Statistics for Arbitrary Distributions

Loren K. Fryxell[1] and Charlotte Siegmann[2]

[1]Department of Economics and Global Priorities Institute, University of Oxford
[2]Department of Economics, MIT

Preliminary Draft
November 30, 2023
(Latest Version Here)

Abstract

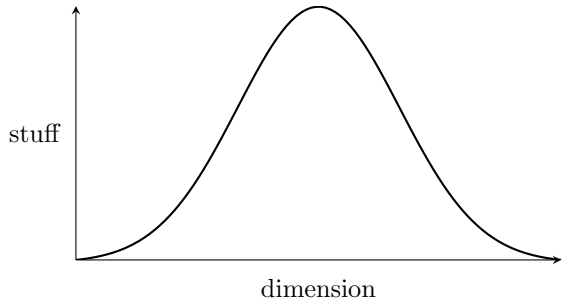
We introduce the concept of an arbitrary distribution and show how to apply descriptive statistics to them. Arbitrary distributions extend the domain of distributions on which statistics can be usefully applied beyond the usual frequency and probability distributions. For example, we can consider the distribution of the benefits of a policy across income levels and over time, and we can compute the center of mass and the spread of such a distribution. The key challenge is that such benefits, or more generally the weights within an arbitrary distribution, can be negative. We propose a method which we call *ironing* as a natural solution to the problem of statistics for arbitrary distributions.

1 Introduction

Many social programs have benefits that are distributed unevenly across income levels and across time. For these social programs, it can be useful to compute the average income level and the average time at which the benefits accrue, along with the standard deviation of the benefits across income levels and across time. This allows researchers and policymakers to communicate the central location of the benefits from the social program, as well as its spread. Unfortunately, many social programs do not benefit everyone across all income levels or at all times, but rather provide benefits to some and impose costs on others. Because such distributions can take negative values, standard summary statistics are not appropriate. We propose a method which we call *ironing* as a natural solution to the problem of statistics for arbitrary distributions, allowing us to apply standard descriptive statistics to a far more general class of distributions than before.

A *distribution* is an allocation of stuff along some dimension. For example, we may have a distribution of probability (stuff) over time (dimension), we may have a

distribution of the benefits of a policy over time, or we may have a distribution of the benefits of a policy across income levels.



When doing statistics, the stuff is interpreted as providing a weight for each point along the dimension. For example, in a distribution of probability over time, each point in time is weighted by the *probability* assigned to that point. The median time by probability is the time at which half of the probability occurs at or below that time and is a measure of the central tendency of probability across time. The average time by probability is a weighted sum across time weighted by probability and is a measure of the center of mass of probability across time. The standard deviation of time by probability is the square root of a weighted sum across time of the squared deviation from the average time weighted by probability and is a measure of the spread of probability across time.

In the very same way, we may compute the median, average, and standard deviation of time by the *benefits of a policy*. The median time by benefits is the time at which half of the benefits occurs at or below that time and is a measure of the central tendency of benefits across time. The average time by benefits is a weighted sum across time weighted by benefits and is a measure of the center of mass of benefits across time. The standard deviation of time by benefits is a measure of the spread of benefits across time.

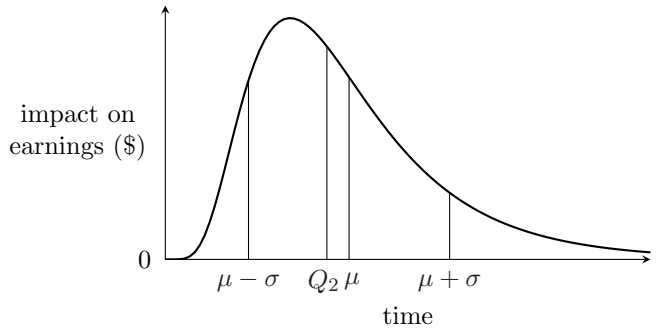


Figure 1: Consider measuring the impact on earnings of an educational reform on children over time. At each moment in time (represented continuously for simplicity), we plot the additional earnings that children who received the educational reform earned over children who did not. μ is the mean, Q_2 is the median (second quartile), and σ is the standard deviation of the distribution.

Similarly, we may take that same policy and compute the median, average, and standard deviation of *income* by the benefits of the policy. The median income level by benefits is the income level at which half of the benefits occurs at or below that income level and is a measure of the central tendency of benefits across income. The average income level by benefits is a measure of the center of mass of benefits across income. The standard deviation of income level by benefits is a measure of the spread of benefits across income.

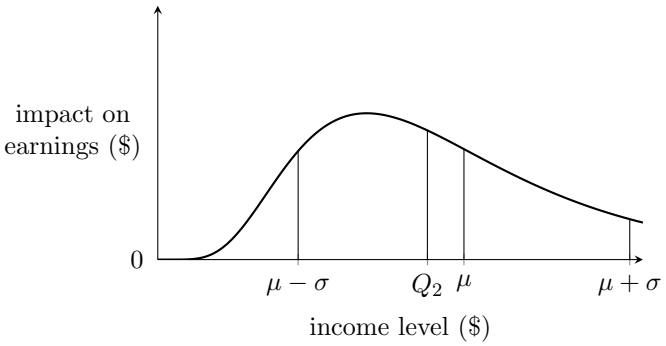
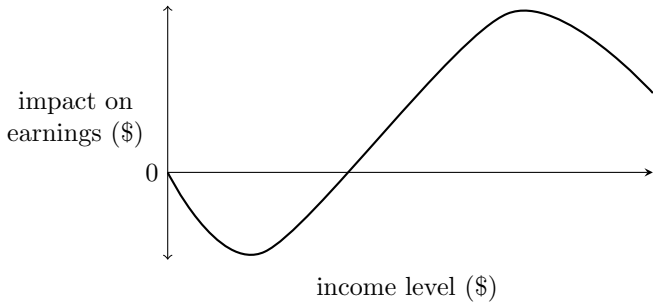


Figure 2: Consider measuring the impact on earnings of an educational reform on children across the income spectrum. At each income level (represented continuously for simplicity), we plot the additional earnings that children at each income level who received the educational reform have over children who did not (per capita).

However, notice that the benefits of a policy at a given time or on those at a given income level can, in fact, be negative.



A central question, and the primary question of this paper, is:

> *How can we apply standard statistics to arbitrary distributions, where weights can meaningfully be negative?*

As we will see, most standard statistics—including the mean and standard deviation—do not behave in intuitive or desirable ways when weights can be negative.[1] We propose a solution, which we call *ironing*,[2] as a method to apply standard statistics to arbitrary distributions. Ironing transforms an arbitrary distribution (which may have negative regions) into a classical distribution (which is everywhere non-negative), upon which we may apply standard statistics.

The key takeaways from this paper are threefold.

1. Arbitrary distributions, beyond probability and frequency distributions, are useful objects.
2. We can usefully apply standard statistics, such as mean and standard deviation, to arbitrary distributions when they are everywhere non-negative. However, applying standard statistics to arbitrary distributions in general fails what we call the *binning principle*—that small changes in the precision level of the dimension of interest should not result in disproportionately large changes in the statistic.
3. We propose a method we call *ironing* as a natural solution to the problem of statistics for arbitrary distributions. Ironing bins observations together in the least obtrusive[3] way such that the resulting distribution is everywhere non-negative. From here, we may apply statistics in the standard way. We show that for all well-behaved distributions, such a solution exists and is unique.

Takeaway 1 was already discussed above. Takeaways 2 and 3 are discussed below, in turn.

[1] See Section 1.1 for a discussion.

[2] The term is inspired by "Myerson ironing", a procedure for transforming a non-monotonic allocation rule into a monotonic (and, in this context, optimal) one. See Myerson (1981).

[3] This is defined formally in Sections 1.2 and 2.

1.1 The Binning Principle

The motivation for ironing rests upon a basic principle, which we call the *binning principle*. Binning is a method of pooling, or smoothing, nearby data points. For example, an individual's date of birth is not usually measured or reported to the hour, but rather to the day. Hence, birthdays are usually binned to the day, spreading the weight of observations uniformly throughout the hours of that day. Birthdays may also be binned to the month or the year depending on context, again spreading the weight uniformly throughout the month or year.

Define binning as *smoothing data points across an interval*. The binning principle states that small changes in bin size should not result in disproportionately large changes in the statistic.

> **Binning Principle.** Small changes in bin size should not result in disproportionately large changes in the statistic (relative to the bin size).[4]

A reflection on binning for arbitrary distributions leads to an important insight. For most weights of interest, *there is nothing qualitatively different between positive and negative weights.* Positive values are simply positive *on net*. Negative values are simply negative *on net*.

Suppose we are interested in the effect of an educational reform on students' earnings over time. Consider plotting the total earnings across time of a student who received some treatment minus their total earnings had they (counterfactually) not received the treatment. Or consider plotting its empirical analogue—the total earnings of a treatment group of students minus the total earnings of a control group over time. In either case, measuring the total earnings gap between the treated and the untreated by day will likely result in significant swings between positive and negative earnings gaps (e.g., due to different paydays), while binning the data by month or year will be more uniform.

[4] A small change to a small bin should lead to no more than a small change in the statistic. On the other hand, a small change to a large bin may lead to a larger change in the statistic, since additional observations are being smoothed across a larger interval.

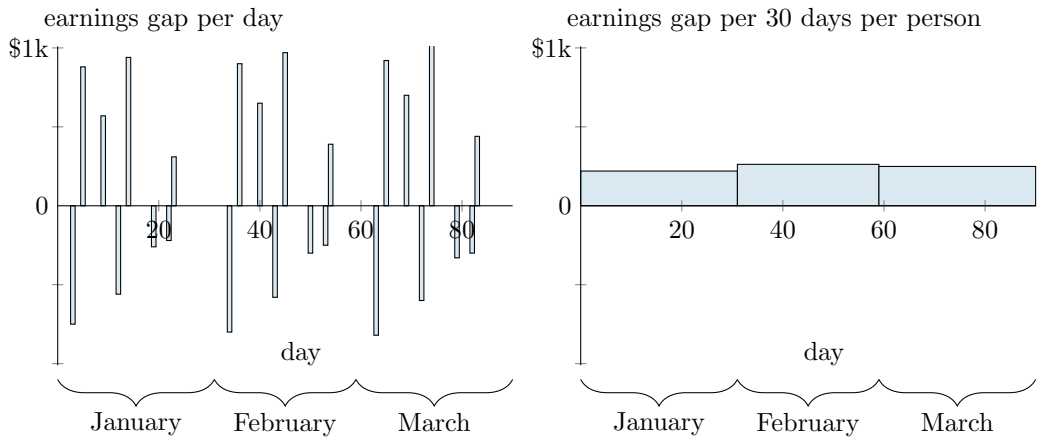


Figure 3: The first chart bins the earnings gap between students with and without a treatment by day, smoothing the earnings gap at each hour in a given day uniformly across the day. The second chart bins the earnings gap by month, smoothing the earnings gap at each day in a given month uniformly across the month.

In particular, \$100 of increased earnings on a particular day d does not mean there were no losses of earnings on that day—it simply means that, on net, there was \$100 of increased earnings in total. For example, one individual might gain \$150 and another lose \$50 on that day. But now suppose we zoom in and measure earnings at the hour it happened. Then at hour h there might be a gain of \$150 and at hour $h + k$ there might be a loss of \$50. The binning principle says that measuring an earnings gap by day or by hour should not significantly change the resulting statistics.

Most standard statistics violate the binning principle when applied to distributions with negative regions. For example, consider the mean. Suppose a policy delays the time at which you are paid by w (say, one minute) but increases your earnings by h (say, \$100). Your previous earnings were x (say, \$50k). Call this benefits distribution f. Suppose we want to compute the mean of f—the time at which the benefits of this policy are centered.

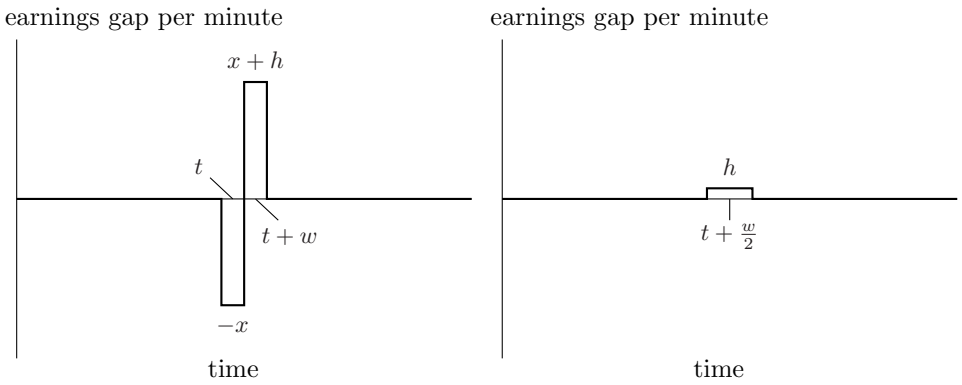


Figure 4: The first chart plots f, binning by minute. The second chart plots $\hat{f}$, binning the two minutes around t and $t+w$ together.

If we create a bin just large enough to smooth the benefits at t and $t+w$, the average time of the benefits of the policy

$$\mathbb{E}(\hat{f}) = t + \frac{w}{2},$$

as one would expect. All the benefits (positive and negative), which net to h units in total, are happening around t and $t+w$, so we would expect the average time of benefits to be around t and $t+w$ as well.

If we do not bin these two minutes together, the average time of the benefits of the policy

$$\mathbb{E}(f) = \frac{(x+h)\cdot(t+w)+(-x)\cdot t}{h} = \frac{xw}{h} + t + w.$$

If $w > 0$, as your previous earnings $x \to \infty$ or your additional earnings $h \to 0$, the average time of benefits $\mathbb{E}(f) \to \infty$. If $w < 0$, as your previous earnings $x \to \infty$ or as your additional earnings $h \to 0$, the average time of benefits $\mathbb{E}(f) \to -\infty$. This is peculiar. All the benefits (positive and negative) are occurring around t and $t+w$, but as we make your previous earnings higher (not changing when it occurs) and/or your additional earnings lower (not changing when it occurs), the average time of benefits shoots off to infinity or negative infinite, depending on whether your old paycheck came before or after your new paycheck.

1.2 Ironing

The binning principle states that small changes in bin size should not result in disproportionately large changes in the statistic. The method we propose for doing statistics on arbitrary distributions, which we call *ironing*, builds on this principle.

The central idea behind ironing is the following. We know that standard statistics work in intuitive and desirable ways for distributions which are everywhere non-negative. The binning principle says that decreasing the bin size slightly—which may introduce negative regions that weren't previously there—shouldn't change the statistic by much. Hence, even if we don't have a good way of computing the statistic directly from a distribution with negative regions, we can deduce that it should be close to the statistic applied to a distribution with slightly larger bins, if that distribution is everywhere non-negative.

This motivates the idea that, when faced with a distribution with negative regions, we should attempt to bin these negative regions in the least obtrusive way (roughly, with the smallest bins), such that the resulting distribution is everywhere non-negative. The binning principle tells us that the statistic applied to this distribution should be relatively close to the ideal statistic applied to the original distribution.

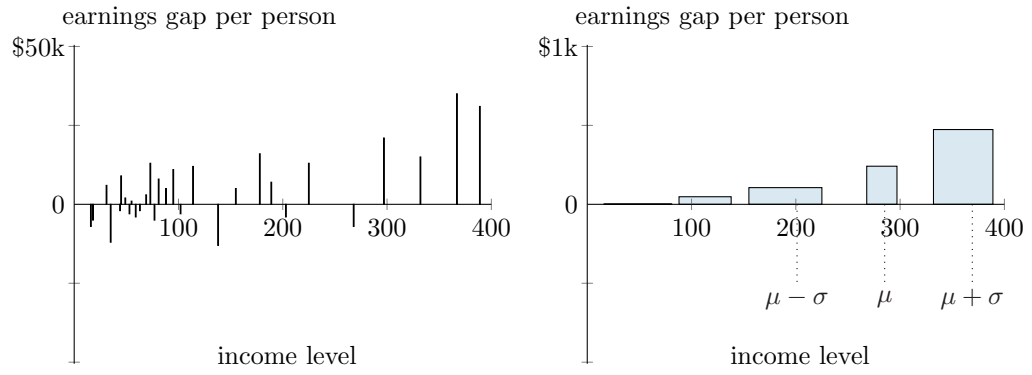


Figure 5: The first chart plots the earnings gap between a treated group and an untreated group across income levels binned to the dollar. The second chart plots the earnings gap with bins sufficiently large to smooth the positive and negative swings across individuals. The mean (μ) and standard deviation (σ) of the latter distribution is 285.20 and 84.03, respectively. By the binning principle, the "ideal" mean and standard deviation of the former distribution should be relatively close to 285.20 and 84.03, respectively.

We propose that a sensible way to construct such bins is to minimize the movement of weights across the dimension of interest. Binning involves smoothing weights across an interval, which involves moving weights away from their initial location. For any two distributions, we can measure the distance between them as the minimum distance required to move the weights from one to achieve the other.[5] Hence, given

[5]Formally, this is known as the earth mover's distance, or the Wasserstein-1 metric. See Section 2 for a formal definition.

any arbitrary distribution f, we seek an "ironed" distribution $\tilde{f}$ which minimizes the distance to f among all distributions which are everywhere non-negative. We show that such a distribution exists and is unique in Section 2.

It turns out that this distance can be measured quite simply as the total absolute deviation between the respective *cumulative* distribution functions. Hence, given any arbitrary cumulative distribution function F (which may be non-monotonic), normalized by the total weight, the ironed cumulative distribution function $\tilde{F}$ is the classical (i.e., non-decreasing) cumulative distribution function which minimizes the total absolute distance to F. We may then apply standard statistics as usual on the ironed distribution $\tilde{F}$.

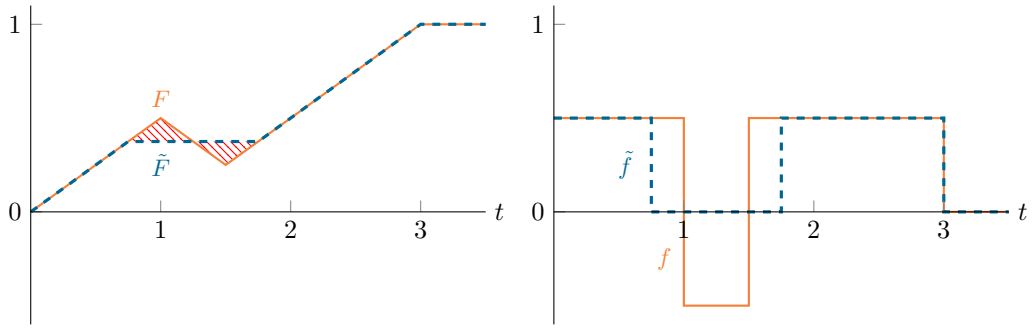


Figure 6: The first chart plots the arbitrary cumulative distribution function F alongside the ironed cumulative distribution function $\tilde{F}$. The second chart plots the associated arbitrary density function f alongside the ironed density function $\tilde{f}$.

Let $\mathcal{F}^{\text{CDF}}$ be the set of all classical (non-decreasing) cumulative distribution functions. The ironing procedure can then be described as follows.

1. Consider any arbitrary cumulative distribution function F normalized by the total weight, i.e. where $\int_{-\infty}^{\infty} \mathrm{d}F(t) = 1$.[6] $F(t)$ is interpreted as the fraction of the total weight assigned to t or below.

2. Iron the distribution, i.e., compute

$$\tilde{F} \in \operatorname*{arg\,min}_{\hat{F} \in \mathcal{F}^{\text{CDF}}} \int_{-\infty}^{\infty} \left| F(t) - \hat{F}(t) \right| \mathrm{d}t.$$

   $\tilde{F}$ exists and is unique by Theorems 1 and 2.

3. Do statistics on the ironed distribution $\tilde{F}$.

[6]The total weight is assumed to be strictly positive. It is not clear what it means to ask, for example, where the center of mass occurs when the total weight is zero. Similarly, if the total weight is negative, we would ask where the center of *negative* mass occurs (i.e., flip the sign and relabel).

2 Formal Results

2.1 Setup

A classical cumulative distribution function is standard from probability theory.

Definition 1. A function $F : \mathbb{R} \to \mathbb{R}$ is a *classical cumulative distribution function*[7] if

1. F is non-decreasing,
2. F is right-continuous,
3. $\lim_{x \to -\infty} F(x) = 0$, and
4. $\lim_{x \to \infty} F(x) = 1$.

Henceforth, we will refer to such functions as classical CDFs, or simply CDFs. Let $\mathcal{F}^{\text{CDF}}$ be the set of all CDFs. Removing properties 1, 3, and 4, we define an arbitrary cumulative distribution function as any function which is right-continuous.

Definition 2. A function $F : \mathbb{R} \to \mathbb{R}$ is an *arbitrary cumulative distribution function* if F is right-continuous.

As discussed in Section 1, we seek to measure the distance between two distributions by the minimum distance required to move the weights from one to achieve the other. Formally, this is known as the earth mover's distance, or the Wasserstein-1 metric. Intuitively, if two distributions are each interpreted as piles of earth, the earth mover's distance represents the minimum cost of transforming one pile of earth into the other, where the cost is the amount of earth moved multiplied by the distance moved.

Definition 3. The *1-Wasserstein distance*[8] between any two real-valued functions F_1 and F_2, denoted $W_1(F_1, F_2)$, is given by

$$W_1(F_1, F_2) = \int_{-\infty}^{\infty} \left| F_1(t) - F_2(t) \right| \, \mathrm{d}t.$$

Finally, we say that a CDF, F^{CDF}, is a best approximation to an arbitrary cumulative distribution function, F, from the set of all CDFs if it minimizes the earth mover's distance to F among all CDFs.

[7] See, e.g., Ash (2008, p. 69).

[8] Note that this is a slight generalization, as the usual definition of the *1-Wasserstein distance* is defined for any two cumulative distribution functions F_1 and F_2, rather than for any two real-valued functions.

Definition 4. A function $F^{\text{CDF}} \in \mathcal{F}^{\text{CDF}}$ is a *best approximation* to $F \in \mathcal{F}$ from $\mathcal{F}^{\text{CDF}}$ if

$$W_1(F, F^{\text{CDF}}) = \inf_{\hat{F} \in \mathcal{F}^{\text{CDF}}} W_1(F, \hat{F}).$$

We now show that such a best approximation exists (Theorem 1) and, for well-behaved arbitrary distributions, is unique (Theorem 2).

2.2 Existence

Lemma 1. $(\mathcal{F}^{\text{CDF}}, W_1)$ *is a complete metric space.*

Proof. Let $(\mathbb{R}, \mathcal{B}, \lambda)$ be a measure space, where $\mathcal{B}$ is the Borel sigma-algebra on $\mathbb{R}$ and λ is the Lebesgue measure on $\mathbb{R}$. The space $L^1(\mathbb{R}, \mathcal{B}, \lambda)$ consists of all real-valued measurable functions on $\mathbb{R}$ that satisfy

$$\|f\|_{L^1} \equiv \int_{-\infty}^{\infty} |f(x)| \, \mathrm{d}x < \infty.$$

Notice that the L^1 norm induces the W_1 metric—i.e., $W_1(F_1, F_2) = \|F_1 - F_2\|_{L^1}$.

We call two functions $F_1, F_2 \in L^1$ equivalent if $F_1 = F_2$ almost everywhere—i.e., $W_1(F_1, F_2) = 0$. By the Riesz-Fischer theorem, L^1 is complete. That is, every Cauchy sequence of functions in L^1 converges to a function in L^1 under the W_1 metric (or, more precisely, every Cauchy sequence of equivalence classes of functions in L^1 converges to an equivalence class of functions in L^1 under the W_1 metric). We would like to show that any Cauchy sequence of equivalence classes containing a CDF converges to an equivalence class of functions containing a CDF under the W_1 metric. For convenience, we will sometimes refer to this as a sequence of CDFs rather than a sequence of equivalence classes that contain a CDF.

Let $\{F_n\}_{n=1}^{\infty}$ be a Cauchy sequence of CDFs (i.e., $F_n \in \mathcal{F}^{\text{CDF}}$ for all n) with respect to the W_1 metric. Let $H_n \equiv F_n - F_1$ for all n. Then for large enough k, $\{H_n\}_{n=k}^{\infty}$ is a Cauchy sequence of L_1 functions, so it converges in the L_1 norm to an L_1 limit function H. Let $F \equiv H + F_1$. Then $\{F_n\}_{n=1}^{\infty}$ converges to F in the W_1 metric.

We first would like to show that F is non-decreasing (i.e., that F_n converges to an equivalence class that contains a non-decreasing function). If $H_n \to H$ in L^1, then there exists a subsequence of H_n which converges pointwise to H almost everywhere,[9] and hence there exists a subsequence of $F_n = H_n + F_1$ which converges pointwise to $F = H + F_1$ almost everywhere. Let $X \subseteq \mathbb{R}$ denote the set of points for which F_n converges to F pointwise. We know that $F_n = H_n + F_1$ is non-decreasing for all n. Hence, F is non-decreasing on X. To see this, suppose by contradiction that there exists $x, y \in X$ such that $x < y$ and $F(x) > F(y)$. Since $F_n \to F$

[9] See Belk (2015, Proposition 7).

pointwise on X, there exists N such that for all $n > N$, $F_n(x) > F_n(y)$, a contradiction. Since F is non-decreasing on X, which contains all but a measure zero of points, there exists a function G which is non-decreasing everywhere in the same equivalence class (i.e., $W(F, G) = 0$). □

We would now like to show that F is right-continuous (i.e., that F_n converges to an equivalence class that contains a non-decreasing, right-continuous function). Every non-decreasing function is equivalent to a right-continuous, non-decreasing function. To see this, suppose that F is non-decreasing and let $G(x) = \lim_{y \to x^+} F(y)$ for all x. G is right-continuous by construction. Moreover, $G(x) = F(x)$ at every point x where F is continuous. Since F is non-decreasing, it has at most a countable set of discontinuity points. Hence, $G(x) = F(x)$ almost everywhere. □

Finally, we would like to show that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. Suppose by contradiction that $\lim_{x \to -\infty} F(x) \neq 0$. Since F is non-decreasing, either $\lim_{x \to -\infty} F(x) = c \neq 0$ or F diverges. In either case, $\int_{-\infty}^{\infty} \left| F_n(t) - F(t) \right| \, \mathrm{d}t$ diverges for any n, since $\lim_{x \to -\infty} F_n(x) = 0$ for each n—a contradiction. Hence, $\lim_{x \to -\infty} F(x) = 0$. A similar argument shows $\lim_{x \to \infty} F(x) = 1$. □ ■

Theorem 1. *For any $F \in \mathcal{F}$, a best approximation to F from $\mathcal{F}^{\mathrm{CDF}}$ exists.*

Proof. $(\mathcal{F}^{\mathrm{CDF}}, W_1)$ is a complete metric space by Lemma 1. Hence, there exists $F \in \mathcal{F}^{\mathrm{CDF}}$ such that

$$W_1(F, F^{\mathrm{CDF}}) = \inf_{\hat{F} \in \mathcal{F}^{\mathrm{CDF}}} W_1(F, \hat{F}).$$

■

2.3 Uniqueness

Lemma 2. *$\mathcal{F}^{\mathrm{CDF}}$ is convex.*

Proof. We would like to show that for any $F, G \in \mathcal{F}^{\mathrm{CDF}}$ and $\alpha \in (0, 1)$, $\alpha F + (1 - \alpha)G \in \mathcal{F}^{\mathrm{CDF}}$.

1. If F and G are non-decreasing, then for any $x < y$, $F(x) \leq F(y)$, $G(x) \leq G(y)$, and hence $\alpha F(x) + (1 - \alpha)G(x) \leq \alpha F(y) + (1 - \alpha)G(y)$. So $\alpha F + (1 - \alpha)G$ is non-decreasing.
2. If F and G are right-continuous, so that for all $c \in \mathbb{R}_+$, $\lim_{x \to c^+} F(x) = F(c)$ and $\lim_{x \to c^+} G(x) = G(c)$, then for all $c \in \mathbb{R}_+$, $\lim_{x \to c^+} \alpha F(x) + (1 - \alpha)G(x) = \alpha F(c) + (1 - \alpha)G(c)$. So $\alpha F + (1 - \alpha)G$ is right-continuous.
3. If $\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} G(x) = 0$, then $\lim_{x \to -\infty} \alpha F(x) + (1 - \alpha)G(x) = 0$.

4. If $\lim_{x\to\infty} F(x) = \lim_{x\to\infty} G(x) = 1$, then $\lim_{x\to\infty} \alpha F(x) + (1-\alpha)G(x) = 1$.

■

Theorem 2. *For any $F \in \mathcal{F}$, if F is continuous and there exists some F^{CDF} such that $W_1(F, F^{\text{CDF}}) < \infty$, then there is a unique best approximation to F from $\mathcal{F}^{\text{CDF}}$.*

Proof. There exists a best approximation to F from $\mathcal{F}^{\text{CDF}}$ by Theorem 1. Suppose by contradiction $F_1 \neq F_2$ are each best approximations to F from $\mathcal{F}^{\text{CDF}}$. Then, by assumption, $W_1(F, F_1) = W_1(F, F_2) \equiv W^* < \infty$. By Lemma 2, $\alpha F_1 + (1-\alpha)F_2 \in \mathcal{F}^{\text{CDF}}$ for any $\alpha \in (0,1)$. Hence,

$$\begin{aligned}
W_1(F, .5(F_1+F_2)) &= \int_{-\infty}^{\infty} \left|F(t) - .5(F_1(t)+F_2(t))\right| \,\mathrm{d}t \\
&= \int_{-\infty}^{\infty} \left|.5(F(t)-F_1(t)) + .5(F(t)-F_2(t))\right| \,\mathrm{d}t \\
&\leq \int_{-\infty}^{\infty} \left|.5(F(t)-F_1(t))\right| + \left|.5(F(t)-F_2(t))\right| \,\mathrm{d}t \\
&= .5\int_{-\infty}^{\infty} \left|(F(t)-F_1(t))\right| \,\mathrm{d}t + .5\int_{-\infty}^{\infty} \left|(F(t)-F_2(t))\right| \,\mathrm{d}t \\
&= W^*.
\end{aligned}$$

Case 1. Suppose $\text{sign}(F(t)-F_1(t)) \times \text{sign}(F(t)-F_2(t)) < 0$ for some t.[10] Then the inequality above is strict, contradicting that F_1 and F_2 are best approximations.

Case 2. Suppose $\text{sign}(F(t)-F_1(t)) \times \text{sign}(F(t)-F_2(t)) \geq 0$ for all t.

If $F_1 = F$, then $F_2 = F$ (since F and F_2 are right-continuous, any $F_2 \neq F$ has $W_1(F, F_2) > 0$), a contradiction with $F_1 \neq F_2$. Suppose $F \neq F_1$ and $F \neq F_2$. Let $T^+ = \{t \in \mathbb{R} : F_1(t), F_2(t) \geq F(t)\}$ and $T^- = \{t \in \mathbb{R} : F_1(t), F_2(t) \leq F(t)\}$. Note that $T^+ \cup T^- = \mathbb{R}$. Let

$$F^*(t) = \begin{cases} \min\{F_1(t), F_2(t)\} & \text{if } t \in T^+ \setminus T^- \\ F_1(t) = F_2(t) & \text{if } t \in T^+ \cap T^- \\ \max\{F_1(t), F_2(t)\} & \text{if } t \in T^- \setminus T^+ \end{cases}.$$

We would now like to show that $F^* \in \mathcal{F}^{\text{CDF}}$.

[10] The sign function is defined by

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}.$$

1. F^* is non-decreasing,

   Consider any $x < y$. We would like to show $F^*(y) \geq F^*(x)$.

   Case 2.1.1. Suppose $F(x) \geq F_1(x) \geq F_2(x)$ and $F(y) \geq F_1(y) \geq F_2(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_1(y)$, so $F^*(y) = F_1(y) \geq F_1(x) = F^*(x)$.

   Case 2.1.2. Suppose $F(x) \geq F_1(x) \geq F_2(x)$ and $F(y) \geq F_2(y) > F_1(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_2(y)$, so $F^*(y) = F_2(y) > F_1(y) \geq F_1(x) = F^*(x)$.

   Case 2.1.3. Suppose $F(x) \geq F_1(x) \geq F_2(x)$ and $F_2(y) > F_1(y) \geq F(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_1(y)$, so $F^*(y) = F_1(y) \geq F_1(x) = F^*(x)$.

   Case 2.1.4. Suppose $F(x) \geq F_1(x) \geq F_2(x)$ and $F_1(y) > F_2(y) \geq F(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_2(y)$.

   First, suppose $F_2(y) \geq F_1(x)$. Then $F^*(y) = F_2(y) \geq F_1(x) = F^*(x)$. Next, suppose by contradiction $F_2(y) < F_1(x)$. Then $F_1(y) \geq F_1(x) > F_2(y) \geq F_2(x)$ and since F_1 and F_2 are non-decreasing, $F_1(t) > F_2(t')$ for all $t, t' \in [x, y]$. Since F is continuous, by the intermediate value theorem, there exists $t^* \in [x, y]$ and $\varepsilon > 0$ such that $F_1(t^* + \varepsilon) \geq F_1(x) = F(t^*) > F(t^* + \varepsilon) > F_2(y) \geq F_2(t^* + \varepsilon)$. Hence, $\text{sign}(F(t^* + \varepsilon) - F_1(t^* + \varepsilon)) \times \text{sign}(F(t^* + \varepsilon) - F_2(t^* + \varepsilon)) < 0$, contradicting Case 2.

   Case 2.1.5. Suppose $F_2(x) > F_1(x) \geq F(x)$ and $F(y) \geq F_1(y) \geq F_2(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_1(y)$, so $F^*(y) = F_1(y) \geq F_1(x) = F^*(x)$.

   Case 2.1.6. Suppose $F_2(x) > F_1(x) \geq F(x)$ and $F(y) \geq F_2(y) > F_1(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_2(y)$, so $F^*(y) = F_2(y) \geq F_2(x) > F_1(x) = F^*(x)$.

   Case 2.1.7. Suppose $F_2(x) > F_1(x) \geq F(x)$ and $F_2(y) > F_1(y) \geq F(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_1(y)$, so $F^*(y) = F_1(y) \geq F_1(x) = F^*(x)$.

   Case 2.1.8. Suppose $F_2(x) > F_1(x) \geq F(x)$ and $F_1(y) > F_2(y) \geq F(y)$. Then $F^*(x) = F_1(x)$ and $F^*(y) = F_2(y)$, so $F^*(y) = F_2(y) \geq F_2(x) > F_1(x) = F^*(x)$.

   This exhausts the sub-cases since a) sub-cases with $F(x)$ in the middle (e.g., with $F_1(x) > F(x) > F_2(x)$) are ruled out by Case 2 and b) switching the order of $F_1(x)$ and $F_2(x)$ is just a matter of relabeling. Hence, F^* is non-decreasing. □

2. F^* is right-continuous,

   Note that $\min\{F_1(t), F_2(t)\}$ and $\max\{F_1(t), F_2(t)\}$ are each right-continuous since F_1 and F_2 are each right-continuous. Fix any $x \in \mathbb{R}$.

Case 2.2.1. Suppose $F(x) < F_1(x)$ or $F(x) < F_2(x)$. Without loss of generality, suppose $F(x) < F_1(x)$. Then there exists $\delta > 0$ such that for all t such that $x < t < x + \delta$, $F(t) < F_1(t)$. Hence, $F^*(t) = \min\{F_1(t), F_2(t)\}$ for all $t \in [x, x + \delta)$, and F^* is right-continuous at x.

Case 2.2.2. Suppose $F(x) > F_1(x)$ or $F(x) > F_2(x)$. The argument follows as in Case 2.2.1.

Case 2.2.3. Suppose $F(x) = F_1(x) = F_2(x)$. For any t, $F^*(t) = \min\{F_1(t), F_2(t)\}$ or $F^*(t) = \max\{F_1(t), F_2(t)\}$. Moreover, $\lim_{t \to x^+} \min\{F_1(t), F_2(t)\} = \lim_{t \to x^+} \max\{F_1(t), F_2(t)\} = F(x) = F^*(x)$. Hence, F^* is right-continuous at x. □

3. $\lim_{x \to -\infty} F^*(x) = 0$ and $\lim_{x \to \infty} F^*(x) = 1$

   Since for every $x \in \mathbb{R}$, $F^*(x) = F_1(x)$ or $F^*(x) = F_2(x)$, and moreover $\lim_{x \to -\infty} F_1(x) = \lim_{x \to -\infty} F_2(x) = 0$ and $\lim_{x \to \infty} F_1(x) = \lim_{x \to -\infty} F_2(x) = 1$, the result follows. □

We would now like to show that $F^* \neq F_1$. Suppose by contradiction $F^* = F_1$. Then F_1 is everywhere weakly closer to F than F_2, and in some places strictly closer. Since F, F_1, and F_2 are right-continuous, this implies $W_1(F, F_1) < W_1(F, F_2)$, a contradiction.

Now, since $F^* \in \mathcal{F}^{\text{CDF}}$ is everywhere weakly closer to F than F_1, and in some places strictly closer, and since F, F_1, and F^* are right-continuous, $W_1(F, F^*) < W_1(F, F_1) = W_1(F, F_2)$, contradicting that F_1 and F_2 are best approximations of F from $\mathcal{F}^{\text{CDF}}$. ■

References

Ash, Robert B. 2008. *Basic Probability Theory.* Dover Publications.

Myerson, Roger B. 1981. "Optimal Auction Design." *Mathematics of Operations Research*, 6(1): 58–73.