A Primer in Econometric Theory

Lecture 3: Foundations of Probability

John Stachurski
Lectures by Akshay Shanker

March 26, 2017

Probability fundamental to statistics and econometrics, but technically demanding:

- set of events we want to assign probabilities to can be very large
- we need ways to manage the complexity

Before we begin:

- a set S is countable if it is finite or can be represented as sequence
- otherwise, a set S is uncountable

Sample Spaces and Events

Sample space can be thought of as a "list" of all possible outcomes in a given random experiment:

- sample space usually denoted by Ω
- sample space can be any non-empty set
- typical element of Ω is denoted ω

A realization of uncertainty will lead to the selection of a particular $\omega \in \Omega$

Example. In a random experiment that involves rolling a die once, the set of possible outcomes is naturally represented by $\Omega := \{1, \ldots, 6\}$

Example. Burton Malkiel's blindfolded monkey throws darts at a dartboard of radius 1

Impose ordinary Cartesian coordinates with origin at the centre of the board

Let (h, v) be a typical location measured on the horizontal and vertical coordinates respectively

A natural sample space is $\Omega := \{(h, v) \in \mathbb{R}^2 : \|(h, v)\| \leq 1\}$ – also called the **unit disk** in $\mathbb{R}^2$

Informally, an **event** is a subset of Ω (we will address some caveats soon)

An event A occurs whenever the individual $\omega \in \Omega$ selected in the random experiment happens to lie in A

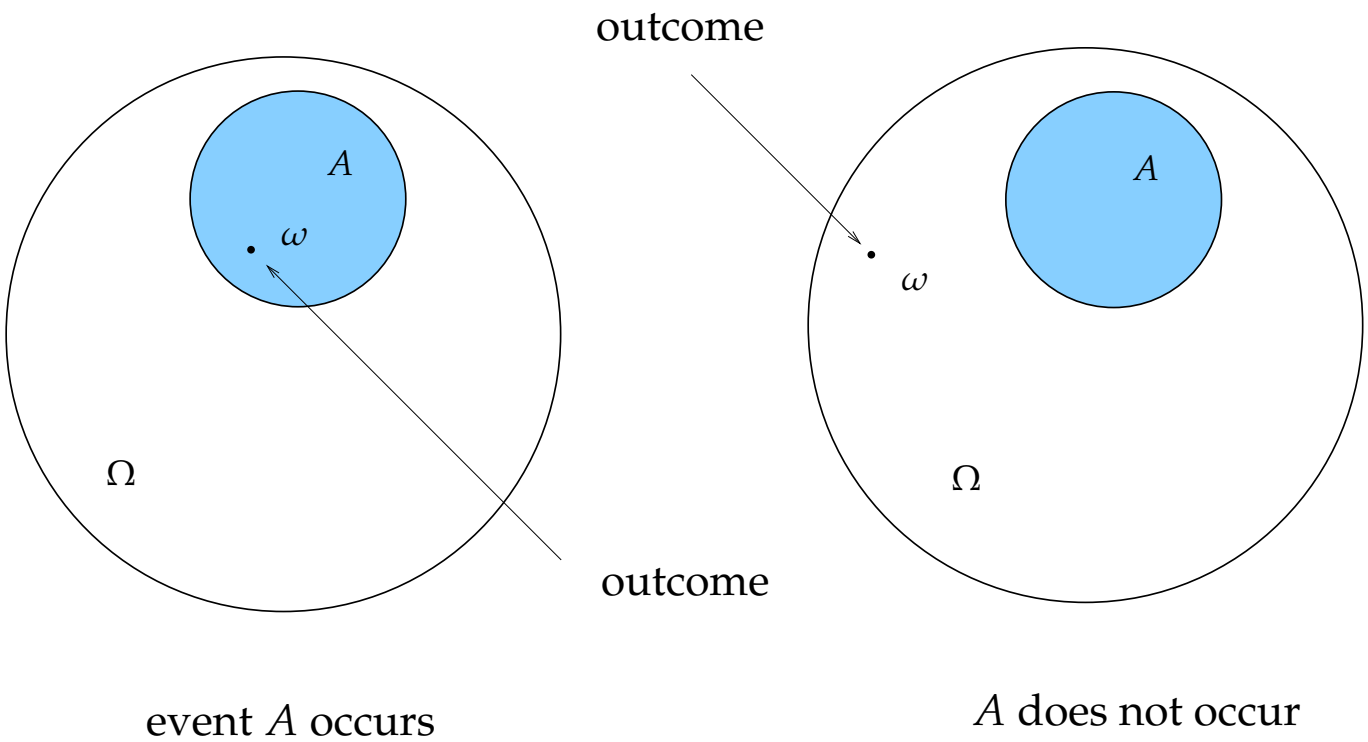

Figure: Outcomes and events

Probabilities and events

Can we assign probabilities to each $\omega \in \Omega$?

Consider dart throwing model where Ω is $\mathbb{R}$:

- for $A \subset \Omega$, probability dart lands in A is proportional to the area of A
- the probability of a point $\omega \in \Omega$ will be less than any area A containing ω
- for any $\epsilon > 0$, we can find an A, containing ω, with area smaller than ϵ

The probability of hitting ω is smaller than ϵ for any $\epsilon > 0$, thus the probability of hitting ω must be zero!

The upshot: when sample space is uncountable, assign probabilities to events (subsets of Ω), not to each $\omega \in \Omega$

But can we assign probabilities to *every* subset of Ω?

In the dart model:

$$\mathbb{P}(A) = \frac{\lambda(A)}{\pi}$$

where $\lambda(A) :=$ area of the set A

Defining area of A, for all $A \subset \Omega$, problematic:

- the space Ω, our dart-board unit disk in $\mathbb{R}^2$, contains many subsets which exhibit strange phenomena
- Banach-Tarski paradox

Solution: do not take the set of events to be all the subsets of Ω

Take the set of events to be certain "well-behaved" subsets of Ω, denoted by $\mathscr{F}$

Assign probabilities only to subsets of Ω in $\mathscr{F}$

Sigma-algebra

How can we ensure $\mathscr{F}$ is large enough? In a sensible probability model, we ideally want:

- the event "not A" to belong to $\mathscr{F}$ if $A \in \mathscr{F}$
- the event "A or B" to belong to $\mathscr{F}$ if $A \in \mathscr{F}$ and $A \in \mathscr{F}$

Formally, $\mathscr{F}$ is a **σ-algebra** on Ω if

1. $A \in \mathscr{F} \implies A^c \in \mathscr{F}$,
2. $A_1, A_2, \ldots \in \mathscr{F} \implies \cup_{n=1}^{\infty} A_n \in \mathscr{F}$, and
3. $\Omega \in \mathscr{F}$

1. - 3. imply that $\emptyset \in \mathscr{F}$ if $\mathscr{F}$ is a σ-algebra

The event $\emptyset$ is called the **impossible event**

The event Ω is called the **certain event**

Example. The set $\{\Omega, \emptyset\}$ is a σ - algebra called the **trivial σ-algebra**

The Borel σ -algebra

The σ-algebra of events varies from problem to problem

In $\mathbb{R}^N$, we use the Borel sets, denoted by $\mathscr{B}(\mathbb{R}^N)$

- the smallest σ-algebra that contains all the rectangles in $\mathbb{R}^N$

Why Borel σ-algebra?

- excludes "strange" sets
- includes day-to-day useful sets (including planes and hyperplanes, circles, spheres, polygons, finite sets, and sequences of points)

Probabilities

For given event $B \in \mathscr{F}$, the symbol $\mathbb{P}(B)$ represents "the probability that event B occurs."

> $\mathbb{P}(B)$ *represents the probability that when uncertainty is resolved and some* $\omega \in \Omega$ *is selected by "nature," the statement* $\omega \in B$ *is true*

We need to place restrictions to make probabilities well-behaved

For example, we want to rule out $\mathbb{P}(B) = -93$ for some B

Let Ω be a nonempty set and let $\mathscr{F}$ be a σ-algebra of subsets of Ω. A **probability** $\mathbb{P}$ on $(\Omega, \mathscr{F})$ is a function from $\mathscr{F}$ to $[0,1]$ that satisfies

1. $\mathbb{P}(\Omega) = 1$ and
2. $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ for any disjoint sequence of sets $A_1, A_2, \ldots \in \mathscr{F}$

$\mathbb{P}$ is also called a **probability measure**; the triple $(\Omega, \mathscr{F}, \mathbb{P})$ is called a **probability space**

Axiom 1.: we require $\mathbb{P}(\Omega) = 1$ because, by construction, every possible ω lies in the set Ω.

Axiom 2. is called **countable additivity**

Disjointness in the statement of axiom (ii) is pairwise: any distinct pair A_i, A_j share no points in common

Countable additivity implies finite **additivity**:

$$\mathbb{P}(A_1 \cup \cdots \cup A_k) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_k) \tag{1}$$

whenever $A_1, \ldots, A_k$ are disjoint

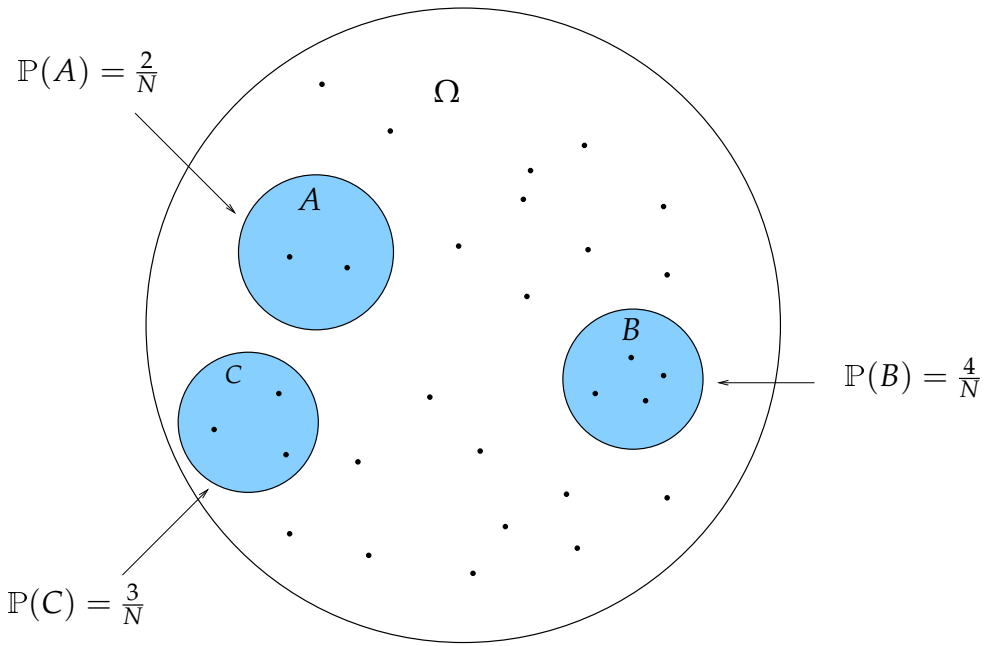

Figure: Each of the N dots occurs with probability $1/N$

$$\mathbb{P}(A \cup B \cup C) = \frac{9}{N} = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$$

Example. Let $\Omega := \{1, \ldots, 6\}$ represent the six different faces of a die, as in example 4.1.1

Since Ω is finite, take $\mathscr{F}$ to be the set of all subsets of Ω

Define a probability $\mathbb{P} \colon \mathscr{F} \to [0, 1]$

$$\mathbb{P}(A) := \frac{|A|}{6} \quad \text{where } |A| := \text{ number of elements in set } A \quad (2)$$

Easy to see $0 \leq \mathbb{P}(A) \leq 1$ for any $A \in \mathscr{F}$, and that $\mathbb{P}(\Omega) = 1$

Example. (cont.) Regarding additivity, suppose A and B are two disjoint subsets of $\{1, \ldots, 6\}$

Then $|A \cup B| = |A| + |B|$, hence

$$\mathbb{P}(A \cup B) = \frac{|A \cup B|}{6} = \frac{|A| + |B|}{6} = \frac{|A|}{6} + \frac{|B|}{6} = \mathbb{P}(A) + \mathbb{P}(B)$$

This proves additivity for pairs of sets. An analogous argument confirms additivity for any finite collection

Finite additivity is in this case equivalent to countable additivity, since the total number of distinct events is finite

Example

A memory chip is made up of billions of tiny switches/bits

- Switches can be off or on (zero or 1)

Random number generator accesses N bits, switching each one on or off

We take

- $\Omega := \{(b_1, \ldots, b_N) : \text{where } b_n \text{ is 0 or 1 for each } n\}$
- $\mathbb{P}(A) := 2^{-N}(\#A)$

Exercise: Show that $\mathbb{P}$ is a probability

Example. Consider again the dartboard model, where Ω is the unit disk in $\mathbb{R}^2$

For the event space, we take $\mathscr{F}$ to be the set of Borel subsets of $\mathbb{R}^2$ that lie inside Ω

For $\mathbb{P}$ we follow the "uniform" probability assignment given

That is, $\mathbb{P}(B) = \lambda(B)/\pi$ for every $B \in \mathscr{F}$

The function λ that assigns area to Borel sets is known to be countably additive, in that $\lambda(\cup_n A_n) = \sum_{n=1}^{\infty} \lambda(A_n)$ whenever these sets are disjoint

Evidently $\mathbb{P}(\Omega) = 1$

Lebesgue Measure

The function λ mapping Borel sets to their "area" is formally known as the **Lebesgue measure**

§15.3.1 in ET provides a brief introduction to this concept

Properties of Probability Measure

Fact. (4.1.1) Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, and let $A, B \in \mathscr{F}$

If $A \subset B$, then

1. $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$,
2. $\mathbb{P}(A) \leq \mathbb{P}(B)$ (**monotonicity**)
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, and
4. $\mathbb{P}(\emptyset) = 0$.

Proof. When $A \subset B$, we have $B = (B \setminus A) \cup A$ and hence

$$\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A)$$

All results follow (why?)

Fact. (4.1.2) If A and B are any (not necessarily disjoint) events, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Proof as exercise 4.4.2 in ET

Fact implies **subadditivity**: for any $A, B \in \mathscr{F}$, we have

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Conditional Probability and Independence

Conditional probability of A given B is

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3}$$

Probability of A, given information that B has occurred

Events A and B called **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

- If A and B independent, then

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

Example. Experiment: roll a dice twice

$$\Omega := \{(i,j) : i,j \in \{1,\ldots,6\}\} \quad \text{and} \quad \mathbb{P}(E) := \#E/36$$

Now consider the events

$$A := \{(i,j) \in \Omega : i \text{ is even}\} \quad \text{and} \quad B := \{(i,j) \in \Omega : j \text{ is even}\}$$

In this case we have

$$A \cap B = \{(i,j) \in \Omega : i \text{ and } j \text{ are even}\}$$

Exercise: Verify that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Hence, A and B are independent under the probability $\mathbb{P}$

Law of Total Probability

The **Law of total probability** states:

Fact. (4.1.3) If $A \in \mathscr{F}$ and $B_1, \ldots, B_M$ is a partition of Ω with $\mathbb{P}(B_m) > 0$ for all m, then

$$\mathbb{P}(A) = \sum_{m=1}^{M} \mathbb{P}(A \mid B_m) \cdot \mathbb{P}(B_m)$$

Proof. Given $A \in \mathscr{F}$ and partition $B_1, \ldots, B_M$:

$$\mathbb{P}(A) = \mathbb{P}[A \cap (\cup_{m=1}^{M} B_m)] = \mathbb{P}[\cup_{m=1}^{M} (A \cap B_m)]$$

$$= \sum_{m=1}^{M} \mathbb{P}(A \cap B_m) = \sum_{m=1}^{M} \mathbb{P}(A \mid B_m) \cdot \mathbb{P}(B_m)$$

Bayes' Law

Bayes law: for any events A and B with positive probability, we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{4}$$

Proof. From the definition of conditional probability:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{and} \quad \mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Hence $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)$

Rearranging yields (4)

Example. Banks use automated systems to detect fraudulent or illegal transactions

Consider test that responds to each transaction with P or N:

- P means "positive" (transaction flagged as fraudulent)
- N means "negative" (transaction flagged as normal)

Let F mean fraudulent, suppose

- $\mathbb{P}(P \mid F) = 0.99$ (the test flags 99% of fraudulent transactions),
- $\mathbb{P}(P \mid F^c) = 0.01$ (rate of false positives), and
- $\mathbb{P}(F) = 0.001$ (prevalence of fraud)

What is the probability of fraud given a positive test?

Note Bayes' law

$$\mathbb{P}(F \mid P) = \frac{\mathbb{P}(P \mid F)\mathbb{P}(F)}{\mathbb{P}(P)}$$

and note the law of total probability

$$\mathbb{P}(P) = \mathbb{P}(P \mid F)\mathbb{P}(F) + \mathbb{P}(P \mid F^c)\mathbb{P}(F^c)$$

Hence

$$\mathbb{P}(F \mid P) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.01 \times 0.999} = \frac{11}{122} \approx \frac{1}{11}$$

Random variables

Informally: A "value that changes randomly"

Formally: A **random variable** x is a function from Ω into $\mathbb{R}$

Interpretation: random variables convert outcomes in sample space into numerical outcomes

General idea:

- "nature" picks out ω in Ω
- random variable reports outcome as $x(\omega) \in \mathbb{R}$

Example. Suppose Ω is set of infinite binary sequences

$$\Omega := \{(b_1, b_2, \ldots) : b_n \in \{0, 1\} \text{ for each } n\}$$

We can create different random variables mapping $\Omega \to \mathbb{R}$:

- number of "flips" till first "heads":

$$x(\omega) = x(b_1, b_2, \ldots) = \min\{n : b_n = 1\}$$

- number of "heads" in first 10 "flips":

$$x(\omega) = x(b_1, b_2, \ldots) = \sum_{n=1}^{10} b_n$$

- number of flips until the first heads:

$$x(\omega) = x(b_1, b_2, \ldots) = \min\{n \in \mathbb{N} : b_n = 1\}$$

- **Binary or Bernoulli random variable** telling us whether any heads occur in the first 10 flips:

$$x(\omega) = y(b_1, b_2, \ldots) := \min\left\{\sum_{n=1}^{10} b_n, 1\right\} \tag{5}$$

Bernoulli Random Variable

Bernoulli or **binaary random variable** RVs x take on values in $\{0, 1\}$

We now a consider generic way to create a Bernoulli RV

Let Q be a statement, such as "a is greater than 3"

Definition: $\mathbb{1}\{Q\}$ equals one if Q true, zero otherwise

Define

$$x(\omega) = \mathbb{1}\{\omega \in A\} \quad \text{where } A \in \mathscr{F}$$

The RV indicates whether or not event C occurs

A common variation on the notation: for arbitrary $A \in \mathscr{F}$:

$$\mathbb{1}_A(\omega) := \mathbb{1}\{\omega \in A\} := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Fact. (4.1.4) If $A_1, \ldots, A_N$ are subsets of Ω, then

1. $\mathbb{1}_{\cap_{n=1}^N A_n} = \prod_{n=1}^N \mathbb{1}_{A_n}$ and
2. $\mathbb{1}_{\cup_{n=1}^N A_n} = \sum_{n=1}^N \mathbb{1}_{A_n}$ whenever the sets are disjoint

See exercise 4.4.5 for proof

Here, equality means evaluated at any $\omega \in \Omega$

Notational Conventions

Common notational convention with RVs:

$$\{x \text{ has some property}\} := \{\omega \in \Omega : x(\omega) \text{ has some property}\}$$

Example.

$$\{x \leq 2\} := \{\omega \in \Omega : x(\omega) \leq 2\}$$

$$\therefore \quad \mathbb{P}\{x \leq 2\} := \mathbb{P}\{\omega \in \Omega : x(\omega) \leq 2\}$$

Example. Given random variable x and $a \leq b$, we claim

$$\mathbb{P}\{x \leq a\} \leq \mathbb{P}\{x \leq b\}$$

This holds because

$$\{x \leq a\} := \{\omega \in \Omega : x(\omega) \leq a\}$$

$$\subset \{\omega \in \Omega : x(\omega) \leq b\} := \{x \leq b\}$$

Now apply monotonicity: $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$

Equalities, inequalities and arithmetic operations should be interpreted *pointwise*:

- $x \leq y \iff x(\omega) \leq y(\omega)$ for all $\omega \in \Omega$,
- $x = y \iff x(\omega) = y(\omega)$ for all $\omega \in \Omega$, and
- $z = \alpha x + \beta y \iff z(\omega) = \alpha x(\omega) + \beta y(\omega)$ for all $\omega \in \Omega$

Random Variables are Measurable Functions

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be any probability space

Let B be any subset of $\mathbb{R}$

Consider the probability

$$\mathbb{P}\{x \in B\} := \mathbb{P}\{\omega \in \Omega : x(\omega) \in B\}$$

Where x is some function from Ω to $\mathbb{R}$

No way of being sure $\{\omega \in \Omega : x(\omega) \in B\}$ is an element of $\mathscr{F}$

- $\mathbb{P}\{x \in B\}$ may not be defined

We need to place restrictions:

- For B, we naturally restrict attention to $\mathscr{B}(\mathbb{R})$, the Borel subsets of $\mathbb{R}$
- For x, we require $\{x \in B\} \in \mathscr{F}$ whenever B is a Borel set

Formal definition of a random variable:

A **random variable** on $(\Omega, \mathscr{F})$ is a function $x\colon \Omega \to \mathbb{R}$ satisfying

$$\{\omega \in \Omega : x(\omega) \in B\} \in \mathscr{F} \quad \text{for all } B \in \mathscr{B}(\mathbb{R}) \tag{6}$$

These kinds of functions are also called $\mathscr{F}$-measurable functions

Pre-image notation: $x^{-1}(B)$ is all $\omega \in \Omega$ such that $x(\omega) \in B$

Rewrite (6) as

$$x^{-1}(B) \in \mathscr{F} \quad \text{for all } B \in \mathscr{B}(\mathbb{R})$$

Thus x "pulls back" Borel sets to events

Measurable Transformations

We want to discuss some transformation of x

For example, $y := e^x$. Is y also a random variable?

Yes, provided transformation satisfies Borel measurability

Formally, $f : \mathbb{R} \to \mathbb{R}$ is called **Borel measurable**, or **$\mathscr{B}$-measurable**, if

$$f^{-1}(B) \in \mathscr{B}(\mathbb{R}) \quad \text{for all } B \in \mathscr{B}(\mathbb{R}) \tag{7}$$

Class of $\mathscr{B}$-measurable functions is vast: any continuous function, any increasing function, etc., etc.

Suppose f is $\mathscr{B}$-measurable and that x is a random variable

We have $\{y \in B\} \in \mathscr{F}$ for all $B \in \mathscr{B}(\mathbb{R})$ because

$$\{y \in B\} = \{f(x) \in B\} = \{x \in f^{-1}(B)\} \tag{8}$$

Thus $y = f(x)$ is a random variable

Expectations

We want to define expectations for an arbitrary RV x

Roughly speaking, $\mathbb{E}[x] :=$ the "sum" of all possible values of x, weighted by their probabilities.

"Sum" in quotes because may be over an infinite number of possibilities

We take a modern, formal and rigorous approach to defining expectations

For finite random variables, given probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and random variable x taking only finitely many distinct values $s_1, \ldots, s_J$, the **expectation** of x is defined as

$$\mathbb{E}\, x = \sum_{j=1}^{J} s_j \, \mathbb{P}\{x = s_j\} \tag{9}$$

Example. Let's apply this definition to the simplest possible case, which is a random variable x satisfying $x(\omega) = \alpha$ for all $\omega \in \Omega$, where α is some constant scalar value. In this case the sum in (9) has only one term, and

$$\mathbb{E}\, x = \alpha \mathbb{P}\{x = \alpha\} = \alpha \mathbb{P}\{\omega \in \Omega : x(\omega) = \alpha\} = \alpha \mathbb{P}(\Omega) = \alpha$$

Example. To evaluate the expectation of a binary random variable x, we apply (9) to obtain

$$\mathbb{E}\, x = 1 \times \mathbb{P}\{x = 1\} + 0 \times \mathbb{P}\{x = 0\} = \mathbb{P}\{x = 1\}$$

Example. Consider N flips of a fair coin

The sample space is $\Omega := \{0,1\}^N$, the events are $\mathscr{F} :=$ all subsets of Ω, and $\mathbb{P}(A) := 2^{-N}|A|$ for all $A \in \mathscr{F}$

Let $x(\omega) = x(b_1, \ldots, b_N) = \sum_{n=1}^{N} b_n$

Observe first that $0 \leq x \leq N$

By the definition of $\mathbb{P}$, for any k we have $\mathbb{P}\{x = k\} = 2^{-N}|A_k|$, where

$$A_k := \{x = k\} = \left\{ (b_1, \ldots, b_N) \in \Omega \ : \ \sum_{n=1}^{N} b_n = k \right\}$$

From combinatorics, $|A_k| = \binom{N}{k}$, where the right-hand side is the so-called **binomial coefficient** for N, k, which satisfies $\sum_{k=0}^{N} k\binom{N}{k} = N2^{N-1}$ for all N

The expectation of x is

$$\mathbb{E}x = \sum_{k=0}^{N} k\, 2^{-N}|A_k| = 2^{-N} \sum_{k=0}^{N} k \binom{N}{k} = \frac{N}{2}$$

For general x, approximate arbitrary random variables with finite random variables

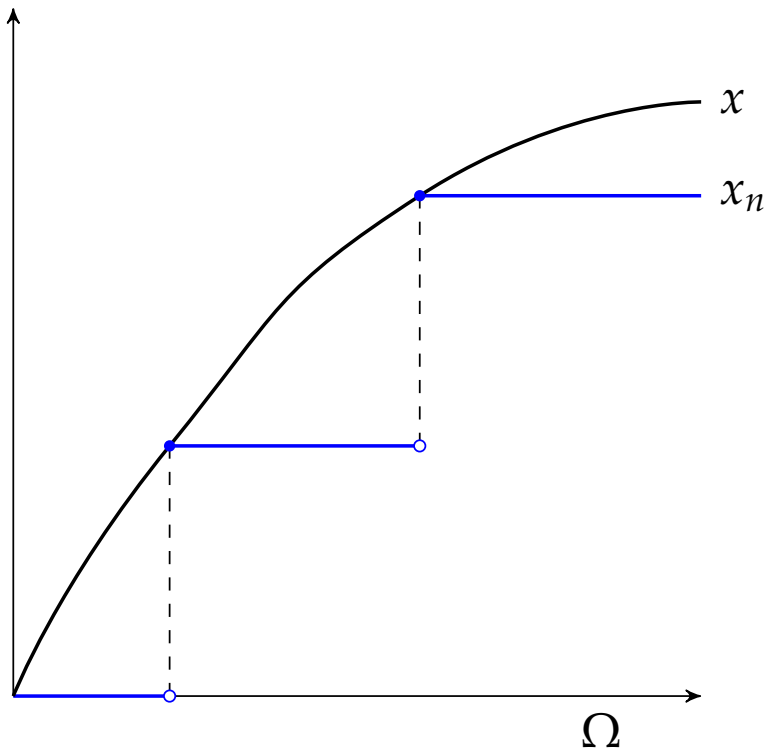

Figure: Finite approximation to a general random variable

We can improve approximation without limit– let x_n take a larger and larger number of distinct values

Process gives sequence of *finite* random variables x_n converging to x

Define the expectation of x as

$$\mathbb{E}\, x := \lim_{n \to \infty} \mathbb{E}\, x_n$$

$\mathbb{E}x$ is also referred to as the **Lebesgue integral** of x with respect to $\mathbb{P}$, with the alternative notation $\mathbb{E}x = \int x(\omega)\mathbb{P}(\mathrm{d}\omega)$

Does a sequence of approximating random variables exist? Yes, See page 94 in ET and Dudley (2002), proposition 4.1.5

If x takes on negative values, then write $x = x^+ - x^{-1}$

Where $x^+ := \max\{x, 0\}$ and $x^- = \min\{x, 0\}$

Define expectation as

$$\mathbb{E}\, x := \mathbb{E}\, x^+ - \mathbb{E}\, x^-$$

Restrict attention to **integrable** random variables: all random variables x such that $\mathbb{E}\,|x| < \infty$

- we have $x^+ \leq |x|$ and $x^- \leq |x|$
- thus, $\mathbb{E}\, x := \mathbb{E}\, x^+ - \mathbb{E}\, x^-$ well-defined (why?)

Properties of Expectation

Fact. (4.1.5) Given any probability space $(\Omega, \mathscr{F}, \mathbb{P})$, there exists a uniquely defined function $\mathbb{E}$ that maps each integrable random variable x on $(\Omega, \mathscr{F}, \mathbb{P})$ into a value

$$\mathbb{E}x = \int x(\omega)\mathbb{P}(\mathrm{d}\omega) \tag{10}$$

in $\mathbb{R}$, called the **expectation of** x under $\mathbb{P}$. The function has the following properties:

1. $\mathbb{E}\alpha = \alpha$ for all $\alpha \in \mathbb{R}$
2. $\mathbb{E}\,\mathbb{1}_A = \mathbb{P}(A)$ for all $A \in \mathscr{F}$
3. $x \leq y \implies \mathbb{E}x \leq \mathbb{E}y$
4. $\mathbb{E}[\alpha x + \beta y] = \alpha\mathbb{E}x + \beta\mathbb{E}y$ for all integrable x, y and constants α, β

To remind ourselves of the underlying probability measure $\mathbb{P}$ we may write $\mathbb{E}_{\mathbb{P}} x$ instead of $\mathbb{E} x$

Note the expression $\mathbb{E} \alpha$ understood as the expectation of a constant random variable equal to α

- Follows from 4. by letting $x = \mathbb{1}_\Omega$ and $\beta = 0$

Exercise: Check 3. for $x(\omega) := \mathbb{1}\{\omega \in A\}$ and $y(\omega) := \mathbb{1}\{\omega \in B\}$

Hint: What does $x \leq y$ imply about A and B?

For further details and references of proofs of the above fact, see page 96 in ET

We now prove that if x is a finite random variable with range $\{s_j\}_{j=1}^J$ and h is any $\mathscr{B}$-measurable function,then

$$\mathbb{E}h(x) = \sum_{j=1}^{J} h(s_j)\mathbb{P}\{x = s_j\} \tag{11}$$

First observe that $\sum_{j=1}^J \mathbb{1}\{x = s_j\} = 1$, and hence we can write $h(x)$ as

$$h(x) = h(x)\sum_{j=1}^{J} \mathbb{1}\{x = s_j\} = \sum_{j=1}^{J} h(s_j)\mathbb{1}\{x = s_j\}$$

Using linearity of expectations:

$$\mathbb{E}\, h(x) = \sum_{j=1}^{J} h(s_j)\mathbb{E}\, \mathbb{1}\{x = s_j\}$$

Applying part 2. of fact 4.1.5 leads to (11)

Chebyshev's inequality: **Fact.** (4.1.6) For any nonnegative random variable x and any $\delta > 0$, we have

$$\mathbb{P}\{x \geq \delta\} \leq \frac{\mathbb{E}\, x}{\delta} \tag{12}$$

A common variation of Chebyshev's inequality is the bound

$$\mathbb{P}\{|x| \geq \delta\} \leq \frac{\mathbb{E}\, x^2}{\delta^2} \tag{13}$$

See Exercise 4.4.29 for proof

Moments and Co-Moments

Let x be a random variable and let $k \in \mathbb{N}$. If x^k is integrable, then

- $\mathbb{E}[x^k]$ is called the **kth moment** of x
- $\mathbb{E}[(x - \mathbb{E}x)^k]$ is called the **kth central moment** of x

If $\mathbb{E}[|x|^k] = \infty$, then the kth moment is said not to exist. For some random variables even the first moment does not exist

For others, every moment exists

Fact. (4.1.7) If the kth moment of x exists, then so does the jth for all $j \leq k$

Proof: Exercise 4.4.24

The **Cauchy–Schwarz inequality for random variables**:

Fact. (4.1.8) If x and y are random variables with finite second moment, then

$$|\,\mathbb{E}\,[xy]\,| \leq \sqrt{\mathbb{E}\,[x^2]\mathbb{E}\,[y^2]} \tag{14}$$

The second central moment of x is called the **variance** of x:

$$\operatorname{var} x := \mathbb{E}\left[(x - \mathbb{E}\, x)^2\right]$$

The **standard deviation** of x:

$$\sigma_x := \sqrt{\operatorname{var} x}$$

The **covariance** of random variables x and y:

$$\operatorname{cov}[x, y] := \mathbb{E}\left[(x - \mathbb{E}\, x)(y - \mathbb{E}\, y)\right]$$

Fact. (4.1.9) If x and y have finite second moments, then

1. $\operatorname{var} x$ and $\operatorname{cov}[x, y]$ are finite
2. $\operatorname{var} x = \mathbb{E}[x^2] - [\mathbb{E}\, x]^2$, and
3. $\operatorname{cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$

Part 1. follows from 2.–3., the Cauchy–Schwarz inequality and fact 4.1.7

Parts 2.–3. follow from linearity of $\mathbb{E}$ and some simple manipulations

Fact. (4.1.10) If $x_1, \ldots, x_N$ are random variables and $\alpha_0, \alpha_1, \ldots, \alpha_N$ are constant scalars, then

$$\operatorname{var}\left[\alpha_0 + \sum_{n=1}^{N} \alpha_n x_n\right] = \sum_{n=1}^{N} \alpha_n^2 \operatorname{var}[x_n] + 2 \sum_{n<m} \alpha_n \alpha_m \operatorname{cov}[x_n, x_m]$$

Some simple implications:

1. $\operatorname{var}[\alpha + \beta x] = \beta^2 \operatorname{var}[x]$ and
2. $\operatorname{var}[\alpha x + \beta y] = \alpha^2 \operatorname{var}[x] + \beta^2 \operatorname{var}[y] + 2\alpha\beta \operatorname{cov}[x, y]$.

The **correlation** of x and y:

$$\operatorname{corr}[x,y] := \frac{\operatorname{cov}[x,y]}{\sigma_x \, \sigma_y}$$

If $\operatorname{corr}[x,y] = 0$, x and y are **uncorrelated**

Positive correlation means that $\operatorname{corr}[x,y]$ is positive and negative correlation means that $\operatorname{corr}[x,y]$ is negative

Fact. (4.1.11) Given any two random variables x, y and positive constants α, β, we have

$$-1 \leq \operatorname{corr}[x,y] \leq 1 \quad \text{and} \quad \operatorname{corr}[\alpha x, \beta y] = \operatorname{corr}[x,y]$$

The first part follows from fact 4.1.8; the second is algebra

Best Linear Predictors

Consider the problem of predicting the value of a random variable y given knowledge of the value of a second random variable x

We seek a function f such that $f(x)$ is close to y on average

To measure the latter, we will use **mean squared error**, which amounts in this case to

$$\mathbb{E}\left[(y - f(x))^2\right]$$

In§5.2.5, to obtain minimizer of the mean squared deviation over all functions of x, we choose

$$f(x) = \mathbb{E}[y \mid x]$$

Here we'll consider finding a good predictor of y within the class of "linear" functions

$$\mathcal{H}_\ell := \{ \text{ all functions of the form } \ell(x) = \alpha + \beta x \}$$

Consider:

$$\min_{\ell \in \mathcal{H}_\ell} \mathbb{E}\left[(y - \ell(x))^2\right] = \min_{\alpha,\beta \in \mathbb{R}} \mathbb{E}\left[(y - \alpha - \beta x)^2\right] \tag{15}$$

If α and β solve (15), then the function

$$\ell^*(x) := \alpha^* + \beta^* x \tag{16}$$

is called the **best linear predictor** of y given x

Example. (4.1.18) Relationship between returns on a given asset R_a and returns on a market benchmark R_m is called the **beta** of the asset

Measures exposure to systemic risk, as opposed to idiosyncratic risk specific to the asset

The beta of R_a is often defined as the coefficient β^* in the best linear prediction (16) when x is returns on the market benchmark and $y = R_a$

To solve (15), expand the square on the right-hand side and use linearity of $\mathbb{E}$ to write the objective function as

$$\psi(\alpha, \beta) := \mathbb{E}[y^2] - 2\alpha\mathbb{E}[y] - 2\beta\mathbb{E}[xy] + 2\alpha\beta\mathbb{E}[x] + \alpha^2 + \beta^2\mathbb{E}[x^2]$$

Computing the derivatives and solving the first-order conditions:

$$\beta^* := \frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}[y] - \beta^*\mathbb{E}[x] \tag{17}$$

See ex. 4.4.27

Distributions

Take a random variable x on probability space $(\Omega, \mathscr{F}, \mathbb{P})$

Probability of x taking on value in Borel set B

$$\mathbb{P}\{x \in B\}$$

In practice, more convenient to represent the probability as a *distribution* over $\mathbb{R}$

Specialise Ω to $\mathbb{R}$ and take the set of events over $\mathbb{R}$ as $\mathscr{B}(\mathbb{R})$

- A probability measure defined over $\mathscr{B}(\mathbb{R})$ is called a **law** or **distribution**

Formally, a distribution P is a map from $\mathscr{B}(\mathbb{R})$ to $[0,1]$ such that

1. $P(\mathbb{R}) = 1$ and
2. $P(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n)$ for any disjoint sequence $\{B_n\}$

If there exists a Borel set S with $P(S) = 1$, then we say that P is **supported** on S

Characterise distributions using a **cumulative distributed function**, or CDF, which is any function $F\colon \mathbb{R} \to [0, 1]$ satisfying

1. monotonicity: $s \leq s'$ implies $F(s) \leq F(s')$,
2. right-continuity: $F(s_n) \downarrow F(s)$ whenever $s_n \downarrow s$, and
3. $\lim_{s \to -\infty} F(s) = 0$ and $\lim_{s \to \infty} F(s) = 1$.

CDFs and distributions on $\mathbb{R}$ can be put in one-to-one correspondence

A distribution P is entirely characterized by the values of the function

$$F(s) := P((-\infty, s]) \qquad (s \in \mathbb{R}) \tag{18}$$

Fact. (4.2.1) The following statements are true:

1. If P is any distribution on $\mathbb{R}$, then the function F in (18) is a CDF.
2. Given any CDF F on $\mathbb{R}$, there exists exactly one distribution P satisfying (18)

For a full proof, see Williams (1991), lemma 1.6, or Dudley (2002), theorem 9.1.1.

Here let's restrict ourselves to showing the function F in (18) satisfies part 1. of the definition of a CDF

- observe $s \leq s'$ implies $(-\infty, s] \subset (-\infty, s']$
- recall $P(A) \leq P(B)$ if $A \subset B$
- we then have $P((-\infty, s]) \leq P((-\infty, s'])$ and $F(s) \leq F(s')$ as claimed

Example. The **univariate normal distributions** or **Gaussian distributions** refer to the class of distributions identified by CDFs of the form

$$F(s) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{s} \exp\left\{ -\frac{(t-\mu)^2}{2\sigma^2} \right\} \mathrm{d}t \qquad (s \in \mathbb{R})$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$

We represent the distribution associated with (μ, σ) by $\mathrm{N}(\mu, \sigma^2)$

The distribution $\mathrm{N}(0,1)$ is called the **standard normal distribution**

We use the symbol Φ for its CDF

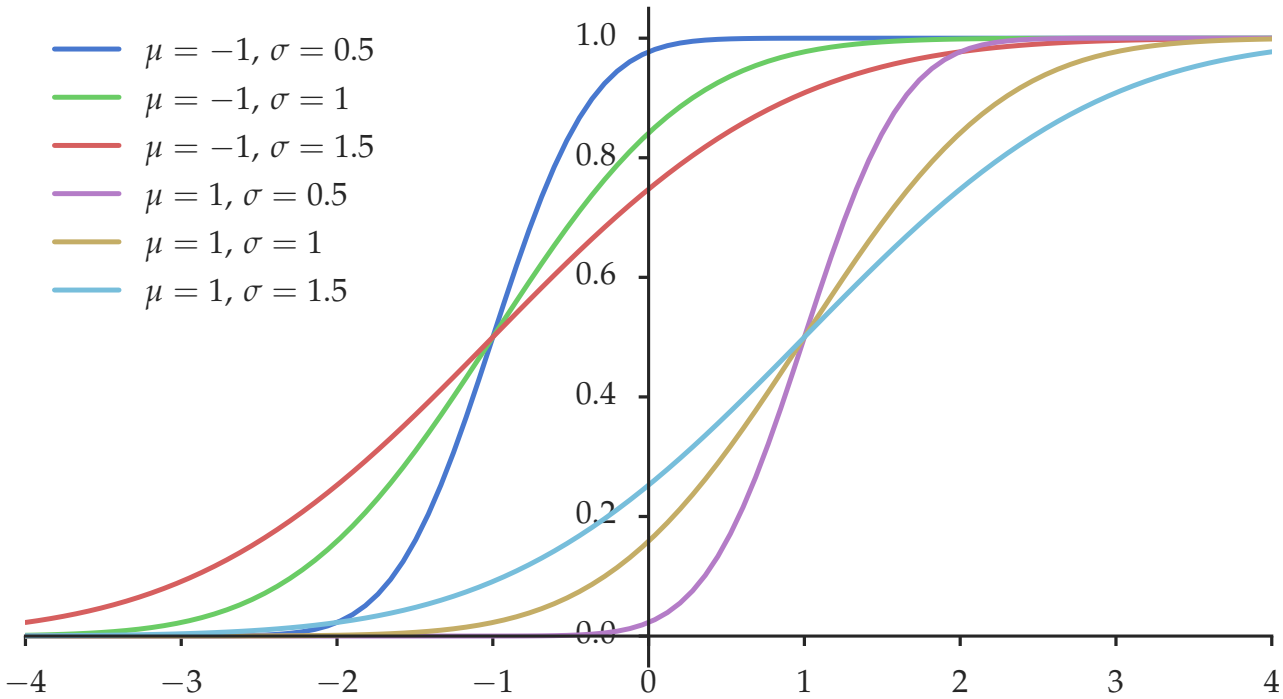

Figure: Normal CDFs

Example. The **Pareto distributions** are the univariate distributions with CDFs of the form

$$F(s) = \begin{cases} 0 & \text{if } s < s_0 \\ 1 - \left(\frac{s_0}{s}\right)^{\alpha} & \text{if } s_0 \leq s \end{cases} \qquad (s \in \mathbb{R},\ s_0,\ \alpha > 0)$$

Pareto distributions often used to model phenomena with heavy right-hand tails, such as the distribution of wealth or income

Example. The class of **beta** CDFs is given by

$$F(s) = \begin{cases} 0 & \text{if } s \leq 0 \\ \frac{1}{B(\alpha,\beta)} \int_0^s u^{\alpha-1}(1-u)^{\beta-1}\,\mathrm{d}u & \text{if } 0 < s < 1 \\ 1 & \text{if } 1 \leq s \end{cases}$$

where $\alpha, \beta > 0$.

In this example $B(\alpha, \beta)$ is the **beta function**

$$B(\alpha,\beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \text{where} \qquad \Gamma(a) := \int_0^\infty u^{a-1}e^{-u}\,\mathrm{d}u$$

The function Γ is called the **gamma function**.

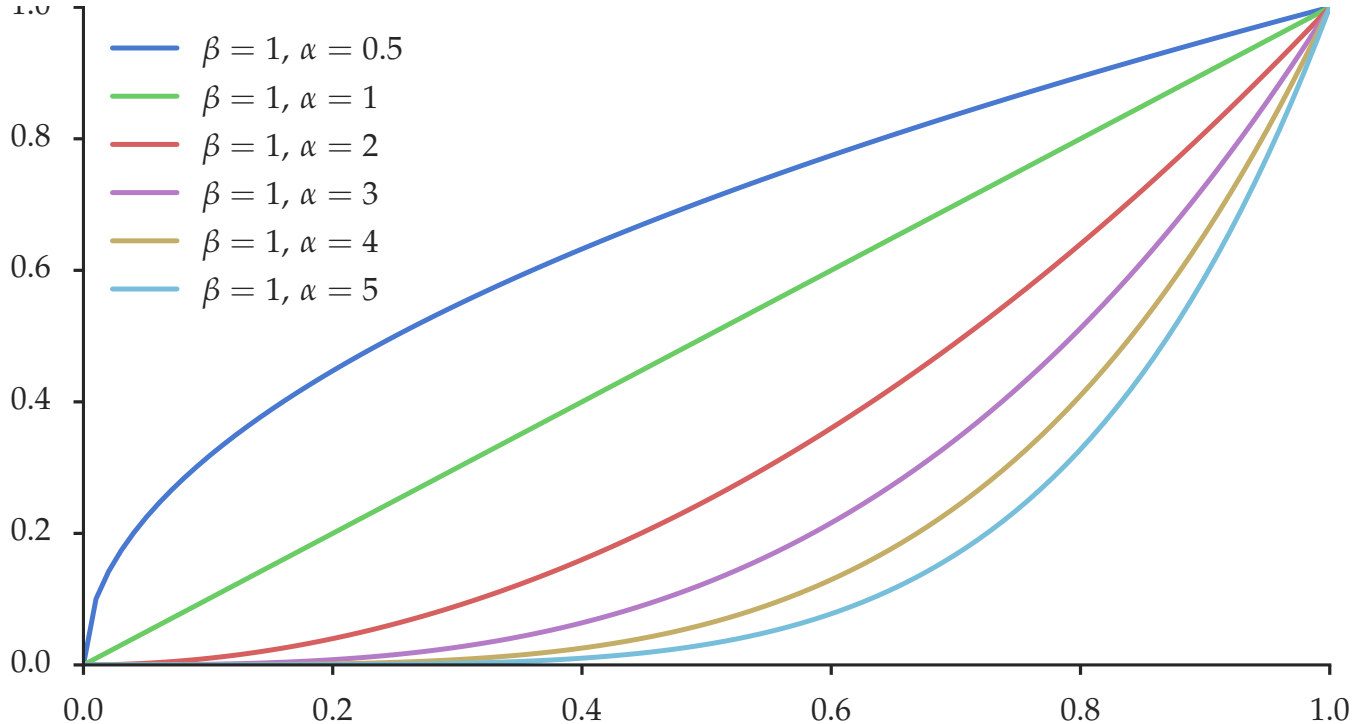

Figure: Beta CDFs

Example. The class of **Cauchy** CDFs is given by

$$F(s) = \frac{1}{\pi} \arctan\left(\frac{s-\tau}{\gamma}\right) + \frac{1}{2} \qquad (s \in \mathbb{R})$$

The parameters $\tau \in \mathbb{R}$ and $\gamma > 0$ are the location and scale parameters respectively

If $\tau = 0$ and $\gamma = 1$, then F is called **standard Cauchy**

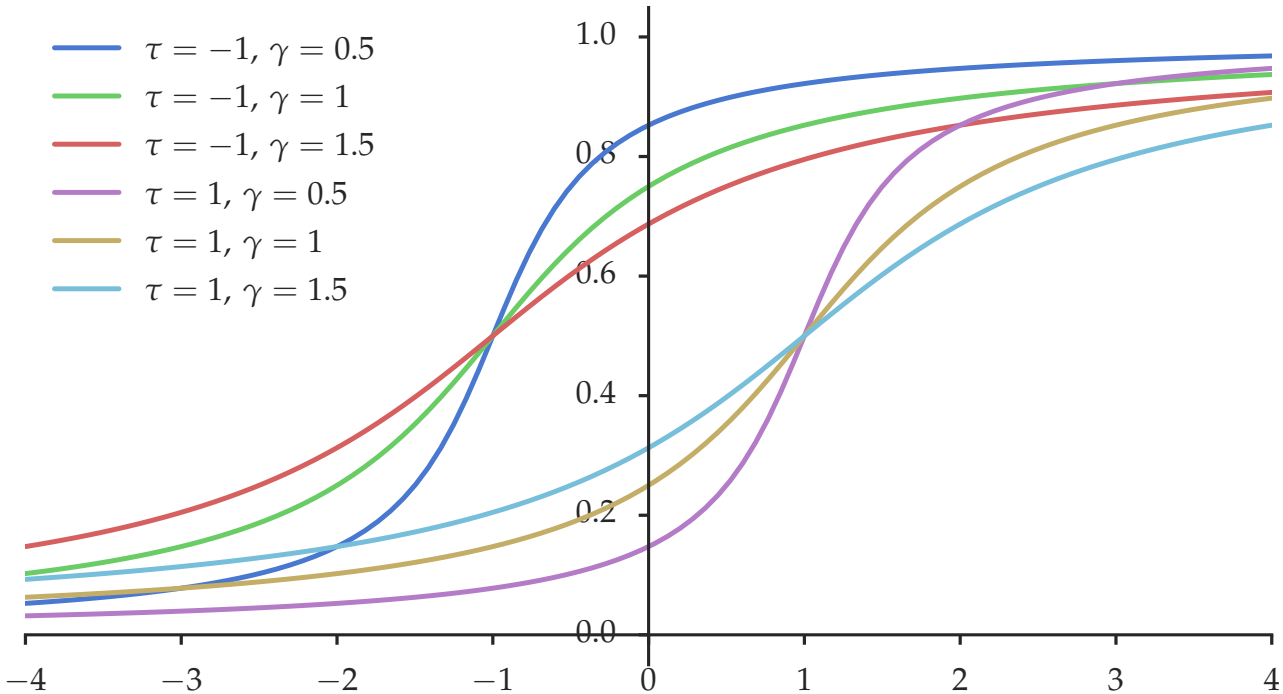

Figure: Cauchy CDFs

Example. Given $a < b$, the **uniform** CDF on $[a, b]$ is the CDF

$$F(s) = \begin{cases} 0 & \text{if } s \leq a \\ \frac{s-a}{b-a} & \text{if } a < s < b \\ 1 & \text{if } b \leq s \end{cases}$$

We represent this distribution symbolically by $U[a, b]$

Densities and Probability Mass Functions

Two convenient special cases

- discrete: CDF is just jumps (a step function)
- absolutley continuous case: CDF smooth with no jumps

Discrete Case

A distribution P is called **discrete** if it is supported on a countable set; that is, if there exists a countable set $\{s_j\}_{j\geq 1}$ with $P(\{s_j\}_{j\geq 1}) = 1$

For such a P let

$$p_j := P\{s_j\} := P(\{s_j\}) = \text{probability mass on the single point } s_j$$

A **probability mass function**, or PMF is any non-negative sequence (finite or infinite) that sums to unity

Exercise: show $\{p_j\}_{j\geq 1}$ is a **probability mass function**

We can express the CDF corresponding to P as:

$$F(s) = \sum_{j \geq 1} \mathbb{1}\{s_j \leq s\} p_j \tag{19}$$

because

$$F_x(s) := \mathbb{P}\{x \leq s\} = \mathbb{P} \bigcup_{j \text{ s.t. } s_j \leq s} \{x = s_j\}$$

$$= \sum_{j \text{ s.t. } s_j \leq s} \mathbb{P}\{x = s_j\} = \sum_{j=1}^{J} \mathbb{1}\{s_j \leq s\} p_j$$

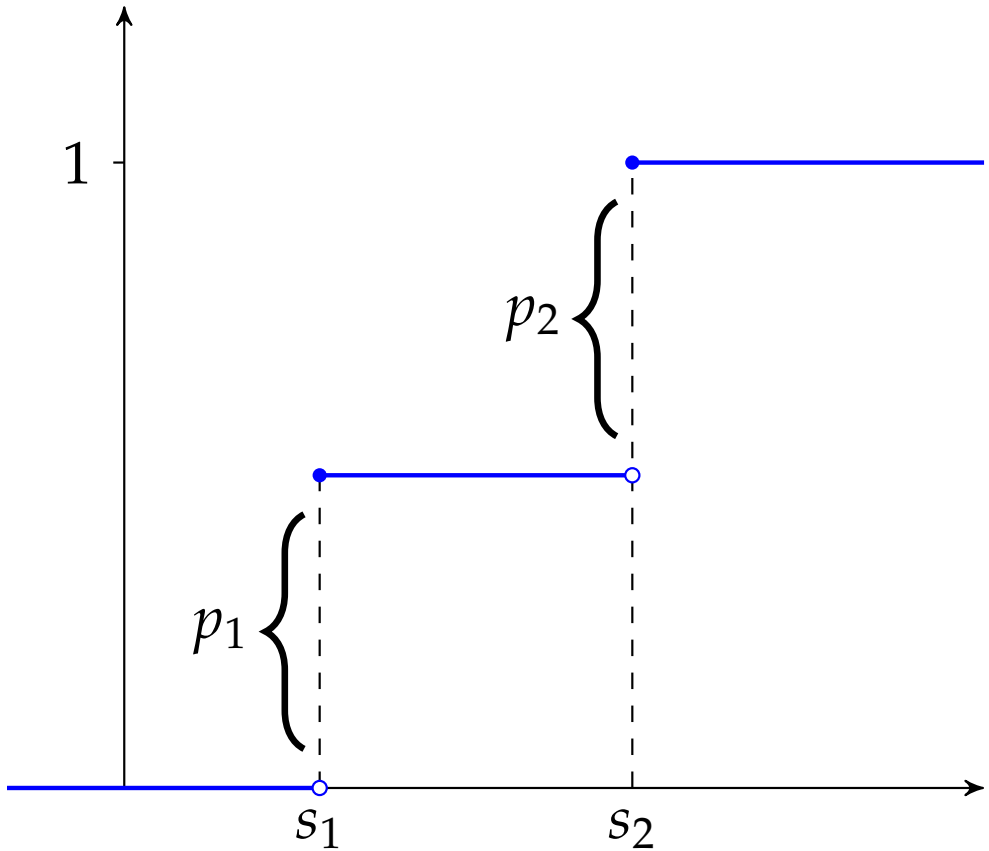

Figure: Discrete CDF

Example. Given $N \in \mathbb{N}$ and $\pi \in (0,1)$, the sequence $\{p_0, \ldots, p_N\}$ defined by

$$p_j = \binom{N}{j} \pi^j (1-\pi)^{N-j}$$

is called the **binomial** PMF

The value p_j is probability of j successes in N independent trials, each having success probability π

Absolutely Continuous Case

A **density** is a nonnegative function p on $\mathbb{R}$ that integrates to 1

A distribution P is **represented by density** p (or **has density** p) if p is a density and

$$P(B) = \int_B p(s)\,\mathrm{d}s \qquad \text{for all } B \in \mathscr{B}(\mathbb{R})$$

Note:

$$\int_B p(s)\,\mathrm{d}s := \int_{-\infty}^{\infty} \mathbb{1}_B(s)p(s)\,\mathrm{d}s$$

An exact necessary and sufficient condition for existence of density representation is absolute continuity

A distribution P on the Borel subsets of $\mathbb{R}$ is called **absolutely continuous** if $P(B) = 0$ whenever B has Lebesgue measure zero (see §15.3.1)

- Any countable subset of $\mathbb{R}$ has Lebesgue measure zero

Fact. (4.2.2) If P is absolutely continuous, then $P(C) = 0$ whenever C is countable

If distribution is absolutely continuous:

- individual points receive no probability mass
- corresponding CDF contains no jumps
- fundamental theorem of calculus says $F(s)$ differentiable at all continuity points of p, and:

$$F'(s) = p(s) \qquad \text{for all } s \in \mathbb{R} \text{ such that } p \text{ is continuous at } s$$

Example. Normal CDFs are differentiable for all μ, σ, with density

$$p(s) = F'(s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\}$$

We reserve the symbol ϕ for the standard normal density

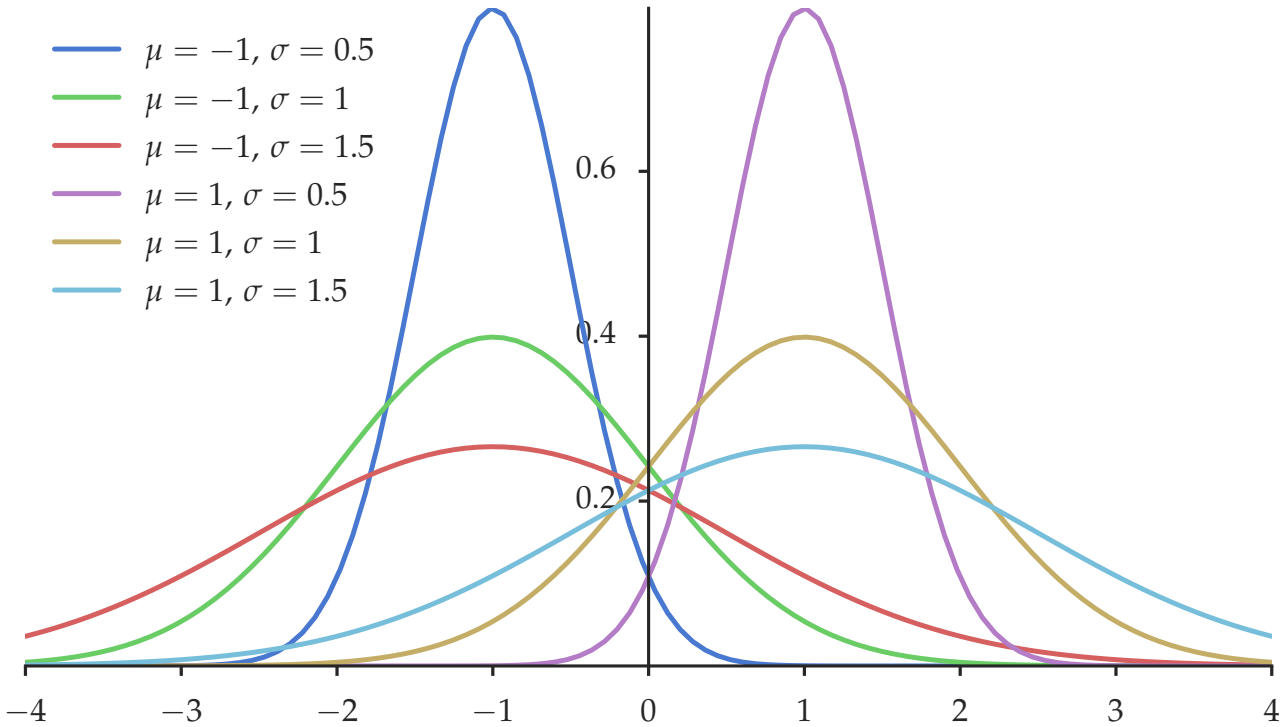

Figure: Normal densities

Example. The Cauchy CDF has density

$$p(s) = \frac{1}{\pi\gamma}\left[1 + \left(\frac{s-\tau}{\gamma}\right)^2\right]^{-1} \qquad (s \in \mathbb{R},\ \gamma > 0,\ \tau \in \mathbb{R})$$

The Cauchy densities are more peaked around their modes and have greater mass in their tails than normal densities

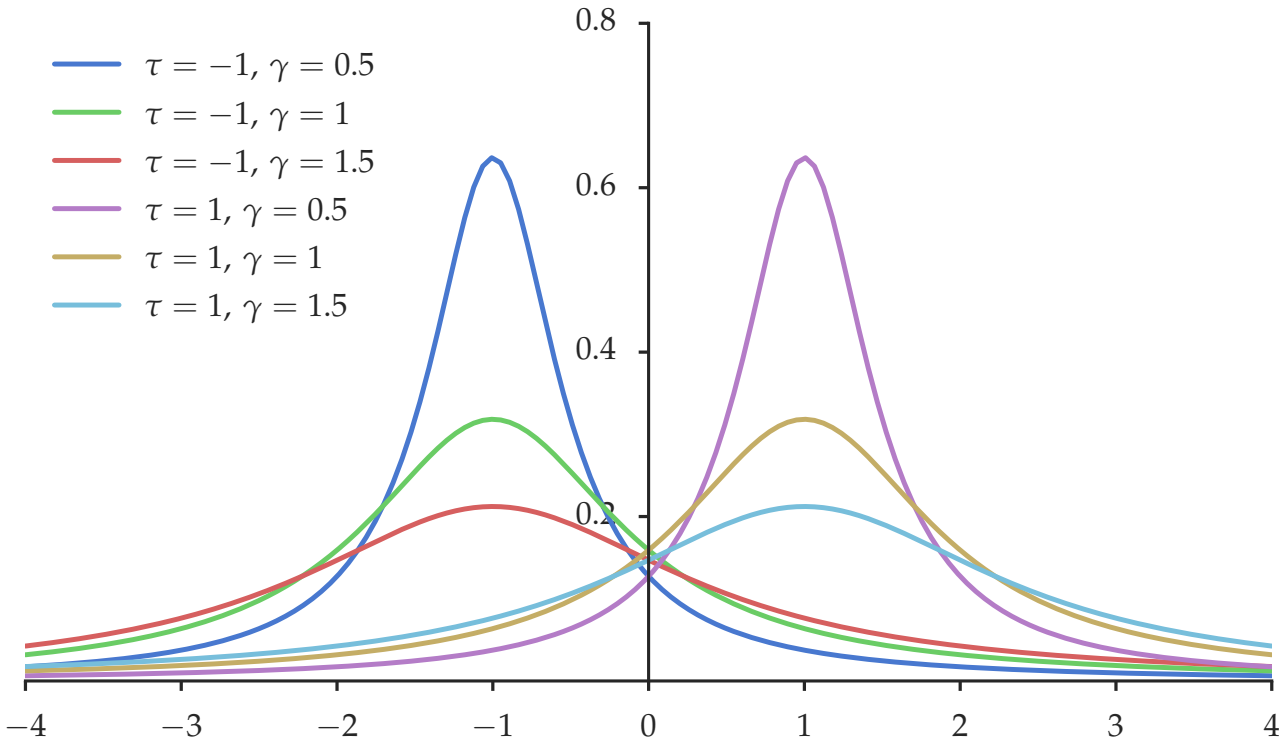

Figure: Cauchy densities

Example. The beta CDFs have densities given by

$$p(s) = \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)} \qquad (\alpha,\beta > 0)$$

when $0 < s < 1$ and zero elsewhere

Example. The $U[a,b]$ distribution represented by the density

$$p(s) = \frac{1}{b-a}\mathbb{1}\{a \leq s \leq b\} \qquad (s \in \mathbb{R},\ a, b \in \mathbb{R},\ a < b)$$

Example. The **gamma distribution** with shape parameter α and scale parameter β is the distribution with density

$$p(s) = \frac{s^{\alpha-1}e^{-s/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \qquad (\alpha, \beta > 0)$$

when $0 < s < 1$ and zero elsewhere

Example. The **chi-squared distribution with k degrees of freedom** is the distribution with density

$$p(s) := \frac{1}{2^{k/2}\Gamma(k/2)} s^{k/2-1} e^{-s/2} \qquad (s > 0,\ k \in \mathbb{N})$$

This distribution is represented by the symbol $\chi^2(k)$

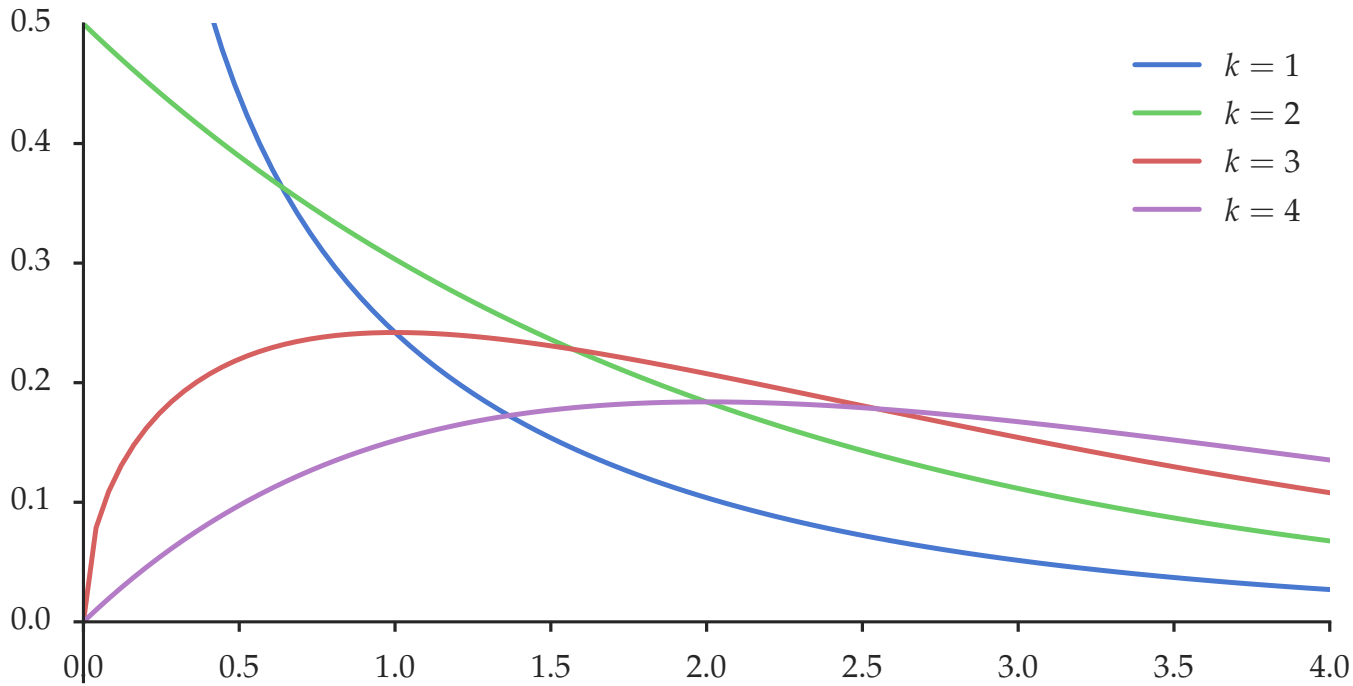

Figure: Chi-squared densities

Example. **Student's t-distribution with k degrees of freedom**, or, more simply, the t-distribution with k degrees of freedom, is the distribution on $\mathbb{R}$ with density

$$p(s) := \frac{\Gamma(\frac{k+1}{2})}{(k\pi)^{1/2}\Gamma(\frac{k}{2})}\left(1+\frac{s^2}{k}\right)^{-(k+1)/2} \qquad (s \in \mathbb{R},\ k > 0)$$

Example. The F-**distribution** with parameters k_1, k_2 is the distribution with the unlikely looking density

$$p(s) := \frac{\sqrt{(k_1 s)^{k_1} k_2^{k_2} / [k_1 s + k_2]^{k_1+k_2}}}{sB(k_1/2, k_2/2)} \qquad (s \geq 0,\ k_1,\ k_2 > 0)$$

The F-distribution arises in a number of hypothesis tests, as discussed below

Integrating with Distribution

Consider ordinary integral $\int_a^b h(s)\,\mathrm{d}s$ of a well-behaved function h on some interval $[a, b]$

Suppose we want to weight this integral, assigning more mass to different regions of $[a, b]$:

$$\int_a^b h(s)p(s)\,\mathrm{d}s$$

For example:

- h as a welfare function and p the density of agents
- p as a density indicating probabilities of outcomes, h as a payoff function

Suppose P does not have a density, but we still want to weight using P

- we want to define $\int h(s)P(\mathrm{d}s)$

Take distribution P on $\mathbb{R}$ and consider $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P)$ as a probability space

$(\mathbb{R}, \mathscr{B}(\mathbb{R}), P)$ has its own expectation operator $\mathbb{E}_P$

Assume h is a random variable on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P)$, then

$$\mathbb{E}_P h :=: \int h(s)P(\mathrm{d}s) := \text{the expectation of } h \text{ under } P$$

Fact. Let $h\colon \mathbb{R} \to \mathbb{R}$ be $\mathscr{B}$-measurable and let P be a distribution on $\mathbb{R}$

If P is discrete, with PMF $\{p_j\}_{j\geq 1}$ and support $\{s_j\}_{j\geq 1}$, then

$$\int h(s)P(\mathrm{d}s) = \sum_{j\geq 1} h(s_j)p_j$$

If P is absolutely continuous with density p, then

$$\int h(s)P(\mathrm{d}s) = \int_{-\infty}^{\infty} h(s)p(s)\,\mathrm{d}s$$

Distributions and Random Variables

Every random variable defines a distribution on $\mathbb{R}$

Let x be a random variable on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$:

- probability $\mathbb{P}\{x \in B\}$ is well-defined for every $B \in \mathscr{B}(\mathbb{R})$ (see (6))
- the set function P defined by

$$P(B) = \mathbb{P}\{x \in B\} \qquad (B \in \mathscr{B}(\mathbb{R})) \tag{20}$$

is the **distribution of** x

The CDF corresponding to the distribution P of x satisfies

$$F(s) = \mathbb{P}\{x \leq s\} \qquad (s \in \mathbb{R}) \tag{21}$$

We write $\mathcal{L}(x) = F$ to indicate that F represents the distribution of x

Fact. (4.2.4) If $\mathcal{L}(x) = F$, then $\mathbb{P}\{a < x \leq b\} = F(b) - F(a)$ for any $a \leq b$

Proof is an exercise (or see page 111 in ET)

For every CDF F, there exists a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a random variable $x\colon \Omega \to \mathbb{R}$ such that $\mathcal{L}(x) = F$; §7.4.1 outlines one construction

If $\mathcal{L}(x) = P$ and P has density p, we say x **has density** p

If the distribution of x is discrete, we'll call x a **discrete random variable**

Fact. If x has a density, then $\mathbb{P}\{x = s\} = 0$ for all $s \in \mathbb{R}$, and for any $a < b$,

$$\mathbb{P}\{a < x < b\} = \mathbb{P}\{a < x \leq b\}$$

$$= \mathbb{P}\{a \leq x < b\} = \mathbb{P}\{a \leq x \leq b\}$$

Distributions of Transformations

Fact. (4.2.6) If $\mathcal{L}(x) = F$ and $y := \psi(x)$, where $\psi\colon \mathbb{R} \to \mathbb{R}$ is strictly increasing, then $\mathcal{L}(y) = G$ where $G(s) := F(\psi^{-1}(s))$.

Proof. Observe under these hypotheses ψ^{-1} exists and is (weakly) increasing. Hence

$$\mathbb{P}\{y \leq s\} = \mathbb{P}\{\psi(x) \leq s\} = \mathbb{P}\{x \leq \psi^{-1}(s)\} = F(\psi^{-1}(s))$$

Note how monotonicity is used in the second equality

Example. If $\mathcal{L}(x) = F$ and $y := \exp(x)$, then the CDF of y is $G(s) := F(\ln(s))$

Fact. (4.2.7) If x has density p on $\mathbb{R}$ and $y := \psi(x)$ where ψ is a diffeomorphism on $\mathbb{R}$, then the distribution of y is absolutely continuous, with density

$$q(s) = p(\psi^{-1}(s)) \left| \frac{\mathrm{d}\psi^{-1}(s)}{\mathrm{d}s} \right| \qquad (s \in \mathbb{R})$$

The term **diffeomorphism** means ψ is a bijection on $\mathbb{R}$ and both ψ and its inverse are differentiable

Example. If x has density p on $\mathbb{R}$ and μ and σ are constants with $\sigma > 0$, then the density of $y := \mu + \sigma x$ is

$$q(s) = p\left(\frac{s-\mu}{\sigma}\right)\frac{1}{\sigma} \qquad (s \in \mathbb{R})$$

When x is standard normal: $y = \mu + \sigma x$ is $\mathrm{N}(\mu, \sigma^2)$

Why?

- Take p to be the standard normal density ϕ
- Recall

$$p(s) = F'(s) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\}$$

Let x be a random variable on probability space $(\Omega, \mathscr{F}, \mathbb{P})$

The distribution of x encodes all information to calculate expectations of x or of any $\mathscr{B}$-measurable transformation $h(x)$

First, let x be finite. Suppose

- $\mathcal{L}(x) = P$
- the function $h\colon \mathbb{R} \to \mathbb{R}$ is any $\mathscr{B}$-measurable function
- P puts all mass on finite set $\{s_j\}_{j=1}^{J}$

Using $\mathbb{P}\{x = s_j\} = P\{s_j\}$ and the definition of expectations:

$$\mathbb{E}h(x) = \sum_{j=1}^{J} h(s_j)\mathbb{P}\{x = s_j\} = \sum_{j=1}^{J} h(s_j)P\{s_j\} = \sum_{j=1}^{J} h(s_j)p_j$$

The expectation of $h(x)$ on $(\Omega, \mathscr{F}, \mathbb{P})$ equal to the expectation of h on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P)$

True also for the infinite case:

Fact. Let x be a random variable on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, let $\mathcal{L}(x) = P$ and let h be a $\mathscr{B}$-measurable function such that $h(x)$ is integrable. The expectation $\mathbb{E}h(x)$ is entirely determined by h and P. In particular,

$$\mathbb{E}h(x) = \int h(s)P(\mathrm{d}s)$$

where $\int h(s)P(\mathrm{d}s)$ is the expectation of h on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P)$

Example. Let x be a random variable whose distribution P is the uniform distribution on $[a, b]$

Apply the definition of the uniform density

$$\mathbb{E}\, x = \int s P(\mathrm{d}s) = \int s p(s)\, \mathrm{d}s = \int_{-\infty}^{\infty} \frac{s}{b-a} \mathbb{1}\{a \leq s \leq b\}\, \mathrm{d}s$$

Solving the integral gives $\mathbb{E}\, x = \mu := (a+b)/2$. The variance is

$$\mathrm{var}[x] = \int (s-\mu)^2 P(\mathrm{d}s)$$

$$= \int_a^b \left(s - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\, \mathrm{d}s = \frac{1}{12}(b-a)^2$$

Example. Suppose $\mathcal{L}(x) = \mathrm{N}(\mu, \sigma)$

If $\sigma > 0$, the mean can be computed via

$$\mathbb{E}\, x = \int_{-\infty}^{\infty} s \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(s-\mu)^2}{2\sigma^2} \right\} \mathrm{d}s = \mu$$

The variance is given by:

$$\mathrm{var}[x] = \int_{-\infty}^{\infty} (s-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(s-\mu)^2}{2\sigma^2} \right\} \mathrm{d}s = \sigma^2$$

Moments from Distributions

Any two random variables with the same distribution share the same moments

Hence moments are best thought of as a property of the distribution, not the random variable

Thus we define

- the **mean** of P as $\mu = \int sP(\mathrm{d}s)$,
- the k**th moment** of P as $\int s^k P(\mathrm{d}s)$,
- the **variance** of P as $\int (s-\mu)^2 P(\mathrm{d}s)$,

and so on

Quantile Function

Let F be a strictly increasing CDF on $\mathbb{R}$

Given $\tau \in (0,1)$, the **τth quantile** of F is the $\xi \in \mathbb{R}$ that solves $F(\xi) = \tau$

Under our assumptions on F, such a ξ exists and is uniquely defined

The 0.5th quantile is called the **median** of F

The **quantile function**:

$$F^{-1}(\tau) := \text{ the unique } \xi \text{ such that } F(\xi) = \tau \qquad (0 < \tau < 1)$$

Example. The quantile function associated with the standard Cauchy distribution is $F^{-1}(\tau) = \tan[\pi(\tau - 1/2)]$

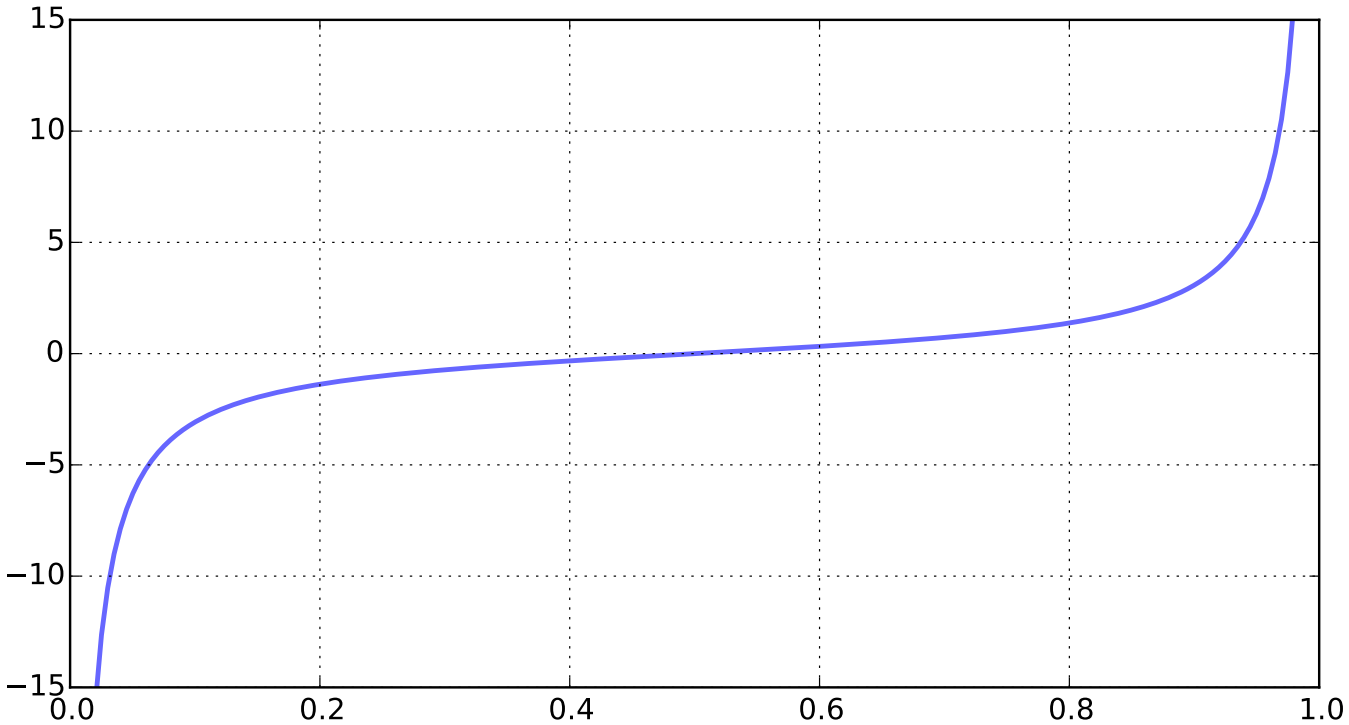

Figure: Cauchy quantile function (The horizontal axis is $\tau \in (0,1)$)

When F is not strictly increasing, F^{-1} not well-defined

We can set:

$$F^{-1}(\tau) := \inf\{s \in \mathbb{R} : F(s) \geq \tau\} \qquad (0 < \tau < 1) \qquad (22)$$

A density p is symmetric if $p(s) = p(-s)$ for all $s \in \mathbb{R}$

A common scenario in hypothesis testing

Fact. (4.2.9) Let x be a random variable with density p. If p is symmetric, then the CDF G of $y := |x|$ is given by

$$G(s) := \mathbb{P}\{y \leq s\} = \begin{cases} 2F(s) - 1 & \text{if } s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Proof as exercise 4.4.18

Fact equivalent to $F(s) = 1 - F(-s)$

Take a random variable x with $\mathcal{L}(x) = F$ and constant $\alpha \in (0,1)$ as given

Consider the c that solves $\mathbb{P}\{-c \leq x \leq c\} = 1 - \alpha$

Fact. If $\mathcal{L}(x) = F$, x has a symmetric density and F is strictly increasing, then

$$c = F^{-1}(1 - \alpha/2) \quad \Longrightarrow \quad \mathbb{P}\{-c \leq x \leq c\} = 1 - \alpha \tag{23}$$

When F is the standard normal CDF Φ, c is usually denoted by $z_{\alpha/2}$:

$$z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2) \tag{24}$$

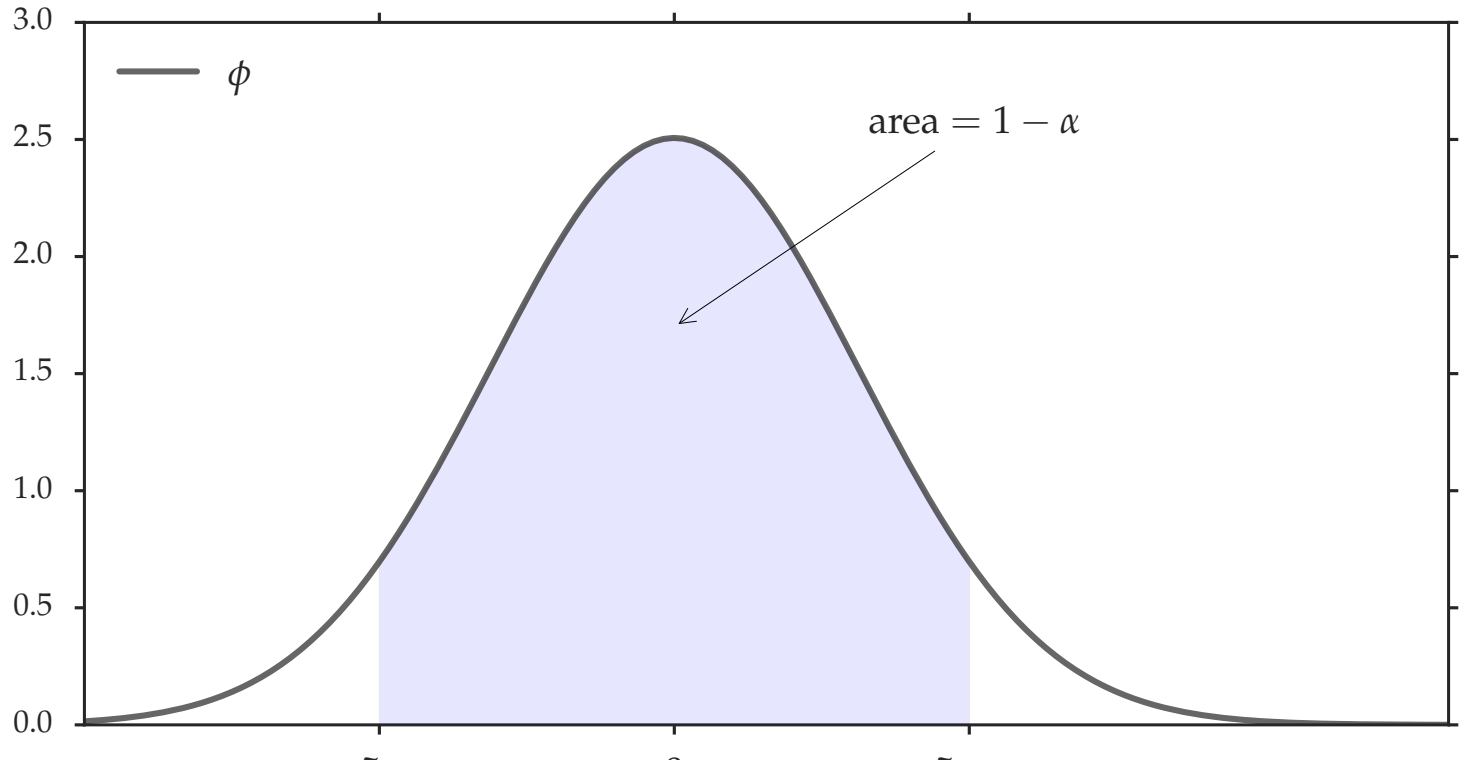

Figure: Critical values for the standard normal density