STA452
Lecture Notes

Yuchen Wang

February 15, 2021

Contents

1 Preface

In our course, you will not see very much data. Our job is to express the ideas that drive the logic (or the logic that drives the ideas). As a consequence, the most of the examples in elementary book appear as pure theory. But they are not defined, they are consequences of abstract mathematical ideas. For example, conditional density is a consequence of conditional expectation, which is an orthogonal projection in a vector space, a pure euclidean geometric idea.

1.1 Preliminary

Definition 1.1. A sequence of sets $A_n \to A$ iff $I(A_n) \to I(A)$.

Remark 1.1. See Appendix 2 of the original notes.

Proposition 1.1. A limit exists when the limsup is equal to the liminf:

$$lim = \overline{lim} = \underline{lim} \tag{1.1}$$

Proof. For $w \in \Omega$,

$$\begin{aligned} \sup_{t \in T} I(A_t)(w) &= I(\cup_{t \in T} A_t)(w) \\ &= 1 \text{ or } 0 \\ \inf_{t \in T} I(A_t)(w) &= I(\cap_{t \in T} A_t)(w) \end{aligned}$$

Therefore,

$$\begin{aligned} \lim_{n \to \infty} I(A_n) &= \overline{\lim_{n \to \infty}} I(A_n) \\ &= \inf_{n=1}^{\infty} \sup_{k=n}^{\infty} I(A_k) \\ &= I\left(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k\right) \\ &= \underline{\lim_{n \to \infty}} I(A_n) \\ &= \sup_{n=1}^{\infty} \inf_{k=n}^{\infty} I(A_k) \\ &= I\left(\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k\right) \end{aligned}$$

∎

Property 1.1. Therefore it is clear that

1. $$A_n \to A \iff A = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k \tag{1.2}$$

2. $$A_n \uparrow \implies A_n \uparrow \cup_{n=1}^{\infty} A_n \tag{1.3}$$

3. $$A_n \downarrow \implies A_n \downarrow \cap_{n=1}^{\infty} A_n \tag{1.4}$$

2 Virtual Dice

2.1 The Law of Large Numbers

Consider a *mechanism/process/system*, W, which generates *outcomes*, w, in a sample space Ω:

$$W : w_1, w_2, \dots, w_n, \dots$$

The outcomes are often referred to as *trials* of the *process* W. w_n is called the nth trial, and the finite sequence $(w_1, w_2, \dots, w_n)$ is the first n trials.
Consider any real-valued function $g : \Omega \to \mathbb{R}$ defined on the *sample space* Ω. Let $X = g(W)$ denote the *extended process* that applies the function g to the outcome w from W to produce the outcome $x = g(w)$. This new process has its own sequence of trial outcomes:

$$\begin{aligned} g(W) &: g(w_1), g(w_2), \dots, g(w_n), \dots \\ \text{or} \quad X \quad &: x_1, x_2, \dots, x_n, \dots \end{aligned}$$

These transformed outcomes are all real values, with which we can do lots of easy arithmetic, while the abstract sample space Ω may not have this property.

Definition 2.1 (sample mean)**.** For each $n \in \mathbb{N}$, the *sample mean* over the first n trials is the *arithmetic average* of the function values over those n trials:

$$\widehat{E}_n X := \frac{g(w_1) + \dots + g(w_n)}{n} = \bar{x}_n \tag{2.1}$$

Definition 2.2 (random variable)**.** A given process W is said to be a *random process / random variable* iff it satisfies the *empirical law of large numbers*, in that, for any real-valued $X = g(W)$, we have

1. *stability*: the sequence of *sample averages* $(\widehat{E}_n g(W), n \in \mathbb{N})$ converges;
2. *invariance*: the limit is independent of any particular realization $(w_n, n \in \mathbb{N})$.

Definition 2.3 (expected value)**.** For each real-valued $X = g(W)$, we obtain a *expected value* in the above limit:

$$EX := \lim_{n\to\infty} \widehat{E}_n g(W) = \lim_{n\to\infty} \hat{x}_n \tag{2.2}$$

Definition 2.4 (indicator function)**.** The indicator function of a subset A of a set X is a function $I_A : X \to \{0, 1\}$ defined as

$$I_A(x) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{2.3}$$

Definition 2.5 (probability)**.** Now *probability* itself is a special case of an *expected value*: for any *indicator function* $g = I_A$ with $A \subset \Omega$ we will get the usual sequence of averages, but now to be referred to as *empirical relative frequencies*. These averages give the proportion of times that A occurs in the first n trials.

$$\widehat{P}_n(W \in A) := \widehat{E}_n I_A(W) = \frac{I_A(w_1) + \dots + I_A(w_n)}{n} \; \forall n \in \mathbb{N} \tag{2.4}$$

As $n \to \infty$, the above equation gives the *long-run frequency*, or *probability*:

$$P_W(A) = P(W \in A) := \lim_{n\to\infty} \widehat{E}_n I_A(W) = \lim_{n\to\infty} \widehat{P}_n(W \in A) \tag{2.5}$$

Notation 2.1. Given a random variable W and a *probability distribution* P_W, we can use the following notation:

$$W \sim P_W \quad \text{on} \quad \Omega$$

to be read as "W is distributed as P_W on Ω" or "W is distributed as P_W".

2.2 Some examples: "virtual dice"

Definition 2.6. For any specific $n \in \mathbb{N}$, the random variable X is said to have a *(finite discrete) uniform distribution* on the sample space $\Omega = \{1, \ldots, n\}$ (denoted $X \sim unif\{1, \ldots, n\}$)
iff

$$P(X = k) = \frac{1}{n}, \quad , k = 1, \ldots, n \tag{2.6}$$

Example 2.1. A ten-sided die: $Y \sim unif\{0, \ldots, 9\}$ Let Y be a 2-stage procedure:
Divide the ten digits $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}$ into two batches

$$A = \{1, 2, 3, 4, 5\} \quad \& \quad B = \{6, 7, 8, 9, 0\} \tag{2.7}$$

and then toss a standard six-sided die twice. On the first toss, if the die shows $1, 2$ or 3 then we go to A, if the die shows $4, 5$ or 6 then we go to B. Thus each batch is selected half the time. On the second toss, ignoring the digit 6, and if the die shows k we take the kth digit in the batch and report the result. It should be clear then we will arrive at each of the ten digits with identical frequency $\frac{1}{10}$.

2.2.1 A higher level of virtuality: 'continuous dice' and random 'real' numbers

Let U denote the hypothetical possibility of generating an infinite decimal expansion of a number between 0 and 1, by performing the physical algorithm outlined in Example (2.1) an infinite number of times. So an outcome u for U entails an infinite number of repetitions $Y_i, i = 1, 2, \ldots$ of the finite procedure Y:

$$U = \sum_{i=1}^{\infty} \frac{Y_i}{10^i} = 0.Y_1 Y_2 Y_3 \ldots \tag{2.8}$$

Example 2.2. If we generate U explicitly to four places $.y_1 y_2 y_3 y_4$, then there are $10,000$ equally likely possibilities, and our 'actual' U is known to be somewhere between $.y_1 y_2 y_3 y_4$ and 0.0001 higher. In other words, the outcome is in one particular of $10,000$ equally likely subintervals of $[0, 1]$:

$$P(.y_1 y_2 y_3 y_4 \leq U \leq .y_1 y_2 y_3 y_4 + 0.0001) = 1/10,000 \quad \forall y_1, y_2, y_3, y_4 \in \Omega \tag{2.9}$$

Thus we can deduce that

$$P(0 \leq U \leq .a_1 a_2 a_3 a_4) = a_1 a_2 a_3 a_4 / 10,000 = .a_1 a_2 a_3 a_4 \forall a_1, a_2, a_3, a_4 \in \Omega \tag{2.10}$$

More generally, if u is an n-place finite decimal in the interval $[0, 1]$ for any $n \in \mathbb{Z}$, then $P(U \leq u) = u$, and for any pair of n-place finite decimals $a, b \in [0, 1]$ with $a \leq b$, we will have the *uniformity condition*

$$P(a \leq U \leq b) = b - a \tag{2.11}$$

Corollary 2.1. The probability of U obtaining any specific value u is zero.

$$P(U = u) = P(u \leq U \leq u) = u - u = 0 \tag{2.12}$$

Definition 2.7 (uniform distribution)**.** The random variable U is said to have a *(continuous) uniform distribution* on the unit interval $[0, 1]$ (denoted $U \sim unif[0, 1]$)
iff

$$P(U \leq u) = u \quad \forall 0 \leq u \leq 1 \tag{2.13}$$

Remark 2.1. This is a mathematical statement, which is different from physical existence as in Definition (2.6).

Corollary 2.2. If $X \sim unif[a, b]$ and $U \sim unif[0, 1]$, then

$$x = a + (b - a) \cdot u \tag{2.14}$$

Example 2.3. Let $V = 1 - U$, then

$$\begin{aligned} P(V \leq u) &= P(1 - U \leq u) = P(U \geq 1 - u) && (2.15)\\ &= 1 - P(U \leq 1 - u) && (2.16)\\ &= 1 - (1 - u) = u = P(U \leq u) && (2.17) \end{aligned}$$

As random variables, U and V behave exactly the same way. They have the same *stochastic behavior*. Accordingly, they are said to be *equal-in-distribution*: $V \stackrel{d}{=} U$.

2.2.2 Equality-in-distribution

Definition 2.8 (equality-in-distribution)**.** Two random variables W_1, W_2 on the same sample space Ω are said to be *identically distributed / stochastically identical* (denoted $W_1 \stackrel{d}{=} W_2$)
iff

$$Eg(W_1) = Eg(W_2) \quad \forall g : \Omega \to \mathbb{R} \tag{2.18}$$

iff

$$P(W_1 \in A) = P(W_2 \in A) \quad \forall A \subset \Omega \tag{2.19}$$

Proposition 2.1 (invariance 1)**.** For any function $\phi : \Omega \to \chi$

$$W_1 \stackrel{d}{=} W_2 \implies \phi(W_1) \stackrel{d}{=} \phi(W_2) \tag{2.20}$$

Proof.

$$Eh(\phi(W_1)) = Eh(\phi(W_2)) \quad \forall h : \chi \to \mathbb{R}$$

■

Proposition 2.2 (invariance 2)**.**

$$W_1 \stackrel{d}{=} W_2 \iff g(W_1) \stackrel{d}{=} g(W_2) \quad \forall g : \Omega \to \mathbb{R} \tag{2.21}$$

2.3 Nature makes them, so can you

2.3.1 Exponential distribution

Let $Z = -\ln U$ with $U \sim unif[0, 1]$. Then it is straightforward to compute that, for any non-negative $0 \leq s \leq t \leq \infty$:

$$P(s \leq Z \leq t) = e^{-s} - e^{-t} \tag{2.22}$$

Proof.

$$\begin{aligned} s \leq Z \leq t &\iff s \leq -\ln U \leq t\\ &\iff -t \leq \ln U \leq -s\\ &\iff e^{-t} \leq U \leq e^{-s} \end{aligned}$$

Therefore

$$\begin{aligned} P(s \leq Z \leq t) &= P(e^{-t} \leq U \leq e^{-s}) \\ &= e^{-s} - e^{-t} \end{aligned}$$

■

Definition 2.9 (standard exponential distribution)**.** The random variable Z is said to have a *standard exponential distribution* on $[0, \infty)$ (denoted $Z \sim \exp(1)$)
iff

$$P(Z \leq z) = 1 - e^{-z} \quad \forall z \geq 0 \tag{2.23}$$

Definition 2.10 (scaled exponential distribution)**.** The random variable X is said to have a *scaled exponential distribution*, with *scale parameter* $\theta > 0$ on $[0, \infty)$ (denoted $X \sim \exp(\theta)$)
iff

$$X \stackrel{d}{=} \theta Z, \quad \text{where } Z \sim \exp(1) \tag{2.24}$$

2.3.2 Consider the generalization

Consider any strictly monotone and C^1 function, g on the interval $[0, 1]$, and let $X \stackrel{d}{=} g(U)$, where $U \sim unif[0, 1]$. Then

$$P(s < X \leq t) = \begin{cases} g^{-1}(t) - g^{-1}(s), & g \uparrow\uparrow \\ g^{-1}(s) - g^{-1}(t), & g \downarrow\downarrow \end{cases} \tag{2.25}$$

Corollary 2.3. Suppose $F : \mathbb{R} \to [0, 1]\, x \mapsto P(X \leq x)$. Then F is certainly *non-decreasing*, and for any $s \leq t$,

$$P(s < X \leq t) = F(t) - F(s) \tag{2.26}$$

Definition 2.11 (distribution function)**.** For any real-valued random variable, X, the *distribution function* of X is given by

$$F(x) \stackrel{or}{=} F_X(x) := P(X \leq x) \quad \forall x \in \mathbb{R} \tag{2.27}$$

Remark 2.2. Let $f(x) = F'(x)$, then we immediately have

$$P(s < X \leq t) = F(t) - F(s) = \int_s^t f(x)dx \quad \forall s, t \tag{2.28}$$

At each $x \in g[0, 1]$,

$$\lim_{s\uparrow x,\, t\downarrow x} \frac{P(s < X \leq t)}{t - s} = \lim_{s\uparrow x,\, t\downarrow x} \frac{F(t) - F(s)}{t - s} = f(x) \tag{2.29}$$

Remark 2.3. $f(x)$ can be interpreted as "amount of probability per unit length at the point x".

Definition 2.12 (probability density function)**.** A real-valued random variable X is said to be *absolutely continuous* (wrt length measure)
iff

$$\exists f : \mathbb{R} \to [0, \infty), P(s < X \leq t) = \int_s^t f(x)\, dx \quad \forall s \leq t \tag{2.30}$$

in which case, the function f (not necessarily unique) is referred to as the *probability density function* of X.

Remark 2.4. For any abs. cont. X,

$$P(X = x) = \int_x^x f(x)\,dx = 0 \quad \forall x \tag{2.31}$$

so there is no discrete contribution to the distribution at any $x \in \mathbb{R}$. Thus,

$$P(s \le X \le t) = P(s < X < t) = P(s < X \le t) = P(s \le X < t)$$

Proposition 2.3. $F : [a, b] \to [0, 1]$ is C^1, iff

$$F(x) = \int_a^x f(s)\,ds \quad \text{with } f = F' > 0 \text{ cont. on } [a, b]$$

Proposition 2.4. If $g = F^{-1}$ and $g \in C^1$, then $F(X) \overset{d}{=} U$

Proof.

$$\begin{aligned} P(F(X) \le u) &= P(X \le g(u)) \\ &= P(X \le g(u)) \\ &= F(g(u)) \\ &= u \\ &= P(U \le u) \end{aligned}$$

■

Definition 2.13 (quantile)**.** For any $0 \le p \le 1$, the value $x_p = g(p) = F^{-1}(p)$ is called the $100 \times p$th *quantile* (or *percentile*) of X. The function g is called the *quantile function.*

$$P(X \le x_p) = p \tag{2.32}$$

2.4 Expected Value

Property 2.1 (finite additivity of probability)**.** If two sets A and B are mutually disjoint, then

$$I(A + B) = I(A) + I(B) \tag{2.33}$$

Therefore

$$\begin{aligned} P(A + B) &= EI(A + B)(W) = E(I(A)(W) + I(B)(W)) && (2.34) \\ &= EI(A)(W) + EI(B)(W) && (2.35) \\ &= P(A) + P(B) && (2.36) \end{aligned}$$

We can prove by induction that

$$P\left(\sum_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) \tag{2.37}$$

Property 2.2 (E is normed on constant random variables)**.** E is *normed* on $g(W) = c$ where $c \in \mathbb{R}$

$$Ec = c \quad \forall c \in \mathbb{R} \tag{2.38}$$

Property 2.3. The indicator function of the whole sample space Ω is 1

$$I_\Omega(W) = 1 \implies P(\Omega) = E1 = 1 \tag{2.39}$$

Property 2.4 (non-negativity of probability)**.**

$$0 \le I(A) \le 1 \implies 0 \le P(A) = EI(A) \le 1 \tag{2.40}$$

2.4.1 Expected Value for an Arbitrary Finite Discrete Distribution

Definition 2.14 (finite scheme)**.** For any finite discrete distribution, we can write a *finite scheme*

$$W \sim \begin{pmatrix} w_1 & \dots & w_N \\ p_1 & \dots & p_N \end{pmatrix} \tag{2.41}$$

to symbolize the *probability mass function*

$$P(W = w_i) = p_i, \quad i \in \{1, 2, \dots, N\}$$

where $\sum_{i=1}^{N} p_i = 1$.

Corollary 2.4. For any real-valued function $g(W), W \in \Omega$, the expected value is

$$Eg(W) = \sum_{i=1}^{N} g(w_i) P(W = w_i) = \sum_{i=1}^{N} g(w_i) p_i \tag{2.42}$$

Proof. $g(W)$ can be explicitly represented as a finite linear combination of simple indicator functions

$$g(W) = \sum_{i=1}^{N} g(w_i) I(W = w_i)$$

So that applying E to both sides gives us the result. ■

2.4.2 Full generality: lebesgue-stieltjes

Suppose we are given a distribution function $F(x) = P(X \leq x), x \in \mathbb{R}$, for a real-valued random variable $X = g(W)$, with $W \sim P$ on sample space Ω. Then consider some discrete approximation to X, for example,

$$X_n = \sum_{i=-n}^{n} \frac{i-1}{\sqrt{n}} I\left(\frac{i-1}{\sqrt{n}} < X < \frac{i}{\sqrt{n}}\right) \tag{2.43}$$

For this particular approximation,

$$|X - X_n| \leq \frac{1}{\sqrt{n}} + |X| I(|X| > \sqrt{n})$$

Thus $X_n \to X$ as $n \to \infty$. Then any continuous real-valued function $h(X_n) \to h(X)$ as $n \to \infty$. If $h(X)$ is bounded, then

$$Eh(X) = \lim_{n \to \infty} \sum_{i=-n}^{n} h(\frac{i-1}{\sqrt{n}}) \left(F(\frac{i}{\sqrt{n}}) - F(\frac{i-1}{\sqrt{n}})\right) \tag{2.44}$$

which is called the *lebesgue-stieltjes integral* of the function $h(x)$. It may be denoted

$$Eh(X) := \int_{-\infty}^{\infty} h(x)\, dF(x) \tag{2.45}$$

2.4.3 Examples

Definition 2.15 (bernoulli trial)**.** The random variable Z is said to be a *bernoulli trial* (denoted $Z \sim bern(p), 0 \leq p \leq 1$) iff

$$Z \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$$

2.4.4 Expected Value for Continuous Functions

2.4.5 Expected Value for C^1 functions

Consider the special case where function g is strictly monotone and C^1.

Proposition 2.5. If $X = g(U), g : [0,1] \to [a,b]$ is strictly monotone and C^1, then for any continuous function $h : \mathbb{R} \to \mathbb{R}$,

$$Eh(X) = \int_a^b h(x) f(x)\, dx \tag{2.46}$$

where

$$F = \begin{cases} g^{-1} & , g \upuparrows \\ 1 - g^{-1} & , g \downdownarrows \end{cases} \quad \text{and} \quad f(x) = F'(x)$$

Proof. For any $0 \le t < 1$, when $g \upuparrows$ and C^1, we have

$$\int_0^t h(g(u))\, du = ???$$

■

complete it later

2.5 Exponential Distribution

notes

2.6 Gamma Distribution

notes

2.7 Continuity Revisited

2.7.1 Sequential Continuity of Probability

Definition 2.16 (σ-additivity)**.** P is said to be *σ-additive / countably additive* iff for any mutually disjoint sequence of events A_n $(n \in \mathbb{N})$

$$P(\sum_1^\infty A_n) = \sum_1^\infty P(A_n) \tag{2.47}$$

Remark 2.5. Equation (2.47) is equivalent to the following pair of equations:

$$\text{finite-additivity: } P(\sum_1^n A_i) = \sum_1^n P(A_n) \tag{2.48}$$

$$\text{continuity: } A_n \to A \implies P(A_n) \to P(A) \tag{2.49}$$

Proposition 2.6. If $A_n \uparrow A$ or $A_n \downarrow A$, then

$$P(A_n) \to P(A)$$

Proof. if If $A_n \uparrow A$ then we have that

$$A = \cup_{n=1}^\infty A_n = \sum_{n=1}^\infty (A_n - A_{n-1})$$

where, for convenience, we have $A_0 = \emptyset$.
Then

$$\begin{aligned} P(A) &= \sum_{n=1}^{\infty}(P(A_n) - P(A_{n-1}) \\ &= \lim_{n\to\infty} \sum_{i=1}^{\infty}(P(A_i) - P(A_{i-1})) \\ &= \lim_{n\to\infty} P(A_n) \end{aligned}$$

On the other hand, $A_n \downarrow A$ is equivalent to $A_n^c \uparrow A^c$. ■

Corollary 2.5 (sequential continuity)**.**

$$A_n \to A \implies P(A_n) \to P(A) \tag{2.50}$$

Proof. Suppose $A_n \to A$, then

$$\begin{aligned} \cup_{n=1}^{\infty} \cap_{k\geq n} A_k &= A = \cap_{n=1}^{\infty} \cup_{k\geq n} A_k \\ \cap_{k\geq n} A_k &\leq A_n, A \leq \cup_{k\geq n} A_k \\ P(\cap_{k\geq n} A_k) &\leq P(A_n), P(A) \leq P(\cup_{k\geq n} A_k) \\ |P(A_n) - P(A)| &\leq P(\cup_{k\geq n} A_k) - P(\cap_{k\geq n} A_k) \\ &\to P(A) - P(A) \\ &= 0 \end{aligned}$$

Therefore,

$$\begin{aligned} &|P(A_n) - P(A)| \to 0 \\ &P(A_n) \to P(A) \end{aligned}$$

■

2.8 Right Continuity of Cumulative Distribution Function

For any $x_n \downarrow x$, simply let $A_n = (-\infty, x_n]$ and $A = (-\infty, x]$.
Then $A_n \downarrow A$, so

$$F(x_n) = P(X \in A_n) \downarrow P(X \in A) = F(x)$$

Denoting the right-limit of F at x by $F(x+) := \lim_{y\downarrow x} F(y)$, and the left-limit $F(x-) := \lim_{y\uparrow x} F(y)$, we get the property of *right-continuity* for CDF

$$F(x+) = F(x) \quad \forall x \in \mathbb{R} \tag{2.51}$$

Remark 2.6. Any distribution function $F(x)$ can actually be discontinuous at no more than a countable number of points, which corresponds to all the jumps on the discrete part of the distribution.

Definition 2.17 (probability mass function)**.** For any real-valued random variable X, the *probability mass function* of X is given by

$$p(x) = p_X(x) = P(X = x) \quad \forall x \in \mathbb{R}$$

Proposition 2.7. Probability mass function

$$p(x) = F(x) - F(x-) \quad \forall x \in \mathbb{R} \tag{2.52}$$

Proof. For any $x_n \uparrow x$, simply let $A_n = (-\infty, x_n]$ and $A = (-\infty, x)$.
Then $A_n \uparrow A$, so

$$F(x-) := \lim_{n\to\infty} P(X \in A_n) = P(X \in A) = P(X < x)$$

Therefore

$$p(x) = P(X \leq x) - P(X < x) = F(x) - F(x-)$$

■

Remark 2.7. The points of continuity C_F of any distribution function correspond perfectly to the points where pmf is zero.

$$\begin{aligned} C_F &= \{x \in \mathbb{R} | F(x-) = F(x+)\} \\ &= \{x \in \mathbb{R} | F(x-) = F(x)\} \\ &= \{x \in \mathbb{R} | p(x) = 0\} = p^{-1}(0) \end{aligned}$$

The complementary region being the discrete part of the distribution

$$D_F = \{x \in \mathbb{R} | p(x) > 0\} = p^{-1}(0)^c$$

Proposition 2.8. D_F is at most countable.

$$\# D_F \leq \# \mathbb{N}$$

Proof. Note that

$$\{x \in \mathbb{R} | p(x) > 0\} = \cup_{n=1}^{\infty} \{x \in \mathbb{R} | p(x) > 1/n\}$$

It is clear that for every $n \in \mathbb{N}$, $\{x \in \mathbb{R} | p(x) > 1/n\}$ has less than n point in it. Otherwise

$$\exists A_n = \{a_1, \ldots, a_n\} \subset \{x \in \mathbb{R} | p(x) > 1/n\} \text{ with } P(A_n) > 1$$

which is a contradiction.
Since a countable union of countable sets is still countable, we have D_F is at most countable. ■

3 Reduction to an Axiomatic System

3.1 The Kolmogorov Axioms

Definition 3.1 (probability space)**.** A *probability space (distribution)* is a triple of objects (Ω, L, E)

1. Ω: any set, called the *sample space*
2. L: any vector space of real-valued functions on Ω that contains the constants, and is closed under taking absolute values ($X \in L \implies |X| \in L$), the elements of which are referred to as *random variables*
3. $E : L \to \mathbb{R}$, any functional that is
   - *normed*: $Ec = c$
   - *non-negative*: $X \geq 0 \implies EX \geq 0$
   - *linear*: $E \sum_1^n a_i X_i = \sum_1^n a_i EX_i$
   - *continuous*: $0 \leq X_n \to X \implies 0 \leq EX_n \to EX$

   referred to as an *expectation operator*, while its value EX at any $X \in L$ is called the *expected value* of that X.

Property 3.1 (continuity)**.** A useful variant of E's *continuous* property is stated as:
If $Z_n \geq 0, n = 1, 2, \ldots$, then

$$E \sum_{i=1}^{\infty} Z_n = \sum_{i=1}^{\infty} EZ_n \tag{3.1}$$

3.1.1 Reducing the Reduction

As understood, probability is a very special case of expected values. Thus we can reduce the definition of a probability space as follows

Definition 3.2 (probability space)**.** A *probability space (distribution)* is a triple of objects $(\Omega, \mathcal{F}, P)$

1. Ω: any set, called the *sample space*
2. $\mathcal{F}$: any σ-algebra of subsets of Ω, which is a non-empty collection closed under countable unions and complements. The elements of $\mathcal{F}$ are referred to as *events*
3. $E : \mathcal{F} \to \mathbb{R}$, any functional that is
   - *normed*: $Ec = c$
   - *non-negative*: $X \geq 0 \implies EX \geq 0$
   - *σ-additive*: $P(\sum_1^\infty A_i) = \sum_1^\infty P(A_i)$

   referred to as *probability measure*, while its value $P(A)$ at any $A \in \mathcal{F}$, is called the *probability* of that A.

Remark 3.1. σ-algebra is identical to σ-field.

Proposition 3.1 (nullity)**.**

$$P(\emptyset = 0)$$

Proof. First we show that $\Omega \in \mathcal{F}$.
If $F \neq \emptyset$, then $\exists A \in \mathcal{F}$ s.t. $A^c \in \mathcal{F}$.
So let $A_1 = A, A_n = A^c \; \forall n \geq 2$.
Then

$$\begin{aligned} \cup_1^\infty A_n &= A \cup A^c \cup A^c \cup \dots \\ &= A \cup A^c \\ &= \Omega \in \mathcal{F} \end{aligned}$$

Define the sequence of mutually disjoint events

$$A_1 = \Omega \quad \& \quad A_n = \emptyset, \quad n \geq 2$$

Then we have $\Omega = \sum_{n=1}^\infty A_n$ and thus

$$1 = 1 + \lim_{n\to\infty} nP(\emptyset)$$

which forces the result. ∎

Proposition 3.2 (finite-additivity)**.**

$$P(A + B) = P(A) + P(B)$$

Corollary 3.1 (complementarity)**.**

$$P(A^c) = 1 - P(A)$$

Corollary 3.2 (negative additivity)**.**

$$P(A - B) = P(A) - P(A \cap B)$$

Proof. Since $A = AB + AB^c = AB + (A - B)$, then $P(A) = P(AB) + P(A - B)$, hence the result. ∎

Corollary 3.3 (monotonicity)**.**

$$A \subset B \implies P(A) \leq P(B)$$

Proof. Since $B = (B - A) \cup A$, then $P(B) - P(A) = P(B - A) \geq 0$, hence the result. ■

Proposition 3.3. Assuming *normed, non-negative* and *σ-additive.* If $A_n \uparrow A$ or $A_n \downarrow A$, then

$$P(A_n) \to P(A)$$