Philstats

Contents

Aims and goals

This is an overview of philosophy of statistics. The main goal is to set out and evaluate the philosophical ideas of founding figures of Neyman-Pearson, Fisher, Savage, and Jeffreys, and their modern proponents in philosophy and statistics like Mayo and Howson-Urbach and Royall. But we begin with an overview and review of relevant bits of probability and statistics and interpretations of probability. And we close by discussing various topics such as causality and simplicity and factor analysis.

Prerequistes

We are going to use the historical development to orientate the discussion. This course does not presuppose any prior familiarity with statistics. In terms of prior knowledge, this is structured so that its topic is to statistics roughly what medical ethics is to medicine. Everyone should be able to get a lot out of it, due in part to the ubiquity of statistics in our everyday lives.

Skip to main content

Acknowledgements

Thanks to the wonderful students in UCLA's Phlios 133C in Spring 2024 for their feedback on this material. This material owes a lot to that of my former colleague Kent Johnson's course. I also learned a lot from Cosmo Grant's philosophy of statistics courses which he ran while he was a post-doc at UCLA a couple of years ago.

Examples and python

Philosophy works best when there are good examples at play to motivate a position or to assess a position. And it is all the better when it is easy for everyone with just a little bit of time and creativity to come up with the examples. In statistics, the recent popularity of python (and R) has made coming up with worked out examples easier than ever. Hence, there are many bits of python code scattered throughout the text, mostly to illustrate key ideas with nice graphics, but also to provide an easy way to work out specific examples. You can just copy and paste the code into your favorite implementation of Python and run it. There are lots of easy ways to run python these days, including the following:

- Programiz (no set-up required; but don't use this if you want/need to easily save)
- UCLA Jupyterhub (virtually no set-up required; but requires UCLA sign on; your institution probably has Juptyerhub)
- Anaconda Navigator (requires about an hour to set up; for use on desktop machines; just watch some youtube videos if you need help getting started)

Outline

The rough plan is as follows, where each Chapter corresponds to a lecture:

Chapters 1-2: sets and events; review and interpretations of probability. Reference: [Galavotti, 2014].

Chapters 3-4: cdfs and pfs, visually and otherwise; independence and what happens in the limit.

Skip to main content

Chapter 7-8: Neyman-Pearson and Mayo. References: [Mayo, 1996], [Howson and Urbach, 2006]

Chapters 9-10: Bayesianism. References: [Mayo, 1996], [Howson and Urbach, 2006], [Savage, 1972]

Chapters 11-12: Likelihood and Fisher. References: [Royall, 2017], [Fisher, 1990]

Chapters 13-14: Objective Bayes. References: [Jeffreys, 1948], [Williamson, 2010]

Chapters 15-16: Causal inference. References: [Hitchcock, 2001] [Hitchcock, 2009]

Chapters 17-18: Model selection. References: [Sober, 2015]

Chapters 19-20 : Factor analysis. References: [Johnson, 2016]

Chapter 1

Some Greek letters

We use the following Greek letters:

- Ω big omega
- ω little omega
- Θ big theta
- θ little theta

We organize our usage of these so that ω ranges over elements of Ω, and θ ranges over elements of Θ

For sets, we also use a sytlized epsilon $\in$ for the *membership symbol*. E.g. $1 \in \{1, 2, 3\}$ and $1 \notin \{2, 3, 4\}$

The following tend to get used to describe certain canonical distributions:

- λ lambda
- μ mu

Skip to main content

- σ sigma

We occasionally also use the following when we run out of latin letters:

- α alpha
- β beta
- γ gamma

Sets rules and examples

Set-theoretic operations

We are usually just working with subsets of a given set Ω, which is fixed by the topic at hand.

There's some operations that we need to remind ourselves of:

- Intersection of A, B, symbol $A \cap B$, definition: $A \cap B = \{x \in \Omega : x \in A \wedge x \in B\}$
- Union of A, B, symbol $A \cup B$, definition: $A \cup B = \{x \in \Omega : x \in A \vee x \in B\}$
- Difference of A from B, symbol $A - B$, definition: $A - B = \{x \in \Omega : x \in A \wedge x \notin B\}$
- Relative complement of A, symbol A^c or $\overline{A}$, definition: $A^c = \overline{A} = \{x \in \Omega : x \notin A\}$

The empty set $\emptyset$ is the subset of Ω that has no elements.

Set-theoretic rules

Suppose that A, B, C are subsets of Ω. Then we have the following rules:

1. $A \cup A^c = \Omega$
2. $(A \cap A^c)^c = \Omega$, equivalently $(A \cap A^c) = \emptyset$
3. $(A^c)^c = A$
4. $A \cap B = B \cap A$
5. $A \cup B = B \cup A$
6. $(A \cap B) \cap C = A \cap (B \cap C)$

Skip to main content

8. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
9. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
10. $(A \cap B)^c = A^c \cup B^c$
11. $(A \cup B)^c = A^c \cap B^c$

Corresponding rules of propositional logic

The previous rules of sets correspond to the following tautologies and equivalences of propositional logic:

1. $p \vee \neg p$ is a tautology, law of excluded middle
2. $\neg(p \wedge \neg p)$, law of non-contradiction
3. $\neg\neg p$ is equivalent to p, law of double-negation
4. $p \wedge q$ is equivalent to $q \wedge p$, law of commutativity of conjunction
5. $p \vee q$ is equivalent to $q \vee p$, law of commutativity of disjunction
6. $(p \wedge q) \wedge r$ is equivalent to $p \wedge (q \wedge r)$, law of associativity of conjunction
7. $(p \vee q) \vee r$ is equivalent to $p \vee (q \vee r)$, law of associativity of disjunction
8. $p \wedge (q \vee r)$ is equivalent to $(p \wedge q) \vee (p \wedge r)$, distribution for conjunction
9. $p \vee (q \wedge r)$ is equivalent to $(p \vee q) \wedge (p \vee r)$, distribution for disjunction
10. $\neg(p \wedge q)$ is equivalent to $\neg p \vee \neg q$, DeMorgan
11. $\neg(p \vee q)$ is equivalent to $\neg p \wedge \neg q$, DeMorgan

Reason for the correspondence

Subsets of a set Ω, and formulas of propositional logic, are both *Boolean algebras*.

The rules of sets described above, and the rules of propositional logic described above, are special cases of rules of Boolean algebras.

For a lot of purposes, these are the most interesting examples of Boolean algebras; which explains why many of us learn the rules in these separate settings.

Skip to main content

Example (of set-theoretic operations)

Suppose that Ω is the set of all possible outcomes of flipping a coin three times.

Suppose we use 0 for tails and 1 for heads. Then Ω is

$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\}$

E.g. 010 is the outcome of: tails, heads, tails.

One also writes $\Omega = \{0, 1\}^3$.

Let A be the set of outcomes with two or more heads.

Let B the set of outcomes with one or more tails.

Then $A = \{011, 101, 110, 111\}$

and $B = \{000, 001, 010, 011, 100, 101, 110\}$

Then one has $A \cap B = \{011, 101, 110\}$.

And one has $B = \Omega \setminus \{111\}$.

Definition (subset)

A set A is a *subset* of B if every element of A is also an element of B.

We write $A \subseteq B$ for A is a subset of B.

For instance $\{0, 1, 2\}$ is a subset of $\{0, 1, 2, 3, 4\}$ since each number in the first set is in the second set.

But we do not have $\{0, 1, 2\}$ is a subset of $\{1, 2, 3, 4\}$ since 0 is in the first set but not in the second set.

Skip to main content

Definition (disjointness)

Sets A, B are *disjoint* if $A \cap B = \emptyset$

For instance, $\{1, 2, 3\}$ and $\{8, 9, 10\}$ are disjoint.

Sets A, B, C are *pairwise disjoint* if A, B are disjoint and A, C are disjoint and B, C are disjoint.

For instance, $\{1, 2\}, \{8, 9\}, \{15, 16\}$ are pairwise disjoint.

Products of sets

If Ω and Θ are two sets, then we can form their product

$\Omega \times \Theta = \{(\omega, \theta) : \omega \in \Omega, \theta \in \Theta\}.$

For instance if $\Omega = \{0, 1\}$ and $\Theta = \{1, 2, 3\}$ then

$\Omega \times \Theta = \{(0, 1), (0, 2), (0, 3), (1, 1), (1, 2), (1, 3)\}$

and

$\Theta \times \Omega = \{(1, 0), (1, 1), (2, 0), (2, 1), (3, 0), (3, 1)\}$

Note that $(0, 2)$ is in $\Omega \times \Theta$ but not in $\Theta \times \Omega$.

n choose k

Consider $\Omega = \{0, 1\}^n$, the set of all possible outcomes of fliping a coin n-times.

This set has size 2^n.

For elementary probability, we need to recall one more fact about sizes of sets.

For $k \leq n$, how many of these outcomes result in one getting k heads?

It is "n choose k" or $\binom{n}{\quad} \quad \frac{n!}{}$

Skip to main content

Thankfully this is a built in macro in most modern computing systems:

```
import math

n = 10

k = 5

math.comb(n, k)
```

```
252
```

Sets and functions in different disciplines

Talking about Ω and its subsets

Different disciplines have different ways of talking about Ω and its relevant subsets:

	Philosophy	**Probability**	**Logic**	**Math**
Ω	Set of worlds	Sample space	Set of models	Underlying set
$\{A : A \subseteq \Omega\}$	Set of propositions	Event space	Sets determined by formulas	Powerset of underlying set

In particular: elements of Ω are called:

- *worlds* in philosophy
- elements of the *sample space* in probability

And subsets of Ω are called:

- *propositions* in philosophy
- *events* in probability

Skip to main content

Some more notes on philosophy terminology and concepts

One should beware philosophers following Davidson use "event" for more singular happenings: this specific touchdown in this specific game. Hence, if you are a philosopher, it is important for our purposes to get on board with the usage of "event" as "set of worlds."

A traditional statement of the propositions as "sets of worlds" is Stalnaker's Inquiry [Stalnaker, 1987]. Of course, philosophers have other concepts of propositions, such as "Russellian propositions," on which propositions are more like interpreted sentences of first-order logic.

Some more notes on logic and math terminology and concepts

If you are in certain parts of logic where a class of models (e.g. groups, rings, fields, partial orders, linear orders) is salient, then it's natural to consider the case where Ω is the class of models and where one restricts attention to subsets which satisfy some formula.

In pure mathematics, one often transitions more rapidly between different sets, and there it is useful to explicitly describe the *powerset operation* taking Ω to the set of all its subsets $\{A : A \subseteq \Omega\}$. The powerset is often written as $\mathcal{P}(\Omega) = \{A : A \subseteq \Omega\}$ or $\wp(\Omega) = \{A : A \subseteq \Omega\}$

A subtle point about when the sample space is infinite

In probability, there is a subtle point:

- When Ω is finite, it is okay to assume that the event space is literally $\{A : A \subseteq \Omega\}$.
- When Ω is infinite, we need to require that the event space is $\{A : A \subseteq \Omega \text{ is Borel }\}$ and that Ω is a well-behaved topological space.

We are largely going to ignore this point in what follows. Not because it is unimportant, but because the Borel sets are ample enough for practical purposes (we can reasonably be sure that we are always in there).

What are the Borel events?

The *Borel* events are the closure of the open sets under countable intersections, union, and relative

The *open sets in Euclidean space* are just the cirlces or balls with their edges removed.

This is the topic of study in measure theory and descriptive set theory.

Often we can ignore it in probability since we restrict ourselves to 'elementary' methods of set formation.

Adding random variables to the table

Finally, we can add to the list the notion of random variable:

	Philosophy	**Probability**	**Logic**	**Math**
Ω	Set of worlds	Sample space	Set of models	Underlying set
$\{A : A \subseteq \Omega\}$	Set of propositions	Event space	Sets determined by formulas	Powerset of underlying set
$X : \Omega \to \mathbb{R}$	na	Random variable	na	Real-valued function

We get to the definition in a moment.

For the moment, recall that $\mathbb{R}$ is just the *real numbers*, the set of all positive and negative numbers, including fractions and including numbers like $e = 2.71\ldots$ and $\pi = 3.14\ldots$.

Random variables

Definition (random variable)

Given sample space Ω, a *random variable* is a function $X : \Omega \to \mathbb{R}$.

When Ω is infinite, one needs to require more, and usually Borel measurability is enough, where this means inverse images of Borel sets are Borel

Skip to main content

Examples of random variables (coin flips)

$\Omega = \{0,1\}^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$

Let $X : \Omega \to \mathbb{R}$ by $X(\omega) = \#$ of heads in ω.

E.g. $X(001) = 1$ and $X(110) = 2$.

Let $Y : \Omega \to \mathbb{R}$ by $Y(\omega) =$ the result of the first flip.

E.g. $Y(001) = 0$ and $Y(110) = 1$.

Examples of random variables (dining experiences)

Often what we know about Ω is reflected in some salient random variables (for at least some small observed portion of it).

In the below example, Ω consists of 244 dining experiences, with seven random variables, corresponding to the features of the dining experience like total bill, the tip, etc.

▶ Show code cell source

```
     total_bill   tip     sex smoker   day    time  size
0         16.99  1.01  Female     No   Sun  Dinner     2
1         10.34  1.66    Male     No   Sun  Dinner     3
2         21.01  3.50    Male     No   Sun  Dinner     3
3         23.68  3.31    Male     No   Sun  Dinner     2
4         24.59  3.61  Female     No   Sun  Dinner     4
..          ...   ...     ...    ...   ...     ...   ...
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2

[244 rows x 7 columns]
```

Skip to main content

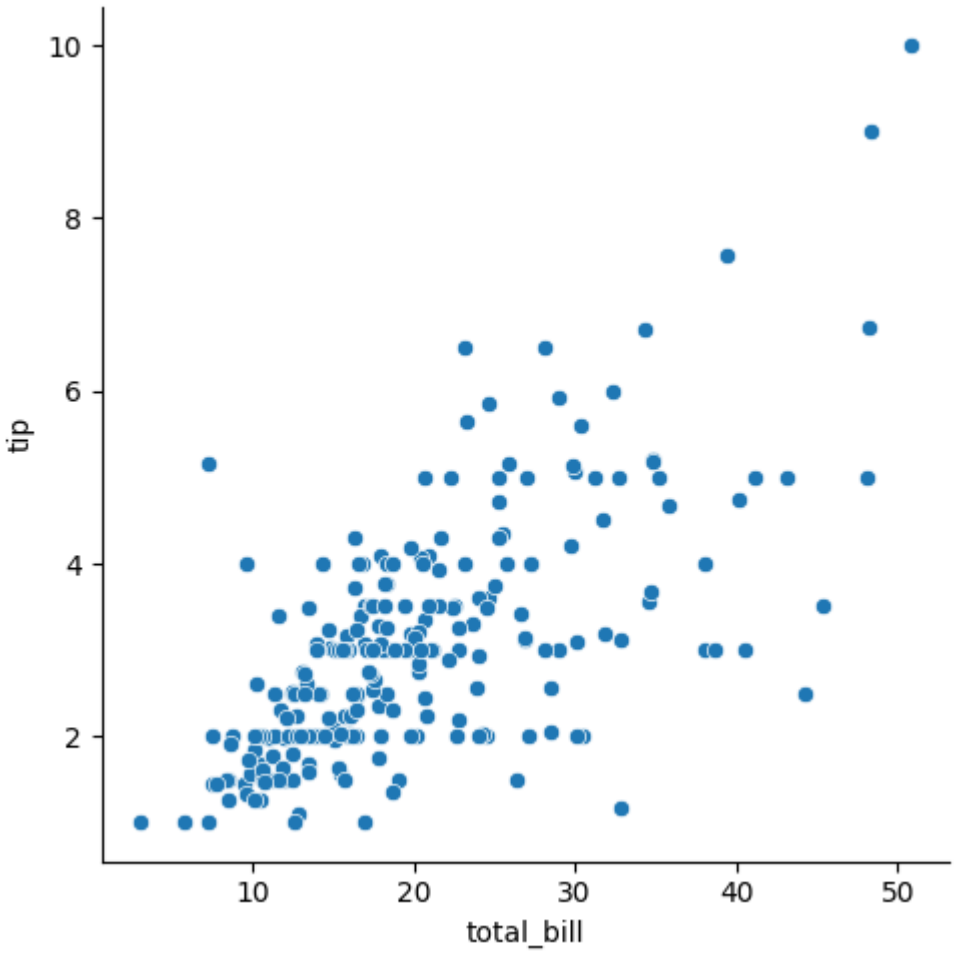
Visualizing two or more random variables

The usual ways of visualizing data sets are ways of visualizing random variables.

This scatterplot shows how to visualize the ‘total bill’ and ‘tip’ random variables

▶ Show code cell source

```
<seaborn.axisgrid.FacetGrid at 0x1580c5e90>
```



Number-theoretic functions applied to random variables

If $X, Y : \Omega \to \mathbb{R}$ are two random variables, then so are $X + Y$ and $X \cdot Y$, and ditto for basically any reasonable function from reals to reals.

Skip to main content

Example of number theoretic functions applied to random variables

$\Omega = \{0,1\}^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$

Let $X : \Omega \to \mathbb{R}$ by $X(\omega) = \#$ of heads in ω.

E.g. $X(001) = 1$ and $X(110) = 2$.

Let $Y : \Omega \to \mathbb{R}$ by $Y(\omega) =$ the result of the first flip.

E.g. $Y(001) = 0$ and $Y(110) = 1$.

Then $(X + Y)(001) = 1$ and $(X + Y)(110) = 3$.

Important idea (inducing propositions/events)

Suppose that $X : \Omega \to \mathbb{R}$ is a random variable.

Suppose that a is a real number.

Then $\{\omega \in \Omega : X(\omega) > a\}$ is a subset of Ω.

For instance, in our tip example, we could take X to be the total bill and a to be 40. We would get:

▶ Show code cell source

	total_bill	tip	sex	smoker	day	time	size
59	48.27	6.73	Male	No	Sat	Dinner	4
95	40.17	4.73	Male	Yes	Fri	Dinner	4
102	44.30	2.50	Female	Yes	Sat	Dinner	3
142	41.19	5.00	Male	No	Thur	Lunch	5
156	48.17	5.00	Male	No	Sun	Dinner	6
170	50.81	10.00	Male	Yes	Sat	Dinner	3
182	45.35	3.50	Male	Yes	Sun	Dinner	3
184	40.55	3.00	Male	Yes	Sun	Dinner	2
197	43.11	5.00	Female	Yes	Thur	Lunch	4
212	48.33	9.00	Male	No	Sat	Dinner	4

Skip to main content

How in general to use a random variable to generate events / propositions

For any subset $R \subseteq \mathbb{R}$, one has that $\{\omega \in \Omega : X(\omega) \in R\}$ is a subset of Ω, i.e. an event or proposition.

But this is kind of awkward looking, and so we often we just write $X \in R$ for this set.

Hence, continuing the earlier examples, we just speak of the event $X > a$, which just means the set $\{\omega \in \Omega : X(\omega) > a\}$.

Likewise, we can talk about the event $a < X + Y \leq b$. This just means the set $\{\omega : a < X(\omega) + Y(\omega) \leq b\}$.

One more note on notation in different disciplines

We can add this event / proposition generating to the table

	Philosophy	**Probability**	**Logic**	**Math**
Ω	Set of worlds	Sample space	Set of models	Underlying set
$\{A : A \subseteq \Omega\}$	Set of propositions	Event space	Sets determined by formulas	Powerset of underlying set
$X : \Omega \to \mathbb{R}$	na	Random variable	na	Real-valued function
$\{\omega \in \Omega : X(\omega) \in R\}$	na	the event $X \in R$	na	the set $X^{-1}(R)$, called the inverse image

In this last one, R is a subset of the reals $\mathbb{R}$

Skip to main content

Chapter 2

Probability axioms

Definition (probability axioms)

A *probability measure* P is a function from subsets of the space Ω to real numbers satisfying the following for all events $A, B \subseteq \Omega$:

- Non-negativity: $P(A) \geq 0$
- Finite additivity: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
- Value of entire space: $P(\Omega) = 1$

Again, if Ω is infinite, one will in general need to restrict attention to Borel events (or some other reasonably defined class of events). And again, we ignore this issue here, not because it is unimportant but because it is a slightly different subject (measure theory and/or descriptive set theory).

For rest of this section, we fix Ω and only consider events (sets) which are subsets of Ω

Proposition (value of complements)

For all events A, we have: $P(A^c) = 1 - P(A)$

Proof:

We always have $\Omega = (\Omega - A) \cup A$ and $\Omega - A$, A are disjoint (Draw the picture)

Then by finite additivity one has:

$$1 = P(\Omega) = P((\Omega - A) \cup A) = P(\Omega - A) + P(A) = P(A^c) + P(A)$$

Then subtract $P(A)$ from both sides.

Skip to main content

Corollary (value for emptyset)

We have $P(\emptyset) = 0$

Proof:

Since $\emptyset^c = \Omega$ we have

$P(\emptyset^c) = 1 - P(\Omega) = 1 - 1 = 0.$

Proposition (monotonicity)

For all events A, B if $A \subseteq B$ then $P(A) \leq P(B)$

Proof:

When $A \subseteq B$ we have $B = A \cup (B - A)$ and $A, B - A$ are disjoint. (Draw the picture)

We then appeal to finite monotoncity as follows:

$P(B) = P(A) + P(B - A) \geq P(A)$

For the last inequality, we appeal to non-negativity.

Corollary (values in the interval 0 to 1)

For all events A, we have $0 \leq P(A) \leq 1.$

Proof:

We always have $A \subseteq \Omega$. Hence by monotonicity $P(A) \leq P(\Omega) = 1.$

Proposition (finite additivity, redeux)

For pairwise disjoint events A, B, C one has

Skip to main content

Proof:

If A, B, C are pairwise disjoint, then $A, B \cup C$ are disjoint too (draw the picture). Then by two applications of finite additivity we have

$$P(A \cup B \cup C) = P(A) + P(B \cup C) = P(A) + P(B) + P(C)$$

Note 1:

In this proof, we are relying on rules of sets, such as associativity of union, which says that $(A \cup B) \cup C = A \cup (B \cup C)$, and so we can "drop parentheses" and just write it as $A \cup B \cup C$.

Note 2:

The same proposition holds for any finite sequence of pairwise disjoint events.

Proposition (inclusion-exclusion)

For all events A, B we have

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

Proof (less detailed):

$$\begin{aligned} P(A \cup B) &= P(A - B) + P(B - A) + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Proof (more detailed):

We have that $A \cup B = (A - B) \cup (B - A) \cup (A \cap B)$ and $A - B, B - A, A \cap B$ are are pairwise disjoint (draw the picture)

Hence by finite additivity we have

(1) $P(A \cup B) = P(A - B) + P(B - A) + P(A \cap B)$

Skip to main content

But we also have $A = (A - B) \cup (A \cap B)$ and $B = (B - A) \cup (A \cap B)$ and the relevant sets are disjoint (draw the picture)

Hence by finite additivity and some subtraction we get:

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

(2′) $P(A - B) = P(A) - P(A \cap B)$

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

(3′) $P(B - A) = P(B) - P(A \cap B)$

We then just chain together (1), (2′), (3′) to get

$$\begin{aligned} P(A \cup B) &= P(A - B) + P(B - A) + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Conditioning and conditionals

Definition (Conditional probability)

Suppose that P is a probability measure and $P(E) > 0$. Then the *conditional probability* $P(H \mid E)$ of H given E is $$P(H \mid E) = \frac{P(H \cap E)}{P(E)}$$

Alterate notation (Subscript notation for conditional probability)

We also write: $P_E(H) = P(H|E)$.

Convention (Assume your interlocutor is not dividing by zero)

Whenever anyone writes $P(H|E)$ or $P_E(H)$, assume that they are restricting their statement to the case where $P(E) > 0$.

Skip to main content

This saves us from needing to write out this hypothesis over and over again.

Proposition (repeated conditioning)

$P_E(H \mid E') = P(H \mid E \cap E')$

Proof:

$$P_E(H \mid E') = \frac{P_E(H \cap E')}{P_E(E')} = \frac{P(H \cap E' \cap E)/P(E)}{P(E \cap E')/P(E)} = \frac{P(H \cap E' \cap E)}{P(E \cap E')} = P(H \mid E \cap E$$

Proposition (conditioning induces a probability measure)

For all E, one has that P_E is also a probability measure.

Proof: Non-negativity: $P_E(A) = \frac{P(A \cap E)}{P(E)} \geq 0$.

Finite additivity: assuming A, B are disjoint, we have

$$P_E(A) + P_E(B) = \frac{P(A \cap E)}{P(E)} + \frac{P(B \cap E)}{P(E)} = \frac{P((A \cap E) \cup (B \cap E))}{P(E)} = \frac{P((A \cup B) \cap E)}{P(E)} = P_E(A \cup B)$$

The first identity is definition, the second is $A \cap E, B \cap E$ disjoint (draw picture); the third identity is distributivity; the last identity is definition.

Value of the whole space: $P_E(\Omega) = \frac{P(\Omega \cap E)}{P(E)} = \frac{P(E)}{P(E)} = 1$ since $\Omega \cap E = \Omega$.

Theorem (Bayes' theorem / formula)

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

Proof:

Bayes' formula can be verified by moving from the right-to-left, with the probabilities of the

$$\frac{P(E \mid H) \cdot P(H)}{P(E)} = \frac{P(E \cap H) \cdot P(H)}{P(H) \cdot P(E)} = \frac{P(H \cap E)}{P(E)} = P(H \mid E)$$

Definition (Likelihood, Prior, Posterior)

In the context of Bayes' Theorem, one has names for the certain quantities:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}, \qquad \text{posterior probability} = \frac{\text{likelihood} \times \text{prior proba}}{\text{probability of evide}}$$

In English: the probability of a hypothesis conditional on evidence is equal to the likelihood of the evidence given the hypothesis times the prior probability associated to the hypothesis, divided by the probability associated to the evidence.

It is useful in situations where one has some sense already of which hypotheses make the evidence at hand more or less probable.

Proposition (a useful formula)

$P(H) = P(H|E) \cdot P(E) + P(H|E^c) \cdot P(E^c)$

Proof:

We have $H = (H \cap E) \cup (H \cap E^c)$ and $H \cap E, H \cap E^c$ are disjoint.

Then $P(H) = P(H \cap E) + P(H \cap E^c) = P(H|E) \cdot P(E) + P(H|E^c) \cdot P(E^c)$

Theorem (Lewis-Stalnaker trivality)

Suppose that conditional probability was *factive*, in that there is a binary operation $\Rightarrow$ on subsets of Ω such that $P(H|E) = P(E \Rightarrow H)$ for all P, H, E with $P(E) > 0$.

Then conditional probability is *trivial* in that $P(H|E) = P(H)$ for all E, H with $P(H \cap E), P(E \cap H^c) > 0$.

Skip to main content

Proof:

$$P(H|E) = P(E \Rightarrow H) = P(E \Rightarrow H \mid H) \cdot P(H) + P(E \Rightarrow H \mid H^c) \cdot P(H^c)$$

$$= P_H(E \Rightarrow H) \cdot P(H) + P_{H^c}(E \Rightarrow H) \cdot P(H^c)$$

$$= P_H(H|E) \cdot P(H) + P_{H^c}(H|E) \cdot P(H^c)$$

$$= P(H|E \cap H) \cdot P(H) + P(H|E \cap H^c) \cdot P(H^c)$$ by repeated conditioning

$$= 1 \cdot P(H) + 0 \cdot P(H^c)$$

$$= P(H)$$

Note: next week we will recognize the 'triviality' as a kind of 'too much independence.'

Interpretations of probability

Laplace and the principle of indifference

From the *Philosophy Essay on Probabilities* of 1825:

> The theory of chances consists in reducing all events of the same kind to a certain number of equally possible cases, that is to say, to cases whose existence we are equally uncertain of, and in determining the number of cases favourable to the event whose probability is sought. The ratio of this number to that of all possible cases is the measure of this probability, which is thus only a fraction whose numerator is the number of favourable cases, and whose denominator is the number of all possible cases ([Dale and Laplace, 1995] p. 4)

It seems he is thus interested in what we would call today the *uniform* measure on a finite space.

If Ω has cardinality n and an event A has m elements, then we define $P(A) = \frac{m}{n}$.

If we use $|\cdot|$ for cardinality (or size), then this can be written as $P(A) = \frac{|A|}{|\Omega|}$.

Virtues: connects our epistemic state to the probabilites, and gets simple cases of fair coin flips

Vices: more often than not, we are not interested in the uniform measure.

von Mises and frequentism

The canonical expression of this is in von Mises' *Probability, Statistics, and Truth* [von Mises, 1957], originally published in 1928. The idea is that given an infinite sequence of observations, the probability of an event can be defined to the limiting relative frequence of the number of times it occured in the sequence.

Formally, given a space Ω, consider the "superspace" $\Omega^{\mathbb{N}} = \{f : \mathbb{N} \to \Omega\}$, i.e. the set of all functions from the natural numbers to Ω.

For instance if $\Omega = \{0, 1\}$, then $\Omega^{\mathbb{N}}$ is *Cantor space*, the space of all infinite sequences of zeros and ones. This is a good initial representation of the space of all infinite sequences of coin flips.

Then for such a function f, define $P_f(A) = \lim_n \frac{|\{m \leq n : f(m) \in A\}|}{n}$.

That is, it is the limiting relative frequency of A's that appear in the sequence f.

Virtues of frequentism

One can derive the elementary laws of probability.

For instance, if A, B are disjoint and the relevant limits exist, then

$$\begin{aligned} P_f(A \cup B) &= \lim_n \frac{|\{m \leq n : f(m) \in A \cup B\}|}{n}. \\ &= \lim_n \frac{|\{m \leq n : f(m) \in A\}| + |\{m \leq n : f(m) \in B\}|}{n} \\ &= \lim_n \frac{|\{m \leq n : f(m) \in A\}| + \lim_n |\{m \leq n : f(m) \in B\}|}{n} \\ &= P_f(A) + P_f(B) \end{aligned}$$

Vices of frequentism

One has to assume that the limits exist and von Mises postualtes this rather than explains it. Some

followed by long stretches where the frequency is $\frac{2}{3}$, repeated ad nauseum.

Subjective Bayesianism

De Finetti described his own view as *subjectivism*, but it is now usually called just *Bayesianism* or perhaps sometimes *subjective Bayesianism*.

The basic idea is that probabilities are just degrees of confidence or degrees of belief: they are reflective of the subjective evaluations of agents.

Different subjective Bayesians will differ in terms of how they understand the "reflectiveness", and can range the gamut from pyschologically real states of agents to the best way to understand from an external point of view the bevahior of an agent.

De Finetti describes these views in his [De Finetti, 1964], originally published in 1937.

de Finetti vs von Mises

Both von Mises and de Finetti were primarily concerned to give a scientifically acceptable account of probability.

Von Mises is often counted as a member of the Vienna circle.

Similarly, de Finetti speaks of his motivation as stemming from a view that "Every notion is only a word without meaning so long as it is not known how to verify practically any statement at all where this notion comes up ..." ({cite}`De_Finetti1964-rr p. 148).

For von Mises, probability was appropriately scientific because "the relative frequency of the repetition is the 'measure' of probability, just as the length of a column of mercury is the 'measure' of temperature" ([von Mises, 1957] p. vi).

For de Finetti, probability was operationalized in terms of human belief.

Virtues of subjective Bayesianism

There are arguments from betting behavior (Dutch books) that one's degrees of belief satisfy the

Similarly, in the finite case, there are arguments from accuracy, to the effect that any violations of the laws of probability would make one's credences less close to the truth than they otherwise could be.

These are really subjects for another course.

Vices of subjective Bayesianism

It is harder to see how one is going to get relative frequences to emerge out of this (if one thought that was a good thing). But Bayesians have some things to say about this (exchangeability)

The Bayesian updating procedure has a hard time with the problem of "old evidence": namely the problem of how to account for how a theory's ability to explain an old already-updated-upon piece of evidence lends credence to the theory ([Glymour, 1980])

Chapter 3

Recalling notation

In what follows, we assume the following:

- Ω is the sample space, that is the set of worlds
- $\{A : A \subseteq \Omega\}$ is the event space, that is the set of propositions
- $X : \Omega \to \mathbb{R}$ is a random variable, that is the means by which we generate propositions of interest to us
- P is a probability measure on the event space

For concreteness, one can take:

- Ω is the set of people under consideration e.g. in the study
- $\{A : A \subseteq \Omega\}$ are the properties of the people under consideration e.g. old/young, tall/short
- $X : \Omega \to \mathbb{R}$ is a measurement we are taking of the people under consideration, e.g. age or height.
- P records the frequency of the properties among the people under consideration (if you are a

consideration are (if you are a Bayesian)

Likewies, for concreteness, one can take:

- Ω is the set of cities under consideration e.g. in the study
- $\{A : A \subseteq \Omega\}$ are the properties of the cities under consideration e.g. big/small, rich/poor, etc.
- $X : \Omega \to \mathbb{R}$ is a measurement we are taking of the people under consideration, e.g. population size or median income
- P records the frequency of the properties among the cities under consideration (if you are a frequentist); or our beliefs about how probable properties among the cities under consideration are (if you are a Bayesian)

Five definitions

In statistics, we rarely interact directly with a probability measure.

Rather everything goes through random avariables.

Definition (cdfs)

The *cumulative density function (cdf)* of random variable X relative to proability measure P is the function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(x) = P(X \leq x)$$

Recall that $X \leq x$ is an abbreviation for the event $\{\omega \in \Omega : X(\omega) \leq x\}$

When X is clear from context, we just write F instead of F_X.

In words: $F(x)$ is the probability that the random variable X takes value $\leq x$

In our people example, it might be the probability that a person has height ≤ 6 ft.

In our city example, it might be the probabiliy that a city has population $\leq$ 50,000

What the units are on the real numbers (feet, number of people) will be made clear by context.

Skip to main content

Main idea (cdfs)

The cdf answers the question: how probable is it that the random variable has outcome less-than-or-equal to a given number?

E.g. how probable is it that a person has height less-than-or-equal to a given number?

E.g. how probable is it that a city has population less-than-or-equal to a given number?

Definition (ccdfs)

The *complementary cumulative density function (cdf)* of random variable X relative to proability measure P is the function $\overline{F}_X : \mathbb{R} \to [0,1]$ given by

$$\overline{F}_X(x) = P(X > x)$$

Recall that $X > x$ is an abbreviation for the event $\{\omega \in \Omega : X(\omega) > x\}$

When X is clear from context, we just write $\overline{F}$ instead of $\overline{F}_X$.

Main idea (ccdfs)

The ccdf answers the question: how probable is it that the random variable has outcome greater than a given number?

E.g. how probable is it that a person has height greater than a given number?

E.g. how probable is it that a city has population greater than a given number?

Proposition (how cdfs and ccdfs relate)

When $F_X(x)$ is close to zero, $\overline{F}_X(x)$ is close to one.

When $F_X(x)$ is close to one, $\overline{F}_X(x)$ is close to zero.

Skip to main content

More generally, both these numbers are in the inteval from 0 to 1, and

$$F_X(x) + \overline{F}_X(x) = 1$$

Proof:

Note that by rules of probability one has

$$F_X(x) + \overline{F}_X(x) = P(X \leq x) + P(X > x) = P(\Omega) = 1$$

Motivating discrete pdfs

Some random variables take only natural number values, e.g. population sizes (you can't have city with 531.5 people)

Suppose that X was one of those cases, and consider $x = 2$.

Then one has $
$\{\omega \in \Omega : X(\omega) \leq 2\} = \{\omega \in \Omega : X(\omega) = 0\} \cup \{\omega \in \Omega : X(\omega) = 1\} \cup \{\omega \in \Omega : X(\omega) = 2\}$
$

Further, these three events are pairwise disjoint. Hence one has

$$F_X(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

More generally, one has the following

$$F_X(x) = \sum_{k \leq x} P(X = k)$$

The only thing special about natural number values here is that once x is fixed, there are only finitely many options $\leq x$.

Skip to main content

Definition (discrete pdfs)

If X takes finite or countably infinite many values, then the *probability density function (pdf)* of X is

$$f_X(x) = P(X = x)$$

Recall that $X = x$ is an abbreviation for the event $\{\omega \in \Omega : X(\omega) = x\}$

When X is clear from context, we just write f instead of f_X

It has the feature that when X takes finitely many values or natural nubmer values one has that:

$$F_X(x) = \sum_{k \leq x} f_X(k)$$

In the discrete case, the pdf is also called the *pmf* (for *probability mass function*)

Main idea (pdf)

The pdf answers the question: how probable is it that the random variable has outcome exactly equal to a given number?

E.g. how probable is it that a person has height exactly equal to a given number?

E.g. how probable is it that a city has population exactly equal to a given number?

Definition (expectation)

If X takes finite or countably infinite many values, then we define its *expectation* to be

$$\mathbb{E}X = \sum_x x \cdot f_X(x) = \sum_x x \cdot P(X = x)$$

Skip to main content

The expectation is also designated as μ, the Greek letter 'm', which is a mnemonic for 'mean' or 'average.'

Example (moral: expectation is generalization of average)

To repeat, the quantity is:

$$\mathbb{E}X = \sum_x x \cdot f_X(x) = \sum_x x \cdot P(X = x)$$

Suppose that $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and ω_1 is 5 feet tall and ω_2 is 6 feet tall and ω_3 is 6 feet tall, and each person is equally likely, and $X(\omega) =$ height of person ω.

Then $P(X = 5) = \frac{1}{3}$ and $P(X = 6) = \frac{2}{3}$

and

$$\mathbb{E}X = \sum_x x \cdot P(X = x) = 5 \cdot \frac{1}{3} + 6 \cdot \frac{2}{3} = 5.67$$

One recognizes this as the average:

$$\frac{5 + 6 + 6}{3} = 5.67$$

Hence *expectation is a generalization of the average*.

Example (expected utility)

To repeat, the quantity is:

$$\mathbb{E}X = \sum_x x \cdot f_X(x) = \sum_x x \cdot P(X = x)$$

Philosophers will recognize this quantity in the case where P reflects beliefs and $X(\omega)$ is the utility

Skip to main content

action.

In this case, the expectation is the *expected utility*: the utility associated with each outcome times one's belief about how probable the outcome is, all added up together.

Main idea (expectation)

It is a generalization of averages.

It is one answer to the question: what should you think that the value of the random variable is going to be?

Definition (variance)

We define the variance of X to be

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$$

A little algebra indicates that it can also be written as

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

Intutively, variance is saying how much one expects the value to differ from its average. We will explore more with this quantity next time, and discuss more of its rules next time.

The variance is also abbreviated as σ^2

We will talk more about the main idea next time.

Bernoulli

Definition (Bernoulli distribution)

We want a high-level compact description of when a random variable is recording the outcome of a

We say random variable X has the *Bernoulli distribution* with probability p, abbreviated $X \sim \mathrm{Bern}(p)$, if $0 \leq p \leq 1$ and X has two outcomes 0 and 1 with probabilities and

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p$$

This is the coin flip sitatuion with a biased coin of probability p.

One has

- expectation: $\mathbb{E}X = \mu = p$
- variance: $\mathrm{Var}(X) = \sigma^2 = p(1-p)$
- pdf: $f_X(0) = 1 - p$, $f_X(1) = p$, in a single formula $f_X(x) = p^x \cdot (1-p)^{1-x}$
- cdf: $F_X(0) = 1 - p$, $F_X(1) = 1$
- ccdf: $\overline{F}_X(0) = p$, $\overline{F}_X(1) = 0$.

Binomial

Definition (Binomial distribution)

We want a high-level compact description of when a random variable is recording the number of successes in n-many independent Bernoulli trials with probability p (we will talk about independence next time).

In this case, we say that X has the *Binomial* (n, p) *distribution*, abbreviated $X \sim \mathrm{Binom}(n, p)$.

One has

- expectation: $\mathbb{E}X = np$
- variance: $\mathrm{Var}(X) = np(1-p)$
- pdf: $f_X(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$
- cdf: no easy formula
- ccdf: no easy formula

The pdf, cdf, and ccdf are best visualized in next slides. But first we do an example.

Skip to main content

Example (procedures)

Again, we say that random variable X has the *Binomial* (n, p) *distribution*, abbreviated $X \sim \text{Binom}(n, p)$, when it records the the number of successes in n-many independent Bernoulli trials with probability p.

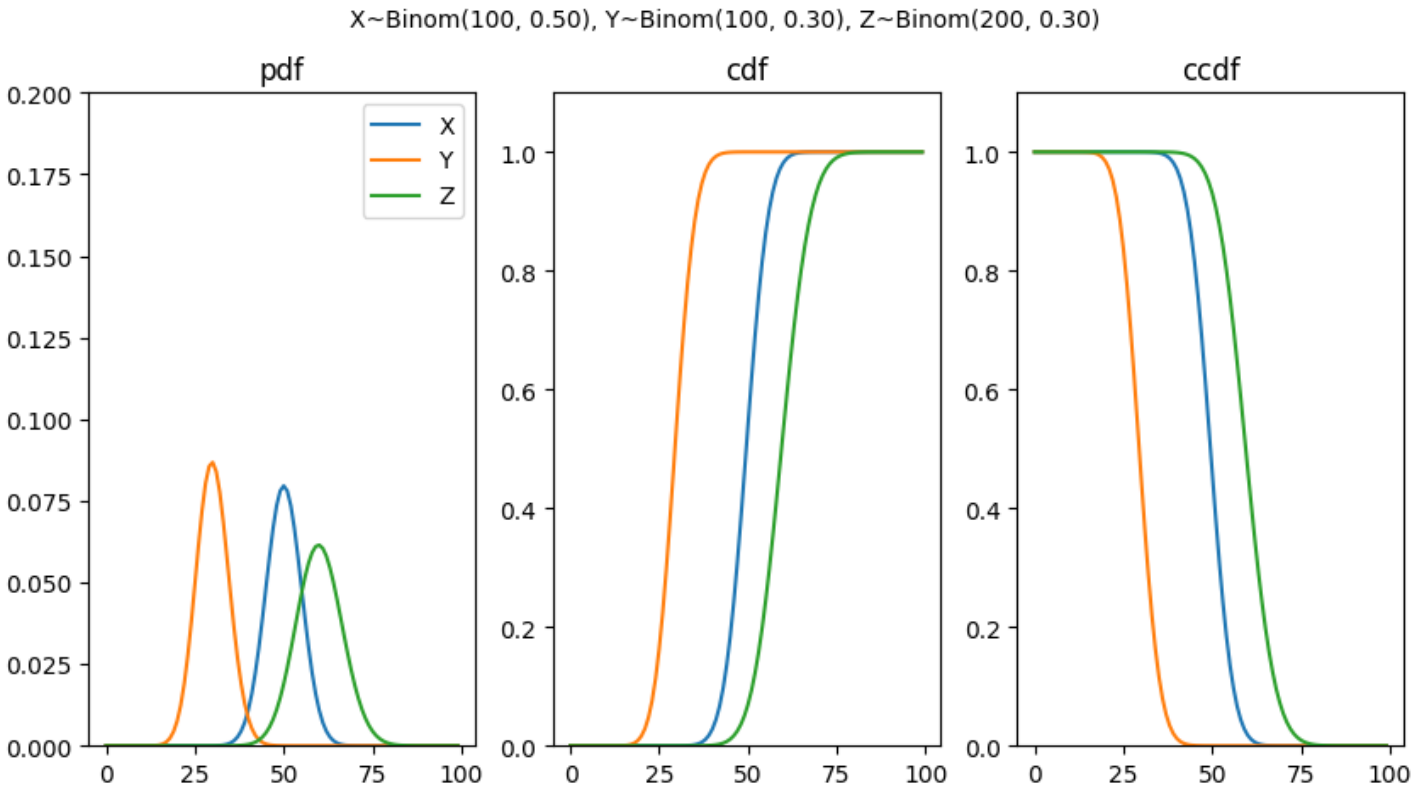
An example is: you have some procedure which has a certain probability p of success each time, and you want to know what the probability is that you will have, say, more than k many successs when you do the procedure n many times. The answer is given by the ccdf $\overline{F}_X(k) = P(X > k)$.

Visualizing Binomial distributions

```
## visualizing X ~ Binom(n,p), Y~Binom(m,q), Z~Binom(l,r)
## pdf, cdf, ccdf

n = 100   # sbet value of n, number of trials for X
p = .5 # set value of p, probability of trial for X

m = 100   # set value of m, number of trials for Y
q = .3 # set value of q, probability of trial for Y

l =  200  # set value of l, number of trials for Z
r = .3 # set value of r, probability of trial for Z
```

▶ Show code cell source

Skip to main content



Computing Binomial distributions

```
## computing X ~ Binom(n,p)
## pdf, cdf, ccdf

n = 100   # set value of n, number of trials

p = .5 # set value of p, probability of trial

k = 50 # set value of number of successes
```

▶ Show code cell source

```
Assuming that
X~Binom(100, 0.50)
k=50
We then obtain
f(k)=0.080
F(k)=0.540
F̄(k)=0.460
```

Skip to main content

Poisson

Motivation (Poisson distribution)

You run a store.

Experience suggests that in $n = 3600$ seconds (aka one hour) you get four customers. Since $p = \frac{4}{3600} \approx .0011$, you could conceive of this with sample space $\Omega = \{1, 2, \ldots, 3600\}$ with random variable $X(\omega)$= number of cumstomers you got up to second ω as $X \sim \text{Binom}(3600, .0011)$.

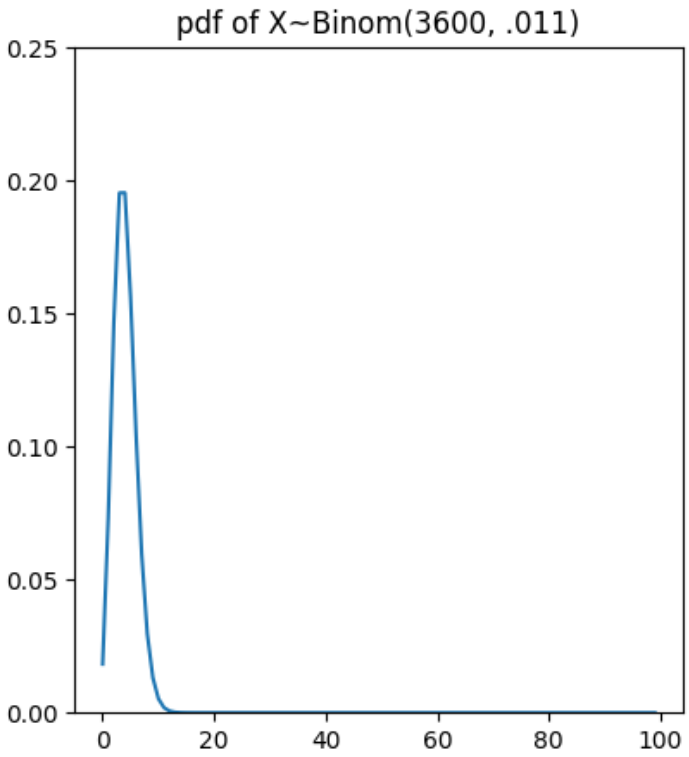
Experience equally suggests that in $n = 60$ minutes (aka one hour) you get four customers. Since $p = \frac{4}{60} \approx .067$, you could conceive of this with sample space $\Omega = \{1, 2, \ldots, 60\}$ with random variable $Y(\omega)$= number of cumstomers you got up to minute ω as $Y \sim \text{Binom}(60, .067)$.

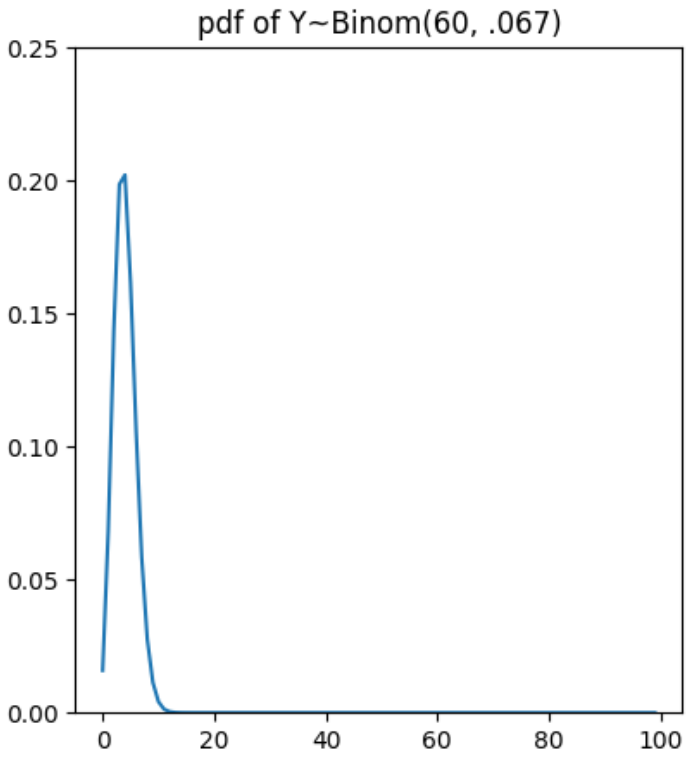
And when you look at $X \sim \text{Binom}(3600, .0011)$ and $Y \sim \text{Binom}(60, .067)$, they don't seem that different.

It would be nice *not* to have to include the first number and just make it a function of the rate.

▶ Show code cell source

Skip to main content





Definition (Poisson distribution)

We want a high-level compact description of when a random variable records the successes in a large number n of Bernoulli trials with small propability p of occuring, where all that matters is the rate $\lambda = np$.

In this case, we say that X has the *Poisson λ distribution*, abbreviated $X \sim \mathrm{Pois}(\lambda)$.

One has

- expectation: $\mathbb{E}X = \lambda$
- variance: $\mathrm{Var}(X) = \lambda$
- pdf: $f_X(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$
- cdf: no easy formula
- ccdf: no easy formula

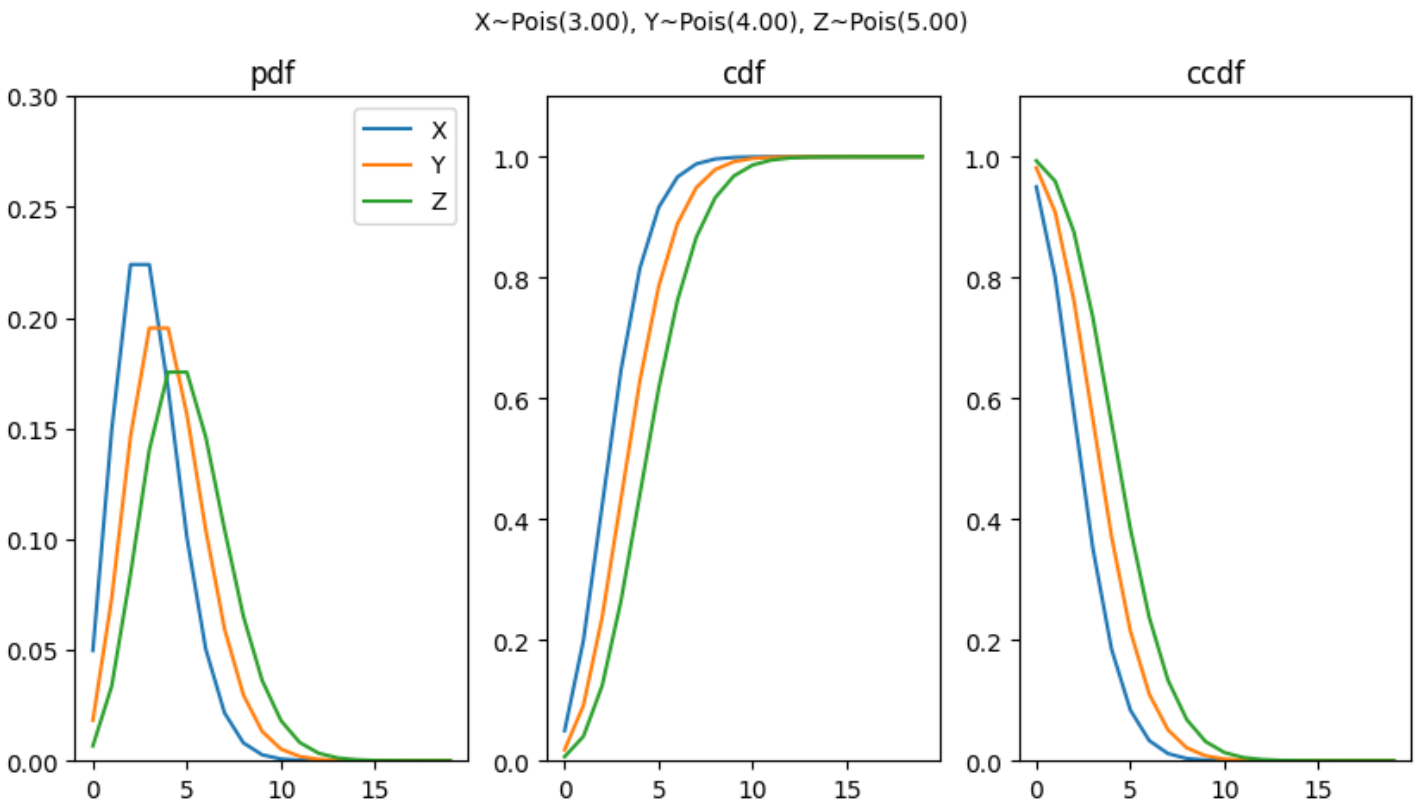
The pdf, cdf, and ccdf are best visualized in next slides.

Skip to main content

Visualizing Poisson distributions

```
# Visualizing X~Pois(λ_1), Y~Pois(λ_2), Z~Pois(λ_3)
# pdf, cdf, ccdf

λ_1 = 3   # set rate of X

λ_2 = 4   # set rate of Y

λ_3 = 5   # set rate of Z
```

▶ Show code cell source



Skip to main content

Computing with Poisson distributions

```
# Computing X~Pois(λ)
# pdf, cdf, ccdf

λ = 3   # set rate

k = 2   # set number of successes
```

▶ Show code cell source

```
Assuming that
X~Pois(3.00)
k=2
We then obtain
f(k)=0.224
F(k)=0.423
F̄(k)=0.577
```

Uniform discrete

Definition (uniform discrete)

We want a high-level compact description of when a random variable is recording outcomes in a finite set of real numbers, each one of which is equally probable.

We may assume that the finite set has n elements and has the form $[a, b] = \{a, a + 1, \ldots, b\}$, where $b = a + n - 1$.

We say random variable X has the *Uniform distribution* in the set $[a, b]$, abbreviated $X \sim \mathrm{U}\{a, b\}$ if for each k in the set one has

$$P(X = k) = \frac{1}{n}$$

One has

Skip to main content

- expectation: $\mathbb{E}X = \frac{a+b}{n}$
- variance: $\mathrm{Var}(X) = \frac{n^2-1}{12}$
- pdf: $f_X(k) = \frac{1}{n}$
- cdf: $F_X(k) = \frac{k-a+1}{n}$
- ccdf: $\overline{F}_X(k) = \frac{b-k}{n}$.

Visualizing uniform discrete distributions

```
# Visualizing X~U{a_1,b_1}, Y~U{a_2,b_2}, Z~U{a_3,b_3},
# pdf, cdf, ccdf

a_1 = 1   # set lower bound a_1 for X
b_1 = 5  # set upper bound b_1 for X
n_1 = b_1-a_1+1   # number of values for X

a_2 = 3   # set lower bound a_2 for Y
b_2 = 6  # set upper bound b_2 for Y
n_2 = b_2-a_2+1   # number of values for Y

a_3 = 5   # set lower bound a_3 for Z
b_3 = 8  # set upper bound b_3 for Z
n_3 = b_3-a_3+1   # number of values for X
```

▶ Show code cell content

Computing with uniform discrete distributions

```
a = 3 # lower bound of X~U{a,b}

b = 11 # upper bound of X~U{a,b}

n = b-a+1 # number of values for X~U{a,b}

k = 6 # set outcome you are interested in
```

▶ Show code cell source

Skip to main content

```
Assuming that
X~U{3, 11}
k=6
We then obtain
f(k)=0.111
F(k)=0.444
F̄(k)=0.556
```

Chapter 4

Rules for expectation

Definition (expectation, redeux)

Recall the following definition from last time, where today we restrict to the finite case:

If X takes finitely many values, then we define its *expectation* to be

$$\mathbb{E}X = \sum_x x \cdot f_X(x) = \sum_x x \cdot P(X = x)$$

The summation is over whatever the values are that the random variable takes.

The expectation is also designated as μ, the Greek letter 'm', which is a mnemonic for 'mean' or 'average.'

The vivification

We need some simple examples which can make these notions and rules vivid to us.

Suppose that the worlds in Ω represent your possible futures.

Suppose that X, Y are two random variables such that:

Skip to main content

- if you choose to act in a Y-like way now, then $Y(\omega)$ represents your earnings in ω.

You decide on a policy of choosing the random variable which has the larger expectation.

Hence you will be in the market for some simple rules to calculate expectations, since you want to actually get to the numbers to be able to compare them.

Example (expectation of Bernoulli)

Let $X \sim \mathrm{Bern}(p)$. Then $\mathbb{E}X = p$.

Proof: by definition (from last time), a random variable X of this kind only two values:

- value 1, with probability p.
- value 0, with probability $1-p$.

Hence we have

$$\mathbb{E}X = \sum_x x \cdot P(X=x) = 1 \cdot P(X=1) + 0 \cdot P(X=0) = 1 \cdot p + 0 \cdot (1-p) = p$$

This was mentioned last time but we prove it now.

Returning to the vivification

Suppose you are choosing between

$X \sim \mathrm{Bern}(.7)$ and $Y \sim \mathrm{Bern}(.9)$.

By previous, one has $\mathbb{E}X = .7$ and $\mathbb{E}Y = .9$.

Hence, you would choose Y.

(This example is less than great since the numbers are tiny; we will revisit it in a few minutes)

The following alternate expressions is occasionally useful for developing the theory (although less than useful in practice):

Skip to main content

Proposition (alternate characterization of expectation)

Suppose that $\Omega = \{\omega_1, \ldots, \omega_n\}$, where we enumerate these without repetition. Then

$\mathbb{E}X = \sum_{i=1}^{n} X(\omega_i) \cdot P(\{\omega_i\})$.

In this expression $\{\omega_i\}$ refers to the event of "being equal to ω_i." That is, it is a singleton set or event.

If you get this illustration, you'll see the proof of the proposition:

Illustration

Suppose that $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and $X(\omega_1) = 5$ and $X(\omega_2) = 6$ and $X(\omega_3) = 5$.

Then $P(X = 5) = P(\{\omega_1, \omega_3\})$ and $P(X = 6) = P(\{\omega_2\})$. Hence one has

$$\begin{aligned}
\mathbb{E}X &= 5 \cdot P(X = 5) + 6 \cdot P(X = 6) \\
&= 5 \cdot P(\{\omega_1, \omega_3\}) + 6 \cdot P(\{\omega_2\}) \\
&= 5 \cdot \big(P(\{\omega_1\}) + \cdot P(\{\omega_3\})\big) + 6 \cdot P(\{\omega_2\}) \\
&= 5 \cdot P(\{\omega_1\}) + 5 \cdot P(\{\omega_3\}) + 6 \cdot P(\{\omega_2\}) \\
&= 5 \cdot P(\{\omega_1\}) + 6 \cdot P(\{\omega_2\}) + 5 \cdot P(\{\omega_3\}) \\
&= X(\omega_1) \cdot P(\{\omega_1\} + X(\omega_2) \cdot P(\{\omega_2\} + X(\omega_3) \cdot P(\{\omega_3\} \\
&= \sum_{i=1}^{3} X(\omega_i) \cdot P(\{\omega_i\})
\end{aligned}$$

Most everything abstract that one wants to do with expectations follow from the following rules:

Proposition (rules for expectation)

Suppose that X, Y are random variables and c is a real number. Then:

Skip to main content

2. $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}X$
3. $\mathbb{E}c = c$ (in writing $\mathbb{E}c$ are thinking of c as the constant function which always outputs c).
4. if $X \geq 0$ everywhere, then $\mathbb{E}X \geq 0$.
5. If $X \geq Y$ everywhere, then $\mathbb{E}X \geq \mathbb{E}Y$.

Returning to the vivification again

Suppose you are choosing between $X = 100 \cdot X'$ and $Y = 100 \cdot Y'$

where $X' \sim \text{Bern}(.7)$ and $Y' \sim \text{Bern}(.9)$.

By previous, one has $\mathbb{E}X = 100 \cdot .7 = 70$ and $\mathbb{E}Y = 100 \cdot .9 = 90$.

Hence, you would choose Y.

Proof of rule 1

Prove: $\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$

Proof: Let $Z = X + Y$ and use the alternate characterization

$$\begin{aligned}\mathbb{E}[X+Y] = \mathbb{E}[Z] &= \sum_{i=1}^{n} Z(\omega_i) \cdot P(\{\omega_i\}) \\ &= \sum_{i=1}^{n} (X(\omega_i) + Y(\omega_i)) \cdot P(\{\omega_i\}) \\ &= \sum_{i=1}^{n} X(\omega_i) \cdot P(\{\omega_i\}) + \sum_{i=1}^{n} Y(\omega_i) \cdot P(\{\omega_i\}) \\ &= \mathbb{E}[X] + \mathbb{E}[Y]\end{aligned}$$

Proof of rule 2

Prove: $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}X$

Proof: use the alternate characterization

$\mathbb{E}[c \cdot X] = \sum_{i=1}^{n} c \cdot X(\omega_i) \cdot P(\{\omega_i\}) = c \cdot \sum_{i=1}^{n} X(\omega_i) \cdot P(\{\omega_i\}) = c \cdot \mathbb{E}[X]$.

Skip to main content

Proof of rule 3

Prove: $\mathbb{E}c = c$

Proof: use the alternate characterization as follows:
$\mathbb{E}c = \sum_{i=1}^{n} c \cdot P(\{\omega_i\}) = c \cdot \sum_{i=1}^{n} P(\{\omega_i\}) = c \cdot P(\Omega) = c$

Proof of rule 4

Prove: if $X \geq 0$ everywhere, then $\mathbb{E}X \geq 0$.

Proof: this is obvious from any characterization since it is a sum of non-negative numbers.

Proof of rule 5

Prove: if $X \geq Y$ everywhere, then $\mathbb{E}X \geq \mathbb{E}Y$.

Proof: use rules 1,4 and the fact that $X - Y \geq 0$ eveywhere to conclude that
$\mathbb{E}X = \mathbb{E}[X - Y + Y] = \mathbb{E}[X - Y] + \mathbb{E}[Y] \geq \mathbb{E}[Y]$.

Example (expectation of Binomial)

Suppose that $X \sim \mathrm{Binom}(n, p)$ Then $\mathbb{E}X = n \cdot p$.

Proof:

Since X is just the number of successes (1's) in n-many independent Bernoulli trials $X_1, \ldots, X_n$, we have

that $X = \sum_{i=1}^{n} X_i$. Then we have $\mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = n \cdot p$.

Returning to the vivification again

Suppose you are choosing between

Skip to main content

$X \sim \text{Binom}(3, .5)$ and $Y \sim \text{Binom}(2, .7)$.

E.g. $X \sim$ "three trials with fair chances" vs. $Y \sim$ "two trials with more favorable chances."

By previous, one has $\mathbb{E}X = 1.5$ and $\mathbb{E}Y = 1.4$.

Hence, you would choose X.

(To make the numbers bigger, multiply each by a big constant).

Computing the expectation of some binomials

Given $X \sim Binom(n, p)$ and $Y \sim Binom(m, q)$, consider
$Z_1 \sim Binom((n+m)/2, (p+q)/w)$ and $Z_2 = (X+Y)/2$.

One can choose X, Y so that Z_1 has smaller expectation than Z_2, such as in the following:

```
# computing expectation of X~Binom(n,p), Y~Binom(m,q),
# and Z_1~Binom((n+m)/2,(p+q)/w), Z_2 = (X+Y)/2

n = 90   # set value of n, number of trials for X
p = .4 # set value of p, probability of trial for X

m = 1000   # set value of m, number of trials for Y
q = .6 # set value of q, probability of trial for Y
```

▶ Show code cell source

```
EX=36.00
EY=600.00
EZ_1=272.50
EZ_2=318.00
```

One can also get the other outcome, where Z_1 has larger expectation than Z_2:

Skip to main content

```
# computing expectation of X~Binom(n,p), Y~Binom(m,q),
# and Z_1~Binom((n+m)/2,(p+q)/w), Z_2 = (X+Y)/2

n = 2000   # set value of n, number of trials for X
p = .4 # set value of p, probability of trial for X

m = 1000   # set value of m, number of trials for Y
q = .6 # set value of q, probability of trial for Y
```

▶ Show code cell source

```
EX=800.00
EY=600.00
EZ_1=750.00
EZ_2=700.00
```

Independence

Definition (independence of events)

Two events H, H' are *independent* if

$$P(H \cap H') = P(H) \cdot P(H')$$

Proposition (when evidence and hypothesis independent, the posterior and prior are same)

If H, E independent of one another, then $P(H|E) = P(H)$

Proof

One has $P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(H) \cdot P(E)}{P(E)} = P(H)$.

Independence of random variables

Two random variables X, Y are *independent* of one another if for all values x, y one has

Skip to main content

$$P(X = x \wedge Y = y) = P(X = x) \cdot P(Y = y)$$

This implies that for all subsets of reals A, B one has

$$P(X \in A \wedge Y \in B) = P(X \in A) \cdot P(Y \in B)$$

For, just write out A, B in terms of the worlds and apply the definition.

Three random variables X, Y, Z are *independent* of one another if for all values x, y, z one has

$$P(X = x \wedge Y = y \wedge Z = z) = P(X = x) \cdot P(Y = y) \cdot P(Z = z)$$

Similarly for longer lists.

Infinite lists $X_1, \ldots, X_n, \ldots$ of random variables are independent of one another if for all $n \geq 1$ one has that $X_1, \ldots, X_n$ are independent of one another.

Proposition (effect of independence on expectation)

If random variables X, Y are independent of one another, then $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Proof:

$\mathbb{E}[X] \cdot \mathbb{E}[Y] = (\sum_{i=1}^{m} x_i \cdot P(X = x_i)) \cdot (\sum_{j=1}^{n} y_j \cdot P(Y = y_j))$

$= \sum_{i=1}^{m} \sum_{j=1}^{n} x_i \cdot y_j \cdot P(X = x_i) \cdot P(Y = y_j)$

$= \sum_{i=1}^{m} \sum_{j=1}^{n} x_i \cdot y_j \cdot P(X = x_i \wedge Y = y_j)$ by independence

$= \sum_{k=1}^{\ell} z_k \cdot P(X \cdot Y = z_k)$

For the last step, given k, one collects together all the pairs (i, j) with $x_i \cdot y_j = z_k$. The associated event is the union of the events $X = x_i \wedge Y = y_j$ over these pairs, which is equal to the event $X \cdot Y = z_k$.

Rules for variance

Skip to main content

Definition (variance)

We define the *variance* of X to be the following

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2]$$

where recall that μ is just another expression for $\mathbb{E}X$.

Intutively, variance is saying how much one expects the value to differ from its expectation.

The variance is also abbreviated as σ^2

The 'square' is coming from the formula for distance in Euclidean space, namely
$d(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Definition (standard deviation)

We define the *standard deviation* of X to be the following:

$$sd(X) = \sqrt{\mathrm{Var}(X)}$$

We also use σ as an abbreviation for the standard deviation.

Hence, you will see some authors write $\sigma(X)$ for the standard deviation of X.

Returning to the vivification again

Earlier, your policy was to choose the random variable which has the larger expectation.

But shouldn't you care about variance as well?

If X has large variance and Y has small variance, then X can vary a lot from its expectation, while Y can vary just a little from its expectation.

Skip to main content

Proposition (simple formula for variance)

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

Proof:

From algebra $(X-\mu)^2 = (X-\mu)\cdot(X-\mu) = X^2 - 2\mu\cdot X + \mu^2$.

Then using the rules we have

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - 2\mu\mathbb{E}X + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

Proposition (how multiplication and addition of reals affects variance)

$$\mathrm{Var}(cX + d) = c^2\mathrm{Var}(X)$$

Proof:

We give the proof when $d = 0$ (the general case is not that much harder):

$$\mathrm{Var}(cX) = \mathbb{E}[c^2X^2] - c^2(\mathbb{E}X)^2 = c^2(\mathbb{E}[X^2] - (\mathbb{E}X)^2) = c^2\cdot\mathrm{Var}(X)$$

Example (variance of Bernoulli)

Let $X \sim \mathrm{Bern}(p)$. Then $\mathrm{Var}(X) = p(1-p)$.

Proof:

by definition (from last time), a random variable X of this kind only two values:

- value 1, with probability p.
- value 0, with probability $1 - p$.

Skip to main content

Since $1^2 = 1$ and $0^2 = 0$, for this random variable we have $X^2 = X$.

Hence we have

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \mathbb{E}[X] - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$$

Returning to the vivification again

Suppose you are choosing between

$X \sim \mathrm{Bern}(.7)$ and $Y \sim \mathrm{Bern}(.9)$.

By earlier, one has $\mathbb{E}X = .7$ and $\mathbb{E}Y = .9$.

By previous, one has $\mathrm{Var}(X) = .21$ and $\mathrm{Var}(Y) = .09$

Hence, Y has higher payoff, and there is a lower level of risk to it.

Proposition (Variance and independence)

Suppose X, Y are independent. Then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Proof

From algebra $(X+Y)^2 = X^2 + 2XY + Y^2$.

Let $\mu = \mathbb{E}[X]$ and $\nu = \mathbb{E}[Y]$.

Then from independence we have $\mathbb{E}(X+Y)^2 = \mathbb{E}X^2 + 2\mu\nu + \mathbb{E}[Y^2]$

From algebra $(\mathbb{E}(X+Y))^2 = (\mu+\nu)^2 = \mu^2 + 2\mu\nu + \nu^2$.

Hence $\mathrm{Var}(X+Y) = \mathbb{E}X^2 - \mu^2 + \mathbb{E}[Y^2] - \nu^2 = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proposition (Root n rule)

Skip to main content

Let $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$

Then the standard deviation of $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$.

Proof

One has

$\operatorname{Var}(\overline{X}) = \sum_{i=1}^n \operatorname{Var}(\frac{X_i}{n})$ by independence

$= \sum_{i=1} \cdot \operatorname{Var}(X_i)$ by earlier proposition

$= \frac{1}{n^2} \sum_{i=1} \operatorname{Var}(X_i)$

$= \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$.

Example (Variance of Binomial)

Let $X \sim \operatorname{Binom}(n, p)$. Then $\operatorname{Var}(X) = n \cdot p(1-p)$

Proof

Suppose that e.g. $n = 2$ and $X = Y + Z$, where $Y, Z \sim \operatorname{Bern}(p)$.

Then one has
$\operatorname{Var}(X) = \operatorname{Var}(Y + Z) = \operatorname{Var}(Y) + \operatorname{Var}(Z) = p(1-p) + p(1-p) = 2 \cdot p(1-p)$

Returning to the vivification

Suppose you are choosing between

$X \sim \operatorname{Binom}(3, .5)$ and $Y \sim \operatorname{Binom}(2, .7)$.

By earlier work, one has $\mathbb{E}X = 1.5$ and $\mathbb{E}Y = 1.4$.

By previous example, one has $\operatorname{Var}(X) = .75$ and $\operatorname{Var}(Y) = .42$

So Y has something to recommend to it: being a little less risky in a certain sense.

Skip to main content

Normal distribution

Definition (normal distribution)

A random variable X has normal distribution $X \sim N(\mu, \sigma^2)$ if it has

- Expectation: μ
- Variance: σ^2
- pdf: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- cdf: $F(x) = \int_{-\infty}^{x} f(x)dx$ (i.e. the area under the curve, for which there is no easy formula)
- ccdf: $\overline{F}(x) = 1 - F(x)$ (for which there is no easy formula)

The cdf of $N(0, 1)$ is often written as $\Phi(x)$

These random variables take as many values as there are real numbers.

Hence the pdf no longer answers the question "what exactly is the probability of a certain outcome."

But the cdf still answers the question "what is the probability of having an outcome less than or equal to a given number."

And using the cdf one can answer questions "what is the probability that we are within a certain range of values."

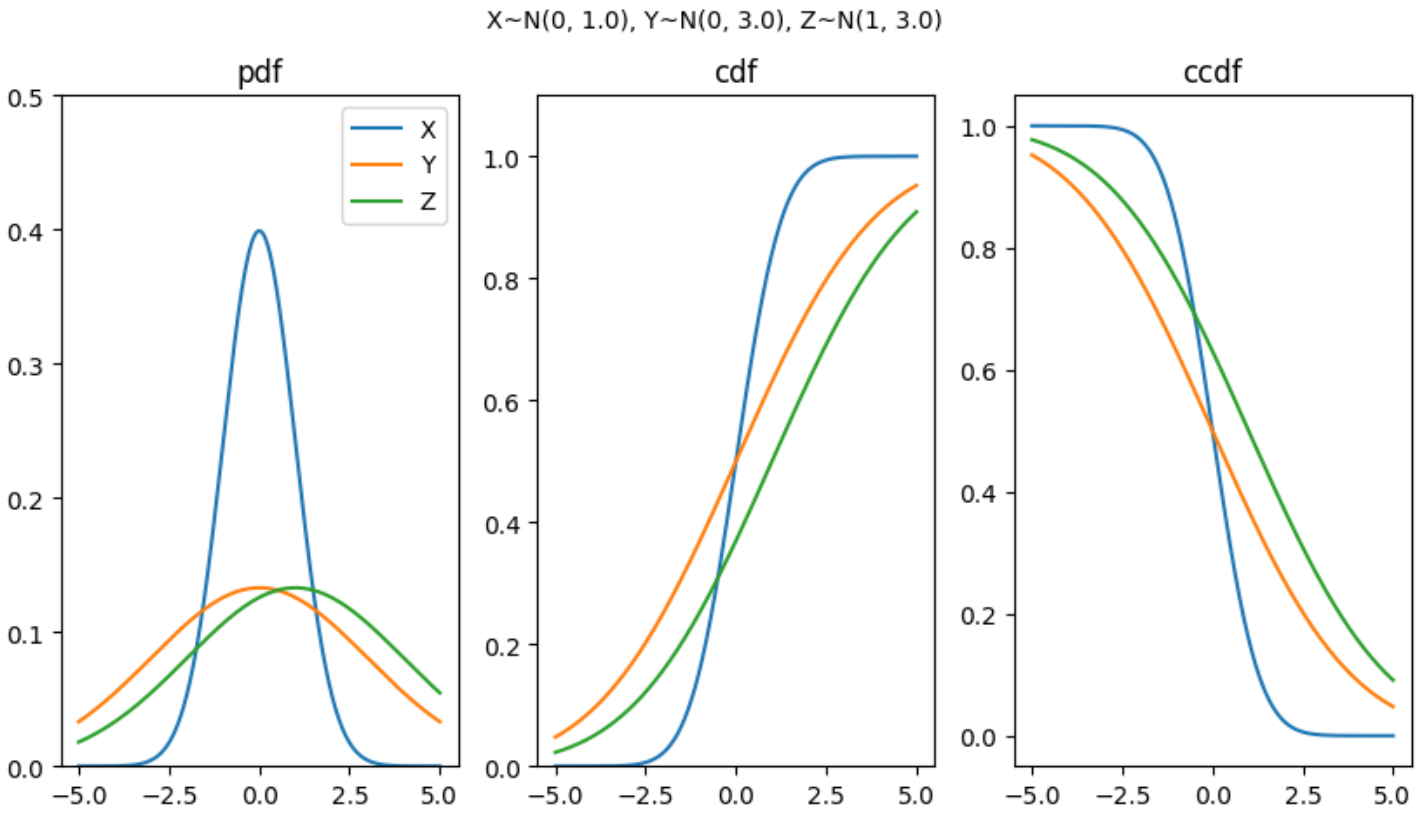
Visualizing the normal distribution

The normal is the familiar bell-shaped curve. Visually, μ is where the center of the bell is, and larger values of σ^2 make the bell a little shorter and spread out.

Skip to main content

```
## visualizing X~N(mu_1,var_1), Y~N(mu_2,var_2), Z~N(mu_3,var_3)
## pdf, cdf, ccdf

mu_1 = 0   # set expectation of X
var_1 = 1 # set variance of X

mu_2 = 0   # set expectation of Y
var_2 = 3 #  set variance of Y

mu_3 =  1  # set expectation of Z
var_3 = 3 # set variance of Y
```

▶ Show code cell source



Skip to main content

Computing the normal distribution

```
## computing X ~ Norm(mu,var)
## P(c<X≤d), P(X≤d), P(X>c)

mu = 3   # set value of expectation

var = 5 # set value of variance

c = 2 # set lower bound of interval you are interested in

d = 2.2 # set upper bound of interval you are interested in
```

▶ Show code cell source

```
Assuming:
X~N(3.0, 5.0)
lower bound c=2.0
upper bound d=2.2
We then obtain:
P(c<X≤d)=0.016
P(X≤d)=Φ(d)=F(d)=0.436
P(X>c)=1−Φ(c)=F̄(c)=0.579
```

Pareto distribution

Definition (Pareto distribution)

A random variable X has *Praeto distribution* with shape b and scale α if $X \sim \text{Pareto}(b, \alpha)$ if $b > 0$ and $\alpha > 2$ and it has

- Expectation: $\alpha \cdot \frac{b}{\alpha - 1}$
- Variance: $\frac{\alpha}{\alpha - 2} \cdot \frac{b}{\alpha - 1}$
- pdf: $f(x) = \frac{\alpha \cdot b^\alpha}{x^{\alpha+1}}$ if $x \geq b$, otherwise $f(x) = 0$.
- cdf: $F(x) = 1 - (\frac{b}{x})^\alpha$ if $x \geq b$, otherwise $F(x) = 1$.
- ccdf: $\overline{F}(x) = (\frac{b}{x})^\alpha$

Skip to main content

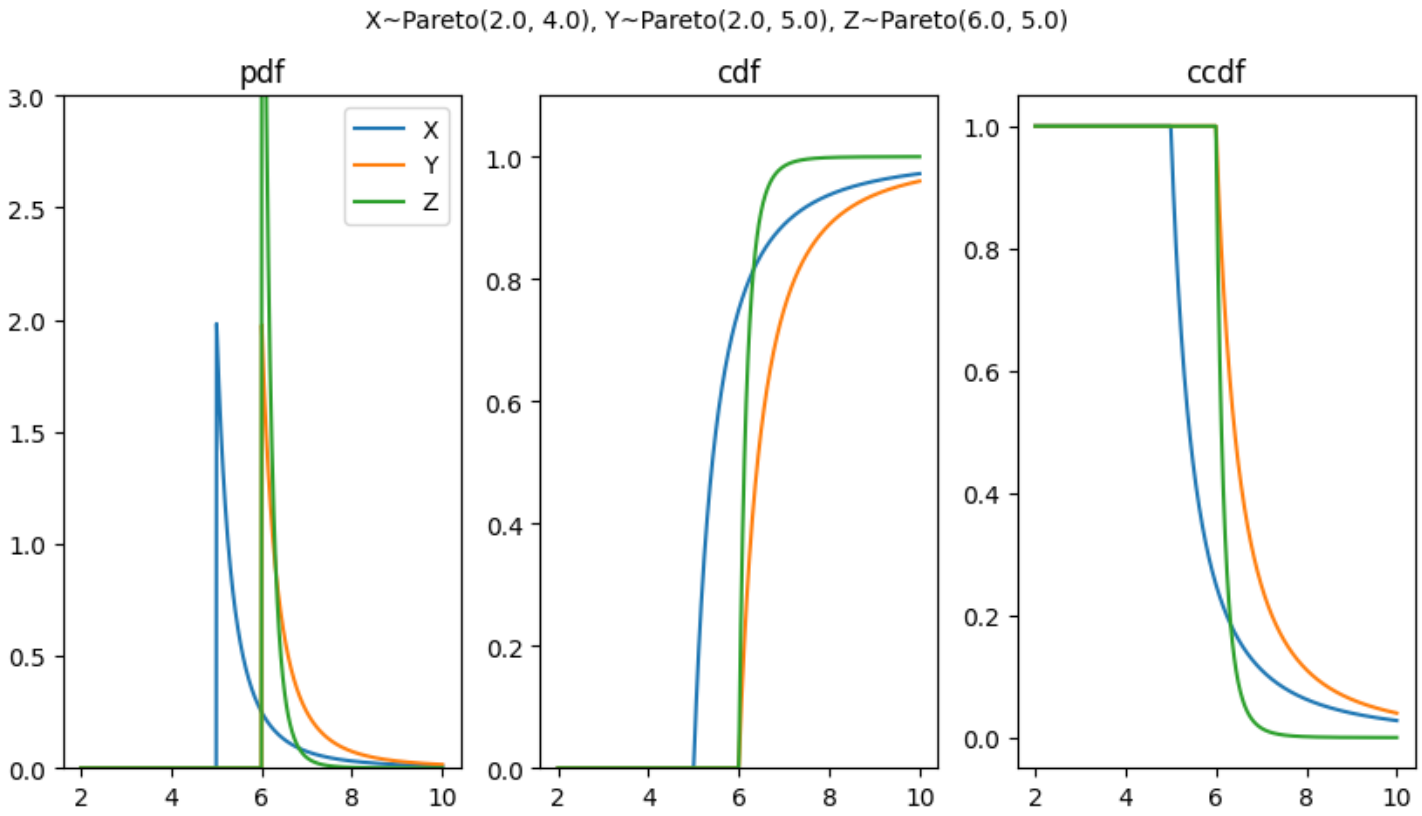
The Praeto distribution is often used to model wealth distributions, where a small group has a large portion of wealth.

It is an example of a *power-law* distribution.

Visualizing a Praeto distribution

```
# visualizing X~Praeto(b_1,alpha_1), Y~Praeto(b_2,alpha_2), Z~Praeto(b_3,alpha_3)

b_1 = 2  # shape parameter of X
alpha_1 = 4  # scale parameter of X

b_2 = 2  # shape parameter of Y
alpha_2 = 5  # scale parameter of Y

b_3 = 6  # shape parameter of Z
alpha_3 = 5  # scale parameter of Z
```

▶ Show code cell source



Skip to main content

Computing a Praeto distribution

```
## computing X ~ Pareto(b,alpha)
## P(c<X≤d), P(X≤d), P(X>c)

b = 3  # set value of shape parameter

alpha = 5 # set value of scale parameter

c = 6 # set lower bound of interval you are interested in

d = 7 # set upper bound of interval you are interested in
```

```
import numpy as np
from scipy.stats import pareto

pdfvalue = pareto.cdf(d, b, alpha) - norm.cdf(c, b, alpha) # P(c<X≤d)

cdfvalue = norm.cdf(d, b, alpha) # P(X≤d)

ccdfvalue = 1-norm.cdf(c, b, alpha) # P(X>c)

print ('Assuming:')

print('X~Pareto(%1.1f, %1.1f)' % (b,alpha))

print('lower bound c=%1.1f' % c)

print('upper bound d=%1.1f' % d)

print('We then obtain:')

print('P(c<X≤d)=%1.3f' % pdfvalue)

print('P(X≤d)=F(d)=%1.3f' % cdfvalue)

print('P(X>c)=F̄(c)=%1.3f' % ccdfvalue)
```

```
Assuming:
X~Pareto(3.0, 5.0)
lower bound c=6.0
upper bound d=7.0
We then obtain:
P(c<X≤d)=0.149
P(X≤d)=F(d)=0.788
```

Skip to main content

References

[DL95] Andrew I Dale and Pierre-Simon Laplace. *Pierre-Simon Laplace Philosophical Essay on Probabilities*. Springer, 1995.

[DF64] Bruno De Finetti. Foresight: its logical laws, its subjective sources. *Studies in subjective probability*, 1964:94–158, 1964.

[EM77] Bradley Efron and Carl Morris. Stein's paradox in statistics. *Sci. Am.*, 236(5):119–127, 1977.

[Fis90] Ronald A Fisher. *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, 1990.

[Gal14] Maria Carla Galavotti. Probability. In Martin Curd and Stathis Psillos, editors, *The Routledge Companion to Philosophy of Science*, pages 458–468. second edition, 2014.

[Gly80] Clark N Glymour. *Theory and Evidence*. Princeton University Press, 1980.

[Hit01] Christopher Hitchcock. The intransitivity of causation revealed in equations and graphs. *J. Philos.*, 98(6):273–299, 2001.

[Hit09] Christopher Hitchcock. Causal modelling. In *The Oxford Handbook of Causation*, pages 299–314. 2009.

[HU06] Colin Howson and Peter Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing, 2006.

[Jef48] Harold Jeffreys. *The Theory of Probability*. Oxford, 1948.

[Joh16] Kent Johnson. Realism and uncertainty of unobservable common causes in factor analysis. *Nous*, 50(2):329–355, June 2016.

[May96] Deborah G Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, August 1996.

[Roy17] Richard Royall. *Statistical Evidence: A Likelihood Paradigm*. Routledge, November 2017.

[Sav72] Leonard J Savage. *The Foundations of Statistics*. Dover, 1972.

[Sob15] Elliott Sober. *Ockham's Razors*. Cambridge University Press, July 2015.

[Spr11] Jan Sprenger. Science without (parametric) models: the case of bootstrap resampling. *Synthese*, 180(1):65–76, May 2011.

[Sta87] Robert C Stalnaker. *Inquiry*. MIT Press, March 1987.

[vM57] Richard von Mises. *Probability, Statistics and Truth*. Macmillan, New York, 1957.

[Wil10] Jon Williamson. *In Defence of Objective Bayesianism*. Oxford University Press, May 2010.