastic equicontinuity condition (3.1) holds. The result now follows from Theorem 3.4.1. □

The asymptotic variance consists of a product of three matrices. Both the inverse of the expected Hessian $\mathbb{E}[\nabla^2 \log f_{\theta_0}(X)]$ and the covariance matrix of the scores $\mathbb{E}[\nabla \log f_{\theta_0}(X)(\nabla f_{\theta_0}(X))^\top]$ are sometimes called *Fisher information matrix*. In a correctly specified model (i.e., $X_i \sim f_{\theta_0}$), the two matrices are the same, which explains the ambiguity in terminology. In this case, the asymptotic variance simplifies to $\Sigma = \mathcal{I}(\theta_0)^{-1} = \mathbb{E}[\nabla^2 \log f_{\theta_0}(X)]^{-1}$. The Fisher information matrix $\mathcal{I}(\theta_0)$ is a measure of the amount of information that an observable random variable X carries about the unknown parameter θ_0. The larger the Fisher information, the more information the data carries about the parameter. Philosophically speaking, it is unlikely that a parametric model is *exactly* correct. So to be safe, we should rather use the general expression for Σ given in the proposition to construct confidence intervals.[1]

Next, we consider sample quantiles. Recall from Example 3.2.3 that the sample α-quantile $\widehat{\theta}_\alpha$ can be written as a Z-estimator with $\psi_\theta(x) = \mathbb{1}\{x \le \theta\} - \alpha$.

Proposition 3.4.7. *Suppose that the density f_X of X has strictly positive, bounded, and continuous density around its α-quantile θ_α. Under the conditions of Proposition 3.3.11, the sample quantile $\widehat{\theta}_\alpha$ satisfies*

$$\sqrt{n}(\widehat{\theta}_\alpha - \theta_\alpha) \to_d \mathcal{N}(0, \alpha(1-\alpha)/f_X(\theta_\alpha)^2).$$

Proof. We check the conditions of Theorem 3.4.1. The map

$$\theta \mapsto \mathbb{E}[\mathbb{1}\{X_i < \theta\} - \alpha] = \mathbb{P}(X < \theta) - \alpha$$

is continuously differentiable at θ_α with derivative $f_X(\theta_\alpha)$, which is bounded away from zero. This also makes the solution well separated, so that Proposition 3.3.11 gives consistency $\widehat{\theta}_\alpha \to_p \theta_\alpha$. Further,

$$\mathbb{E}[(\mathbb{1}\{X < \theta_\alpha\} - \alpha)^2] = \mathbb{P}(X < \theta_\alpha) - 2\alpha\mathbb{P}(X < \theta_\alpha) + \alpha^2 = \alpha - \alpha^2 = \alpha(1-\alpha).$$

[1] This expression is sometimes called *sandwich formula* for the asymptotic variance. In this analogy, the inverse Hessian's on the left and right are the bread, and the covariance of scores in between is the filling.

The stochastic equicontinuity condition (3.1) is satisfied by the same arguments as in Example 3.4.5. The result now follows from Theorem 3.4.1. □

We see that the sample quantiles are less precise when the density is low around the true quantile. This makes intuitive sense: if there are very few data near the quantile, the estimator may pick a value rather far away from the true quantile, just because it's the only one around. The density is typically small in the tails of the distribution, so when α is close to 0 or 1. This is offset only a little by the term $\alpha(1-\alpha)$, which is largest for $\alpha = 1/2$. For example, the normal distribution with $\alpha \to 0$ satisfies $\theta_\alpha = O(-\sqrt{\ln(1/\alpha^2)})$ and $f_X(\theta_\alpha) = O(\alpha^2)$, so the asymptotic variance is of order $O(1/\alpha^3)$.

Interestingly, if f_X is uniform on an interval $[a, b]$, the extreme quantiles are easier to estimate than the median. Taking $\alpha \to 1$, gives $\theta_\alpha \to b$ and the asymptotic variance converges to 0. In fact, one may show $\max\{X_1, \dots, X_n\} - b = O_p(1/n)$, which is much faster than the $1/\sqrt{n}$ rate for the median. This is a special property of bounded random variables, however.

3.4.5 Confidence intervals

Asymptotic normality of estimators is often used to construct confidence intervals. The theoretical result gives us the limiting distribution of the estimator. This limiting distribution depends on the unknown distribution P of X, however, so we need to estimate it. Recall our general result from Theorem 3.4.1. The asymptotic variance is

$$\Sigma = \dot\Psi(\theta_0)^{-1}\mathbb{E}[\psi_{\theta_0}(X)\psi_{\theta_0}(X)^\top]\dot\Psi(\theta_0)^{-1}, \quad \dot\Psi(\theta_0) = \mathbb{E}[\nabla\psi_\theta(X)].$$

The expectations with respect to P can be estimated by sample averages, and unknown value θ_0 can be replaced by the estimator $\widehat\theta$:

$$\widehat{\dot\Psi} = \frac{1}{n}\sum_{i=1}^n \nabla\psi_{\widehat\theta}(X_i), \quad \widehat\Sigma = \widehat{\dot\Psi}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \psi_{\widehat\theta}(X_i)\psi_{\widehat\theta}(X_i)^\top\right)\widehat{\dot\Psi}^{-1}. \tag{3.3}$$

Here, we assume that ψ_θ is differentiable for simplicity. Intuitively, $\widehat\theta$ converges to θ_0 and the sample average converge to the expectation. So the estimated asymptotic variance should converge to the true value. This is indeed the case, as shown by the following result.

Proposition 3.4.8. *Suppose that the conditions of Theorem 3.4.1 hold and ψ_θ is twice continuously differentiable and for some $\varepsilon > 0$,*

$$\mathbb{E}\left[\sup_{\|\theta-\theta_0\|<\varepsilon}|\partial_{\theta_j}\psi_\theta(X)|\right] < \infty, \quad \mathbb{E}\left[\sup_{\|\theta-\theta_0\|<\varepsilon}|\partial_{\theta_j}\partial_{\theta_k}\psi_\theta(X)|\right] < \infty.$$

Then the estimator in (3.3) satisfies $\widehat\Sigma \to_p \Sigma$.

Proof. We first show $\hat{\dot{\Psi}} \to_p \dot{\Psi}(\theta_0)$. It holds

$$\begin{aligned}\hat{\dot{\Psi}} &= \frac{1}{n}\sum_{i=1}^{n} \nabla\psi_{\widehat{\theta}}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \nabla\psi_{\theta_0}(X_i) + \frac{1}{n}\sum_{i=1}^{n} \nabla^2\psi_{\widetilde{\theta}}(X_i)(\widehat{\theta}-\theta_0).\end{aligned}$$

The first term converges to $\dot{\Psi}(\theta_0)$ because of the law of large numbers. Because $\|\widehat{\theta}-\theta_0\| < \varepsilon$ with probability going to one, the second can be bounded by

$$\sum_{j=1}^{p}\sum_{k=1}^{p}\frac{1}{n}\sum_{i=1}^{n}\sup_{\|\widehat{\theta}-\theta_0\|<\varepsilon} |\partial_{\theta_j}\partial_{\theta_k}\psi_{\widetilde{\theta}}(X_i)| \times O_p(n^{-1/2}) = O_p(1)\times O_p(n^{-1/2}) = o_p(1),$$

because the averages converge to a finite value by the law of large numbers. We have shown $\hat{\dot{\Psi}} \to_p \dot{\Psi}(\theta_0)$. A similar argument shows

$$\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{\widehat{\theta}}(X_i)\psi_{\widehat{\theta}}(X_i)^\top\right) \to_p \mathbb{E}[\psi_{\theta_0}(X)\psi_{\theta_0}(X)^\top].$$

Now the result follows from the continuous mapping theorem. □

Now we can piece things together. We know the limit is normal, and we can estimate the asymptotic variance. This allows us to construct confidence sets S by finding a solution to

$$\int_S \phi_{\widehat{\theta},\widehat{\Sigma}/\sqrt{n}}(x)\,dx = 1-\alpha,$$

where $\phi_{\mu,\Sigma}$ is the $\mathcal{N}(\mu,\Sigma)$ density. The typical choice for a single parameter is

$$\widehat{CI} = (\widehat{\theta} + \Phi^{-1}(\alpha/2)\widehat{\sigma}/\sqrt{n},\ \widehat{\theta} + \Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{n}), \tag{3.4}$$

where Φ is the standard normal cdf. Most often, we use $\alpha = 0.05$, so that $\Phi^{-1}(0.05/2) \approx 1.96$. For multiple parameters, we usually construct the confidence set as an ellipsoid centered at $\widehat{\theta}$ with axes given by the eigenvectors of $\widehat{\Sigma}$ and lengths given by the square roots of the corresponding eigenvalues. We may now show that the confidence sets have the correct coverage probability $1-\alpha$. For simplicity, we restrict ourselves to the one-dimensional case.

Proposition 3.4.9. *If $\sqrt{n}(\widehat{\theta}-\theta_0) \to_d Z \sim \mathcal{N}(0,\sigma^2)$ and $\widehat{\sigma} \to_p \sigma$, then the confidence interval in (3.4) has asymptotically exact coverage:*

$$\mathbb{P}(\theta_0 \in \widehat{CI}) \to 1-\alpha.$$

Proof. We have

$$\mathbb{P}(\theta_0 \in \widehat{CI}) = \mathbb{P}\left(\frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}/\sqrt{n}} \in (\Phi^{-1}(\alpha/2), \Phi^{-1}(1-\alpha/2))\right).$$

By the continuous mapping theorem (for $\widehat{\sigma} \to_p \sigma$) and asymptotic normality, $\sqrt{n}(\widehat{\theta} - \theta_0)/\widehat{\sigma} \to_d Z \sim \mathcal{N}(0,1)$. Thus,

$$\begin{aligned}\mathbb{P}(\theta_0 \in \widehat{CI}) &\to \mathbb{P}(Z \in (\Phi^{-1}(\alpha/2), \Phi^{-1}(1-\alpha/2))) \\ &= \Phi(\Phi^{-1}(1-\alpha/2)) - \Phi(\Phi^{-1}(\alpha/2)) \\ &= 1 - \alpha.\end{aligned}$$

□

3.4.6 Significance tests

Another common application of normality results is hypothesis testing. We will keep the arguments in this section informal to not waste too much time. Suppose we want to test the hypothesis $H_0 \colon \theta_0 = \theta^*$ against the alternative $H_1 \colon \theta_0 \neq \theta^*$. For example, in regression models, it is common to test whether a coefficient is zero. Consider the test statistic

$$\widehat{T} = n(\widehat{\theta} - \theta^*)^\top \widehat{\Sigma}_*^{-1}(\widehat{\theta} - \theta^*),$$

where $\widehat{\Sigma}_*$ is the estimator from (3.3), but with $\widehat{\theta}$ replaced by θ^*. By the law of large numbers and continuous mapping $\widehat{\Sigma}_* \to_p \Sigma$ under the null-hypothesis. Further, the asymptotic normality result above gives

$$\sqrt{n}\Sigma^{-1/2}(\widehat{\theta} - \theta^*) \to_d Z \sim \mathcal{N}(0, I_p).$$

And by the continuous mapping theorem,

$$n(\widehat{\theta} - \theta^*)^\top \Sigma^{-1}(\theta_0 - \theta^*) \to_d Z^\top Z \sim \chi^2(p).$$

We may now construct a test of size α by rejecting the null hypothesis if $\widehat{T} > \chi^2_{1-\alpha}(p)$, where $\chi^2_{1-\alpha}(p)$ is the $1-\alpha$ quantile of the $\chi^2(p)$ distribution. For maximum likelihood estimation, we call this a *Wald test*, but it works more generally.

For M-estimators, another way to test the hypothesis is to compare the values of the objective function. Consider the statistic

$$\widehat{T} = 2n(M_n(\widehat{\theta}) - M_n(\theta^*)).$$

Assuming sufficient regularity, a Taylor expansion gives

$$2n(M_n(\theta^*) - M_n(\widehat{\theta})) = 2n\nabla M_n(\widehat{\theta})(\theta^* - \widehat{\theta}) + n(\theta^* - \widehat{\theta})^\top \nabla^2 M_n(\widetilde{\theta})(\theta^* - \widehat{\theta}),$$

and $\nabla M_n(\widehat{\theta}) = 0$, because $\widehat{\theta}$ is the minimizer. We may show that, under the null,

$\nabla^2 M_n(\widetilde{\theta}) = \nabla^2 M(\theta^*) + o_p(1)$, so that

$$n(\theta^* - \widehat{\theta})^\top \nabla^2 M_n(\widetilde{\theta})(\theta^* - \widehat{\theta}) \to_d Q^\top Q,$$

where

$$Q \sim \mathcal{N}(0, (\nabla^2 M(\theta^*))^{-1/2} \, \Sigma \, (\nabla^2 M(\theta^*))^{-1/2}).$$

This limiting distribution is generally complicated. In the special case of the MLE, the test is called the *likelihood ratio test*, because

$$\widehat{T} = 2 \log \left(\frac{\prod_{i=1}^n f_X(X_i; \widehat{\theta})}{\prod_{i=1}^n f_X(X_i; \theta^*)} \right),$$

is essentially the log of the likelihood ratio. Under the null, it holds $\nabla^2 M(\theta^*) = \Sigma$, the Fisher information. Then the limiting distributions simplify to $Q \sim \mathcal{N}(0, I)$, and again $Q^\top Q \sim \chi^2(p)$.

3.5 Efficiency

Asymptotic normality results tell us a lot about the quality of an estimator. Estimators with small asymptotic variance (and no asymptotic bias) are good estimators, because we can be quite certain about the value of the estimator. Such estimators make very efficient use of the data.

3.5.1 The idea

To compare the efficiency of two estimators of the same quantity θ_0, we can compare their asymptotic variances. For simplicity, we will focus on the case of scalar θ_0. Suppose we have two estimators $\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}$ satisfying

$$\sqrt{n}(\widehat{\theta}^{(j)} - \theta_0) \to_d N(0, \sigma_j(\theta_0)^2), \quad j = 1, 2.$$

The *asymptotic relative efficiency* at θ_0 is defined as

$$\mathrm{ARE}(\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}) = \frac{\sigma_1(\theta_0)^2}{\sigma_2(\theta_0)^2}.$$

Note that the numerical value of this quantity can be different for different values of θ_0. The quantity has two interpretations:

- By comparing a ratio of variances, we compare how concentrated the two estimators are around the true value. More concentration means less uncertainty, so the estimator $\widehat{\theta}^{(1)}$ is more efficient than $\widehat{\theta}^{(2)}$ if $\mathrm{ARE}(\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}) < 1$, and less efficient if $\mathrm{ARE}(\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}) > 1$. With this interpretation, the ARE should rather be called the 'asymptotic relative uncertainty'.
- The alternative interpretation explains the name 'efficiency'. Suppose we want to know how many observations we need to achieve a certain level of (asymptotic)

precision $\sigma_j(\theta_0)^2/n \leq \varepsilon$. Let $n_{\varepsilon,j}, j = 1,2$ be this number. It holds $n_{\varepsilon,j} = \sigma_j(\theta_0)^2/\varepsilon$, so that the ratio of the two sample sizes is

$$\frac{n_{\varepsilon,1}}{n_{\varepsilon,2}} = \frac{\sigma_1(\theta_0)^2}{\sigma_2(\theta_0)^2} = \mathrm{ARE}(\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}).$$

Thus, the ARE quantifies (asymptotically) how much more/fewer data we need to achieve the same precision with $\widehat{\theta}^{(1)}$ compared to $\widehat{\theta}^{(2)}$.

Example 3.5.1 (Mean vs median for estimating location)**.** *Let $X_1, \ldots, X_n$ be an* iid *sample from a density $f_X(x;\theta_0) = g(x-\theta_0)$, with g symmetric around 0. This is called a* location model*. We want to estimate the location parameter θ_0. Because of the symmetry, θ_0 is both the mean and the median of X, so we can use both the sample mean $\widehat{\theta}^{(1)}$ and the sample median $\widehat{\theta}^{(2)}$ as estimators. From earlier results, we know*

$$\sqrt{n}(\widehat{\theta}^{(1)} - \theta_0) \to_d N(0, \mathbb{V}\mathrm{ar}[X]) \quad \textit{and} \quad \sqrt{n}(\widehat{\theta}^{(2)} - \theta_0) \to_d N(0, 1/4 f_X(\theta_0)^2).$$

The asymptotic relative efficiency is

$$ARE(\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}) = \frac{\mathbb{V}\mathrm{ar}[X]}{1/[4 f_X(\theta_0)^2]},$$

whose value depends on the density f_X. For example, the ARE computes to $2/\pi \approx 0.64$ if $X \sim \mathcal{N}(\theta_0, 1)$. This means that the sample mean is more efficient than the sample median in this case. From an efficiency point of view, using the median instead of the sample mean is similar to throwing away about a third of the data. On the other hand, if $X \sim \mathrm{Laplace}(\theta_0, 1)$, i.e., $g(x) = \frac{1}{2}e^{-|x|}$, the ARE is 2. Now the median is more efficient, and using the mean amounts to throwing away half of the data.

3.5.2 Superefficiency

Now consider the *Hodges* estimator for the location parameter θ_0 in the location model from Example 3.5.1:

$$\widehat{\theta}_H = \overline{X}_n \times \mathbb{1}\left\{|\overline{X}_n| > n^{-1/4}\right\}.$$

If $\theta_0 \neq 0$, $\overline{X}_n = \theta_0 + O_p(n^{-1/2})$, so $\mathbb{1}\{|\overline{X}_n| > n^{-1/4}\} = 1$ with probability tending to 1. Thus, $\widehat{\theta}_H = \overline{X}_n$ with probability tending to 1, and the Hodges estimator is asymptotically equivalent to the sample mean. If on the other hand $\theta_0 = 0$, we have $\overline{X}_n = O_p(n^{-1/2})$, so that $\mathbb{1}\{|\overline{X}_n| > n^{-1/4}\} = 0$ with probability tending to 1. Thus $\widehat{\theta}_H = 0$ with probability tending to 1, and

$$n^{\alpha}(\widehat{\theta}_H - \theta_0) \to_d 0, \quad \text{for every } \alpha > 0 \text{ and } \theta_0 = 0.$$

The estimator converges 'infinitely fast' to the true value $\theta_0 = 0$. This is called *superefficiency*.

From the analysis above it may seem that we should always use the Hodges estimator instead of the sample mean. It is asymptotically the same as the sample mean if $\theta_0 \neq 0$, and it is infinitely better if $\theta_0 = 0$. However, this would be very bad advice resulting from a misuse of asymptotics. To see this, suppose the data form a triangular array such that $X_1, \ldots, X_n \sim \mathcal{N}(\theta_n, 1)$ are *iid* with mean $\theta_n = n^{-1/4}/2$. For the sample mean, the central limit theorem for triangular arrays gives

$$\sqrt{n}(\overline{X}_n - \theta_n) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta_n\right) \to_d N(0, 1).$$

For the Hodges estimator, we have $\overline{X}_n = n^{-1/4}/2 + O_p(n^{-1/2})$, so that $\mathbb{1}\{|\overline{X}_n| > n^{-1/4}\} = 0$ with probability tending to 1. Thus, $\widehat{\theta}_H = 0$ with probability tending to 1, and

$$\sqrt{n}(\widehat{\theta}_H - \theta_n) = -\sqrt{n} \times n^{-1/4}/2 + O_p(1) \to_p -\infty.$$

This tells us that, for finite samples, the Hodges estimator is terrible if θ_0 is in a close neighborhood of 0.

One can show that superefficiency can only occur at a set of values θ_0 of Lebesgue measure 0. In practice, it is very unlikely that the true value of the parameter is exactly one of these values. Further, we have seen that superefficiency comes at a cost when we move away from point-wise convergence. In fact, the superefficiency phenomenon tells us that point-wise comparisons of estimators are not the right tool to think about efficiency, but that we have to study the behavior of the estimators for data $X_1, \ldots, X_n$, in which the true parameter θ_n is in a shrinking neighborhood of θ_0.

3.5.3 Comments

There's much more to say about the efficiency of estimators. You probably already know about the Cramér-Rao lower bound on the asymptotic variance of an estimator. Another interesting result is that the normal distribution is optimal as a limit of $\sqrt{n}(\widehat{\theta} - \theta_0)$. While such results are deep and fundamental, they aren't immediately useful when constructing or studying new statistical methods. We thus leave our discourse to efficiency here and refer to the relevant chapters in Van der Vaart (2000) for more details.

3.6 Model selection

Formally speaking, a statistical model $\mathcal{M}$ for *iid* data $X_1, \ldots, X_n$ is a collection of possible distributions P for X_1 (typically indexed by some parameter $\theta \in \Theta$). In applications, we often have several plausible models $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(K)}$ to choose from. Models can differ in many ways. Two common examples are differences in the shape of the distribution (e.g., Normal vs Laplace distribution) or differences in the number of parameters (e.g., linear models with differing sets of covariates). The goal of model selection is to find the model that best describes the data.

3.6.1 AIC and BIC

To simplify our discussion, let us assume that all models under considerations are parametric and absolutely continuous, i.e,

$$\mathcal{M}^{(k)} = \{f_k(\cdot;\theta)\colon \theta \in \Theta^{(k)} \subseteq \mathbb{R}^{p_k}\},$$

where the f_X are probability density functions. The two most common criteria for model selection are *Akaike's Information Criterion (AIC)* and the *Bayesian Information Criterion (BIC)*:

$$\begin{aligned}\mathrm{AIC}(\mathcal{M}^{(k)}) &= -2\ell_{n,k}(\widehat{\theta}^{(k)}) + 2p,\\ \mathrm{BIC}(\mathcal{M}^{(k)}) &= -2\ell_{n,k}(\widehat{\theta}^{(k)}) + p\log n,\end{aligned}$$

where $\ell_{n,k}(\theta) = \sum_{i=1}^{n} \ln f_k(X_i;\theta)$ is the log-likelihood of the data under model $\mathcal{M}^{(k)}$, and $\widehat{\theta}^{(k)}$ is the corresponding MLE.

The motivation/derivation of these criteria is outlined in, for example, Kauermann et al. (2021, Chapter 9). The AIC takes a purely predictive view, simply trying to minimize the KL divergence. The penalty $2p$ in AIC attempts to correct for the bias stemming from using the same data for estimating the parameter and approximating the KL-divergence by the negative likelihood. The BIC is derived from a Bayesian perspective. It (approximately) selects the model with the highest posterior probability, treating all models equally likely *a priori*.

3.6.2 What is the 'right' model?

AIC and BIC are supposed to help us find the 'right' model. But what is 'right'? A natural way to measure quality of a model is the *Kullback-Leibler divergence* between the true distribution $f_X(\cdot)$ of X and the model distribution $f_k(\cdot;\theta^{(k)})$:

$$\mathrm{KL}[f_X||f_k(\cdot;\theta^{(k)})] = \mathbb{E}\left[\ln \frac{f_X(X)}{f_k(X;\theta^{(k)})}\right] = \mathbb{E}\left[\ln f_X(X)\right] - \mathbb{E}\left[\ln f_k(X;\theta^{(k)})\right].$$

The KL divergence is a measure of how well $f_k(\cdot;\theta^{(k)})$ approximates the true distribution $f_X(\cdot)$. Note that the first term on the right is independent of the model, so minimizing KL-divergence is equivalent to maximizing the expected log-likelihood. The model $\mathcal{M}^{(k)}$ allows for many different parameter values. For the KL divergence to be an adequate measure of model quality, we define the optimal (within $\mathcal{M}^{(k)}$) parameter $\theta_*^{(k)}$ as the one that minimizes the KL divergence:

$$\theta_*^{(k)} = \operatorname*{arg\,min}_{\theta\in\Theta^{(k)}} \mathrm{KL}[f_X||f_k(\cdot;\theta)].$$

Now it is natural to define the 'best' model among $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(K)}$ as the one where $\theta_*^{(k)}$ gives the smallest divergence:

$$k^* = \underset{k=1,\ldots,K}{\arg\min}\, \mathrm{KL}[f_X || f_k(\cdot; \theta_*^{(k)})].$$

This choice is not feasible in practice because we know neither the true distribution $f_X(\cdot)$ nor the true parameter $\theta_*^{(k)}$. It also isn't guaranteed to be unique and quite often it is not. For example, suppose $\mathcal{M}^{(1)}$ is nested in $\mathcal{M}^{(2)}$, i.e., $\mathcal{M}^{(1)} \subseteq \mathcal{M}^{(2)}$. Then if $k^* = 1$ is optimal, $k^* = 2$ has to be optimal as well. An example is a linear regression model with and without an intercept term, or with and without a certain covariate. If this is the case, we often prefer the more *parsimonious* model (the one with smaller p_k) and call this the 'right' model. Parsimonious models are less likely to overfit, so that might help on finite samples.

3.6.3 Selection consistency

Both AIC and BIC are derived from reasonable principles. But what mathematical guarantees do they provide? The most important property of a model selection criterion is *selection consistency*. A model selection criterion is called *selection consistent* if it selects the 'right' model with probability tending to 1 as the sample size n tends to infinity.

Let us now turn to the question of selection consistency. By standardizing the AIC and BIC by $1/2n$, we can cast them into the form

$$\mathrm{IC}(\mathcal{M}^{(k)}) = -\frac{1}{n}\ell_{n,k}(\widehat{\theta}^{(k)}) + \lambda_{n,k},$$

where $\lambda_{n,k} \in \mathbb{R}_{\geq 0}$ is some penalty term, and select the model with index

$$\widehat{k} = \underset{k=1,\ldots,K}{\arg\min}\, \mathrm{IC}(\mathcal{M}^{(k)}).$$

The AIC has $\lambda_{n,k} = p_k/n$ and the BIC has $\lambda_{n,k} = (p_k/n)\ln n$.

There is nothing special about the likelihood when it comes to model selection. It is appropriate when parameters have been estimated by the MLE. In fact, their motivation/derivation and theoretical guarantees do not apply when prameters are estimated differently. It will be both simpler and more general to consider general M-estimators where

$$\widehat{\theta}^{(k)} = \underset{\theta \in \Theta^{(k)}}{\arg\min}\, M_{n,k}(\theta), \quad \theta_*^{(k)} = \underset{\theta \in \Theta^{(k)}}{\arg\min}\, M_k(\theta).$$

The criterion $M_{n,k}(\theta)$ is a generalization of the negative averaged log-likelihood. The set of optimal models is defined as

$$\mathcal{K}^* = \underset{k=1,\ldots,K}{\arg\min}\, M_k(\theta_*^{(k)}).$$

For this to make sense, the criteria $M_1, \ldots, M_K$ should be comparable, for example the square loss applied to different regression models. In practice, we pick the model with index

$$\widehat{k} = \underset{k=1,\ldots,K}{\arg\min}\, M_{n,k}(\widehat{\theta}^{(k)}) + \lambda_{n,k}.$$

While this is not guaranteed to be unique, it typically is in practice. We shall just assume that it is unique from now on.

Theorem 3.6.1 (Weak selection consistency)**.** *Suppose that:*

(i) The conditions of Theorem 3.3.2 hold for every model $\mathcal{M}^{(k)}$*.*

(ii) The penalties satisfy $\max_k \lambda_{n,k} \to 0$*.*

Then $\mathbb{P}(\widehat{k} \in \mathcal{K}^*) \to 1$ *as* $n \to \infty$*.*

Proof. The statement $\widehat{k} \notin \mathcal{K}^*$ is equivalent to

$$\min_{k \in \mathcal{K}^*} M_{n,k}(\widehat{\theta}^{(k)}) + \lambda_{n,k} \geq \min_{k \notin \mathcal{K}^*} M_{n,k}(\widehat{\theta}^{(k)}) + \lambda_{n,k},$$

which implies

$$\begin{aligned} &\max_{k \in \mathcal{K}^*} [M_{n,k}(\widehat{\theta}^{(k)}) - M_k(\widehat{\theta}^{(k)})] + \max_{k \in \mathcal{K}^*} M_k(\widehat{\theta}^{(k)}) \\ &\geq \min_{k \notin \mathcal{K}^*} [M_{n,k}(\widehat{\theta}^{(k)}) - M_k(\widehat{\theta}^{(k)})] + \min_{k \notin \mathcal{K}^*} M_k(\widehat{\theta}^{(k)}) + o(1). \end{aligned}$$

Observe that

$$\begin{aligned} \left| \max_{k \in \mathcal{K}^*} [M_{n,k}(\widehat{\theta}^{(k)}) - M_k(\widehat{\theta}^{(k)})] \right| &\leq \max_k \sup_{\theta \in \Theta^{(k)}} |M_{n,k}(\theta) - M_k(\theta)| = o_p(1), \\ \left| \min_{k \in \mathcal{K}^*} [M_{n,k}(\widehat{\theta}^{(k)}) - M_k(\widehat{\theta}^{(k)})] \right| &\leq \max_k \sup_{\theta \in \Theta^{(k)}} |M_{n,k}(\theta) - M_k(\theta)| = o_p(1), \end{aligned}$$

by the uniform convergence condition and the fact that a finite maximum is a continuous map. The preceding display then implies

$$\max_{k \in \mathcal{K}^*} M_k(\widehat{\theta}^{(k)}) \geq \min_{k \notin \mathcal{K}^*} M_k(\widehat{\theta}^{(k)}) + o_p(1).$$

Using continuity of M_k and the fact that $\widehat{\theta}^{(k)} \to_p \theta_*^{(k)}$, we get

$$\max_{k \in \mathcal{K}^*} M_k(\theta_*^{(k)}) \geq \min_{k \notin \mathcal{K}^*} M_k(\theta_*^{(k)}) + o_p(1).$$

This event has probability tending to 0 since the definition of $\mathcal{K}^*$ implies

$$\max_{k \in \mathcal{K}^*} M_k(\theta_*^{(k)}) < \min_{k \notin \mathcal{K}^*} M_k(\theta_*^{(k)}) + \eta$$

for some $\eta > 0$. □

Condition (ii) is clearly satisfied for both AIC and BIC, so both are consistent in the sense of the theorem. Note that the proof even applies to $\lambda_{n,k} = 0$; penalties are not even necessary for this type of consistency. It is considered weak, because it does not distinguish between models that are equally close to the true model. For example, if $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ are nested, the theorem does not tell us which one is selected. Take for example the problem of selecting the covariates in a linear model. If the response is only related to the first covariate. Weak consistency does *not* tell us that the model with only the first covariate is selected. Any model that includes the first covariate may be selected. This is quite unsatisfactory.

In practice, we prefer models that are parsimonious, i.e., have fewer parameters. It may look like both AIC and BIC do this, but that's not the case. As we shall see, the AIC does not necessarily select the optimal model with the fewest parameters asymptotically, but the BIC does. The key difference is that $n\lambda_{n,k} \to \infty$ for the BIC, but not for the AIC.

Theorem 3.6.2 (Selection consistency)**.** *Suppose that, in addition to the conditions of Theorem 3.6.1, the following holds:*

(i) The index $k^* = \arg\min_{k \in \mathcal{K}^*} M_k(\theta_*^{(k)}) + \lambda_{n,k}$ *is unique (i.e., there is a unique most parsimoniuos model).*

(ii) It holds $M_{n,k}(\theta_*^{(k)}) = M_{n,k'}(\theta_*^{(k')})$ *almost surely for every* $k, k' \in \mathcal{K}^*$.

(iii) The penalties satisfy $n(\lambda_{n,k^*} - \lambda_{n,k}) \to -\infty$ *for all* $k \in \mathcal{K}^* \setminus \{k^*\}$.

(iv) It holds $\max_k \|\widehat{\theta}^{(k)} - \theta_*^{(k)}\| = O_p(n^{-1/2})$.

(v) Each $M_{n,k}$ *is twice continuously differentiable and* $\sup_{\theta \in \Theta^{(k)}} \|\nabla^2 M_{n,k}(\theta)\| = O_p(1)$.

Then $\mathbb{P}(\widehat{k} = k^*) \to 1$ *as* $n \to \infty$.

Condition (ii) implies that the set of all optimal models are all equivalent when evaluated at the optimal parameter. In GLMs, for example, that's the case when there is a unique optimal parameter θ_* in the model that includes all covariates. Condition (iii) requires that the penalties $\lambda_{n,k}$ don't vanish too fast. Condition (iv) asks for the typical convergence rate of the M-estimator. Condition (v) is a technical condition, similar to a uniform law of large numbers for the Hessian matrix of the criterion.

Proof. Because of Theorem 3.6.1 it suffices to show that $\widehat{k}$ selects the right model when restricted to the set $\mathcal{K}^*$. The union bound gives

$$\mathbb{P}(\widehat{k} \in \mathcal{K}^* \setminus \{k^*\}) \leq \sum_{k \in \mathcal{K}^* \setminus \{k^*\}} \mathbb{P}(\widehat{k} = k),$$

so it is enough to show that $\mathbb{P}(\widehat{k} = k) \to 0$ for any $k \in \mathcal{K}^* \setminus \{k^*\}$. Let k be one such index. Then the event $\widehat{k} = k$ implies

$$\begin{aligned} & M_{n,k}(\widehat{\theta}^{(k)}) + \lambda_{n,k} \leq M_{n,k^*}(\widehat{\theta}^{(k^*)}) + \lambda_{n,k^*} \\ \Leftrightarrow \quad & n[M_{n,k}(\widehat{\theta}^{(k)}) - M_{n,k^*}(\widehat{\theta}^{(k^*)})] \leq n[\lambda_{n,k^*} - \lambda_{n,k}]. \end{aligned}$$

For the terms on the left-hand side, a Taylor expansion and assumptions (iv) and (v) give

$$\begin{aligned} & M_{n,k}(\widehat{\theta}^{(k)}) \\ &= M_{n,k}(\theta_*^{(k)}) + \underbrace{\nabla M_{n,k}(\widehat{\theta}^{(k)})}_{=0}(\widehat{\theta}^{(k)} - \theta_*^{(k)}) + \frac{1}{2}(\widehat{\theta}^{(k)} - \theta_*^{(k)})^\top \nabla^2 M_{n,k}(\widetilde{\theta}^{(k)})(\widehat{\theta}^{(k)} - \theta_*^{(k)}) \\ &= M_{n,k}(\theta_*^{(k)}) + O_p(n^{-1}), \end{aligned}$$

and similarly for $M_{n,k^*}(\widehat{\theta}^{(k^*)})$. By assumption (ii), we thus have $n[M_{n,k}(\widehat{\theta}^{(k)}) - M_{n,k^*}(\widehat{\theta}^{(k^*)})] = O_p(1)$. Since $n[\lambda_{n,k^*} - \lambda_{n,k}] \to -\infty$ by assumption, the event $\widehat{k} = k$ has probability tending to 0. □

The AIC does not meet condition (iii) and it is not selection consistent in the sense of the theorem. The BIC does meet condition (iii) and is selection consistent.

3.6.4 Error probabilities

We now have a more fine-grained look at the behavior of model selection procedures. While it's great that the probability of selecting the wrong model goes to zero, asymptotically, it may do so very slowly. We can get a clearer picture by exploiting distributional limits.

We consider two types of error probabilities corresponding to the two types of consistency. The first is the probability of selecting a model $k \notin \mathcal{K}^*$.

Proposition 3.6.3. *Let $k \notin \mathcal{K}^*$. Let the conditions of the previous theorem hold, define $\eta_k = M_k(\theta_*^{(k)}) - M_{k*}(\theta_*^{(k*)}) > 0$ and suppose*

$$[M_{n,k^*}(\theta_*^{(k^*)}) - M_{n,k}(\theta_*^{(k)})] - \eta_k \to_d \mathcal{N}(0, \sigma_k^2), \quad \sqrt{n}(\lambda_{n,k^*} - \lambda_{n,k}) \to 0.$$

Then

$$\mathbb{P}(\widehat{k} = k) \leq \Phi\left(\frac{-\sqrt{n}\eta_k}{\sigma_k}\right) + o(1),$$

where Φ is the standard normal CDF.

Proof. Using arguments as in the previous proof, we may show that $\widehat{k} = k$ implies

$$\begin{aligned} & \sqrt{n}[M_{n,k}(\theta_*^{(k)}) - M_{n,k^*}(\theta^{(k^*)})] + O_p(n^{-1/2}) \leq \sqrt{n}[\lambda_{n,k^*} - \lambda_{n,k}] \\ \Leftrightarrow \quad & \sqrt{n}[M_{n,k}(\theta_*^{(k)}) - M_{n,k^*}(\theta^{(k^*)}) - \eta_k] + o_p(1) \leq -\sqrt{n}\eta_k. \end{aligned}$$

Using the central limit theorem and Slutsky's lemma, the left-hand side converges in distribution to a $\mathcal{N}(0, \sigma_k^2)$ random variable. The result follows. □

The proposition is not much more helpful in a precise sense, since we don't know how fast the $o(1)$ term goes to zero. How fast this remainder vanishes depends on how many moments of the $M_{n,k}$ exist. Let's ignore this for now and take the normal approximation as exact. The error probability $\Phi(-\sqrt{n}\eta_k/\sigma_k)$ decays exponentially fast in n, so selecting a specific model $k \notin \mathcal{K}^*$ is very unlikely. However, there potentially are many such models, so the error probabilities over all $k \notin \mathcal{K}^*$ accumulates. For exhaustively search through GLMs with p covariates, for example, there are $2^p - 1$ such models. For the error probability to remain small, we then need $p \ll n$.

We move on to the second error: selecting a model $k \in \mathcal{K}^*$ that isn't maximally parsimonious.

Proposition 3.6.4. *Let $k \in \mathcal{K}^*$ but $k \neq k^*$. Let the conditions of the previous theorem hold, and suppose for all k,*

$$\sup_{\theta \in \Theta^{(k)}} \|\nabla^2 M_{n,k}(\theta) - \nabla^2 M_k(\theta)\| = o_p(1),$$

and

$$\sqrt{n}[\nabla^2 M_k(\theta_*^{(k)})]^{1/2}(\widehat{\theta}^{(k)} - \theta_*^{(k)}) \to_d Z_k.$$

Then

$$\mathbb{P}(\widehat{k} = k) \leq \mathbb{P}\left(\|Z_{k^*}\|^2 - \|Z_k\|^2 > -2n[\lambda_{n,k*} - \lambda_{n,k}]\right) + o(1).$$

Proof. Define

$$Q_{n,k} = (\widehat{\theta}^{(k)} - \theta_*^{(k)})^\top \nabla^2 M_k(\theta_*^{(k)})(\widehat{\theta}^{(k)} - \theta_*^{(k)}).$$

Following the arguments from the proof of Theorem 3.6.2, the event $\widehat{k} = k$ implies

$$n[Q_{n,k^*} - Q_{n,k^*}] + o_p(1) \leq -2n[\lambda_{n,k*} - \lambda_{n,k}].$$

The left-hand side converges in distribution to $\|Z_{k^*}\|^2 - \|Z_k\|^2$ by the continuous mapping theorem and Slutsky's lemma. The result follows. □

The approximate error probability is somewhat unwieldy. To simplify the discussion, observe that $\|Z_{k^*}\|^2 - \|Z_k\|^2 \leq \|Z_{k^*}\|^2 + \|Z_k\|^2 := W_k$, and that the right-hand side is a weighted sum of $\chi^2(1)$ random variables. Such variables satisfy the tail bound

$$\mathbb{P}(|W_k| > t) \leq \exp(a - bt),$$

for some $a, b > 0$. Hence, the error probability decays exponentially fast in $n|\lambda_{n,k*} - \lambda_{n,k}|$. For the AIC, we have $n|\lambda_{n,k*} - \lambda_{n,k}| = p_{k^*} - p_k$ which is constant. So indeed the

probability for selecting a model in $k \in \mathcal{K}^*$ that isn't parsimonious enough does not vanish. For the BIC, we have $n|\lambda_{n,k*} - \lambda_{n,k}| = (p_{k^*} - p_k) \ln n$, so

$$\mathbb{P}(|W_k| > 2n|\lambda_{n,k*} - \lambda_{n,k}|) \exp(a - 2b|p_{k^*} - p_k| \ln n) \leq e^a n^{-2b|p_{k^*} - p_k|}.$$

This decay is polynomial in n. This is still small enough when comparing only a finite number of models $k \in \mathcal{K}^*$ and n is sufficiently large. If the selection criterion is used to search exhaustively through many models, stronger penalties are needed. Taking $\lambda_{n,k} = p_k n^{1/4}$ for example, the error probabilities decay similar to $\exp(-n^{1/4})$.

4 Estimators as functionals

We now take a different perspective on statistical methods, which often proves useful. It certainly will in later chapters.

4.1 Introduction

Recall that most statistical methods can be described as functions that map data to estimates. In this framework, the quantity we want to estimate is somehow detached from the estimator. Now observe that the data can be equivalently be represented by the empirical measure $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_x denotes the Dirac measure[1] at x. We can now think of the estimator as a *functional* T that maps the empirical distribution to an estimate. Quite often, the target quantity is then the same functional T applied to the true distribution. Let us first defined what a functional is.

Definition 4.1.1 (Functional)**.** *A* ***statistical functional*** *is a map $T : \mathcal{P} \to \mathbb{R}^p$, where $\mathcal{P}$ is a set of probability measures.*

Let's see some examples.

Example 4.1.2 (Sample mean)**.** *The sample mean can be written as*

$$T(\mathbb{P}_n) = \int x \, d\mathbb{P}_n(x) = \frac{1}{n} \sum_{i=1}^n X_i,$$

and the true mean as

$$T(P) = \int x \, dP(x) = \mathbb{E}_P[X].$$

Example 4.1.3 (M-estimator)**.** *An M-estimator can be written as*

$$T(\mathbb{P}_n) = \arg\min_{\theta \in \Theta} \int m_\theta(x) \, d\mathbb{P}_n(x) = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m_\theta(X_i),$$

and the true parameter as

$$T(P) = \arg\min_{\theta \in \Theta} \int m_\theta(x) \, dP(x) = \arg\min_{\theta \in \Theta} \mathbb{E}_P[m_\theta(X)].$$

[1] The Dirac measure δ_x is defined as $\delta_x(A) = \mathbb{1}(x \in A)$ and represents a point-mass at x.

You can work out a few more examples for yourself as an exercise. A common property among reasonable statistical functionals is *Fisher consistency*.

Definition 4.1.4 (Fisher consistency)**.** *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model indexed by the parameter θ. A statistical functional T is called* ***Fisher consistent*** *if $\theta = T(P_\theta)$ for all $P_\theta \in \mathcal{P}$.*

We can always write the target quantity as a functional of the true distribution P. Estimators that are constructed by plugging in the empirical distribution $\mathbb{P}_n$ in the same functionals are so common that they have a name.

Definition 4.1.5 (Plug-in estimators)**.** *Let T be a statistical functional. An estimator of $T(P)$ is called a* ***plug-in estimator*** *if it can be written as $T(\mathbb{P}_n)$.*

4.2 von Mises calculus

Now let's see, first informally, how this view on statistical methods can be useful. By the Glivenko-Cantelli theorem

$$\|\mathbb{P}_n - P\|_\infty := \sup_{x \in \mathbb{R}^d} |\mathbb{P}_n((-\infty, x]) - P((-\infty, x])| \to_p 0,$$

so we know that the empirical measure converges to the true distribution in a certain sense. This suggests that plug-in estimators should be consistent if the functional T is continuous with respect to the Kolmogorov distance $\|\cdot\|_\infty$. But we can say more. Because $\mathbb{P}_n$ is close to P, we may hope that a Taylor-like expansion of T around P tells us something about the behavior of the plug-in estimator $T(\mathbb{P}_n)$. This is the idea of *von Mises calculus*. A first order expansion would look something like

$$T(\mathbb{P}_n) - T(P) = T'_P(\mathbb{P}_n - P) + o(\|\mathbb{P}_n - P\|_\infty),$$

so the first order behavior is the same as that of $T'_P(\mathbb{P}_n - P)$. This term needs some explanation.

Recall that the derivative of a function $f(x)$ with respect to a vector x is the gradient $\nabla f(x)$, a function mapping x to another vector. The first order Taylor-expansion of f around x is

$$f(x + h) = f(x) + \nabla f(x)h + o(\|h\|),$$

where $h \mapsto \nabla f(x)h$ is a linear function of h that depends on x. The situation is similar for functionals. The first-order term $T'_P(h)$ is a linear functional T'_P of the difference h that depends on P. As we shall see, linear functionals in this context are maps of the form

$$T'_P(\mathbb{P}_n - P) = \int \rho_P(x)\, d\mathbb{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \rho_P(X_i),$$

for some function ρ_P with $\mathbb{E}_P[\rho_P(X)] = 0$.

The first order expansion of $T(\mathbb{P}_n)$ around $T(P)$ is a sample average with zero mean, which we understand well. In particular, the estimator should be consistent, and asymptotically normal with asymptotic variance $\mathbb{V}\mathrm{ar}[\rho_P(X)]$. For the remainder of the expansion to be negligible for $\sqrt{n}$-asymptotic normality, we require $\|\mathbb{P}_n - P\|_\infty = O_P(n^{-1/2})$, which will be proven later in this chapter (Lemma 4.4.1). In that case, the estimator is *asymptotically linear*:

$$T(\mathbb{P}_n) - T(P) = \frac{1}{n}\sum_{i=1}^{n} \rho_P(X_i) + o_p(n^{-1/2}).$$

The function ρ_P is called the *influence function*, because $\rho_P(X_i)$ characterizes the influence of the observation X_i on the estimator (to first order). It will play a bigger role in robust statistics later in the course.

4.3 Derivatives of functionals

The above considerations were only conceptual, because we haven't yet defined what the derivative of a functional is. It's a bit more involved than for functions, but the idea is the same. There are several generalizations of a derivative on general metric spaces. They usually agree when applied to Euclidean spaces, but differ in more general spaces. We shall only consider the space of probability measures here.

Definition 4.3.1 (Gateaux derivative)**.** *The* ***Gateaux derivative*** *of T at P in direction Q is*

$$T'_P(Q) = \lim_{t\to 0} \frac{T((1-t)P + tQ) - T(P)}{t} = \left[\frac{d}{dt} T((1-t)P + tQ)\right]_{t=0}.$$

The map T is called ***Gateaux differentiable*** *at P if the limit exists for all Q.*

This is indeed a linear map, for $T'_P(a_1Q_1 + a_2Q_2) = a_1T'_P(Q_1) + a_2T'_P(Q_2)$ by the usual rules of the derivative with respect to t. The Gateaux derivative is nice, because it makes calculations easy.

Example 4.3.2 (Expectation)**.** *To warm up, consider the expectation functional $T(P) = \mathbb{E}_P[g(X)] = \int g(x)dP(x)$ for some g. We have*

$$\begin{aligned}\frac{T((1-t)P + tQ) - T(P)}{t} &= \frac{(1-t)\int g(x)\,dP(x) + t\int g(x)\,dQ(x) - \int g(x)\,dP(x)}{t} \\ &= \int g(x)\,dQ(x) - \int g(x)\,dP(x) = \mathbb{E}_Q[g(X)] - \mathbb{E}_P[g(X)].\end{aligned}$$

Thus

$$T'(\mathbb{P}_n - P) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}_P[g(X)] = \frac{1}{n}\sum_{i=1}^{n} \rho_P(X_i),$$

with $\rho_P(x) = g(x) - \mathbb{E}_P[g(X)]$. *In this case, the Taylor expansion is in fact an identity,*

$$T(\mathbb{P}_n) - T(P) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}_P[g(X)],$$

so this was a nice exercise, but not very enlightening.

Example 4.3.3 (Variance)**.** *Now consider the variance functional*

$$T(P) = \mathbb{E}_P[X^2] - \mathbb{E}_P[X]^2 = \int x^2 dP(x) - \left(\int x dP(x)\right)^2.$$

We have

$$\begin{aligned}\frac{d}{dt}\int x^2 d[(1-t)P(x) + tQ(x)] &= \frac{d}{dt}(1-t)\int x^2\, dP(x) + \frac{d}{dt}t\int x^2\, dQ(x)\\ &= \int x^2\, dQ(x) - \int x^2\, dP(x).\end{aligned}$$

Further, using $df(x)^2/dx = 2f(x)f'(x)$,

$$\begin{aligned}&\frac{d}{dt}\left(\int x d[(1-t)P(x) + tQ(x)]\right)^2\\ &= 2\left(\int x\, d[(1-t)P(x) + tQ(x)]\right)\left(\int x\, dQ(x) - \int x\, dP(x)\right).\end{aligned}$$

Together this gives

$$\begin{aligned}&\left[\frac{d}{dt}T((1-t)P + tQ)\right]_{t=0}\\ &= \int x^2 dQ(x) - \int x^2 dP(x) - 2\left(\int x\, dP(x)\right)\left(\int x\, dQ(x) - \int x\, dP(x)\right)\\ &= \mathbb{E}_Q[X^2] - \mathbb{E}_P[X^2] - 2\mathbb{E}_P[X](\mathbb{E}_Q[X] - \mathbb{E}_P[X]).\end{aligned}$$

Hence

$$T'_P(\mathbb{P}_n - P) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mathbb{E}_P[X^2] - 2\mathbb{E}_P[X]\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}_P[X]\right) = \frac{1}{n}\sum_{i=1}^{n}\rho_P(X_i),$$

where $\rho_P(x) = x^2 - \mathbb{E}_P[X^2] - 2\mathbb{E}_P[X](x - \mathbb{E}_P[X])$ *has* $\mathbb{E}_P[\rho_P(X)] = 0$. *As in Example 2.3.12, we may assume that* $\mathbb{E}_P[X] = 0$. *Then we expect* $T(\mathbb{P}_n)$ *to be asymptotically normal with asymptotic variance*

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\rho_P(X)] &= \mathbb{E}\left[(X^2 - \mathbb{E}_P[X^2])^2\right]\\ &= \mathbb{E}[X^4] - 2\mathbb{E}[X^2]\mathbb{E}[X^2] + \mathbb{E}[X^2]^2\\ &= \mathbb{E}[X^4] - \mathbb{E}[X^2]^2,\end{aligned}$$

as already shown in Example 2.3.12.

Example 4.3.4 (Z-estimators)**.** *Consider the functional $T(P) = \theta_P$, where θ_P is the unique solution of $\int \psi_\theta(x) dP(x) = 0$ for some function ψ. The Z-estimator can be written as the plug-in estimator: $T(\mathbb{P}_n) = \widehat{\theta}$, where $\widehat{\theta}$ solves $\int \psi_\theta(x)\, d\mathbb{P}_n(x) = 0$. Define θ_t as the solution of $\int \psi_{\theta_t}(x)\, d[(1-t)P(x) + tQ(x)] = 0$. Let's take the derivative of this identity:*

$$\begin{aligned}
0 &= \frac{d}{dt} \int \psi_{\theta_t}(x)\, d[(1-t)P(x) + tQ(x)] \\
&= \frac{d}{dt}(1-t) \int \psi_{\theta_t}(x)\, dP(x) + \frac{d}{dt} t \int \psi_{\theta_t}(x)\, dQ(x) \\
&= (1-t)\frac{d}{dt} \int \psi_{\theta_t}(x)\, dP(x) - \int \psi_{\theta_t}(x)\, dP(x) + t\frac{d}{dt} \int \psi_{\theta_t}(x)\, dQ(x) + \int \psi_{\theta_t}(x)\, dQ(x),
\end{aligned}$$

which evaluated at $t = 0$ gives

$$0 = \left[\frac{d}{dt} \int \psi_{\theta_t}(x)\, dP(x)\right]_{t=0} + \int \psi_P(x)\, d[Q(x) - P(x)].$$

Rewriting $d/dt = (d/d\theta_t) \times (d\theta_t/dt)$, we get

$$0 = \nabla_\theta \int \psi_{\theta_P}(x)\, dP(x) \times \left.\frac{d\theta_t}{dt}\right|_{t=0} + \int \psi_P(x)\, d[Q(x) - P(x)].$$

Now observe that

$$\left.\frac{d\theta_t}{dt}\right|_{t=0} = \lim_{t\to 0} \frac{\theta_t - \theta_0}{t} = \lim_{t\to 0} \frac{T((1-t)P + tQ) - T(P)}{t},$$

is the quantity we are interested in. Noting that $\int \psi_{\theta_P}(x)\, dP(x) = 0$ and solving for $d\theta_t/dt$ we get

$$\lim_{t\to 0} \frac{T((1-t)P + tQ) - T(P)}{t} = -(\nabla \mathbb{E}_P[\psi_{\theta_P}(X)])^{-1} \mathbb{E}_Q[\psi_{\theta_P}(X)],$$

so in particular,

$$T'_P(\mathbb{P}_n - P) = -(\nabla \mathbb{E}_P[\psi_{\theta_P}(X)])^{-1} \frac{1}{n} \sum_{i=1}^n \psi_{\theta_P}(X_i) = \frac{1}{n} \sum_{i=1}^n \rho_P(X_i),$$

with $\rho_P(x) = -\mathbb{E}[\nabla \psi_{\theta_P}(X)]^{-1} \psi_{\theta_P}(x)$. This suggests that Z-estimators are asymptotically normal with the same asymptotic variance (of course) that we had already derived in Theorem 3.4.1.

I intentionally used careful wording. The calculations *suggest* asymptotic normality with a certain asymptotic variance. The calculations are not a proof, but they give us a quick idea of the estimator's asymptotic behavior. To make the calculations more formal, we need a stronger notion of a derivative. The Gateaux derivative assumes takes perturbations of P in a fixed direction Q. However, our first-order expansions are in the direction of $Q_n = \mathbb{P}_n$, which is a sequence of different directions. This motivates

the following definition.

Definition 4.3.5 (Fréchet derivative)**.** *The functional T is called* ***Fréchet differentiable*** *at P if there is a continuous, linear map $h \mapsto T'_P(h)$ such that for every sequence Q_n with $\|Q_n - P\|_\infty \to 0$,*[a]

$$T(Q_n) - T(P) - T'_P(Q_n - P) = o(\|Q_n - P\|_\infty).$$

[a] We shall work with the Kolmogorov metric $d(P_1, P_2) = \|P_1 - P_2\|_\infty$ here, even though it is slightly stronger than necessary for our purposes.

It is easy to verify that Fréchet differentiability implies Gateaux differentiability, but not vice versa. A common strategy is to compute T'_P as the Gateaux derivative, and then show that it is also a Fréchet derivative. If we succeed, the informal derivations above can be made rigorous. Establishing Fréchet differentiability is often challenging and technical. Unless the influence function and higher order terms are bounded, it often requires stronger regularity conditions and more effort than our arguments in the previous sections. Things may get a little easier by considering Hadamard derivatives, which are weaker than Fréchet derivatives, but stronger than Gateaux derivatives. We will not go into the details here, but take Fréchet differentiability for granted in the formal results ahead.

4.4 Formal results

We start by proving an important intermediate result.

Lemma 4.4.1. *It holds $\|\mathbb{P}_n - P\|_\infty = O_p(n^{-1/2})$.*

Proof. Observe that

$$\|\mathbb{P}_n - P\|_\infty = \sup_{x \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le x\} - \mathbb{E}[\mathbb{1}\{X \le x\}] \right|.$$

We may show as in the proof of the Glivenko-Cantelli theorem that the class of functions $\mathcal{F} = \{\mathbb{1}(\cdot \le x) \colon x \in \mathbb{R}^d\}$ has envelope $F(x) \equiv 1$ and bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) = O(\varepsilon^{-d})$. Let $[\underline{f}, \overline{f}]$ be one of the brackets. Because any $f \in \mathcal{F}$ satisfies $0 \le f \le 1$, we may assume that the bracket satisfies $|\overline{f} - \underline{f}| \le 1$. Then

$$\mathbb{E}[|\overline{f}(X) - \underline{f}(X)|^2] \le \mathbb{E}[|\overline{f}(X) - \underline{f}(X)|] \le \varepsilon.$$

Hence, the $L_2(P)$-size of the brackets is at most $\sqrt{\varepsilon}$. It follows that $N_{[]}(\varepsilon, \mathcal{F}, L_2(P)) = O(\varepsilon^{-2d})$ and by Lemma 3.4.2, $\|\mathbb{P}_n - P\|_\infty = O_p(n^{-1/2})$. □

Next we justify the form of the Fréchet derivative as a sample average over the influence function.

Lemma 4.4.2. *Suppose that T is Fréchet differentiable at P. Then the derivative takes the form*

$$T'_P(Q - P) = \mathbb{E}_Q[\rho_P(X_i)]$$

for some function ρ_P with $\mathbb{E}_P[\rho_P(X)] = 0$.

Proof. Let $T'_P(Q)$ be the Fréchet derivative of T at P. Define $\rho_P(x) = T'_P(\delta_x) - T'_P(P)$, where δ_x is the Dirac measure at x. First suppose that Q can be written as $Q = Q_N = \sum_{i=1}^N a_i \delta_{x_i}$ with $\sum_{i=1}^n a_i = 1$ and some $x_1, \dots, x_N \in \mathcal{X}$. Because the map $h \to T'_P(h)$ is linear, we have

$$T'_P(Q_n - P) = \sum_{i=1}^N a_i T'_P(\delta_{x_i}) - T'_P(P) = \sum_{i=1}^N a_i \rho_P(x_i) = \mathbb{E}_{Q_N}[\rho_P(x_i)].$$

Because we can approximate any probability measure Q by a sequence of measures of the form Q_N arbitrarily well, continuity of T'_P and the expectation functional $Q \mapsto \int \rho_P(x) dQ(X)$, implies that the same holds for any Q. Finally, observe $\mathbb{E}_P[\rho_P(X)] = T'_P(P - P) = T'_P(0) = 0$ by linearity of T'_P. □

We can now state the main theorem of this chapter.

Theorem 4.4.3. *Suppose that T is Fréchet differentiable at P and that the influence function $\rho_P(X)$ has finite variance. Then the plug-in estimator $T(\mathbb{P}_n)$ is consistent,*

$$T(\mathbb{P}_n) \to_p T(P),$$

and asymptotically normal,

$$\sqrt{n}(T(\mathbb{P}_n) - T(P)) \to_d \mathcal{N}(0, \mathbb{V}\mathrm{ar}[\rho_P(X)]).$$

Proof. By Fréchet differentiability, we have

$$T(\mathbb{P}_n) - T(P) = T'_P(\mathbb{P}_n - P) + o_p(\|\mathbb{P}_n - P\|_\infty).$$

By Lemma 4.4.1, we have $\|\mathbb{P}_n - P\|_\infty = O_p(n^{-1/2})$, so the remainder term is $o_p(n^{-1/2})$. By Lemma 4.4.2, the first order term is a sample average of the influence function, which gives

$$T(\mathbb{P}_n) - T(P) = \frac{1}{n} \sum_{i=1}^n \rho_P(X_i) + o_p(n^{-1/2}).$$

Because ρ_P has finite variance and mean zero, the law of large numbers give consistency $T(\mathbb{P}_n) - T(P) = o_p(1)$. Further, the central limit theorem and Slutsky's lemma (Lemma 2.3.9) imply asymptotic normality. □

4.5 Higher-order expansions

Rarely, a higher-order expansion is useful. A second-order expansion would look like

$$T(\mathbb{P}_n) - T(P) = T'_P(\mathbb{P}_n - P) + \frac{1}{2}T''_P(\mathbb{P}_n - P, \mathbb{P}_n - P) + o(\|\mathbb{P}_n - P\|_\infty^2),$$

where $(h_1, h_2) \mapsto T''_P(h_1, h_2)$ is a *bilinear* map, explained in a second. On the one hand, this expansion reveals the higher-order behavior of the estimator, beyond the asymptotically linear part. If the first order term is zero, however, the second order term is the leading term, and required to understand the behavior of the estimator. Bilinear maps in this context are of the form

$$T''_P(\mathbb{P}_n - P, \mathbb{P}_n - P) = \int\int \tilde{\rho}_P(x, y) d\mathbb{P}_n(x) d\mathbb{P}_n(y) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \tilde{\rho}_P(X_i, X_j),$$

for some function $\tilde{\rho}_P(x, y)$ that is symmetric ($\tilde{\rho}_P(x, y) = \tilde{\rho}_P(y, x)$) and satisfies $\mathbb{E}[\tilde{\rho}_P(x, X)] = 0$. A double sum over a kernel ρ_P is called a *V-statistic* and closely related to *U-statistics* discussed later in the course. The statistic is called degenerate if $\mathbb{E}[\rho_P(x, X)] = 0$ as above. Degenerate V-statistics converge to non-normal limits with much more complicated distribution. We will leave it at this for now.

5 Robust statistics

The field of *robust statistics* developed around the 1970s, but is still active today. Here, robustness refers to the resilience of a statistical method against contamination of the data through *outliers*. An outlier is an observation that comes from a different distribution than the bulk of the data. Outliers can be caused by measurement errors, data entry errors, or they can be genuine observations that are just very different from the rest of the data.

5.1 Motivation

To motivate the concepts ahead, let us start with simple examples. Consider a sample $X_1, \dots, X_n \in \mathbb{R}$ of size n from a distribution P. We are interested in estimating the mean $\theta = \mathbb{E}[X_1]$. The sample mean is the natural estimator for θ: $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$. Now suppose that our sample has been contaminated by an outlier. That is, one of the observations X_i is not from the distribution P but from a different distribution P'. For simplicity, let's assume that X_n has been replaced by $X_n' \sim P'$. The sample mean of the contaminated sample is

$$\widehat{\theta}' = \frac{1}{n}\sum_{i=1}^{n-1} X_i + \frac{1}{n}X_n'.$$

Now ask yourself a **first question**: how much does $\widehat{\theta}'$ differ from $\widehat{\theta}$ if X_n' is chosen maximally far away from the rest of the data? The answer is: a lot. In fact, taking $X_n' \to \infty$, we have $|\widehat{\theta} - \widehat{\theta}'| \to \infty$. So if we're unlucky, a single outlier can completely ruin our estimate! We conclude that the sample mean is not robust against outliers.

Now consider the cumulative probability $\theta = F(x)$ for some $x \in \mathbb{R}$ and the corresponding estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}.$$

The estimator on the contaminated sample is

$$\widehat{\theta}' = \frac{1}{n}\sum_{i=1}^{n-1} \mathbb{1}\{X_i \le x\} + \frac{1}{n}\mathbb{1}\{X_n' \le x\}.$$

Stop for a moment and ask yourself: For which values of X_n, X_n' does $\widehat{\theta}$ differ most from $\widehat{\theta}'$ and by how much? First observe that if $X_n, X_n' \le x$ or $X_n, X_n' > x$, the estimate does not change at all: $\widehat{\theta}' = \widehat{\theta}$. A difference only arises when $X_n \le x < X_n'$ or

$X'_n \leq x < X_n$. In this case, the estimated probability changes at most by $1/n$, which is less than the typical statistical $O_p(1/\sqrt{n})$ error. The estimator here is robust against the outlier.

Now let's ask a **second question**. How many of the observations can be outliers before the estimate is completely ruined? For the sample mean, a single estimate was enough. For the cumulative probability, suppose for simplicity that $\widehat{\theta} = 1/2$. Note that we have $|\widehat{\theta} - \widehat{\theta}'| \leq 1/2$ since $\widehat{\theta}' \in [0, 1]$. The estimator is completely ruined when this bound is attained. This is the case if we shift all $n/2$ observations to the left of x to the right or vice versa. That is, half of the entire sample must be contaminated to ruin the estimate. Again, the probability estimator is much more robust than the sample mean.

As a potential **third question**, we may ask: If an ε-fraction of the sample has been contaminated by outliers, how much does the estimate change in the worst case? For the sample mean, the answer is still ∞. For the probability estimator, the answer is $\min(\lceil \varepsilon n \rceil / n, 1/2)$.[1] Yet again, we find that the probability estimator is more robust than the sample mean.

In what follows we will formalize these ideas in a way that allows us to study more general estimators.

5.2 Contamination models

We approach the study of robustness through the lens of estimators as statistical functionals. Let $\theta_0 = T(P)$ be the parameter of interest and $\widehat{\theta} = T(\mathbb{P}_n)$ the corresponding estimator. From this perspective, robustness of an estimator is primarily a property of the functional T. This allows us to gain deep insights from studying the functional T itself, rather than dealing with a specific realization of the sample.

There are multiple ways to define quantitative measures of robustness. They all follow similar ideas and lead to qualitatively similar interpretations. It all starts with the concept of *contamination*. Let $\mathcal{P}(\mathcal{X})$ be the set of all probability measures on the set $\mathcal{X}$ and $d(P, P')$ be some measure of distance between two distributions $P, P' \in \mathcal{P}(\mathcal{X})$. Most generally, the *contamination neighborhood* of $P \in \mathcal{P}(\mathcal{X})$ is defined as the set

$$\mathcal{P}_\varepsilon(P, \mathcal{X}) = \{P' \in \mathcal{P}(\mathcal{X}) : d(P, P') \leq \varepsilon\}. \tag{5.1}$$

This collects the set of distributions P' that are ε-close to P in the distance d. We may now ask how much the value of $T(P)$ changes when P is replaced by the worst-case distribution in the contamination neighborhood. We may also ask, how much ε needs to be to make $T(P)$ attain the boundary of its range of potential values (often $\pm\infty$).

To simplify our study, we focus on a slightly simpler notion of contamination.

[1] The quantity $\lceil x \rceil$ is defined as the smallest integer at least as large as x. Colloquially: 'x rounded up'.

Definition 5.2.1. *For given $\varepsilon > 0$, we define the* ***contamination neighborhood*** *of $P \in \mathcal{P}(\mathcal{X})$ as*

$$\mathcal{P}_\varepsilon(P, \mathcal{X}) = \{P' = (1-\varepsilon)P + \varepsilon\delta_x \colon x \in \mathcal{X}\}$$

where δ_x is the Dirac measure (point mass) at x.

The contaminated distributions P' in the above model are a mixture of the original distribution P and a point mass at x. The mixture weights are given by $(1-\varepsilon)$ and ε. In the motivating examples above, we can interpret this as replacing an ε-fraction of the sample by outliers with value x. Compared to the general definition (5.1) of contamination neighborhoods, we (i) only allow for mixtures of the original distribution and point masses, and (ii) insist on all outliers to take the same value x. This simplifies both interpretation and mathematical analysis while still capturing the essence of the problem.

5.3 Measures of robustness

We shall discuss several measures of robustness that are commonly used in the literature. They correspond to the three different questions we asked in the motivating section.

5.3.1 Influence function and error sensitivity

The *influence function* is a measure of how much the value of $T(P)$ changes when the distribution P is replaced by an infinitesimally contaminated distribution $P' \in \mathcal{P}_\varepsilon(P, x)$.

Definition 5.3.1. *Let $T : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be a functional. The* ***influence function*** *of T at $x \in \mathcal{X}$ and $P \in \mathcal{P}(\mathcal{X})$ is defined as*

$$\mathrm{IF}_T(x, P) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)P + \varepsilon\delta_x) - T(P)}{\varepsilon}.$$

We used the term 'influence function' before and this is not a coincidence. Because Fréchet differentiability implies Gateaux differentiability, the function $\mathrm{IF}_T(x, P)$ is equal to $\mathbb{E}_{X\sim\delta_x}[\rho_P(X)] = \rho_P(x)$ defined in Lemma 4.4.2. In the robustness context, the interpretation is as follows. If the distribution P is replaced by a (slightly) contaminated distribution $P' = (1-\varepsilon)P + \varepsilon\delta_x$, the value of the functional T changes by approximately $\varepsilon\,\mathrm{IF}_T(x, P)$. The influence function measures the sensitivity of the functional T to an infinitesimal contamination of P by a point mass at x.

The influence functions depends on the value of x of the contamination. Not all values of x are necessarily problematic. For example, the sample mean isn't very sensitive to contaminations at $x = \mathbb{E}[X]$. In fact, this contamination makes the estimate better! When we speak about robustness, we always look for the worst-case contamination.

Definition 5.3.2. *The* ***gross error sensitivity*** *of T at P is defined as*

$$\gamma_T(P) = \sup_{x \in \mathcal{X}} |\operatorname{IF}_T(x, P)|.$$

The gross error sensitivity answers the question: how much can the value of $T(P)$ change when P is replaced by the worst-case (infinitesimally) contaminated distribution P'. If the value is small, the functional T is robust. If the value is large, the functional is not robust. Another interpretation is as follows. If $\varepsilon \ll 1$, the value of $T(P)$ changes by approximately $\varepsilon\gamma_T(P)$ in the worst-case. In the finite sample setting from the motivating examples, we may choose $\varepsilon = 1/n$ which corresponds to 'replacing one of the observations'. The gross error sensitivity thus formalizes the answer to our first question from the motivation.

Example 5.3.3 (Sample mean)**.** *Consider the sample mean $T(P) = \mathbb{E}_P[X]$ and the influence function $\operatorname{IF}_T(x, P) = x - \mathbb{E}_P[X]$. Hence, the gross error sensitivity is*

$$\gamma_T(P) = \sup_{x \in \mathcal{X}} |\operatorname{IF}_T(x, P)| = \infty.$$

The sample mean is infinitely sensitive to outliers.

Example 5.3.4 (Cumulative probability)**.** *Consider the cumulative probability $T(P) = \mathbb{E}_p[\mathbb{1}\{X \leq x\}] = \mathbb{P}(X \leq x) = F(x)$. The influence function is $\operatorname{IF}_T(x, P) = \mathbb{1}\{X \leq x\} - F(x)$, which yields*

$$\gamma_T(P) = \sup_{x \in \mathcal{X}} |\operatorname{IF}_T(x, P)| \leq 1.$$

As in the motivating example, this suggests that replacing an $\varepsilon = 1/n$ fraction of the sample by a worst-case outlier changes the estimate by at most $1/n$.

The gross error sensitivity indeed seems to formalize our intuition from the motivating examples and leads to the exactly same conclusions. The more abstract definition given here has some benefits. First, we do not have to reason about a finite sample and distinguish between different realizations of the sample. Second, we can apply the same reasoning to more complex functionals T. Whenever we know the influence function, we can compute the gross error sensitivity. In particular, in Example 4.3.4 we derived the influence function of a Z-estimator as

$$\rho_P(x) = -\mathbb{E}[\nabla\psi_{\theta_P}(X)]^{-1}\psi_{\theta_P}(x),$$

which gives

$$\gamma_T(P) = \sup_{x \in \mathcal{X}} \left|\mathbb{E}[\nabla\psi_{\theta_P}(X)]^{-1}\psi_{\theta_P}(x)\right|.$$

This makes it easier to study different M- or Z-estimators of the same parameter θ_P.

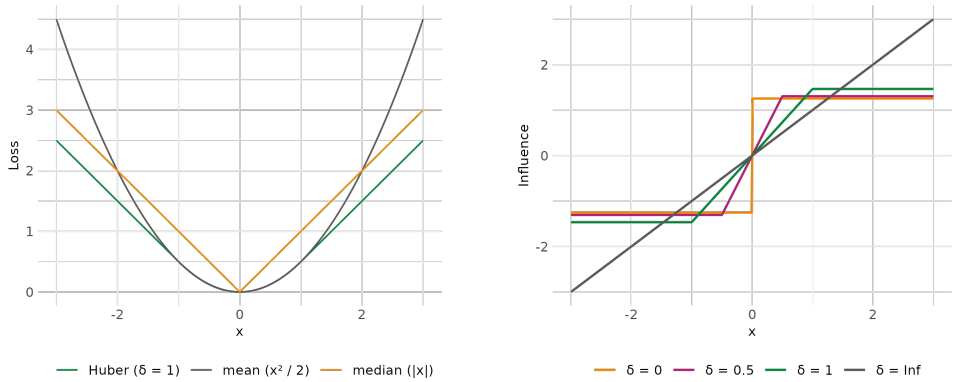

Figure 5.1: The Huber loss function (left) and its influence function (right) for P the standard normal distribution; $\delta = 0$ corresponds to the median, and $\delta = \infty$ to the mean.

For example, in the location model from Example 3.5.1 the sample mean and median are estimators of the same parameter. Reasoning about the sample median from a finite-sample perspective is somewhat awkward. In the abstract setting considered here, this is quite easy.

Example 5.3.5 (Sample median)**.** *Consider the sample median* $\theta_P = T(P) = F^{-1}(1/2)$ *which has influence function*

$$\mathrm{IF}_T(x, P) = \frac{\mathbb{1}\{x \leq \theta_P\} - 1/2}{2 f_X(\theta_P)}.$$

The gross error sensitivity is

$$\gamma_T(P) = \sup_{x \in \mathcal{X}} |\mathrm{IF}_T(x, P)| = \frac{1}{4 f_X(\theta_P)}.$$

Compared the sample mean, the sample median is much more robust against outliers. In particular, if $f_X(\theta_P)$ *is large, many observations fall close to the median* θ_P *and the sample median is very robust. The median is less robust if* $f_X(\theta_P)$ *is small. This makes sense: there are few observations close to* θ_P*, so replacing the observations that is closest to* θ_P *changes the median to the next closest observation, which may be far away. Nevertheless, the median is always more robust than the sample mean whose gross error sensitivity is infinite.*

In the location model, the sample median trades some potential loss in efficiency for robustness. This is a common theme in robust statistics. The mean and median are in a way extremes of robustness. The mean is very efficient but not robust at all. The median is very robust but not very efficient. Many estimators provide a trade-off between these two extremes. One example is given in the following.

Example 5.3.6 (Huber loss)**.** *Consider again the location model from Example 3.5.1.*

An alternative estimator is the minimizer of the Huber loss*:*

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} h_\delta(X_i - \theta), \quad \textit{where} \quad h_\delta(x) = \begin{cases} x^2/2 & \textit{if } |x| \le \delta, \\ \delta|x| - \delta^2/2 & \textit{if } |x| > \delta. \end{cases}$$

The Huber loss function is shown in the left panel of Fig. 5.1. It behaves like the square loss (used for the sample mean) at small values of $|x-\theta|$ and like the absolute value loss (used for the median) for large values. The parameter δ controls where the transition happens. The Huber loss is a compromise between the two and is often used in robust statistics. Let us compute the IF of the Huber loss. We have

$$h'_\delta(x) = \psi_\delta(x) = \begin{cases} x & \textit{if } |x| \le \delta, \\ \delta \operatorname{sign}(x) & \textit{if } |x| > \delta. \end{cases},$$

and

$$\mathbb{E}[\psi_\delta(X-\theta)] = \mathbb{E}[(X-\theta_0)\mathbb{1}\{|X-\theta| \le \delta\}] + \delta\mathbb{E}[\operatorname{sign}(X-\theta)\mathbb{1}\{|X-\theta| > \delta\}].$$

If f_X is symmetric around θ_0, both terms are indeed zero if $\theta = \theta_0$. For the influence function, we need the derivative of this expression with respect to θ evaluated at θ_0. It holds

$$\begin{aligned} \frac{\partial}{\partial\theta_0}\mathbb{E}[X\mathbb{1}\{|X-\theta_0| \le \delta\}] &= \frac{\partial}{\partial\theta_0}\int_{\theta_0-\delta}^{\theta_0+\delta} x f_X(x)\,\mathrm{d}x \\ &= (\theta_0+\delta)f_X(\theta_0+\delta) - (\theta_0-\delta)f_X(\theta_0-\delta) \\ &= 2\delta f_X(\theta_0+\delta), \end{aligned}$$

where we used symmetry of f_X around θ_0, i.e., $f_X(\theta_0+\delta) = f_X(\theta_0-\delta)$, in the last step. Next,

$$\begin{aligned} \frac{\partial}{\partial\theta_0}\mathbb{E}[-\theta_0\mathbb{1}\{|X-\theta_0| \le \delta\}] &= -\frac{\partial}{\partial\theta_0}\theta_0[F_X(\theta_0+\delta) - F_X(\theta_0-\delta)] \\ &= -[F_X(\theta_0+\delta) - F_X(\theta_0-\delta)], \end{aligned}$$

using the product rule for derivatives and symmetry. Finally,

$$\begin{aligned} \frac{\partial}{\partial\theta_0}\mathbb{E}[\delta\operatorname{sign}(X-\theta_0)\mathbb{1}\{|X-\theta_0| > \delta\}] &= \delta\frac{\partial}{\partial\theta_0}\int_{\theta_0+\delta}^{\infty} f_X(x)\,\mathrm{d}x - \delta\frac{\partial}{\partial\theta_0}\int_{-\infty}^{\theta_0-\delta} f_X(x)\,\mathrm{d}x \\ &= -\delta f_X(\theta_0+\delta) - \delta f_X(\theta_0-\delta) \\ &= -2\delta f_X(\theta_0+\delta). \end{aligned}$$

Taking everything together, the influence function of the Huber loss is given by

$$\begin{aligned}\mathrm{IF}_T(x, P) &= \frac{\psi_\delta(x-\theta_0)}{\frac{\partial}{\partial\theta_0}\mathbb{E}[\psi_\delta(X-\theta_0)]} \\ &= \frac{(x-\theta_0)\mathbb{1}\{|x-\theta_0| \le \delta\} + \delta\,\mathrm{sign}(x-\theta_0)\mathbb{1}\{|x-\theta_0| > \delta\}}{-[F_X(\theta_0+\delta) - F_X(\theta_0-\delta)]}.\end{aligned}$$

The influence function is shown in the right panel of Fig. 5.1. The influence function approaches that of the mean as $\delta \to \infty$, since $\lim_{\delta\to\infty}[F_X(\theta_0+\delta) - F_X(\theta_0-\delta)] = 1$. It approaches the IF of the median as $\delta \to 0$, since $\lim_{\delta\to 0}[F_X(\theta_0+\delta) - F_X(\theta_0-\delta)]/\delta = 2f_X(\theta_0)$. The gross error sensitivity is

$$\gamma_T(P) = \sup_{x\in\mathcal{X}} |\,\mathrm{IF}_T(x, P)| = \frac{\delta}{F_X(\theta_0+\delta) - F_X(\theta_0-\delta)}.$$

This approaches $1/2f_X(\theta_0)$ for $\delta \to 0$ and ∞ for $\delta \to \infty$, again corresponding to the error sensitivities of the mean and median, respectively. So the Huber loss indeed interpolates between the two extreme cases of robustness and non-robustness.

Example 5.3.7 (Trimmed mean)**.** *Another robust estimator of location is the trimmed mean. An α-trimmed mean can be defined as*

$$\widehat{\theta} = \frac{\sum_{i=1}^n X_i \mathbb{1}\{X_i \in [F_X^{-1}(\alpha/2), F_X^{-1}(1-\alpha/2)]\}}{\sum_{i=1}^n \mathbb{1}\{X_i \in [F_X^{-1}(\alpha/2), F_X^{-1}(1-\alpha/2)]\}},$$

which can be rewritten as an M-estimator

$$\widehat{\theta} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n m_\theta(X_i), \quad \text{with } m_\theta(x) = (x-\theta)^2\mathbb{1}\{x \in [F_X^{-1}(\alpha/2), F_X^{-1}(1-\alpha/2)]\},$$

or Z-estimator

$$\frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i - \widehat{\theta}) = 0, \quad \text{with } \psi_\theta(x) = (x-\theta)\mathbb{1}\{x \in [F_X^{-1}(\alpha/2), F_X^{-1}(1-\alpha/2)]\}.$$

The trimmed mean discards all observations smaller than the $\alpha/2$-quantile and larger than the $(1-\alpha/2)$-quantile. It then computes the mean of only the remaining observations. In practice, the true quantiles are not known and replaced by sample quantiles. Deriving the influence function and error sensitivity of the trimmed mean is left as an exercise.

Example 5.3.8 (Winsorized mean)**.** *A related estimator is the Winsorized mean. The Winsorized mean replaces all observations smaller than the $\alpha/2$-quantile by the $\alpha/2$-quantile and all observations larger than the $(1-\alpha/2)$-quantile by the $(1-\alpha/2)$-quantile. It then computes the mean of the modified sample. The Winsorized mean is a compromise between the sample mean and the trimmed mean. It is more robust than*

the sample mean but less robust than the trimmed mean.

5.3.2 Maximum bias and breakdown point

The gross error sensitivity only considers relatively mild forms of contamination, because ε is infinitesimally small. We now introduce two concepts that assess the robustness of a functional T against more severe forms of contamination.

Definition 5.3.9. *The **maximum bias** of T at P is defined as*

$$b_T(\varepsilon, P) = \sup_{P' \in \mathcal{P}_\varepsilon(P, \mathcal{X})} |T(P') - T(P)|.$$

The maximum bias measures how much the value of $T(P)$ can change when P is replaced by the worst-case distribution in an ε-contamination neighborhood. Here, we no longer require ε to be infinitesimally small. The maximum bias is often studied under the general contamination model (5.1). We shall continue to use the simpler model from Definition 5.2.1, however.

At $\varepsilon = 1$, we replace the actual distribution P entirely by another distribution that may be completely different. Consider the boundary value

$$\bar{b}_T = \lim_{\varepsilon \to 1} b_T(\varepsilon, P).$$

Depending on the functional under study, this quantity may be finite or infinite. For example, if $T(P) \in [0, 1]$ for all measures P, it holds $\bar{b}_T \leq 1$. For the mean functional $T(P) = \mathbb{E}_P[X]$, it holds $\bar{b}_T = \infty$. We consider a contaminated version $T(P')$ of $T(P)$ to be ruined as soon as $|T(P) - T(P')| = \bar{b}_T$. The breakdown point is the fraction of the sample that can be contaminated before the estimate is ruined. This formalizes the second question from our motivating examples.

Definition 5.3.10. *The **breakdown point** of T at P is defined as*

$$\varepsilon_T^*(P) = \sup\{\varepsilon : b_T(\varepsilon, P) < \bar{b}_T\}.$$

Example 5.3.11 (Sample mean)**.** *Consider the sample mean $T(P) = \mathbb{E}_P[X]$. The maximum bias is*

$$b_T(\varepsilon, P) = \sup_{P' \in \mathcal{P}_\varepsilon(P, \mathcal{X})} |\mathbb{E}_{P'}[X] - \mathbb{E}_P[X]| = \varepsilon \sup_{x \in \mathcal{X}} |x - \mathbb{E}_P[X]| = \varepsilon\infty.$$

The breakdown point is $\varepsilon_T^*(P) = 0$, indicating that the most tiny contamination is enough to ruin the estimate. This is in line with our previous findings from studying the gross error sensitivity.

Example 5.3.12 (Sample median)**.** *Consider the sample median $T(P) = F_P^{-1}(1/2)$. Let P be absolutely continuous with strictly positive density everywhere on $\mathbb{R}$. Let $P' = (1-\varepsilon)P = \varepsilon\delta_x$. As $\varepsilon \to 1$ and $x \to \infty$, we have $T(P') \to \infty$, so the boundary value is $\bar{b}_T = \infty$. To make the median functional attain this value, at least half of a sample has to be changed. More precisely, let $x = \infty$ and note that*

$$\begin{aligned} 1/2 = P'\{X \le F_{P'}^{-1}(1/2)\} &= (1-\varepsilon)P\{X \le F_{P'}^{-1}(1/2)\} + \varepsilon \mathbb{1}\{x \le F_{P'}^{-1}(1/2)\} \\ &= (1-\varepsilon)F_P(F_{P'}^{-1}(1/2)). \end{aligned}$$

Solving for $F_{P'}^{-1}(1/2)$ gives

$$F_{P'}^{-1}(1/2) = F_P^{-1}\left(\frac{1}{2-2\varepsilon}\right).$$

This is finite if $\varepsilon < 1/2$ and infinite if $\varepsilon \ge 1/2$. Hence, $\varepsilon_T^*(P) = 1/2$.

5.3.3 Comments

The measures of robustness introduced above are only concerned with the population level. There are finite-sample versions of the concepts. These are defined by simply replacing the probability measure P by the (random) empirical measure $\mathbb{P}_n$ in the definitions. Additionally, it is common to insist on ε being of the form j/n for some $j \in \mathbb{N}$. With the tools developed in this course, we may now show that the finite sample versions converge to the population versions as $n \to \infty$. Hence, the population versions are often called *asymptotic* (as in: asymptotic breakdown point etc.) in the literature.

5.4 Further examples of robust estimators

We have already seen a few examples of robust estimators of location. There are robust estimators for basically any estimation target you can imagine. The two most common other targets are scale and regression. We shall discuss robust estimators for these targets in the following.

5.4.1 Robust regression

Consider the linear regression model $Y = X^\top\beta + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$. The least squares estimator is the most common estimator for β in this model. It minimizes the empirical risk with respect to the square loss $L(y, x^\top\beta) = (y - x^\top\beta)^2$:

$$\widehat{\beta}_{OLS} = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top\beta)^2.$$

The same principle is used also for nonlinear regression models $Y = f(X) + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$. Now take $\mathcal{F}$ as a class of possibly nonlinear functions (like splines or

neural networks), and estimate the regression function f by

$$\widehat{f}_{LS} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

The ideas of robust location estimation directly translate to this general setting. For example, we can use the Huber loss from Example 5.3.6 and estimate the regression function by

$$\widehat{f}_{\delta} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} h_{\delta}(Y_i - f(X_i)).$$

A crucial question is how to choose the parameter δ. Intuitively, this parameter should somehow relate to the average scale of the residuals $Y_i - f(X_i)$, which needs to be estimated, again non-robustly.

5.4.2 Robust estimation of dispersion/scale

The most common estimator of dispersion or scale is the sample standard deviation (SD):

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

The sample standard deviation measures dispersion of the sample around an estimate of location (here: the sample mean). You should already see that this estimator is not robust. A robust version of this estimator is the median absolute deviation (MAD). Let $\widehat{m} = \text{med}(X_1, \ldots, X_n)$ be the sample median. The MAD is defined as

$$\widehat{\sigma} = \text{med}(|X_1 - \widehat{m}|, \ldots, |X_n - \widehat{m}|).$$

The double use of the median is necessary to make the estimator robust. Using sample averages either to estimate location or spread around the estimated location would break robustness. Another robust alternative is the *interquartile range* (IQR) defined as the difference between the 75% and 25% quantiles of the sample.

The SD, MAD, and IQR all somehow quantify dispersion of the distribution, but they do so in different ways. For $X \sim \mathcal{N}(\mu, \sigma)$, one can show that SD estimates σ, MAD estimates $2\phi(0)\sigma$, and IQR estimates $2\Phi^{-1}(0.75)\sigma$. It is common to standardize MAD and IQR by $2\phi(0) \approx 0.8$ and $2\Phi^{-1}(0.75) \approx 1.35$ to make them valid estimators of σ, at least for the normal distribution. For other distributions, the standardization factors would be different, and are often unknown. We simply have to live with the fact that these estimators target different population quantitites. All of them are sensible estimators of dispersion in the sense that they satisfy the following properties:

- *shift invariance*: $Q(X_1, \ldots, X_n) = Q(X_1 + c, \ldots, X_n + c)$ for all $c \in \mathbb{R}$,
- *scale equivariance*: $Q(aX_1, \ldots, aX_n) = |a|Q(X_1, \ldots, X_n)$ for all $a \in \mathbb{R} \setminus \{0\}$.

Any statistic satisfying these properties is called *dispersion estimate*.

6 U-statistics

U-statistics are a generalization of sample means and are used to estimate population quantities of the form $\mathbb{E}[h(X_1, \dots, X_m)]$, involving multiple *iid* copies $X_1, \dots, X_m \sim P$. To estimate such quantities, U-statistics average over all m-tuples of the data. The 'U' in U-statistics stands for *unbiased*. The concept (and terminology) was introduced 1948 by Wassily Hoeffding and became a hot research topic in nonparametric statistics in the late 1980s and early 1990s. U-statistics have many interesting examples, including the (unbiased) sample variance, Kendall's dependence measure, the Wilcoxon signed rank statistic, and the Cramér-von Mises statistic. We shall see that, while U-statistics average over dependent random variables, they admit a law of large numbers and central limit theorem.

6.1 Definitions

Let $X_1, \dots, X_m$ be *iid* copies of a random variable X with distribution P. Suppose we are interested in estimating a parameter $\theta = \mathbb{E}[h(X_1, \dots, X_m)]$, where $h : \mathcal{X}^m \to \mathbb{R}$ is some measurable function, called *kernel*. Given a data set $X_1, \dots, X_n$, a natural estimator for θ is the average over all m-tuples of the data.

Definition 6.1.1. *A quantity of the form*

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} h(X_{i_1}, \dots, X_{i_m})$$

*is called an m-th order **U-statistic** with kernel h.*

U-statistics are natural generalizations of sample means, which are U-statistics of order $m = 1$. The statistic U_n is indeed unbiased for θ:

$$\mathbb{E}[U_n] = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} \mathbb{E}[h(X_{i_1}, \dots, X_{i_m})] = \mathbb{E}[h(X_1, \dots, X_m)] = \theta.$$

Observe that the statistic U_n averages over all m-tuples of the data, but keeps the indices in increasing order. This only makes sense if the kernel function h is symmetric in its arguments. Otherwise, averaging over unordered m-tuples would increase efficiency. Any non-symmetric kernel $\widetilde{h}$ can be symmetrized by defining

$$h(x_1, \dots, x_m) = \frac{1}{m!} \sum_{\pi \in \Pi(1, \dots, m)} \widetilde{h}(X_{\pi(1)}, \dots, X_{\pi(m)}),$$

where $\Pi(1,\dots,m)$ is a set of all permutations of $(1,\dots,m)$. For example, for $m=2$, this simply reads

$$h(x_1,x_2) = \frac{1}{2}\widetilde{h}(x_1,x_2) + \frac{1}{2}\widetilde{h}(x_2,x_1).$$

This is indeed without loss of generality since $\mathbb{E}[h(X_1,\dots,X_m)] = \mathbb{E}[\widetilde{h}(X_1,\dots,X_m)]$ and

$$\frac{1}{n(n-1)\cdots(n-m)} \sum_{1\le i_1,\dots,i_m\le n} \widetilde{h}(X_{i_1},\dots,X_{i_m}) = U_n.$$

For convenience, we shall always assume that the kernel h is symmetric.

Example 6.1.2 (Sample variance)**.** *The (unbiased) sample variance is a second-order U-statistic with kernel* $h(x_1,x_2) = \frac{1}{2}(x_1-x_2)^2$*:*

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 = \frac{1}{n(n-1)}\sum_{1\le i<j\le n}(X_i-X_j)^2 = \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\frac{1}{2}(X_i-X_j)^2.$$

It is indeed an unbiased estimator for the parameter

$$\theta = \mathbb{V}\mathrm{ar}[X_1] = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \frac{1}{2}\mathbb{E}[(X_1-X_2)^2].$$

U-statistics are closely related to *V-statistics*, which are defined as

$$V_n = \frac{1}{n^m}\sum_{1\le i_1,\dots,i_m\le n} h(X_{i_1},\dots,X_{i_m}).$$

The main difference is that V-statistics also average over terms where the same observation X_{i_j} appears multiple times in the kernel.[1] This usually leads to biased estimates $\mathbb{E}[V_n] \neq \theta$. To see this, consider the simple kernel $h(x_1,x_2) = x_1x_2$ and assume $X_1,\dots,X_n \overset{iid}{\sim} \mathcal{N}(0,1)$. Then $\mathbb{E}[h(X_1,X_2)] = \mathbb{E}[X_1X_2] = 0$, but $\mathbb{E}[h(X_1,X_1)] = \mathbb{E}[X_1^2] = 1$. The 'V' in the name relates to 'variance', since the biased version of the sample variance $n^{-1}\sum_{i=1}^n(X_i-\overline{X}_n)^2$ can be written as

$$V_n = \frac{1}{n^2}\sum_{1\le i,j\le n}\frac{1}{2}(X_i-X_j)^2.$$

In most cases, V-statistics are asymptotically equivalent to U-statistics.

Lemma 6.1.3. *If* $\mathbb{E}[|h(X_{i_1},\dots,X_{i_m})|] < \infty$ *for all* $1 \le i_1,\dots,i_m \le n$*, it holds*

$$V_n = U_n + O_p\left(\frac{1}{n}\right).$$

[1] The extra terms are sometimes called 'diagonal' of the V-statistic.

Proof. For clarity, we give the proof only for $m = 2$. Observe

$$\begin{aligned} V_n &= \frac{1}{n^2} \sum_{1 \le i,j \le n} h(X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i \neq j} h(X_i, X_j) + \frac{1}{n^2} \sum_{i=j} h(X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i \neq j} h(X_i, X_j) + \frac{1}{n^2} \sum_{i=1}^{n} h(X_i, X_i) \\ &= \frac{1}{n^2} \sum_{i \neq j} h(X_i, X_j) + O_p\left(\frac{1}{n}\right), \end{aligned}$$

since $h(X_i, X_i) = O_p(1)$ by Markov's inequality. Further, observe

$$\frac{1}{n^2} = \frac{1}{n(n-1)} \frac{(n-1)}{n} = \frac{1}{n(n-1)} \left(1 + \frac{1}{n}\right) = \frac{1}{n(n-1)} + O\left(\frac{1}{n^3}\right).$$

Since there only $O(n^2)$ pairs with $i \neq j$, and $h(X_i, X_j) = O_p(1)$, it then follows

$$\frac{1}{n^2} \sum_{i \neq j} h(X_i, X_j) = \frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j) + O_p\left(\frac{n^2}{n^3}\right) = U_n + O_p\left(\frac{1}{n}\right). \qquad \square$$

The lemma implies that if $\sqrt{n}(U_n - \theta) \not\to_p 0$, then V_n and U_n are asymptotically equivalent. We'll see that this is usually the case, but not always.

6.2 Examples

Let's see some further.

Example 6.2.1 (Wilcoxon's signed rank statistic)**.** *The Wilcoxon test is a nonparametric test for centrality of a distribution. It does so by estimating the probability* $\mathbb{P}(X_1 + X_2 > 0)$*, where* $X_1, X_2 \sim P$ *are iid copies of a random variable* X*. If* $\mathbb{P}(X_1 + X_2 > 0) \neq 1/2$*, then the distribution* P *is not centered around 0. The Wilcoxon statistic is a second-order* U*-statistic with kernel* $h(x_1, x_2) = \mathbb{1}_{\{x_1 + x_2 > 0\}}$ *estimating this probability:*

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathbb{1}_{\{X_i + X_j > 0\}}.$$

Example 6.2.2 (Kendall's τ)**.** *Kendall's* τ *is a measure of dependence between two random variables* $(X, Y) \sim P$ *that fixes some deficiencies of the Pearson correlation. In particular, it is an adequate measure for non-linear (but monotonic) relationships. Kendall's* τ *is defined as the difference between the probabiliy of concordance and discordance of two pairs of observations:*

$$\tau = \mathbb{P}\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - \mathbb{P}\{(X_1 - X_2)(Y_1 - Y_2) < 0\}.$$

The first term is the probability of concordance, i.e., the probability of 'X and Y are large/small at the same time'. The second term is the probability of discordance, i.e., the probability of 'X is large and Y is small' or vice versa. The probabilities quantify this by comparing two pairs of observations (X_1, Y_1) *and* (X_2, Y_2)*. The event* $(X_1 - X_2)(Y_1 - Y_2) > 0$ *means that* X_1 *and* Y_1 *are both larger than* X_2 *and* Y_2 *or both smaller (concordance). The event* $(X_1 - X_2)(Y_1 - Y_2) > 0$ *means that only one of* X_1 *and* Y_1 *is larger than* X_2 *and* Y_2 *(discordance). It holds* $\tau \in [-1, 1]$*,* $\tau = 0$ *if* X *and* Y *are independent,* $\tau = 1$ *if* $X = f(Y)$ *for a strictly increasing function* f*, and* $\tau = -1$ *if* $X = f(Y)$ *for a strictly decreasing function* f*. Assuming that* P *is absolutely continuous, Kendall's* τ *can be rewritten as*

$$\tau = 2\mathbb{P}\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - 1.$$

This is naturally estimated by a second-order U*-statistic with kernel* $h((x_1, y_1), (x_2, y_2)) = \mathbb{1}_{\{(x_1 - x_2)(y_1 - y_2) > 0\}}$*:*

$$\widehat{\tau} = \frac{2}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathbb{1}_{\{(X_i - X_j)(Y_i - Y_j) > 0\}} - 1.$$

Example 6.2.3 (Cramér-von Mises statistic)**.** *The Cramér-von Mises statistic is used to measure the goodness-of-fit test of a reference distribution* F^* *to a sample* $X_1, \ldots, X_n \sim F$*. The population quantity of interest is the expected squared difference between the two distribution functions* $F*$ *and* F*:*

$$\theta = \mathbb{E}[(F^*(X) - F(X))^2] = \int (F^*(x) - F(x))^2 dF(x).$$

The true distribution F *is of course unknown, but the quantity above can be estimated It is defined as the integral of the squared difference between the empirical distribution function* F_n *and the true distribution function* F*:*

$$CvM = \int (\hat{F}_n(x) - F(x))^2 dF(x).$$

Expanding the square and rearringing, we have

$$CvM = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int (\mathbb{1}_{X_i \le x} - F(x))(\mathbb{1}_{X_j \le x} - F(x)) dF(x),$$

a second-order V*-statistic. This one, we shall see, is* not *asymptotically equivalent to the corresponding* U*-statistic.*

Another way that U-statistics arise naturally is when estimating a parameter from nonparametric *pseudo-samples*. The general setup is as follows: We want to estimate a

parameter α from a sample $X_1, \dots, X_n$ through an estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \psi_\alpha(g(X_i)) = 0,$$

where ψ_α is a known identifying function, but the function g is not known. Instead, we estimate g by a nonparametric estimator $\widehat{g}$ from the sample, and then solve the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{\widehat{\alpha}}(\widehat{g}(X_i)) = 0.$$

Now assume that ψ_α is sufficiently differentiable and $\widehat{g}$ is asymptotically linear in the sense

$$\sup_x \left| \widehat{g}(x) - g(x) - \frac{1}{n}\sum_{i=1}^{n} \gamma(x, X_i) \right| = o_p(1/\sqrt{n}),$$

for some function γ. Then $\widehat{\alpha}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{\widehat{\alpha}}(g(X_i)) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \psi'_{\widehat{\alpha}}(g(X_i))\gamma(X_i, X_j) + o_p(1/\sqrt{n}) = 0,$$

and we now have to deal with a V-statistic to analyze the asymptotic properties of $\widehat{\alpha}$.

6.3 Consistency

U-statistics are consistent estimators for $\theta = \mathbb{E}[h(X_1, \dots, X_m)]$ under mild conditions.

Theorem 6.3.1. *If $\mathbb{E}[|h(X_1, \dots, X_m)|^2] < \infty$, the U-statistic U_n is a consistent estimator for $\theta = \mathbb{E}[h(X_1, \dots, X_m)]$:*

$$U_n \xrightarrow{p} \theta.$$

Proof. We already know that $\mathbb{E}[U_n] = \theta$. To show convergence in probability, we use Chebyshev's inequality:

$$\mathbb{P}(|U_n - \theta| > \varepsilon) \leq \frac{\mathbb{V}\mathrm{ar}[U_n]}{\varepsilon^2}.$$

It thus remains to show that $\mathbb{V}\mathrm{ar}[U_n] \to 0$. We have

$$\mathbb{V}\mathrm{ar}[U_n] = \frac{1}{\binom{n}{m}^2} \sum_{1 \leq i_1 < \cdots < i_m \leq n} \sum_{1 \leq j_1 < \cdots < j_m \leq n} \mathbb{C}\mathrm{ov}[h(X_{i_1}, \dots, X_{i_m}), h(X_{j_1}, \dots, X_{j_m})].$$

Define

$$\xi_k = \mathbb{C}\mathrm{ov}[h(X_{i_1}, \dots, X_{i_m}), h(X_{j_1}, \dots, X_{j_m})], \quad \text{for } |\{i_1, \dots, i_m\} \cap \{j_1, \dots, j_m\}| = k.$$

By symmetry of the kernel, this covariance only depends on the number of overlapping indices, but not on the specific indices. Since $X_1, \dots, X_m$ are independent, $\xi_0 = 0$. To count how many terms we have with $|\{i_1, \dots, i_m\} \cap \{j_1, \dots, j_m\}| = k$, do the following: 1) choose $i_1 < \cdots < i_m$ arbitrarily (there are $\binom{n}{m}$ ways to do this); 2) choose k overlapping indices from $\{i_1, \dots, i_m\}$ (there are $\binom{m}{k}$ ways to do this); 3) choose the remaining $m-k$ indices from the remaining $n-m$ indices (there are $\binom{n-m}{m-k}$ ways to do this). In total we have

$$\mathbb{V}\mathrm{ar}[U_n] = \frac{1}{\binom{n}{m}^2} \sum_{k=1}^{m} \binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}\xi_k = \sum_{k=1}^{m} O\left(\frac{1}{n^k}\right)\xi_k = O\left(\frac{1}{n}\right). \qquad \square$$

From the proof, we actually learn a little more. Markov's inequality gives

$$U_n - \theta = O_p\left(\mathbb{V}\mathrm{ar}[U_n]^{1/2}\right) = O_p(1/\sqrt{n}),$$

so we recover the 'usual' rate of convergence. This suggests that we may hope to find a non-trivial limiting distribution for the quantity $\sqrt{n}(U_n - \theta)$.

6.4 Normality via Hoeffding's decomposition

To see that the limit should be normal, we first give a direct proof for the case $m = 2$. Define $h_1(x) = \mathbb{E}[h(x, X_j)]$ and consider the following decomposition

$$\begin{aligned} U_n - \theta &= \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j) - \theta \\ &= \underbrace{\frac{2}{n}\sum_{i=1}^{n}[h_1(X_i) - \theta]}_{:=A_n} + \underbrace{\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \left(h(X_i, X_j) - h_1(X_i) - h_1(X_j) + \theta\right)}_{=:B_n}. \end{aligned}$$

This is called *Hoeffding decomposition* of the U-statistic. Observe that

$$\mathbb{E}[h_1(X_i)] = \mathbb{E}[h(X_i, X_j)] = \theta,$$

so both terms in the decomposition have mean zero. The first term is a sample mean and converges to a normal limit by the CLT:

$$\sqrt{n}A_n \xrightarrow{d} \mathcal{N}(0, 4\mathbb{V}\mathrm{ar}[h_1(X_1)]).$$

The term B_n is also a U-statistic, but asymptotically negligible. To see this, define

$$\eta(X_i, X_j) = h(X_i, X_j) - h_1(X_i) - h_1(X_j) + \theta,$$

and observe

$$\mathbb{V}\mathrm{ar}\left[B_n\right] = \frac{1}{n^2(n-1)^2} \sum_{1 \le i_1 \ne i_2 \le n} \sum_{1 \le j_1 \ne j_2 \le n} \mathbb{E}[\eta(X_{i_1}, X_{i_2})\eta(X_{j_1}, X_{j_2})].$$

Since

$$\mathbb{E}[\eta(X_i, X_j) \mid X_i] = \mathbb{E}[\eta(X_i, X_j) \mid X_j] = 0,$$

all terms with $\{i_1, i_2\} \neq \{j_1, j_2\}$ are in fact zero. For example, if $i_1 = j_1$, but $i_2 \neq j_2$, we have

$$\begin{aligned}
\mathbb{E}[\eta(X_{i_1}, X_{i_2})\eta(X_{j_1}, X_{j_2})] &= \mathbb{E}\left[\mathbb{E}[\eta(X_{i_1}, X_{i_2})\eta(X_{j_1}, X_{j_2}) \mid X_{j_1}, X_{j_2}]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\eta(X_{i_1}, X_{i_2}) \mid X_{j_1}, X_{j_1}]\, \eta(X_{j_1}, X_{j_2})\right] \\
&= \mathbb{E}\left[\mathbb{E}[\eta(X_{i_1}, X_{i_2}) \mid X_{i_1}]\, \eta(X_{j_1}, X_{j_2})\right] \\
&= 0.
\end{aligned}$$

Since there are only $O(n^2)$ terms with $\{i_1, i_2\} = \{j_1, j_2\}$, we have $\mathbb{V}\mathrm{ar}[B_n] = O(1/n^2)$, so $\sqrt{n}B_n \to_p 0$.

We have shown the following result.

Proposition 6.4.1. *Let $m = 2$ and $\xi_1 = \mathbb{V}\mathrm{ar}[h_1(X_1)] \in (0, \infty)$ and $\mathbb{E}[h(X_1, X_2)^2] < \infty$. Then*

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, 4\xi_1).$$

The argument can be generalized to higher order U-statistics, but the notation and computations become tedious. We will see a more elegant approach in the next section.

6.5 Normality via Hájek's projection principle*

Normality of complicated statistics like higher-order U-statistics can be established by a general approach called *Hájek's projection principle*. The idea is somewhat abstract but extremely elegant.

Random variables can be seen as elements of the vector space

$$\mathcal{L}^2 = \{X : \mathbb{E}[X^2] < \infty\} = \left\{X : \Omega \to \mathbb{R},\ s.t. \int X^2(\omega) dP(\omega) < \infty\right\}$$

of square-integrable random variables equipped with the inner product $\langle X, Y\rangle = \mathbb{E}[XY]$. Let $\mathcal{S}$ be a linear subspace of $\mathcal{L}^2$, meaning that $aX + bY \in \mathcal{S}$ for all $X, Y \in \mathcal{S}$ and $a, b \in \mathbb{R}$. We can now define the *projection* of an arbitrary random variable T onto $\mathcal{S}$ by the random variable $\Pi_{\mathcal{S}} T$ that minimizes the distance in $\mathcal{L}^2$:

$$\mathbb{E}[(T - \Pi_{\mathcal{S}} T)^2] = \min_{S \in \mathcal{S}} \mathbb{E}[(T - S)^2].$$

A standard result for Hilbert spaces says that $\widehat{S}$ is a projection if and only if the following orthogonality condition holds:

$$\mathbb{E}[(T - \widehat{S})S] = 0 \quad \text{for all } S \in \mathcal{S}.$$

This view is now exploited as follows. Let T_n be some statistic with complicated form (like a U-statistic). We project this estimator onto a subspace of random variables that are easier to analyze. The hope is then that the projection is so close to the original estimator that the difference between T_n and its projection $\widehat{S}_n$ is asymptotically negligible.

Lemma 6.5.1. *Let $\mathcal{S}_n \subseteq \mathcal{L}^2$ be linear spaces of random variables containing the constants. Then if $\mathbb{V}\mathrm{ar}[T_n]/\mathbb{V}\mathrm{ar}[\Pi_{\mathcal{S}_n} T_n] \to 1$, it holds*

$$\frac{T_n - \mathbb{E}[T_n]}{\sqrt{\mathbb{V}\mathrm{ar}[T_n]}} - \frac{\Pi_{\mathcal{S}_n} T_n - \mathbb{E}[\Pi_{\mathcal{S}_n} T_n]}{\sqrt{\mathbb{V}\mathrm{ar}[\Pi_{\mathcal{S}_n} T_n]}} = o_p(1).$$

Proof. We will show that the variance of the term in question goes to zero. Write $\widehat{S}_n = \Pi_{\mathcal{S}_n} T_n$. Since $\mathcal{S}_n$ contains constants, the orthogonality condition

$$\mathbb{E}[(T_n - \widehat{S}_n)S] = 0 \quad \text{for all } S \in \mathcal{S}_n,$$

implies $\mathbb{E}[T_n] = \mathbb{E}[\widehat{S}_n]$. Since both T_n and $\widehat{S}_n$ are centered in the expression in question, we can assume without loss of generality that $\mathbb{E}[T_n] = \mathbb{E}[\widehat{S}_n] = 0$. We have

$$\mathbb{V}\mathrm{ar}\left[\frac{T_n}{\sqrt{\mathbb{V}\mathrm{ar}[T_n]}} - \frac{\widehat{S}_n}{\sqrt{\mathbb{V}\mathrm{ar}[\widehat{S}_n]}}\right] = 2 - 2\frac{\mathbb{C}\mathrm{ov}[T_n, \widehat{S}_n]}{\sqrt{\mathbb{V}\mathrm{ar}[T_n]}\sqrt{\mathbb{V}\mathrm{ar}[\widehat{S}_n]}}.$$

and

$$\mathbb{C}\mathrm{ov}[T_n, \widehat{S}_n] = \mathbb{E}[T_n \widehat{S}_n] = \mathbb{E}[(T_n - \widehat{S}_n)\widehat{S}_n] + \mathbb{E}[\widehat{S}_n^2] = 0 + \mathbb{E}[\widehat{S}_n^2] = \mathbb{V}\mathrm{ar}[\widehat{S}_n].$$

Thus, using the assumption $\mathbb{V}\mathrm{ar}[T_n]/\mathbb{V}\mathrm{ar}[\Pi_{\mathcal{S}_n} T_n] \to 1$, we get

$$2 - 2\frac{\mathbb{C}\mathrm{ov}[T_n, \widehat{S}_n]}{\sqrt{\mathbb{V}\mathrm{ar}[T_n]}\sqrt{\mathbb{V}\mathrm{ar}[\Pi_{\mathcal{S}_n} T_n]}} \to 0. \qquad \square$$

So which linear space of random variables is suitable to establish normality? A natural choice is a space of sample averages. Let

$$\mathcal{S}_n = \left\{ \frac{1}{n} \sum_{i=1}^{n} g(X_i) \colon g(X_i) \in \mathcal{L}^2 \right\}. \tag{6.1}$$

We have the following result.

Lemma 6.5.2. *The projection of a random variable $T \in \mathcal{L}^2$ with $\mathbb{E}[T] = 0$ onto the space $\mathcal{S}_n$ defined in (6.1) is*

$$\widehat{S}_n = \sum_{i=1}^{n} \mathbb{E}[T \mid X_i].$$

Proof. The term on the right is an element of $\mathcal{S}_n$ with the choice $g(X_i) = n\mathbb{E}[T \mid X_i]$. To show that it is a projection, we can verify the orthogonality condition. For an

arbitrary element

$$S_n = \frac{1}{n}\sum_{i=1}^{n} \overline{g}(X_i) \in \mathcal{S}_n,$$

it holds

$$\mathbb{E}[(T - \widehat{S}_n)S_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(T - \widehat{S}_n)\overline{g}(X_i)].$$

Further

$$\mathbb{E}[(T - \widehat{S}_n)\overline{g}(X_i)] = \mathbb{E}\left[\mathbb{E}[(T - \widehat{S}_n)\overline{g}(X_i) \mid X_i]\right] = \mathbb{E}\left[\mathbb{E}[(T - \widehat{S}_n) \mid X_i]\, \overline{g}(X_i)\right]$$

and

$$\mathbb{E}[(T - \widehat{S}_n) \mid X_i] = \mathbb{E}[T \mid X_i] - \sum_{j=1}^{n} \mathbb{E}[\mathbb{E}[T \mid X_j] \mid X_i] = 0,$$

because $\mathbb{E}[\mathbb{E}[T \mid X_j] \mid X_i] = \mathbb{E}[T] = 0$ for $i \neq j$ and $\mathbb{E}[\mathbb{E}[T \mid X_j] \mid X_i] = \mathbb{E}[T \mid X_i]$ for $i = j$. We have shown

$$\mathbb{E}[(T - \widehat{S}_n)S_n] = 0, \quad \text{for all } S_n \in \mathcal{S}_n,$$

so $\widehat{S}_n$ is indeed a projection. □

We now apply Hájek's projection principle to a general m-th order U-statistic U_n.

Lemma 6.5.3. *Let U_n be an m-th order U-statistic with symmetric kernel h, $\mathbb{E}[h(X_{i_1}, \ldots, X_{i_m})^2] < \infty$, and $\mathcal{S}_n$ be the space defined in (6.1). Define*

$$h_1(x) = \mathbb{E}[h(x, X_2, \ldots, X_m)].$$

Then the projection of U_n onto $\mathcal{S}_n$ defined in (6.1) is

$$\widehat{S}_n = \frac{m}{n}\sum_{i=1}^{n} h_1(X_i).$$

Proof. We have

$$\mathbb{E}[U_n - \theta \mid X_i] = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} \mathbb{E}[h(X_{i_1}, \ldots, X_{i_m}) \mid X_i] - \theta$$

It holds $\mathbb{E}[h(X_{i_1}, \ldots, X_{i_m}) \mid X_i] = \theta$ if $i \notin \{i_1, \ldots, i_m\}$. This happens for $\binom{n-1}{m}$ terms. For the remaining

$$\binom{n}{m} - \binom{n-1}{m} = \binom{n-1}{m-1}$$

terms, $\mathbb{E}[h(X_{i_1}, \dots, X_{i_m}) \mid X_i] = h_1(X_i)$. This gives

$$\mathbb{E}[U_n - \theta \mid X_i] = \frac{\binom{n-1}{m-1}}{\binom{n}{m}} h_1(X_i) = \frac{m}{n} h_1(X_i),$$

so the projection of U_n onto $\mathcal{S}_n$ is

$$\widehat{S}_n = \sum_{i=1}^{n} \mathbb{E}[U_n \mid X_i] = \frac{m}{n} \sum_{i=1}^{n} h_1(X_i). \qquad \square$$

To obtain a general normality result for U-statistics, it only remains to put things together.

Theorem 6.5.4. *Let U_n be an m-th order U-statistic with symmetric kernel h and define*

$$h_1(x) = \mathbb{E}[h(x, X_2, \dots, X_m)].$$

If $\mathbb{E}[h(X_1, \dots, X_m)^2] < \infty$ and $\xi_1 = \mathbb{V}\mathrm{ar}[h_1(X_1)] \in (0, \infty)$, it holds

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, m^2 \xi_1).$$

Proof. We apply Lemma 6.5.1 with $T_n = U_n$ and $\mathcal{S}_n$ as defined in (6.1). By Lemma 6.5.3, the projection of U_n onto $\mathcal{S}_n$ is given by

$$\widehat{S}_n = \sum_{i=1}^{n} \mathbb{E}[U_n \mid X_i] = \frac{m}{n} \sum_{i=1}^{n} h_1(X_i).$$

By the CLT and Slutsky's lemma, we have

$$\frac{\widehat{S}_n - \mathbb{E}[\widehat{S}_n]}{\sqrt{\mathbb{V}\mathrm{ar}[\widehat{S}_n]}} \to_d \mathcal{N}(0, 1),$$

where $\mathbb{V}\mathrm{ar}[\widehat{S}_n] = \frac{m^2}{n} \xi_1$. Furthermore,

$$\mathbb{V}\mathrm{ar}[U_n] = \frac{m^2}{n} \xi_1 + o\left(\frac{1}{n}\right),$$

by the computations in the proof of Theorem 6.3.1, so $\mathbb{V}\mathrm{ar}[U_n] / \mathbb{V}\mathrm{ar}[\widehat{S}_n] \to 1$. Now the claim follows from Lemma 6.5.1. $\square$

6.6 Further topics

We briefly discuss to a few more advanced topics in the theory of U-statistics. More details can be found in, e.g., Van der Vaart (2000, Chapter 12)

6.6.1 Degenerate U-statistics and non-normal limits

Normality of U-statistics is not always guaranteed. In our result, we made the seemingly innocuous assumption $\xi_1 = \mathbb{V}\mathrm{ar}[h_1(X_1)] > 0$. We have actually already encountered an example of a U-statistic where this is not the case, B_n in Section 6.4. Another example is the Crámer-von Mises statistic from Example 6.2.3, which is a V-statistic with kernel

$$h(x_1, x_2) = \mathbb{E}[(\mathbb{1}_{\{x_1 \leq X\}} - F(X))(\mathbb{1}_{\{x_2 \leq X\}} - F(X))].$$

It holds

$$\begin{aligned}
\mathbb{E}[h(x_1, X_2)] &= \mathbb{E}[(\mathbb{1}_{\{x_1 \leq X\}} - F(X))(\mathbb{1}_{\{X_2 \leq X\}} - F(X))] \\
&= \mathbb{E}\left[\mathbb{E}[(\mathbb{1}_{\{x_1 \leq X\}} - F(X))(\mathbb{1}_{\{X_2 \leq X\}} - F(X)) \mid X]\right] \\
&= \mathbb{E}\left[(\mathbb{1}_{\{x_1 \leq X\}} - F(X))\mathbb{E}[\mathbb{1}_{\{X_2 \leq X\}} - F(X) \mid X]\right] \\
&= \mathbb{E}\left[(\mathbb{1}_{\{x_1 \leq X\}} - F(X)) \times 0\right] = 0.
\end{aligned}$$

U-statistics with $h_1(X_1) = \mathbb{E}[h(X_1, \ldots, X_m) \mid X_1] = 0$ almost surely are called *degenerate*. For degenerate U-statistics, the first-order term vanishes and our CLT argument no longer works. Instead, the asymptotic behavior is determined by higher-order terms, which can no longer be represented by a sample average. Additionally, the limiting behavior of V- and U-statistics is no longer the same, since their difference is of the same order as the (now dominant) second-order term in the U-statistic.

Dealing with degenerate U-statistics is messy, and the limiting distributions are complicated. These distributions are called *Gaussian chaos* and often resemble a (potentially infinite) weighted sum of χ^2 random variables. To determine the weights, one has to find the eigenvalues and -functions of an operator derived from the kernel h.

6.6.2 Multi-sample U-statistics

Another extension of U-statistics relates to multi-sample problems. Let's consider the two-sample case briefly. We have two independentamples $X_1, \ldots, X_n \overset{iid}{\sim} P$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} Q$ and want to estimate the parameter $\theta = \mathbb{E}[h(X_{i_1}, \ldots, X_{i_r}, Y_{j_1}, \ldots, Y_{j_s})]$ for some kernel h. A natural estimator is the two-sample U-statistic of order (r, s):

$$U_n = \frac{1}{\binom{n}{r}\binom{m}{s}} \sum_{1 \leq i_1 < \cdots < i_r \leq n} \sum_{1 \leq j_1 < \cdots < j_s \leq m} h(X_{i_1}, \ldots, X_{i_r}, Y_{j_1}, \ldots, Y_{j_s}). \tag{6.2}$$

Example 6.6.1. *An example for a two-sample U-statistic is the Mann-Whitney statistic, which is used to test for equality of distributions. The parameter of interest is*

$$\theta = \mathbb{P}(X \leq Y) = \mathbb{E}[\mathbb{1}_{\{X \leq Y\}}],$$

where $X \sim P$ and $Y \sim Q$ are independent. If $P = Q$, then $\theta = 1/2$. A deviation from $1/2$ indicates a difference in the distributions: if $\theta > 1/2$, X is stochastically smaller than Y, and if $\theta < 1/2$, X is stochastically larger than Y. The parameter θ can be

estimated a two-sample U-statistic of order $(1,1)$ *with kernel* $h(x,y) = \mathbb{1}_{\{x \le y\}}$*:*

$$U_n = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{\{X_i \le Y_j\}}.$$

To analyze the behavior of such a statistic, we can use similar arguments as for one-sample statistics. Assume both m and n tend to ∞ such that $n/(n+m) \to \lambda \in (0,\infty)$. Projecting U_n onto the space of random variables of the form

$$\frac{1}{n} \sum_{i=1}^{n} g_1(X_i) + \frac{1}{m} \sum_{j=1}^{m} g_2(Y_j),$$

gives

$$U_n \approx \frac{r}{n} \sum_{i=1}^{n} h_{1,1}(X_i) + \frac{s}{m} \sum_{j=1}^{m} h_{1,2}(Y_j),$$

where

$$h_{1,1}(x) = \mathbb{E}[h(x, X_2, \dots, X_r, Y_1, \dots, Y_s)], \quad h_{1,2}(y) = \mathbb{E}[h(X_1, \dots, X_r, y, Y_2, \dots, Y_s)].$$

Now the central limit theorem gives

$$\sqrt{n+m}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, r^2 \xi_{1,1}/\lambda + s^2 \xi_{1,2}/(1-\lambda)),$$

where

$$\xi_{1,1} = \mathbb{V}\mathrm{ar}[h_{1,1}(X_1)], \quad \xi_{1,2} = \mathbb{V}\mathrm{ar}[h_{1,2}(Y_1)].$$

6.6.3 U-processes

Just as sample averages can be extended to empirical processes, U-statistics can be extended to *U-processes*

$$\left\{ U_n(h) = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} h(X_{i_1}, \dots, X_{i_m}) - \mathbb{E}[h(X_1, \dots, X_m)] \colon h \in \mathcal{H} \right\},$$

where $\mathcal{H}$ is a class of symmetric kernel functions. Using Hoeffding's decomposition and bracketing/covering arguments, one can show that U-processes are often asymptotically equivalent to the corresponding empirical process

$$\left\{ \frac{m}{n} \sum_{i=1}^{n} h_1(X_i) - \mathbb{E}[h_1(X_i)] \colon h \in \mathcal{H} \right\},$$

where $h_1(x) = \mathbb{E}[h(x, X_2, \dots, X_m)]$ as before.

7 Dependent data

Our results so far rely heavily on the assumption that data are independent. This assumption is so common in statistics that we sometimes forget to question it. In many domains, however, dependent data are the norm, not the exception. The most common situation is that data are recorded over time, leading to time series. Another common example is spatial data, where observations are taken at different locations. For example, today's air temperature is correlated with yesterday's temperature, and the temperature in Munich is correlated with the temperature in Augsburg.

If the dependence is not too strong, we can still obtain meaningful results, but some care is in order. We will formalize this focusing on the most basic results underlying our theory: the law of large numbers and the central limit theorem. Everything else we built on top can be generalized as well, but we will not go into the details here.

To simplify the exposition, we focus our discussion on time series data: a sequence of random variables $(X_t)_{t\in\mathbb{N}}$, where X_t is the observation at time t. For spatial data or other general forms of dependent data, we would need tuples of indices, which only complicates the notation. The concepts should be clear, though, and can be extended to other types of dependent data.

7.1 Some preliminary considerations

It is instructive to see how statistical inference can fail when data are dependent. We consider the situation where $X_1, \dots, X_n \sim P$ have the same distribution (we call this *stationary*) but are dependent and ask how the sample average

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

behaves.

Perfect dependence First assume the extreme case of perfect dependence, where $X_1 = \cdots = X_n$ with probability 1. Clearly, $\overline{X}_n = X_1$ with probability 1. Specifically, $\overline{X}_n$ converges in probability to a random variable that is equal to X_1. Contrast this to the *iid* case where $\overline{X}_n$ converges in probability to the deterministic value $\mathbb{E}[X_1]$. Making inference about $\mathbb{E}[X_1]$ under this strong form of dependence is virtually impossible, just as it would be if we only had a single observation X_1.

Exchangeable dependence Next, let's consider a weaker form of dependence. Assume that

$$(X_1, \dots, X_n) \sim \mathcal{N}(0, \Sigma), \qquad \Sigma_{ij} = \begin{cases} 1, & i = j \\ \rho \in (0,1), & i \neq j. \end{cases}$$

By the properties of the multivariate normal, we know that

$$\overline{X}_n = \frac{1}{n}(X_1, \dots, X_n)\mathbf{1} \sim \mathcal{N}\left(0, \frac{\mathbf{1}^\top \Sigma \mathbf{1}}{n^2}\right),$$

where $\mathbf{1} = (1, \dots, 1)^\top$. Further

$$\frac{\mathbf{1}^\top \Sigma \mathbf{1}}{n^2} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \Sigma_{ij} = \frac{1}{n} + \frac{n(n-1)}{n^2}\rho \quad \overset{n\to\infty}{\longrightarrow} \rho.$$

It follows that $\overline{X}_n \to_d \mathcal{N}(0, \rho)$. The limit of $\overline{X}_n$ is still random. But the smaller ρ (= the weaker the dependence), the more information $\overline{X}_n$ reveals about $\mathbb{E}[X_1]$. The normality assumption for the data isn't even necessary for this conclusion. If we just assume that $\mathbb{E}[X_i] = 0$ and $\mathbb{C}\text{ov}(X_i, X_j) = \mathbb{1}(i = j) + \rho\mathbb{1}(i \neq j)$, we have

$$\mathbb{E}[\overline{X}_n] = 0, \quad \mathbb{V}\text{ar}[\overline{X}_n] = \frac{\mathbf{1}^\top \Sigma \mathbf{1}}{n^2} \to \rho.$$

However, we have no idea about the shape of the distribution of $\overline{X}_n$ since the central limit theorem does not apply.

m-dependence The last example should give us some indication that dependence is not a complete dealbreaker. As long as the covariance is such that $\mathbf{1}^\top \Sigma \mathbf{1}/n^2 \to 0$, a law of large numbers would hold! For example, consider the structure

$$\Sigma_{ij} = \begin{cases} 1, & i = j \\ \rho \in (0,1), & |i-j| = 1 \\ 0, & |i-j| > 1, \end{cases}$$

in which two subsequent observations are dependent, but observations at least two time steps apart are uncorrelated. Such a covariance structure could be generated, e.g., from the *moving average* model

$$X_t = \varepsilon_t + \rho\varepsilon_{t-1},$$

where ε_t is a sequence of *iid innovations* with variance $1/(1+\rho^2)$. It holds

$$\frac{\mathbf{1}^\top \Sigma \mathbf{1}}{n^2} = \frac{1}{n} + \frac{\rho}{n} \to 0,$$

so $\overline{X}_n \to_p \mathbb{E}[X_1]$; a law of large numbers. Furthermore, assuming $(X_1, \dots, X_n) \sim \mathcal{N}(0, \Sigma)$, we have the central limit theorem

$$\sqrt{n}\overline{X}_n \sim \mathcal{N}(0, 1 + \rho).$$

We see that positive dependence ($\rho > 0$) can lead to a larger variance of the sample mean than in the *iid* case, but negative dependence ($\rho < 0$) can lead to a smaller variance. The above is a special case of what is called *m-dependence* with $m = 1$. In general, a sequence $X_1, X_2, \dots$ of random variables is called m-dependent if $(X_s)_{s \le t}$ and $(X_s)_{s > t+m}$ are independent for all $t \in \mathbb{N}$. This is a simple form of asymptotic independence: the future is independent of the past provided there are m observations in between.

Markovian/autoregressive dependence Another common type of *weak depedence* arises from Markovian models. For example, take the *autoregressive model* of order 1:

$$X_t = \rho X_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, 1 - \rho^2).$$

This is a special case of Markovian dependence, which asserts that $\mathbb{P}(X_t \mid X_{t-1}, X_{t-2}, \dots) = \mathbb{P}(X_t \mid X_{t-1})$. This is equivalent to independence of X_t and $(X_{t-k})_{k>2}$ conditional on X_{t-1}. One can show that for all t,

$$\mathbb{E}[X_t] = 0, \qquad \mathbb{C}\text{ov}(X_t, X_{t-h}) = \begin{cases} 1, & h = 0 \\ \rho^{|h|}, & h \neq 0. \end{cases}$$

Then,

$$\mathbb{V}\text{ar}[X_n] = \frac{\mathbf{1}^\top \Sigma \mathbf{1}}{n^2} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho^{|i-j|} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{h=1-i}^{n-i} \rho^{|h|}.$$

Since by the geometric series formula,

$$\sum_{h=i-n}^{n-i} \rho^{|h|} \to \sum_{h=-\infty}^{\infty} \rho^{|h|} = 1 + 2 \sum_{h=1}^{\infty} \rho^h = 1 + \frac{2\rho}{1-\rho} = \frac{1+\rho}{1-\rho},$$

we get

$$\mathbb{V}\text{ar}[X_n] \approx \frac{1}{n} \frac{1+\rho}{1-\rho} \to 0,$$

so the law of large numbers $\overline{X}_n \to_p 0$ holds. Further, we have the central limit theorm

$$\sqrt{n}\overline{X}_n \sim \mathcal{N}\left(0, \frac{1+\rho}{1-\rho}\right).$$

Again, the variance of the sample mean can be larger or smaller than in the *iid* case, depending on the sign of ρ.

Conclusion The above examples show that the law of large numbers and the central limit theorem can hold for dependent data. The key is that the dependence is not too strong. More specifically, the examples that we've seen exhibit a form of *mixing* behavior, where the dependence between observations decays as the distance between them increases. This is a common property of many time series models, and it is what will allow us to extend our results to dependent data. We also saw that laws of large numbers generalizes quite easily, since only the covariance structure of the data matters. Another important observation is that the asymptotic variance is generally different from the independent case, so blindly applying inference tools from the independent world will lead to incorrect inferences. Central limit theorems we could only obtain under the assumption that the data are normal in the first place. We will see that this assumption is not necessary, but requires more sophisticated assumptions on the dependence.

7.2 Law of large numbers

Let's now make our preliminary considerations more formal. To simplify the discussion a bit, we shall make the following assumption.

Definition 7.2.1. *A sequence of random variables $(X_t)_{t\in\mathbb{N}}$ is called **stationary** if*

$$(X_{t_1},\dots,X_{t_k}) \stackrel{d}{=} (X_{t_1+h},\dots,X_{t_k+h}) \quad \text{for all} \quad t_1,\dots,t_k,h\in\mathbb{N}.$$

Stationarity is a common assumption in time series analysis. It says that the distribution of the data does not change over time. In particular $\mathbb{E}[X_t]$ and $\mathbb{C}\text{ov}(X_t,X_{t+h})$ do not depend on t or the sign of h. It can be relaxed in essentially everything that follows, but we will not bother with this here. For simplicity we also set $X_t=0$ for all $t\leq 0$.

Generalizing the law of large numbers is straightforward.

Theorem 7.2.2. *Let $(X_t)_{t\in\mathbb{N}}$ be a stationary sequence of random variables such $\sum_{h=-\infty}^{\infty}|\mathbb{C}\text{ov}(X_t,X_{t+h})|<\infty$. Then, $\overline{X}_n\to_p\mathbb{E}[X_1]$.*

Proof. The proof is essentially the same as for the *iid* case. We have $\mathbb{E}[\overline{X}_n]=\mu$ and

$$\begin{aligned}\mathbb{V}\text{ar}[\overline{X}_n]=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\mathbb{C}\text{ov}(X_i,X_j)&=\frac{1}{n^2}\sum_{t=1}^n\sum_{j=1}^n\mathbb{C}\text{ov}(X_t,X_j)\\&=\frac{1}{n^2}\sum_{t=1}^n\sum_{h=1-t}^{n-t}\mathbb{C}\text{ov}(X_t,X_{t+h})\\&\leq\frac{1}{n}\sum_{h=-\infty}^{\infty}|\mathbb{C}\text{ov}(X_t,X_{t+h})|\to 0.\end{aligned}$$

Now the result follows form Markov's inequality. □

The main assumption is that the dependence between observations is not too strong. Specifically, we assume that the covariances $\mathbb{C}\text{ov}(X_t, X_{t+h})$ are *absolutely summable*. If $\mathbb{C}\text{ov}(X_t, X_{t+h}) = \sigma^2\rho(h)$, the summability condition becomes $\sigma^2 \sum_{h=0}^{\infty} \rho(h) < \infty$, which is equivalent to the existence of a second moment and some form of asymptotic uncorrelatedness. In particular, the condition is satisfied for $\rho(h) \sim h^{-\alpha}$ for some $\alpha > 1$, but violated for $\rho(h) \sim h^{-\alpha}$ with $\alpha \leq 1$. The summability condition thus requires the correlations to decay fast enough.

7.3 Mixing conditions

We have seen that a certain level of asymptotic uncorrelatedness is enough to establish a LLN, because its proof merely relies on the summability of covariances. The central limit theorem, however, requires more. The standard proof of the CLT (which I now added in Section 2.8) relies more heavily on independence (as opposed to uncorrelatedness) of observations. To establish a CLT for dependent data, we thus need to strengthen our asymptotic uncorrelatedness condition to a form of asymptotic independence. This is commonly expressed through *mixing conditions* involving *mixing coefficients* that measure the strength of dependence as deviations from independence.[1]

There are several coefficients that can be used to measure mixing. To simplify the notation a bit, we write $X_{\leq t} = (X_s)_{s \leq t}$ and $X_{\geq t} = (X_s)_{s \geq t}$ for the past and future of X_t, respectively. Here are some common mixing coefficients:

- α-mixing or *strong mixing coefficient*:

$$\alpha(h) = \sup_{A,B} |\mathbb{P}(X_{\geq t+h} \in A, X_{\leq t} \in B) - \mathbb{P}(X_{\geq t+h} \in A)\mathbb{P}(X_{\leq t} \in B)|,$$

- ϕ-mixing or *uniform mixing coefficient*:

$$\phi(h) = \sup_{A,B} |\mathbb{P}(X_{\geq t+h} \in A \mid X_{\leq t} \in B) - \mathbb{P}(X_{\geq t+h} \in A)|,$$

- β-mixing or *absolute regularity coefficient*:

$$\beta(h) = \sup_{A} \mathbb{E}_{X_{\leq t}} \left[|\mathbb{P}(X_{\geq t+h} \in A \mid X_{\leq t}) - \mathbb{P}(X_{\geq t+h} \in A)| \right],$$

The suprema are taken over all sets where the corresponding probabilities are well-defined. Each coefficient measures deviation from independence in its own way. Indeed, we have $\alpha(h) = \phi(h) = \beta(h) = 0$ whenever the future $X_{\geq t+h}$ is independent of the past $X_{\leq t}$ for all t, and they are non-zero when the past carries some information about the future. We call a process X_t $\alpha/\phi/\beta$-mixing if $\alpha(h)/\phi(h)/\beta(h) \to 0$ as $h \to \infty$.

Although not immediately obvious, we have the following relation

$$2\alpha(h) \leq \beta(h) \leq \phi(h), \tag{7.1}$$

[1] The definition of mixing coefficients and following results usually involve some measure-theoretic technicalities that I deliberately omit to keep things simple.

so that α-mixing imposes the weakest condition on dependence (i.e., it allows for stronger dependence). In fact, ϕ-mixing is too strict a requirement for many applications; for example, the autoregressive process from the preliminary considerations is not ϕ-mixing. Most common time series models are both α- and β-mixing, although proving this is often hard. It is commonly accepted to assume that one of the conditions holds. CLTs can be established under all three forms of mixing. Since β-mixing is often the most convenient, we will prove a CLT only for this condition in the following section. (This also implies a CLT for ϕ-mixing processes by (7.1).)

First we state, without proof, a useful lemma relating the mixing coefficients to the covariance structure of the data (Rio et al., 2017, eq. (1.12b) and Theorem 6.3).

Lemma 7.3.1. *Let $(X_t)_{t\in\mathbb{N}}$ be a stationary sequence of random variables with $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[|X_t|^q] \leq K$ for some $q > 2, K < \infty$.*

(i) It holds

$$\mathbb{C}\mathrm{ov}(X_t, X_{t+h}) \leq \alpha(h)^{1-2/q} K^{2/q}$$

(ii) If $\alpha(h) = O(h^{-\gamma})$ for some $\gamma > q/(q-2)$, there is $q' > 2$ such that

$$\mathbb{E}\left[\left|\sum_{t=1}^{n} X_t\right|^{q'}\right] = O(n^{q'/2}).$$

The lemma is stated in terms of strong mixing coefficients $\alpha(h)$, but the same result holds for the other two coefficients by the relation (7.1).

The first part of the lemma shows that the lag-h covariance is bounded by (a power of) the strong mixing coefficient. This is intuitive because the mixing coefficient measures strength of dependence. And if there is little dependence, there should be little correlation. How 'direct' this relation is depends on the moment condition. If q is small (close to 2), the bound is very weak, because $\alpha(h)$ is raised to a power very close to 0. If q is large (close to ∞), the bound is strong, because $\alpha(h)$ is raised to a power close to 1. If the mixing coefficients decay quickly enough, this inequality can be used to ensure summability of the covariances.

The second part of the lemma is a bit harder to interpret. It relates to the moments of a sum over the sequence. By Jensen's inequality, it also implies $\mathbb{E}[|\sum_{t=1}^{n} X_t|^2] = O(n)$, which in turn implies

$$\mathbb{V}\mathrm{ar}\left[\sqrt{n}\,\overline{X}_n\right] = O(1),$$

indicating that $\sqrt{n}$ is the right scaling in a CLT. The importance of the result comes from the fact that we can use a power q' slightly larger than 2, which will allow us to

check Lyapunov's condition (Lemma 2.6.5) in the proof of the CLT. In particular,

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{t=j}^{j+k_n} X_t\right|^{q'}\right] = O((k_n/n)^{q'/2}) = o(1),$$

whenever $k_n/n \to 0$.

7.4 Coupling

Mixing coefficients quantify the dependence between the past and the distant future. As the distance between the past and the future increases, the dependence should decay. The idea is now to compare the original process with a process where the past and (distant) future are exactly independent. The construction of such a second process is called *coupling*. We give here, without proof, a coupling result for β-mixing processes.

Lemma 7.4.1 (Berbee's maximal coupling)**.** *Let $(Y_s)_{s\in\mathbb{N}}$ be an $\mathbb{R}^d$-valued process with mixing coefficients $\beta(h)$. Then for every $t \in \mathbb{Z}, h \geq 1$, there exists a process $(Y_s^*)_{s\in\mathbb{N}}$ with the following properties:*

(i) $Y_{\geq t+h} \stackrel{d}{=} Y^*_{\geq t+h}$,

(ii) $Y^*_{\geq t+h}$ *is independent of* $Y_{\leq t}$,

(iii) $\mathbb{P}(Y_{\geq t+h} \neq Y^*_{\geq t+h}) = \beta(h)$.

Let's read this in plain language. We start with an arbitrary time series Y_s, whose dependence we measure by the β-mixing coefficient. We now fix some time point t and a gap h between what we consider past and (distant) future. The result says that we can construct a new process Y_s^* whose future values $Y^*_{\geq t+h}$ ...

(i) behave exactly like those of the original process,

(ii) are independent of the past of the original process,

(iii) differ from the original process only on an event with probability $\beta(h)$.

Now suppose the dependence in the original process is weak, i.e., $\beta(h) \to 0$ quickly as $h \to \infty$. We can now replace the future of the original process, without error and high probability, by something entirely independent of the past. That's quite remarkable! It's also a central piece in proving a CLT for β-mixing processes.

7.5 Central limit theorem for mixing variables

Proving a CLT for β-mixing processes relies on the combination of two clever ideas. Bernstein's blocking technique and coupling. The idea behind the blocking technique is illustrated in Fig. 7.1. We partition the data $X_1, \ldots, X_n$ into an alternating sequence

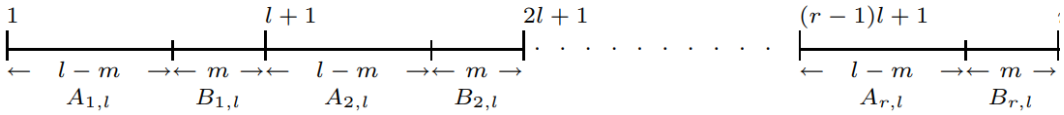

Figure 7.1: Illustration of the blocking technique: The data is divided into alternating blocks of large and small size.

of large blocks $A_{k,l}$ and small blocks $B_{k,l}$ size respectively. If we sum the data blockwise and make the large blocks large enough, the small blocks become negligible asymptotically. The key observation is the following: if the process is mixing, the large blocks are almost independent of another because there are many observations between them. Now we use coupling to replace the large blocks by independent ones. The (independent) sum of large block sums thus satisfies a (triangular array) CLT. This strategy gives us an important insight: short-term dependence is irrelevant for whether a CLT holds or not, but the long-run dependence in the data needs to fade out sufficiently fast.

Theorem 7.5.1. *Let $(X_t)_{t\in\mathbb{N}}$ be a stationary sequence of random variables. Suppose that there is a $q > 2$ and $\gamma > q/(q-2)$, such that*

$$\mathbb{E}[|X_t|^q] < \infty \quad \text{and} \quad \beta(h) = O(h^{-\gamma}).$$

Then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \sigma^2 = \sum_{h=-\infty}^{\infty} \mathbb{C}\mathrm{ov}(X_t, X_{t+h}).$$

Before preceding to a formal proof, let's discuss the key insights the theorem provides. It holds,

$$\sigma^2 = \sum_{h=-\infty}^{\infty} \mathbb{C}\mathrm{ov}(X_t, X_{t+h}) = \mathbb{V}\mathrm{ar}[X_1] \left(1 + 2\sum_{h=1}^{\infty} \mathrm{Corr}(X_t, X_{t+h})\right).$$

If all observations are uncorrelated, we indeed have $\sigma^2 = \mathbb{V}\mathrm{ar}[X_1]$ as in the independent setting. If the observations are correlated, however, the asymptotic variance is generally different. It can be both larger than under independence (if correlations are predominately positive) or smaller (if correlations are predominately negative).

The second difference lies in the more assumptions. First, we require the existence of moments of order q for some $q > 2$, while $q = 2$ was enough under independence. The larger we choose q, the more restrictive the condition becomes. The second condition constrains the decay of $\beta(h)$, which should be interpreted as a constraint on the asymptotic strength of dependence. The larger we choose q, the stronger the dependence can be. If q is close to 2, we need to have $\beta(h) = O(h^{-\gamma})$ for γ close to ∞—the dependence must fade out very quickly. The large we make q, the slower the decay can be, approaching $\beta(h) \sim h^{-1}$ in the limit $q \to \infty$. This reflects a fundamental

trade-off in dependent data: the stronger the dependence, the more moments we need to exist (= random variables are less likely to take on extreme values/have lighter tails).

Proof of Theorem 7.5.1. Without loss of generality, we assume $\mathbb{E}[X_1] = 0$.

Step 1: Bernstein's blocking strategy. We divide the data into r_n blocks of size l_n. Each of the r_n blocks we split again into one large block of size $l_n - m_n$ followed by one small block of size m_n. We choose

$$r_n \sim \ln n, \quad l_n \sim \frac{n}{\ln n} - n^{1/4}, \quad m_n \sim n^{1/4}$$

in what follows. Let A_{k,l_n} and $B_{k,l}$ be the sums over the large and small blocks, respectively:

$$A_{k,l_n} = \sum_{t=(k-1)l_n+1}^{kl_n-m_n} X_{t,l_n}, \quad B_{k,l_n} = \sum_{t=kl_n-m_n+1}^{kl_n} X_{t,j}.$$

We have

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{r_n} A_{k,l_n} + \frac{1}{n}\sum_{k=1}^{r_n} B_{k,l_n}.$$

Step 2: Negligibility of small blocks. Observe that $\mathbb{E}[B_{k,l_n}] = 0$ and, by Lemma 7.3.1 (ii) (and the following discussion), $\mathbb{V}\mathrm{ar}[B_{k,l_n}] = O(m_n)$. Thus,

$$\mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{k=1}^{r_n} B_{k,l_n}\right] = \frac{1}{n^2}\sum_{k=1}^{r_n} \mathbb{V}\mathrm{ar}[B_{k,l_n}] = O\left(\frac{r_n m_n}{n^2}\right) = O\left(\frac{n^{1/4}\ln n}{n^2}\right) = o\left(\frac{1}{n}\right).$$

This implies

$$\sqrt{n}\,\overline{X}_n = \sum_{k=1}^{r_n} \frac{A_{k,l_n}}{\sqrt{n}} + o_p(1).$$

Step 3: Coupling. The idea is now to turn the large block sums $A_{1,l_n}, \ldots, A_{r_n,l_n}$ into an independent sequence using coupling. We start with setting $A^*_{1,l_n} = A_{1,l_n}$ and then iteratively apply Berbee's lemma (Lemma 7.4.1) to construct $A^*_{2,l_n}, \ldots, A^*_{r_n,l_n}$ such that A^*_{k,l_n} is independent of $A^*_{1,l_n}, \ldots, A^*_{k-1,l_n}$. By Lemma 7.4.1 it holds

$$\mathbb{P}(A_{k,l_n} \neq A^*_{k,l_n} \text{ for some } k) = \sum_{k=1}^{r_n} \mathbb{P}(A_{k,l_n} \neq A^*_{k,l_n}) = r_n \beta(m_n) = n^{-\gamma/4}\ln n \to 0.$$

We thus have

$$\sqrt{n}\,\overline{X}_n = \sum_{k=1}^{r_n} \frac{A^*_{k,l_n}}{\sqrt{n}} + o_p(1),$$

with probability tending to 1.

Step 4: Applying the Lindeberg-Feller CLT. The coupled large block sums $A^*_{1,l_n}, \dots, A^*_{r_n,l_n}$ form a triangular array of independent observations, so we can apply the Lindeberg-Feller CLT (Theorem 2.6.4) to prove our claim. It remains to check the conditions of Theorem 2.6.4. We first check the converging covariance conditions. Because the blocks are *iid*,

$$\sum_{k=1}^{r_n} \mathbb{V}\mathrm{ar}[A^*_{k,l_n}/\sqrt{n}] = \frac{r_n}{n}\mathbb{V}\mathrm{ar}[A_{1,l_n}] = \frac{1}{l_n}\sum_{t=1}^{l_n-m_n}.$$

It holds $l_n/(l_n - m_n) \to 1$ and

$$\begin{aligned}\frac{1}{l_n - m_n}\mathbb{V}\mathrm{ar}[A^*_{k,l_n}] = \frac{1}{l_n - m_n}\mathbb{V}\mathrm{ar}[A_{1,l_n}] &= \frac{1}{l_n - m_n}\sum_{t=1}^{l_n-m_n}\sum_{s=1}^{l_n-m_n}\mathbb{C}\mathrm{ov}(X_{t,l_n}, X_{s,l_n}) \\ &= \frac{1}{l_n - m_n}\sum_{t=1}^{l_n-m_n}\sum_{h=1-t}^{l_n-m_n-t}\mathbb{C}\mathrm{ov}(X_{t,l_n}, X_{t+h,l_n}).\end{aligned}$$

We have,

$$\sum_{h=1-t}^{l_n-m_n-t}\mathbb{C}\mathrm{ov}(X_{t,l_n}, X_{t+h,l_n}) \to \sum_{h=1-t}^{\infty}\mathbb{C}\mathrm{ov}(X_{t,l_n}, X_{t+h,l_n}).$$

Since we are averaging infinitely many of these (bounded) terms, any finite number of them can be discarded without changing the result. The only terms that matter are those where we also take $t \to \infty$, so

$$\frac{1}{l_n - m_n}\sum_{t=1}^{l_n-m_n}\sum_{h=1-t}^{l_n-m_n-t}\mathbb{C}\mathrm{ov}(X_{t,l_n}, X_{t+h,l_n}) \to \sum_{h=-\infty}^{\infty}\mathbb{C}\mathrm{ov}(X_{t,l_n}, X_{t+h,l_n}) = \sigma^2.$$

We have shown,

$$\sum_{k=1}^{r_n}\mathbb{V}\mathrm{ar}[A^*_{k,l_n}/\sqrt{n}] \to \sigma^2,$$

verifying the converging covariance condition of Theorem 2.6.4

Next, we check Lyapunov's condition. By Lemma 7.3.1 (ii), there is $q' > 2$ such that

$$\mathbb{E}[|A_{k,l_n}/\sqrt{n}|^{q'}] = \frac{1}{n^{q'/2}}\mathbb{E}\left[\left|\sum_{t=(k-1)l_n+1}^{kl_n-m_n} X_{t,l_n}\right|^{q'}\right] = O\left(\frac{l_n^{q'/2}}{n^{q'/2}}\right).$$

Thus, setting $\delta = q' - 2$,

$$\sum_{k=1}^{r_n}\mathbb{E}[|A_{k,l_n}/\sqrt{n}|^{2+\delta}] = O\left(\frac{r_n l_n^{1+\delta/2}}{n^{1+\delta/2}}\right) = O\left(\frac{l_n^{\delta/2}}{n^{\delta/2}}\right) = o(1),$$

which implies Lindeberg's condition by Lemma 2.6.5. □

7.6 Extension of our core theory

For our core theory (M-estimators, functionals) to continue working under dependence, we need a uniform law of large numbers. That's still possible. For example, the following is a generalization of Lemma 3.4.2 to β-mixing data.

Lemma 7.6.1. *Let $\mathcal{F}$ by a class of functions with* envelope *F, i.e., $|f(x)| \le F(x)$ for all $f \in \mathcal{F}$. Suppose there is $q > 2$ and $\gamma > q/(q-2)$, such that*

$$\mathbb{E}[F(X_t)^q] < \infty \quad \text{and} \quad \beta(h) = O(h^{-\gamma}).$$

Then if for some $C < \infty$, $\alpha \in (0, 2)$,

$$\ln N_{[]}(\varepsilon \|F\|_{L_q(P)}, \mathcal{F}, L_q(P)) \le C\varepsilon^{-\alpha},$$

it holds

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| = O_p\left(\frac{\|F\|_{L_q(P)}}{\sqrt{n}} \right).$$

We see that the main difference is a change of norm (from L_2 to L_q) and the additional conditions on moments and mixing rate from Theorem 7.5.1.

Bibliography

Kauermann, G., Küchenhoff, H., and Heumann, C. (2021). *Statistical Foundations, Reasoning and Inference*. Springer.

Rio, E. et al. (2017). *Asymptotic theory of weakly dependent random processes*, volume 80. Springer.

van der Vaart, A. and Wellner, J. A. (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Nature.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press.