Elements of probability theory

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Motivation

Add some motivation or summary of what was done last time

Today's outline

Elements of probability theory

Outline

Some vocabulary

We want to model (describe, analyze, work with) outcomes that are not deterministic in nature, or at least not tractably deterministic in nature.

- We call such outcomes random, or subject to chance,
- and the situation generating them is known as **random experiment**.

Definition

A set, $\mathcal{S}$, that contains all possible outcomes of a random experiment is called sample space.

Some examples

Example (Dice)

If the experiment consists of tossing a die, the sample space contains six possible outcomes given by $\mathcal{S} = \{$⚀,⚁,⚂,⚃,⚄,⚅$\}$.

Example (Traffic deaths)

If the experiment consists of recording the number of traffic deaths in Germany next year, the sample space would contain all positive integers, $\mathcal{S} = \{0, 1, 2, ...\}$.

Example (Light bulbs)

If the experiment consists of observing the length of life of a light bulb, the sample space would contain all positive real numbers, $\mathcal{S} = (0, \infty)$.

A taxonomy with consequences

The sample space $\mathcal{S}$, as all sets, can be classified according to whether the number of elements in the set are

- finite (discrete sample space), e.g., $\mathcal{S} = \{0, 1, 2, .., 6\}$
- countably infinite (discrete sample space), e.g., $\mathcal{S} = \mathbb{N} = \{0, 1, 2, ...\}$
- uncountably infinite (continuous sample space), e.g., $\mathcal{S} = \mathbb{R}$.

But all these possible outcomes cannot occur at the same time:

- We'll use the term probability to talk about relative likelihoods of occurrence.
- So we need to be able to assign probabilities to various outcomes.

Outcomes are not enough

Example

Let the experiment consist of tossing a die, but let the possible outcomes of interest be

1. either 1 or 2 dots,
2. either 3 or 4 dots, and
3. either 5 or 6 dots.

To deal with this, we may either

- construct a new sample space to reflect these outcomes, or
- re-use the "raw" sample space $\mathcal{S} = \{$⚀,⚁,⚂,⚃,⚄,⚅$\}$ in some suitable way.

Putting outcomes together

Definition

An event, say A, is a subset of the sample space $\mathcal{S}$ (including $\mathcal{S}$ itself).

- Let A be an event, a subset of $\mathcal{S}$. We say the event A occurs if the outcome of the experiment is in the set A.
- An event consisting of a single element or outcome is called elementary event.
- The event $\mathcal{S}$ is called the sure or certain event.
- Events whose intersection is the empty set $\emptyset$ are mutually exclusive.

We want to assign probabilities to **events**; in other words,

What is the probability that event A occurs?

Tossing dice...

The experiment consists of tossing a die and counting the number of dots facing up. The sample space is defined to be $\mathcal{S} = \{1, 2, ..., 6\}$. Let then

$$A_1 = \{1, 2, 3\}, \qquad A_2 = \{2, 4, 6\}, \qquad A_3 = \{6\}.$$

1. A_1 is an event whose occurrence means that the number of dots is less than four.
2. A_2 is an event for which the number of dots is even.
3. A_3 is an elementary event.
4. Note that $A_1 \cap A_3 = \emptyset$ so they are mutually exclusive events.
5. Also, if the outcome is 2, both A_1 and A_2 occur.

Probability

For each event A in $\mathcal{S}$ we associate a number between 0 and 1 that will be called the probability of A.

For this purpose we will use an appropriate set function, say $\mathrm{P}(\cdot)$, with the set of all events as domain.

Irrespective of what you understand under "probability", e.g.

- classical probability Laplace
- relative frequency based probabilities, or
- subjective probability (as many bayesians claim to use)

Probability calculus deals with probabilities in a coherent manner for all events!

Event spaces

Definition (Event space)

The set of all events in the sample space $\mathcal{S}$ is called the event space $\mathcal{Y}$.

We will use collections of subsets of S which satisfy minimal conditions...

Definition

A collection of subsets of $\mathcal{S}$ is called a sigma algebra, denoted by $\mathcal{B}$, if it satisfies the following conditions:

(i) $\emptyset \in \mathcal{B}$ (empty set is an element of $\mathcal{B}$);

(ii) If $A \in \mathcal{B}$, then $\overline{A} \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation);

(iii) If $A_1, A_2, \ldots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

The countable case

A typical sigma algebra used as event space $\mathcal{Y}$ if the sample space $\mathcal{S}$ is finite or countable is

$$\mathcal{B} = \{\text{all subsets of } \mathcal{S}, \text{ including } \mathcal{S}\}.$$

Note that if $\mathcal{S}$ has n elements there are 2^n sets in $\mathcal{B}$.

Example (All in)

If $\mathcal{S} = \{1, 2, 3\}$, then the sigma algebra consisting of all subsets of $\mathcal{S}$ is the following collection of $2^3 = 8$ sets:

$$\{1\},\ \{2\},\ \{3\},\ \{1,2\},\ \{1,3\},\ \{2,3\},\ \{1,2,3\},\ \{\emptyset\}.$$

The uncountable case

A bit of trouble: although the power set (even of an uncountable set) is a σ-algebra, the real line is sometimes unfriendly to probabilities, and we may need restrictions when dealing with uncountable sample spaces.

Example (σ-algebras on the real line)

A typical sigma algebra used as event space $\mathcal{Y}$ if the sample space is an interval on the real line (i.e. $\mathcal{S} \subset \mathbb{R}$) is

$\mathcal{B}$ containing all sets of all closed, open and half-open intervals:

$$[a, b], \ (a, b], \ [a, b), \ (a, b), \quad \forall a, b \in \mathcal{S},$$

as well as all sets that can be formed by taking (possibly countably infinite) unions and intersections of these intervals[a].

[a] This special sigma algebra is usually referred to as a collection of Borel sets (see, e.g., Mittelhammer, 1996, p.21).

Axiomatic Probability Definition

▸ Kolmogorov

Definition (Probability function)

Given a sample space $\mathcal{S}$ and an associated event space $\mathcal{Y}$ (a sigma algebra on $\mathcal{S}$), a probability (set) function is a set function P with domain $\mathcal{Y}$ s.t.

1 (non-negativity) $\mathrm{P}(A) \geq 0$ for all $A \in \mathcal{Y}$.

2 (standardization) $\mathrm{P}(\mathcal{S}) = 1$.

3 (additivity) If $A_1, A_2, \ldots \in \mathcal{Y}$ is a sequence of disjoint events ($A_i \cap A_j = \emptyset$ for $i \neq j$; $i, j \in \mathbb{N}$), then $\mathrm{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}\left(A_i\right)$.

- Wisely, this definition makes no attempt to tell what particular set function P to choose.

Dice (& Laplace)

Let $\mathcal{S} = \{1, 2, ..., 6\}$ be the sample space for rolling a fair die and observing the number of dots facing up. The set function

$$\mathrm{P}(A) = N(A)/6 \quad \text{for} \quad A \subset \mathcal{S}$$

(where $N(A)$ is the size of set A) represents a probability set function on the events of S:

- the value of the function $\mathrm{P}(A) \geq 0$ for all $A \subset \mathcal{S}$ (non-negativity);
- the value of the function for the set $\mathcal{S}$ is $P(\mathcal{S}) = N(\mathcal{S})/6 = 1$ (standardization);
- for any collection of disjoint sets $A_1, A_2, ..., A_n$ we have

$$\mathrm{P}(\cup_{i=1}^{n} A_i) = \frac{N(\cup_{i=1}^{n} A_i)}{6} = \frac{\sum_{i=1}^{n} N(A_i)}{6} = \sum_{i=1}^{n} \mathrm{P}(A_i) \quad \text{(additivity)}.$$

Now with numbers

Take the sample space $\mathcal{S} = \{1, 2, ...\} = \mathbb{N} \setminus \{0\}$, together with

$$\mathrm{P}(A) = \sum_{x \in A} \left(\frac{1}{2}\right)^x \quad \text{for} \quad A \subset \mathcal{S}.$$

This set function represents a probability set function since

1. the value of the function $\mathrm{P}(A) \geq 0$ for all $A \subset \mathcal{S}$, because P is defined as the sum of non-negative numbers (non-negativity);
2. the value of the function for the set $\mathcal{S}$ is (standardization)

$$\mathrm{P}(\mathcal{S}) = \sum_{x \in S} \left(\frac{1}{2}\right)^x = \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \underbrace{\sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^x}_{geom.\ series} - 1 = \frac{1}{1 - \frac{1}{2}} - 1 = 1$$

3. for any collection of disjoint sets $A_1, A_2, \ldots, A_n, \ldots$ we have

$$\mathrm{P}(\cup_{i=1}^{\infty} A_i) = \sum_{x \in (\cup_{i=1}^{\infty} A_i)} \left(\frac{1}{2}\right)^x = \sum_{i=1}^{\infty} \left[\sum_{x \in A_i} \left(\frac{1}{2}\right)^x\right] = \sum_{i=1}^{\infty} \mathrm{P}(A_i) \quad \text{(additivity)}.$$

And uncountable sets

Let $\mathcal{S} = [0, \infty)$ be the sample space for an experiment consisting of observing the length of life of a light bulb and consider the set function

$$\mathrm{P}(A) = \int_{x \in A} \frac{1}{2} e^{-\frac{x}{2}} \mathrm{d}x \quad \text{for} \quad A \in \mathcal{Y}.$$

This set function represents a probability set function since

1. the value of the function $\mathrm{P}(A) \geq 0$ for all $A \subset \mathcal{S}$, because P is defined as an integral with a non-negative integrand (non-negativity);
2. the value of the function for the set $\mathcal{S}$ is
$$\mathrm{P}(\mathcal{S}) = \int_{x \in S} \frac{1}{2} e^{-\frac{x}{2}} \mathrm{d}x = \int_{0}^{\infty} \frac{1}{2} e^{-\frac{x}{2}} \mathrm{d}x = 1 \quad \text{(standardization);}$$
3. for any disjoint sets $A_1, A_2, \ldots, A_n, \ldots$ we have (additivity)

$$\mathrm{P}(\cup_{i=1}^{\infty} A_i) = \underbrace{\int_{x \in (\cup_{i=1}^{\infty} A_i)} \frac{1}{2} e^{-\frac{x}{2}} \mathrm{d}x = \sum_{i=1}^{\infty} \left[\int_{x \in A_i} \frac{1}{2} e^{-\frac{x}{2}} \mathrm{d}x \right]}_{\textit{addititivity property of Riemann integrals}} = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

We have what we need

- Once we have defined the 3-tuple $\{\mathcal{S}, \mathcal{Y}, \mathrm{P}\}$ (called probability space) for an experiment of interest,
- ... we have all information needed to assign probabilities to various events.

It is the choice of an appropriate probability set function P *that represents the major challenge in statistical real-life applications.*

Either way, the three axioms imply many properties of the probability function.

We list implications...

Theorem (1.1)

Let A be an event in $\mathcal{S}$. Then $\mathrm{P}(A) = 1 - \mathrm{P}\left(\bar{A}\right)$.

Theorem (1.2)

$\mathrm{P}(\emptyset) = 0$.

Theorem (1.3)

Let A and B be events in $\mathcal{S}$ such that $A \subset B$. Then $\mathrm{P}(A) \leq \mathrm{P}(B)$ *and* $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$.

Theorem (1.4)

Let A and B be events in $\mathcal{S}$. Then $\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap \bar{B})$.

… and go on …

Theorem (1.5)

Let A and B be events in $\mathcal{S}$. Then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Corollary (1.1, Boole's Inequality)

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

Theorem (1.6)

Let A be an event in $\mathcal{S}$. Then $\mathrm{P}(A) \in [0, 1]$.

Theorem (1.7, Bonferroni's Inequality)

Let A and B be events in $\mathcal{S}$. Then $\mathrm{P}(A \cap B) \geq 1 - \mathrm{P}(\bar{A}) - \mathrm{P}(\bar{B})$.

... like the Duracell bunny

Theorem (1.8)

Let $A_1, \ldots, A_n$ *be events in* $\mathcal{S}$*. Then* $\mathrm{P}\left(\cap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} \mathrm{P}(\bar{A}_i)$ *and* $\mathrm{P}\left(\cup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathrm{P}(A_i)$

Theorem (1.9, Classical probability)

Let $\mathcal{S}$ *be the finite sample space for an experiment having* $n = N(\mathcal{S})$ *equally likely outcomes, say* $E_1, ..., E_n$ *, and let* $A \subset \mathcal{S}$ *be an event containing* $N(A)$ *elements. Then the probability of the event* A *is given by* $N(A)/N(\mathcal{S})$*.*

Outline

Let's talk about what we know

So far, we have considered probabilities of events on the assumption that no information was available about the experiment other than the sample space $\mathcal{S}$.

Sometimes, however, it is known that an event B has happened.

- Can we use this information in making a statement concerning the outcome of another event A?
- I.e. (how) can we update the probability calculation for the event A based on the information that B has happened?

Tricks with coins

Example (Knowledge is power)

Consider tossing two fair coins. The sample space is $\mathcal{S} = \{(\mathsf{H,H}), (\mathsf{H,T}), (\mathsf{T,H}), (\mathsf{T,T})\}$ (H= Head, T=Tail).

Examine the events

$$A = \{\text{both coins show same face}\}, \quad B = \{\text{at least one coin shows H}\}.$$

Then $\mathrm{P}(A) = 2/4 = 1/2$.

If B is known to have happened, we know for sure that the outcome (T,T) cannot happen. This suggest that

$$\mathrm{P}(A \text{ conditional on } B \text{ having happened}) = 1/3.$$

A different probability?

Focus on sub-algebras

Definition (Conditional probability)

Let A and B be any two events in a sample space $\mathcal{S}$. If $\mathrm{P}(B) \neq 0$, then the conditional probability of event A, given event B, is given by $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) / \mathrm{P}(B)$.

Example

The experiment consists of tossing two fair coins. The sample space is $\mathcal{S} = \{(\text{H,H}), (\text{H,T}), (\text{T,H}), (\text{T,T})\}$. The conditional probability of the event obtaining two heads

$$A = \{(\text{H,H})\},$$

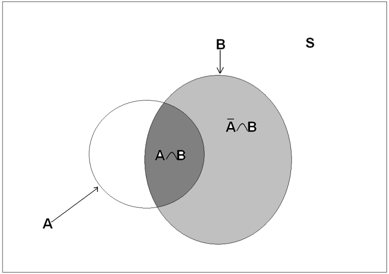
given the first coin toss results in heads, $B = \{(\text{H,H}), (\text{H,T})\}$ is

$$\mathrm{P}(A|B) = \mathrm{P}(A \cap B) / \mathrm{P}(B) \overset{(class.\ prob.)}{=} (1/4)/(1/2) = 1/2.$$

Still a probability

Theorem (1.10)

Given a probability space $\{\mathcal{S}, \mathcal{Y}, \mathrm{P}\}$ *and an event* B *for which* $\mathrm{P}(B) \neq 0$, $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) / \mathrm{P}(B)$ *defines a probability set function with domain* $\mathcal{Y}$.



- A occurs conditionally on B iff $A \cap B$ occurs.
- Hence, $\mathrm{P}(A \mid B) \propto \mathrm{P}(A \cap B)$.
- B plays the role of the sample space
- $\mathrm{P}(B|B) \overset{Def.}{=} \mathrm{P}(B \cap B)/\mathrm{P}(B) = \mathrm{P}(B)/\mathrm{P}(B) = 1$.

Undo the conditioning

The multiplication rule allows one to factorize the joint probability for the events A and B into

- the conditional probability for event A, given event B and
- the unconditional probability of B.

Theorem (1.11, Multiplication Rule)

Let A and B be any two events in $\mathcal{S}$ for which $\mathrm{P}(B) \neq 0$. Then $\mathrm{P}(A \cap B) = \mathrm{P}(A \mid B)\,\mathrm{P}(B)$.

Example: No dice, no coins

A test facility conducts blood tests to find some disease. The tested person is sent to a hospital if (and only if) the test is positive.

- The prevalence of the disease in the population is 2%, so a person picked at random has probability 0.02 of suffering from that disease (say event D such that $\mathrm{P}(D) = 0.02$).
- The probability that a test is positive (event A) if the tested person is actually ill (that is given event D) is $\mathrm{P}(A|D) = 0.95$.

The probability that a tested person is sent to the hospital (A) and is actually ill (D) is $\mathrm{P}(A \cap D) = \mathrm{P}(A \mid D)P(D) = 0.95 \cdot 0.02 = 0.019$.

We can do better

Theorem (1.12, Extended Multiplication Rule)

Let $A_1, A_2, \ldots, A_n, n \geq 2$, be events in $\mathcal{S}$. Then if all of the conditional probabilities exist,

$$\begin{aligned} \mathrm{P}\left(\cap_{i=1}^{n} A_i\right) &= \mathrm{P}(A_1) \cdot \mathrm{P}(A_2|A_1) \cdot \ldots \cdot \mathrm{P}(A_n|A_{n-1} \cap A_{n-2} \cap \ldots \cap A_1) \\ &= \mathrm{P}\left(A_1\right) \prod_{i=2}^{n} \mathrm{P}\left(A_i \mid \cap_{j=1}^{i-1} A_j\right). \end{aligned}$$

This is important

Definition (Independence of events, 2–event case)

Let A and B be two events in $\mathcal{S}$. Then A and B are independent iff $\mathrm{P}(A \cap B) = \mathrm{P}(A)P(B)$. If A and B are not independent, A and B are said to be dependent events.

Independence of A and B implies

$$\begin{aligned}\mathrm{P}(A|B) &= \mathrm{P}(A \cap B)/\,\mathrm{P}(B) = \mathrm{P}(A)\,\mathrm{P}(B)/\,\mathrm{P}(B) = \mathrm{P}(A), \text{ if } \mathrm{P}(B) > 0 \\ \mathrm{P}(B|A) &= \mathrm{P}(B \cap A)/\,\mathrm{P}(A) = \mathrm{P}(B)\,\mathrm{P}(A)/\,\mathrm{P}(A) = \mathrm{P}(B), \text{ if } \mathrm{P}(A) > 0.\end{aligned}$$

Thus the probability of event A occurring is unaffected by the occurrence of event B, and vice versa.

There is more

Independence of A and B implies independence of the complements also. In fact we have the following theorem:

Theorem (1.13)

If events A and B are independent, then events A and $\bar{B}$, $\bar{A}$ and B, and $\bar{A}$ and $\bar{B}$ are also independent.

More events

Definition (Independence of events, n–event case)

Let $A_1, A_2, \ldots, A_n$, be events in the sample space $\mathcal{S}$. The events $A_1, A_2, \ldots, A_n$ are (jointly) independent iff

$$\mathrm{P}(\cap_{j\in J} A_j) = \prod_{j\in J} \mathrm{P}\left(A_j\right), \quad \text{for all subsets} \quad J \subset \{1, 2, \ldots, n\}$$

for which $N(J) \geq 2$. If the events $A_1, A_2, \ldots, A_n$ are not independent, they are said to be dependent events.

Note: Pairwise independence is not enough! E.g. in the case of $n = 3$ events, joint independence requires:

$\mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\,\mathrm{P}(A_2),\ \mathrm{P}(A_1 \cap A_3) = \mathrm{P}(A_1)\,\mathrm{P}(A_3),\ \mathrm{P}(A_3 \cap A_2) = \mathrm{P}(A_3)\,\mathrm{P}(A_2),$

(pairwise independence) and

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\,\mathrm{P}(A_2)\,\mathrm{P}(A_3).$$

A counterexample

Let the sample space $\mathcal{S}$ consists of all permutations of the letters a, b, c along with three triples of each letter, that is,

$$\mathcal{S} = \{\text{aaa}, \text{bbb}, \text{ccc}, \text{abc}, \text{bca}, \text{cba}, \text{acb}, \text{bac}, \text{cab}\}.$$

Furthermore, let each element of $\mathcal{S}$ have probability $1/9$. Consider the events

$$A_i = \{i\,\text{th place in the triple is occupied by a}\}.$$

According to the classical probability we obtain for all $i = 1, 2, 3$

$$\mathrm{P}(A_i) = 3/9 = 1/3, \quad \text{and} \quad \mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1 \cap A_3) = \mathrm{P}(A_2 \cap A_3) = 1/9,$$

so A_1, A_2, A_3 are pairwise independent. But they are not jointly independent since

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = 1/9 \neq \mathrm{P}(A_1)\,\mathrm{P}(A_2)\,\mathrm{P}(A_3) = 1/27.$$

Outline

First glue partitions together

Bayes's rule provides an alternative representation of conditional probabilities. It turns out to be very useful though...

In fact, it is a simple consequence of the total probability rule established in the following theorem:

Theorem (1.14, Law of Total Probability)

Let the events $B_i, i \in I$, be a finite or countably infinite partition of $\mathcal{S}$, so that $B_j \cap B_k = \emptyset$ for $j \neq k$, and $\cup_{i \in I} B_i = \mathcal{S}$. Let $\mathrm{P}(B_i) > 0 \; \forall \; i$. Then the "total" probability of event A is

$$\mathrm{P}(A) = \sum_{i \in I} \mathrm{P}(A \mid B_i)\, \mathrm{P}(B_i).$$

"Just" a corollary

Corollary (1.2, Bayes's Law)

Let the events $B_i, i \in I$, be a finite or countably infinite partition of $\mathcal{S}$, so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\cup_{i\in I} B_i = \mathcal{S}$. Let $\mathrm{P}(B_i) > 0 \; \forall \; i \in I$. Then, provided $\mathrm{P}(A) \neq 0$,

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)}{\sum_{i\in I} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)}, \quad \forall \; j \; \in \; I.$$

Hence, Bayes's law provides the means for updating the probability of the event B_j, given the "signal" that the event A occurs.

QUITE! useful

Do e.g. a test for some disease. Let A be the event that the test result is positive and B be the event that the individual has the disease.

- The test detects the disease with prob. 0.98 if the disease is, in fact, in the individual being tested: $\mathrm{P}(A|B) = 0.98$.
- The test yields a 'false positive' result for 1 percent of the healthy subjects: $\mathrm{P}(A|\bar{B}) = 0.01$.

Finally, 0.1 percent of the population has the disease, $\mathrm{P}(B) = 0.001$.

If the test result is positive, what is the actual probability that a randomly chosen person to be tested actually has the disease?

The application of Bayes's rule yields

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \mid B)\,\mathrm{P}(B)}{\mathrm{P}(A \mid B)\,\mathrm{P}(B) + \mathrm{P}(A \mid \bar{B})\,\underbrace{\mathrm{P}(\bar{B})}_{1-\mathrm{P}(B)}} = \frac{.98 \cdot .001}{.98 \cdot .001 + .01 \cdot .999} = .089.$$

Outline

Coming up

Random variables and their distribution