Chapter 5

Fundamentals of Estimation

5.1 Introduction

Now that we have been introduced to two notions of convergence for RV sequences, we can begin to appreciate the basic limit theorems used in statistical inference. The problem of estimation is of fundamental importance in statistical inference and learning. We will formalise the general estimation problem here. There are two basic types of estimation. In point estimation we are interested in estimating a particular point of interest that is supposed to belong to a set of points. In (confidence) set estimation, we are interested in estimating a set with a particular form that has a specified probability of "trapping" the particular point of interest from a set of points. Here, a point should be interpreted as an element of a collection of elements from some space.

5.2 Point Estimation

Point estimation is any statistical methodology that provides one with a "**single best guess**" of some specific quantity of interest. Traditionally, we denote this **quantity of interest as** θ^*. This quantity of interest, which is usually unknown, can be:

- an **integral** $\vartheta^* := \int_A h(x)\,dx \in \boldsymbol{\Theta}$. If ϑ^* is finite, then $\boldsymbol{\Theta} = \mathbb{R}$. The risk is a prime example, or
- a **parameter** θ^* which is an element of the **parameter space** $\boldsymbol{\Theta}$, denoted $\theta^* \in \boldsymbol{\Theta}$,
- a **distribution function (DF)** $F^* \in \mathbb{F} :=$ the set of all DFs
- a **density function (pdf)** $f \in \{$"not too wiggly Sobolev functions"$\}$, or

- a **regression function** $g^* \in \mathbb{G}$, where $\mathbb{G}$ is a class of regression functions in a regression experiment, or
- a **classifier** $g^* \in \mathbb{G}$.

Definition 5.1 (Data)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, assume that $X = (X_1, \ldots, X_n)$ is a sequence of $\mathbb{R}^m$ valued random variables taking values in the* **data space** $\mathbb{X}$*:*

$$X(\omega) : \Omega \to \mathbb{X} \ .$$

Note that $\mathbb{X} \subset (\mathbb{R}^m)^{\otimes n}$. The realisation of the RV X when an experiment is performed is the observation or data $x \in \mathbb{X}$. That is, when the experiment is performed once and it yields a specific $\omega \in \Omega$, the data $X(\omega) = x \in \mathbb{X}$ is the corresponding realisation of the RV X.

Definition 5.2 (Statistic)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, assume that $X : \Omega \to \mathbb{X}$ is a random variable (sequence of $\mathbb{R}^m$ valued) taking values in the data space $\mathbb{X}$, then a* **statistic** *T is any Borel (see Definition 2.42) function on the data space:*

$$T(x) : \mathbb{X} \to \mathbb{T} \ .$$

Remark 5.3. *Thus, given a statistic T, we can associate with it a RV $T(X)$ that takes values in the space $\mathbb{T}$. Sometimes we use $\mathbb{X}_n$, $T_n(X)$ and $\mathbb{T}_n$ to emphasise that X is a sequence of n random variables, i.e. $\mathbb{X}_n \subset (\mathbb{R}^m)^{\otimes n}$*

Definition 5.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let*

$$\mathcal{E} = \{F(x; \lambda) : \mathbb{X} \to [0,1] : \lambda \in \mathbf{\Lambda}, \ F \text{ is a DF}\}$$

be a statistical model of distribution functions. Let a parameter map be given $\theta : \mathbf{\Lambda} \to \mathbf{\Theta}$. Consider the sequence $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} F(\cdot; \lambda^*) \in \mathcal{E}$ be $\mathbb{R}^m$-valued RVs. A **point estimator** *of $\theta^* := \theta(\lambda^*) \in \mathbf{\Theta}$ is a statistic, i.e.*

$$\hat{\Theta} : \mathbb{X} \to \mathbf{\Theta},$$

sometimes we denote it as $\hat{\Theta}_n$ to highlight that it depends on n values.

The bias *of an estimator $\widehat{\Theta}_n$ of $\theta^* \in \mathbf{\Theta}$ is:*

$$\mathsf{bias}(\widehat{\Theta}_n(X)) := \mathbb{E}(\widehat{\Theta}_n(X)) - \theta^* = \int \widehat{\Theta}_n(x)\, dF(x; \lambda^*) - \theta(\lambda^*) \ . \tag{5.1}$$

Some comments are in order to connect these concepts to the risk minimization problems in supervised learning, as see in Chapter 4. Let us be

given a statistical model $\mathcal{E}$ of distribution functions $F_{X,Y}$, and let us consider the regression example, i.e. we wish to estimate

$$r(x) = \int y dF_{Y|X}(y \mid x)$$

this means that for each fix x the parameter map is $\theta_x(F_{X,Y}) = r(x)$, and if given $X_1, \ldots, X_n$ we come up with a proposal function from $\mathcal{M}$, i.e. $g_n(X_1, \ldots, X_n) \in \mathcal{M}$ then $g_n(X_1, \ldots, X_n; x)$ becomes a point estimator of $r(x)$. Sometimes however, in regression experiments you assume that the model space and the statistical model are the same and parametric (finite dimensional). In this case the parameter becomes easy to define and much of this simplifies.

As we shall see later, we are often not so concerned with the statistical properties of the specific estimator of the regression function but we are interested in some functional of it. Usually we are interested in the Risk, and in this case it is as simple as an expectation, see Example 5.6.

5.2.1 Some Properties of Point Estimators

Given that an estimator is merely a function from the data space to the parameter space, we need a way to define what a good estimator is. Recall that a point estimator $\widehat{\Theta}_n$, being a statistic, has a corresponding RV $\widehat{\Theta}_n(X)$ which has a probability distribution over its range $\boldsymbol{\Theta}$. This distribution over $\boldsymbol{\Theta}$ is called the **sampling distribution** of $\widehat{\Theta}_n(X)$.

Definition 5.5 (Bias of a Point Estimator)**.** *We say that the estimator* $\widehat{\Theta}_n$ *is* **unbiased** *if*

$$\mathsf{bias}(\widehat{\Theta}_n(X)) = 0,$$

for every n*. If*

$$\lim_{n\to\infty} \mathsf{bias}_n(\widehat{\Theta}_n) = 0,$$

we say that the estimator is **asymptotically unbiased**.

Since the expectation of the sampling distribution of the point estimator $\widehat{\Theta}_n$ depends on the unknown λ^*, we emphasise the λ^*-dependence by $\mathbb{E}_{\lambda^*}(\widehat{\Theta}_n(X))$.

Example 5.6. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let*

$$\mathcal{E} = \{F(x;\lambda) : \mathbb{X} \to [0,1] : \lambda \in \boldsymbol{\Lambda},\ F \text{ is a DF}\}$$

Let the parameter map $\theta(\lambda) := \int x dF(\lambda)$ *be the expectation. Let* $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F(\cdot, \lambda^*)$*, that is, with some unknown parameter* λ^* *and hence some unknown mean* $\theta^* = \theta(\lambda^*)$*. Consider the* **sample mean** *estimator*

$$\widehat{\Theta}_n(X) := \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then

$$\mathsf{bias}(\widehat{\Theta}_n(X)) = \mathbb{E}_{\lambda^*}\left[\overline{X}_n\right] - \theta(\lambda_*) = 0,$$

hence the sample mean estimator is unbiased in our statistical model $\mathcal{E}$ *with respect to the parameter map* θ*.*

Remark 5.7. *In the above example our statistical model is parametrized by* Λ *which is infinite dimensional, we would thus say that this is a nonparametric model. Another example of estimation would be that we assume that the statistical model is that of normal distributions with mean* μ *and variance* σ^2*. In this case we could take* $\mathbf{\Lambda}$ *to be two dimensional and the parameter map be the identity map from* $\mathbb{R}^2 \to \mathbb{R}^2$*.*

Remark 5.8. *The bias of an estimator is a term that we use to theoretically study the properties of the estimator. In a real setting,* θ^* *is unknown so we could never compute the bias, but perhaps we can get bounds for it. In certain cases we can prove that an estimator is asymptotically unbiased without knowing* θ^**.*

For instance if $X_1, \ldots, X_n$ *as in our example above, if we furthermore assumed in our statistical model that* $X_i \in L^2(\mathbb{P})$ *then the law of large numbers Theorem 3.26 implies the asymptotic unbiasedness.*

Definition 5.9 (Standard Error of a Point Estimator)**.** *The standard deviation of the point estimator* $\widehat{\Theta}_n(X)$ *of* $\theta^* \in \mathbf{\Theta}$ *is called the* **standard error***:*

$$\mathsf{se}(\widehat{\Theta}_n(X)) := \sqrt{\mathbb{V}_{\lambda^*}(\widehat{\Theta}_n)} := \sqrt{\int \left(\widehat{\Theta}_n(x) - \mathbb{E}_{\lambda^*}(\widehat{\Theta}_n)\right)^2 dF(x;\lambda^*)}\ . \quad (5.2)$$

Since the variance of the sampling distribution of the point estimator $\widehat{\Theta}_n$ depends on the fixed and possibly unknown λ^*, as emphasised by $\mathbb{V}_{\lambda^*}$ in (5.2), the $\mathsf{se}(\widehat{\Theta}_n(X))$ is also a possibly unknown quantity and may itself be estimated from the data.

Example 5.10 (Standard Error of our Estimator of θ^*)**.** *Consider the sample mean estimator* $\widehat{\Theta}_n := \overline{X}_n$ *of* θ^**, from* $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$*. Observe that the statistic:*

$$T_n((X_1, X_2, \ldots, X_n)) := n\,\widehat{\Theta}_n((X_1, X_2, \ldots, X_n)) = \sum_{i=1}^{n} X_i$$

is the Binomial(n, θ^*) *RV. The standard error* se_n *of this estimator is:*

$$\begin{aligned}\mathsf{se}(\widehat{\Theta}_n) &= \sqrt{\mathbb{V}_{\lambda^*}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right)} = \sqrt{\left(\sum_{i=1}^{n} \frac{1}{n^2}\,\mathbb{V}_{\lambda^*}(X_i)\right)} \\ &= \sqrt{\frac{n}{n^2}\,\mathbb{V}_{\lambda^*}(X_i)} = \sqrt{\theta^*(1-\theta^*)/n}\ .\end{aligned}$$

Another reasonable property of an estimator is that it converge to the "true" parameter θ^* – here "true" means the supposedly fixed and possibly unknown θ^*, as we gather more and more IID data from a θ^*-specified DF $F(x; \theta^*)$. This property is stated precisely next.

Definition 5.11 (Asymptotic Consistency of a Point Estimator)**.** *A point estimator* $\widehat{\Theta}_n$ *of* $\theta^* \in \Theta$ *is said to be* **asymptotically consistent** *if:*

$$\widehat{\Theta}_n \xrightarrow{\mathbb{P}} \theta^*.$$

Definition 5.12 (Mean Squared Error (MSE) of a Point Estimator)**.** *Often, the quality of a point estimator* $\widehat{\Theta}_n$ *of* $\theta^* \in \Theta$ *is assessed by the* **mean squared error** *or* MSE *defined by:*

$$\mathsf{MSE}_n(\widehat{\Theta}_n(X)) := \mathbb{E}_{\lambda^*}\left((\widehat{\Theta}_n(X) - \theta^*)^2\right)\ .$$

The following proposition shows a simple relationship between the mean square error, bias and variance of an estimator $\widehat{\Theta}_n$ of θ^*.

Proposition 5.13 (The $\sqrt{\mathsf{MSE}_n} : \mathsf{se}_n : \mathsf{bias}_n$–Sided Right Triangle of an Estimator)**.** *Let* $\widehat{\Theta}_n$ *be an estimator of* $\theta^* \in \Theta$*. Then:*

$$\mathsf{MSE}_n(\widehat{\Theta}_n) = (\mathsf{se}_n(\widehat{\Theta}_n))^2 + (\mathsf{bias}_n(\widehat{\Theta}_n))^2\ . \tag{5.3}$$

Proof. Consider the mean squared error

$$\begin{aligned}\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \theta^*)^2\right] &= \mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \theta^* - \mathbb{E}\left[\widehat{\Theta}_n(X)\right] + \mathbb{E}\left[\widehat{\Theta}_n(X)\right])^2\right] \\ &= \mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right])^2\right] \\ &\quad + \mathbb{E}_{\lambda^*}\left[(\theta^* - \mathbb{E}\left[\widehat{\Theta}_n(X)\right])^2\right] \\ &\quad + 2\,\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right])(\mathbb{E}\left[\widehat{\Theta}_n(X)\right] - \theta^*)\right]\end{aligned}$$

the first expectation on the RHS is just the standard error, the second is the bias and the last expectation is 0 because

$$\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right](\mathbb{E}\left[\widehat{\Theta}_n(X)\right] - \theta^*)\right] \\ = \mathsf{bias}(\widehat{\Theta}_n(X))\,\mathbb{E}\left[\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right]\right] = 0.$$

□

Proposition 5.14 (Asymptotic consistency of a point estimator)**.** *Let $\widehat{\Theta}_n$ be an estimator of $\theta^* \in \Theta$. Then, if $\mathsf{bias}_n(\widehat{\Theta}_n) \to 0$ and $\mathsf{se}_n(\widehat{\Theta}_n) \to 0$ as $n \to \infty$, the estimator $\widehat{\Theta}_n$ is asymptotically consistent:*

$$\widehat{\Theta}_n \xrightarrow{\mathbb{P}} \theta^* \ .$$

Proof. If $\mathsf{bias}(\widehat{\Theta}_n) \to 0$ and $\mathsf{se}(\widehat{\Theta}_n) \to 0$, then by (5.3), $\mathsf{MSE}(\widehat{\Theta}_n) \to 0$, i.e.

$$\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n - \theta^*)^2\right] \to 0.$$

That is, $\widehat{\Theta}_n(X) \to \theta^*$ in $L^2(\mathbb{P})$ which implies convergence in probablity, see Section 3.2.1. □

We want our estimator to be unbiased with small standard errors as the sample size n gets large.

Example 5.15 (Asymptotic consistency of our Estimator of θ^*)**.** *Consider the sample mean estimator $\widehat{\Theta}_n(X) := \overline{X}_n$ of θ^*, from $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim}$ Bernoulli(θ^*). Since $\mathsf{bias}_n(\widehat{\Theta}_n) = 0$ for any n and $\mathsf{se}_n = \sqrt{\theta^*(1-\theta^*)/n} \to 0$, as $n \to \infty$, by Proposition 5.14, $\widehat{\Theta}_n \xrightarrow{\mathbb{P}} \theta^*$. That is $\widehat{\Theta}_n$ is an* **asymptotically consistent estimator** *of θ^*.*

We saw in Section 3.1 that the concentration inequalities gives us quite some control, let us see an application to the mean square error of the sample mean estimator.

Lemma 5.16. *Let Y be a RV satisfying the estimate for fixed $c_0 \geq 1$ and for all $\epsilon > 0$*

$$\mathbb{P}(|Y| \geq \epsilon) < 2e^{-c_0\epsilon^2}. \tag{5.4}$$

Then

$$\mathbb{E}\left[|Y|^2\right] \leq \frac{5}{c_0}$$

Before we proceed with the proof let us take an example

Example 5.17. *Let us revisit the problem of estimating the mean of an L^2 RV. Let $X_1, \ldots, X_n$ be i.i.d. RVs in $L^2(\mathbb{P})$ that are also sub-Gaussian with parameter σ, then using Theorem 3.6 and Lemma 5.16 we get*

$$\mathsf{MSE}(\overline{X}_n) = \mathbb{E}\left[|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]|^2\right] \leq \frac{10\sigma^2}{n}$$

So for sub-Gaussian RVs we have almost the same standard error as if the random variables where Gaussian. Optimizing the proof of Lemma 5.16 we can eek out a smaller constant.

Proof. Let $\delta > 0$, we will choose it later

$$\mathbb{E}\left[|Y|^2\right] \leq \mathbb{E}\left[\sum_{k=2}^{\infty} \delta^2 k^2 \mathbb{1}_{\delta(k-1)\leq Y<\delta k}\right] + \mathbb{E}\left[Y^2 \mathbb{1}_{Y\leq\delta}\right] = I + II$$

We first estimate II by noting that

$$\mathbb{E}\left[Y^2 \mathbb{1}_{Y\leq\delta}\right] \leq \delta^2$$

To estimate I note that according to (5.4) we get

$$\begin{aligned} I = \mathbb{E}\left[\sum_{k=2}^{\infty} \delta^2 k^2 \mathbb{1}_{\delta(k-1)\leq Y<\delta k}\right] &\leq \mathbb{E}\left[\sum_{k=2}^{\infty} \delta^2 k^2 \mathbb{1}_{\delta(k-1)\leq Y}\right] \\ &\leq 2\sum_{k=2}^{\infty} \delta^2 k^2 e^{-c_0\delta^2(k-1)^2} \end{aligned}$$

Now choose $\delta^2 = 1/c_0$, from this we get

$$\sum_{k=2}^{\infty} \delta^2 k^2 e^{-c_0\delta^2(k-1)^2} \leq \frac{1}{c_0}\sum_{k=2}^{\infty} k^2 e^{-(k-1)^2} \leq \frac{2}{c_0}.$$

Putting it all together we get

$$\mathbb{E}\left[|Y|^2\right] \leq \frac{1}{c_0} + \frac{4}{c_0} \leq \frac{5}{c_0}.$$

□

Exercise 5.18. *If you use the equality*

$$\mathbb{E}\left[X\right] = \int_0^{\infty} \mathbb{P}(X > t)dt$$

(valid for non-negative RV's) can you improve upon the constant in Lemma 5.16?

Exercise 5.19. *Advanced!! What happens in Lemma 5.16 if we replace sub-Gaussian with sub-exponential?*

5.3 Non-parametric DF Estimation

So far, we have been interested in some estimation problems where the parameter map has a finite dimensional range. For instance, in the mean estimation problem of $\mathbb{R}$ valued RVs, the space Θ is of dimension 1. Similarly, if we are estimating the mean and variance the space Θ is of dimension 2.

Next we consider a non-parametric experiment in which n IID samples are drawn according to some fixed and possibly unknown DF F^* from the space of **All Distribution Functions**: That is, our statistical model is

$$\mathcal{E} := \{\text{All DFs}\} := \{\, F(x; F) : F \text{ is a DF} \,\}$$

and we assume that there is an $F^* \in \mathcal{E}$, which is the DF for an i.i.d. sequence $X_1, \ldots, X_n$. Here the parameter space is $\mathcal{E}$ itself, which is infinite dimensional and the parameter map θ is the identity map, so $\Theta = \mathcal{M}$.

Consider now a Model space which is

$$\mathcal{M}_0 := \{F(x) = \int_{-\infty}^{x} p(x; p)dx, \quad p \text{ is a PDF}\}$$

that is the space of all DFs from all continuous random variables. Let $F^* \in \mathcal{M}_0$ be a DF, let $X = (X_1, \ldots, X_n)$ be i.i.d. from DF F^*, and for simplicity of presentation assume that X is continuous and $f^* = (F^*)'$. For any distribution function $G \in \mathcal{M}_0$ with density g, we define the relative entropy loss functional as

$$L(x, G) = \ln\left(\frac{f^*(x)}{g(x)}\right)$$

and the relative entropy risk becomes

$$R(G) = \int \ln\left(\frac{f^*(x)}{g(x)}\right) f^*(x)dx.$$

The relative entropy risk is just the relative entropy between F and G, it can also be identified with the Kullbach-Leibler divergence between F and G. We would like to minimize $R(G)$ over all distribution functions G, we know from Section 4.2 that F is the minimizer. However we only have access to the empirical risk, namely

$$\widehat{R}_n(p; X) = \frac{1}{n}\sum_i \ln\left(\frac{f^*(X_i)}{g(X_i)}\right)$$

Exercise 5.20. *Show that the relative entropy risk is the same risk as we saw in Section 4.2, it only differs by a constant.*

If the law of large numbers is applicable we know that

$$\lim_{n\to\infty} \frac{1}{n} \sum_i \ln\left(\frac{f^*(X_i)}{g(X_i)}\right) = R(G).$$

The infimum of the empirical risk over $\mathcal{M}_0$ is given by a CDF $F_n \notin \mathcal{M}_0$ of a discrete RV for which PMF that puts weight $1/n$ on each X_i, i.e. the PMF

$$p_n(x; X) = \frac{1}{n} \sum_i \mathbb{1}_{x=X_i}.$$

Exercise 5.21. *Prove that the minimizer is necessarily a discrete measure.*

Exercise 5.22. *Prove that among all the discrete distributions with support on the X_i the uniform one is minimizing the risk.*

> We know that in good cases the LLN implies that the empirical risk for a fixed DF converges to the risk.
>
> Question: What happens to the minimum of the empirical risk, does it converge to the minimum of the risk as $n \to \infty$?

Definition 5.23. *Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sequence of RVs with DF F. We denote*

$$\widehat{F}_n(x; X) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le x}$$

we call $\widehat{F}_n(x; X)$ the empirical distribution function. Often we will supress the X, and just write $\widehat{F}_n(x)$, but we must never forget X.

For each fixed $x \in \mathbb{R}$, $\widehat{F}_n(x; X)$ is a statistic and is thus a RV. We can say that $\widehat{F}_n(x; X) : \mathbb{R} \times \mathbb{X} \to [0, 1]$ is a random function. It has the following properties:

Lemma 5.24. *Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sequence of RVs with DF F. Let $\widehat{F}_n(x; X)$ be the empirical distribution function, then*

1.

$$\mathbb{E}\left[\widehat{F}_n(x; X)\right] = F(x)$$

2.

$$\mathbb{V}\left[\widehat{F}_n(x; X)\right] = \frac{F(x)(1 - F(x))}{n}$$

Proof. To compute the expectation, note that

$$\mathbb{E}\left[\widehat{F}_n(x;X)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\leq x}\right] = F(x).$$

To compute the variance we simply compute

$$\begin{aligned}\mathbb{E}\left[\widehat{F}_n(x;X)^2\right] &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{1}_{X_i\leq x}\mathbb{1}_{X_j\leq x}\right] \\ &= \frac{1}{n^2}\sum_{i\neq j}F(x)^2 + \frac{1}{n^2}\sum_{i=j}F(x) = F(x)^2 - \frac{n}{n^2}F(x)^2 + \frac{n}{n^2}F(x)\end{aligned}$$

So,

$$\mathbb{V}\left[\widehat{F}_n(x;X)^2\right] = \frac{1}{n}F(x)(1-F(x)).$$

□

Remark 5.25. *We see from the above that the empirical distribution function is unbiased and asymptotically consistent. Note: We dont even need to know anything about the integrability of X as the empirical distribution function is always a bounded RV, i.e. takes values between* $[0,1]$.

If we use Hoeffdings inequality we get the following concentration:

Lemma 5.26. *Let $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$. Then, for any $\epsilon > 0$ and x;*

$$\mathbb{P}(|\widehat{F}_n(x;X) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

This is perhaps weaker than we would like, actually we can prove that the same estimate holds but over all x at the same time. The next proposition is often referred to as the **fundamental theorem of statistics** and is at the heart of non-parametric inference, empirical processes, and computationally intensive bootstrap techniques.

Theorem 5.27 (The Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)**.** *Let $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$. Then, for any $\epsilon > 0$:*

$$P\left(\sup_x |\widehat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}. \tag{5.5}$$

You have now seen the first example of the empirical risk minimization framework. This lies at the heart of machine learning, we will see this later in the pattern recognition problem, where, for certain not overly complex model spaces $\mathcal{M}_0$ we can guarantee estimates like the DKW above. This is coined the uniform convergence of empirical means (ECEMP).

5.4 Plug-in Estimators of Statistical Functionals: Direct estimation

A **statistical functional** is simply any function of the DF F. For example, the median $T(F) = F^{[-1]}(1/2)$ is a statistical functional. Thus, $T(F) : \{\text{All DFs }\} \to \mathbb{T}$, being a map or function from the space of DFs to its range $\mathbb{T}$, is a functional. The idea behind the plug-in estimator for a statistical functional is simple: just plug-in the point estimate $\widehat{F}_n$ instead of the unknown DF F^* to estimate the statistical functional of interest.

Definition 5.28 (Plug-in estimator)**.** *Suppose, $X_1, \ldots, X_n \overset{IID}{\sim} F^*$. The plug-in estimator of a statistical functional of interest, namely, $T(F^*)$, is defined by:*

$$\widehat{T}_n := \widehat{T}_n(X_1, \ldots, X_n) = T(\widehat{F}_n) \ .$$

Definition 5.29 (Linear functional)**.** *If $T(F) = \int r(x)dF(x)$ for some function $r(x) : \mathbb{X} \to \mathbb{R}$, then T is called a* **linear functional**. *Thus, T is linear in its arguments:*

$$T(aF + a'F') = aT(F) + a'T(F') \ .$$

Proposition 5.30 (Plug-in Estimator of a linear functional)**.** *The plug-in estimator for a linear functional $T = \int r(x)dF(x)$ is:*

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

Furthermore, if $r(X) \in L^2(\mathbb{P})$ then the estimator $T(\widehat{F}_n)$ is unbiased and asymptotically consistent.

Proof. That $T(\widehat{F}_n)$ is asymptotically consistent follows from Theorem 3.26. □

Remark 5.31. *This means that any plug in estimator of a linear functional is actually the sum of independent RVs. If $r(X)$ is nice enough, say sub-Gaussian or sub-exponential, then we can utilize the concentration inequalities in Section 3.1.*

Remark 5.32. *However, there are non-linear functionals that one is often interested in. For instance the median $T(F) = F^{-1}(\frac{1}{2})$.*

Definition 5.33. *The influence function is defined as*

$$L_F(y) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon \mathbb{1}_{x \geq y}) - T(F)}{\epsilon}$$

If T is "nice enough" we can basically use the above definition of derivative to cook up a first order approximation which would look like

$$T(\widehat{F}_n) - T(F) \approx \int L_F(y) d\widehat{F}_n = \frac{1}{n}\sum L_F(X_i).$$

The point here being that the linear term is of leading order for large n, i.e. the quadratic term is quadratic in $1/n$.

If $L_F(X_i) \in L^2(\mathbb{P})$ then according to the central limit theorem Theorem 3.28 we have that $\frac{1}{\sqrt{n}}\sum L_F(X_i)$ is asymptotically normal. Keep this in mind when we later go in to **the bootstrap**.

Lemma 5.34. *Let F be a DF and $a \in (0,1)$ a quantile, then if F is differentiable at $m = F^{[-1]}(x)$ with positive derivative (actually F is invertible at this point). Then the influence function for the quantile a is*

$$L_F(y) = \frac{a - \mathbb{1}_{m \geq y}}{\frac{dF}{dx}(m)}$$

where $m = F^{[-1]}(a)$. We thus see that L_F is bounded if the density at the median is non-zero, and as such $L_F(X)$ is sub-Gaussian and we get good concentration for the first order term.

Remark 5.35. *Note that even though L_F is bounded, we cannot know this bound as it depends on the density at the quantile. It is fairly easy to construct a density which is zero at the quantile a. Think of the median in a symmetric bimodal distribution for which the density is 0 at the middle between the two modes.*

This is actually more problematic than it seems! If we define the median as

$$F^{[-1]}(1/2)$$

then in some cases this is a set and F is not invertible at $1/2$. Thus demanding consistency does not really make much sense. For this, there is a notion of weak consistency

Proof. Let y be given and consider

$$q = T((1-\epsilon)F + \epsilon \mathbb{1}_{x \geq y})$$

Let $m = F^{[-1]}(a)$ then if $y > m$ we get

$$(1-\epsilon)F(q) = a$$
$$q = F^{[-1]}(\frac{a}{1-t}).$$

In the case that $y \leq m$ we get

$$(1-\epsilon)F(q) + \epsilon = a$$
$$q = F^{[-1]}(\frac{a-t}{1-t})$$

We now know that

$$\left.\frac{dq}{dt}\right|_{t=0} = \frac{a - \mathbb{1}_{m \geq y}}{\frac{dF}{dx}(m)}$$

□

Let us dig deeper into the quantiles, let us define the quantile function

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

then F^{-1} is a left-continuous function with range equal to the support of F and is hence unbounded. Note the subtle difference between the set-valued formal inverse $F^{[-1]}$ and the quantile function F^{-1}. Let us record some interesting properties of F^{-1}.

Lemma 5.36. *For every $0 < p < 1$ and $x \in \mathbb{R}$,*

1. *$F^{-1}(p) \leq x$ if and only if $p \leq F(x)$*
2. *$F(F^{-1}(p)) \geq p$ with equality iff p is in the range of F.*
3. *$F^{-1}(F(x)) \leq x$, where equality fails iff x is in the interior or at the right end of a flat piece of F.*
4. *$F^{-1}(F(F^{-1})) = F^{-1}$, and $F(F^{-1}(F)) = F$.*

Exercise 5.37. *Prove this lemma!*

Recall that a uniform distribution function on the interval $[0,1]$ is given by $F_{unif}(x) = \min(\max(x,0),1)$. Let $U \sim F_{unif}$ then (1) above implies that

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = \min(\max(F(x),0),1) = F(x)$$

in other words $F^{-1}(U) \sim F$. Let us collect that as a theorem

Theorem 5.38 (Inversion sampling)**.** *If $U \sim$ Uniform($[0,1]$) and F is a DF, then $F^{-1}(U) \sim F$.*

Some specific examples of statistical functionals we have already seen include:

1. The **mean** of RV $X \sim F$ is a function of the DF F:
$$T(F) = \mathbb{E}(X) = \int x \, dF(x) \ .$$

2. The **variance** of RV $X \sim F$ is a function of the DF F:
$$T(F) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int (x - \mathbb{E}(X))^2 \, dF(x) \ .$$

3. The **value of DF at a given** $x \in \mathbb{R}$ of RV $X \sim F$ is also a function of DF F:
$$T(F) = F(x) \ .$$

4. The $q^{\textbf{th}}$ **quantile** of RV $X \sim F$:
$$T(F) = F^{[-1]}(q) \ \text{ where } q \in [0, 1] \ .$$

5. The **first quartile** or the $0.25^{\textbf{th}}$ **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.25) \ .$$

6. The **median** or the **second quartile** or the $0.50^{\textbf{th}}$ **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.50) \ .$$

7. The **third quartile** or the $0.75^{\textbf{th}}$ **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.75) \ .$$