

Advanced Statistics I

Probability calculus

Prof. Dr. Matei Demetrescu

It all starts with data

Month	Day	Store 1		Store 2		Store 3		Store 4		Store 5					
1	2	Price	Quantity	M	N	O	P								
3	1	22	16,46	553	19,20	138	18,01	519	20,00	351	17,85	81			
4	1	25	16,48	471	15,60	107	16,52	157	16,91	167	16,29	88			
5	1	3	16,16	434	18,87	212	16,67	432	17,99	201	16,94	86			

Data have been around for some time in one form or another;

... one would like to make use of them.

This takes us to quantitative methods:

- reduce the amount of data to a presentable form
- construct models
- draw conclusions
- make predictions

Today's outline

Kick-off

- 1 Probability and statistics
- 2 The course
- 3 Q & A
- 4 Up next

Outline

1 Probability and statistics

2 The course

3 Q & A

4 Up next

Quantitative methods?

We focus here on building and analyzing models with **uncertainty**.¹

The basic premise is the presence of some apparent or intrinsic randomness:

- ① Sampling variability
- ② Variables not under control
- ③ ... the works.

¹Related, not irrelevant topics: measuring and communicating uncertainty, making decisions under uncertainty, design of experiments, etc.

Randomness and statistics

Say you want to draw conclusions from a data sample.

Will your conclusions hold for more than the given sample?

Typically yes, if the sample is **representative**.

- **Random** samples are as representative as it gets without prior knowledge.
- But then we need to understand how randomness affects our analyses.

More generally, we aim to construct and analyze models including random components.

No magic!

In particular, we are concerned with the distribution of possible outcomes of nondeterministic phenomena, and

... rely heavily on probability theory to do the job.

Spelled out:

- We can't forecast single outcomes!
- But we may be able to tell how likely they are.

Why probability theory?

Actually, it is not that obvious that we should resort to probability theory.²

But:

- Distributions arise naturally in a sampling framework
- Probability calculus also delivers the tools to manipulate distributions.
- There are several interpretations for the notion of probability, but there is only one probability calculus.
- This allows for a transparent/comprehensible use of probability theory, which is essential for the acceptance of statistical analyses.

²Fuzzy logic also allows for uncertainty.

Outline

1 Probability and statistics

2 The course

3 Q & A

4 Up next

Aim of the course

This course is, together with AdvStat II, essential for deploying quantitative methods.

After successfully participating, you will become ...

- ... quite familiar with probability calculus,
and, in particular,
- ... able to work with models having (fully characterized) stochastic components.

Outline I

① Elements of Probability Theory:

- ① Sample Space and Events
- ② Probability
- ③ Properties of the Probability Function
- ④ Conditional Probability
- ⑤ Independence
- ⑥ Total Probability Rule and Bayes' Law

② Random Variables and their Probability Distributions:

- ① Univariate Random Variables
- ② Univariate Cumulative Distribution Functions
- ③ Multivariate Random Variables
- ④ Marginal Distributions
- ⑤ Conditional Distributions
- ⑥ Independence of Random Variables

Outline II

③ Moments of Random Variables:

- ① Expectation of a Random Variable
- ② Expectation of a Function of Random Variables
- ③ Conditional Expectation
- ④ Moments of a Random Variable
- ⑤ Moment-Generating Functions
- ⑥ Joint Moments and Moments of Linear Combinations
- ⑦ Means and Variances of Linear Combinations of Random Variables

④ Parametric Families of Density Functions:

- ① Discrete Density Functions
- ② Continuous Density Functions
- ③ Normal Family of Densities
- ④ Exponential Class of Distributions
- ⑤ (Multivariate) Extensions

Outline III

⑤ Basic Asymptotics:

- ① Convergence of Number and Function Sequences
- ② Convergence Concepts for Sequences of Random Variables
- ③ Weak Laws of Large Numbers
- ④ Central Limit Theorems
- ⑤ Asymptotic Distributions of Functions for Asymptotically Normally Distributed Random Variables

Textbooks

Main ones

- Mood, A. M., Graybill, F. A. and D.C. Boes (1974, 3rd ed.).
Introduction to the Theory of Statistics. McGraw-Hill.
- Mittelhammer, R. C. (1996). Mathematical Statistics for Economics and Business. Springer.

Other useful ones

- Wassermann, L. (2004). All of Statistics: A Concise Course in Statistical Inference. Springer.
- Casella, G. and R. Berger (2002, 2nd ed.). Statistical Inference. Duxbury.
- Rohatgi, V. K. und A. K. Saleh (2001, 2nd ed.). An Introduction to Probability Theory and Mathematical Statistics. John Wiley & Sons.
- Linton, O. (2017). Probability, Statistics and Econometrics. Academic Press.

Materials & workflow

- ① Detailed lecture notes via OLAT
... but use wisely: <http://pss.sagepub.com/content/25/6/1159>
- ② Weekly slides & videos will be available in due time (OLAT)
- ③ Flipped classroom:
 - Prepare for class using screencasts
 - In class we discuss details and have Q&As
(on site and via BigBlueButton/OLAT)

Tutorial

- ① Mariia Okuneva, Uliana Zaspa
- ② Problem sets will be available in due time via OLAT
- ③ You have the choice between three different time slots
- ④ *non-compulsory* PC tutorial (**R**): every second week (tutored by Mariia Okuneva and Uliana Zaspa, online)
- ⑤ May earn bonus points

Why R

- Open source and free – check out <https://cran.r-project.org>
- Packages for almost anything
- High interest
- ...

Exam & grades

- Written exam (pass with at least 50 out of 100 pts)
problems similar to, or even the same as the pen & paper tutorial
- probably online
- either way you're allowed to use a formulary (OLAT in due time)
- don't forget the bonus points (max 15 on top of written exam)
- Extra: collection of problems and old exams (OLAT, around December)

Office hours

When you have questions outside class:

- Try first the OLAT forum
- Then the weekly Q&A sessions
- Also ask directly, mdeme@stat-econ.uni-kiel.de,
mokuneva@stat-econ.uni-kiel.de or
uzaspa@stat-econ.uni-kiel.de
- otherwise by appointment

Outline

1 Probability and statistics

2 The course

3 Q & A

4 Up next

Have I forgotten anything?

Any specific questions?

Outline

1 Probability and statistics

2 The course

3 Q & A

4 Up next

Coming up

Events and Probability

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

Elements of probability theory

- 1 The probability function
- 2 Conditional probability and independence
- 3 Total probability rule and Bayes's rule
- 4 Up next

Outline

- 1 The probability function
- 2 Conditional probability and independence
- 3 Total probability rule and Bayes's rule
- 4 Up next

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 $\mathcal{S} = \{\square, \square\cdot, \square\cdot\cdot, \square\square, \square\square\cdot, \square\square\cdot\cdot\}$.

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, $\mathcal{S} = \{0, 1, 2, \dots\}$.

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., $\mathcal{S} = \{0, 1, 2, \dots, 6\}$
- countably infinite (discrete sample space), e.g., $\mathcal{S} = \mathbb{N} = \{0, 1, 2, \dots\}$
- 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 $\mathcal{S} = \{\square, \bullet\square, \circ\square, \square\square, \square\circ, \square\bullet\}$ 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 $\mathcal{S} = \{1, 2, \dots, 6\}$. Let then

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

- 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 $P(\cdot)$, 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 \mathcal{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 $A_1, A_2, \dots \in \mathcal{B}$, 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.

^aThis 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) $P(A) \geq 0$ for all $A \in \mathcal{Y}$.
- 2 (standardization) $P(\mathcal{S}) = 1$.
- 3 (additivity) If $A_1, A_2, \dots \in \mathcal{Y}$ is a sequence of disjoint events ($A_i \cap A_j = \emptyset$ for $i \neq j$; $i, j \in \mathbb{N}$), then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

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

Dice (& Laplace)

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

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

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

- the value of the function $P(A) \geq 0$ 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, \dots, A_n we have

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

Now with numbers

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

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

This set function represents a probability set function since

- ① the value of the function $P(A) \geq 0$ 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**)

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

- ③ for any collection of disjoint sets $A_1, A_2, \dots, A_n, \dots$ we have

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

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

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

This set function represents a probability set function since

- ① the value of the function $P(A) \geq 0$ 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

$$P(\mathcal{S}) = \int_{x \in S} \frac{1}{2} e^{-\frac{x}{2}} dx = \int_0^\infty \frac{1}{2} e^{-\frac{x}{2}} dx = 1 \quad (\text{standardization});$$

- ③ for any disjoint sets $A_1, A_2, \dots, A_n, \dots$ we have (**additivity**)

$$P(\bigcup_{i=1}^{\infty} A_i) = \underbrace{\int_{x \in (\bigcup_{i=1}^{\infty} A_i)} \frac{1}{2} e^{-\frac{x}{2}} dx}_{\text{additivity property of Riemann integrals}} = \sum_{i=1}^{\infty} \left[\int_{x \in A_i} \frac{1}{2} e^{-\frac{x}{2}} dx \right] = \sum_{i=1}^{\infty} P(A_i).$$

We have what we need

- Once we have defined the 3-tuple $\{\mathcal{S}, \mathcal{Y}, P\}$ (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 $P(A) = 1 - P(\bar{A})$.

Theorem (1.2)

$P(\emptyset) = 0$.

Theorem (1.3)

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

Theorem (1.4)

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

... and go on ...

Theorem (1.5)

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

Corollary (1.1, Boole's Inequality)

$P(A \cup B) \leq P(A) + P(B)$.

Theorem (1.6)

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

Theorem (1.7, Bonferroni's Inequality)

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

... like the Duracell bunny

Theorem (1.8)

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

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, \dots, 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

- 1 The probability function
- 2 Conditional probability and independence
- 3 Total probability rule and Bayes's rule
- 4 Up next

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

$$\mathcal{S} = \{(H,H), (H,T), (T,H), (T,T)\} \quad (H = \text{Head}, T = \text{Tail}).$$

Examine the events

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

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

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

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

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 $P(B) \neq 0$, then the conditional probability of event A , given event B , is given by $P(A | B) = P(A \cap B) / P(B)$.

Example

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

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

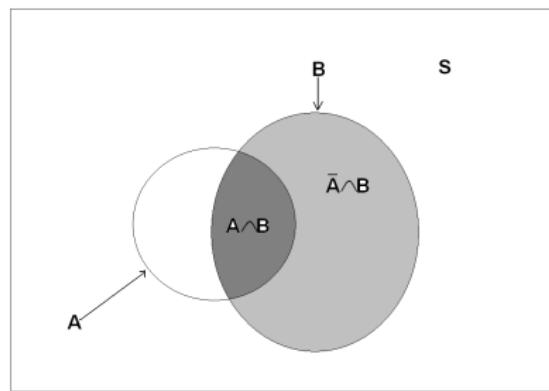
given the first coin toss results in heads, $B = \{(H,H), (H,T)\}$ is

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

Still a probability

Theorem (1.10)

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



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

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 $P(B) \neq 0$. Then
$$P(A \cap B) = P(A | B) P(B).$$

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 $P(D) = 0.02$).
- The probability that a test is positive (event A) if the tested person is actually ill (that is given event D) is $P(A|D) = 0.95$.

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

We can do better

Theorem (1.12, Extended Multiplication Rule)

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

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

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 $P(A \cap B) = P(A)P(B)$. If A and B are not independent, A and B are said to be dependent events.

Independence of A and B implies

$$P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A), \text{ if } P(B) > 0$$

$$P(B|A) = P(B \cap A)/P(A) = P(B)P(A)/P(A) = P(B), \text{ if } P(A) > 0.$$

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

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

for which $N(J) \geq 2$. If the events A_1, A_2, \dots, 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:

$P(A_1 \cap A_2) = P(A_1) P(A_2)$, $P(A_1 \cap A_3) = P(A_1) P(A_3)$, $P(A_3 \cap A_2) = P(A_3) P(A_2)$,
 (pairwise independence) and

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) P(A_2) P(A_3).$$

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 $i = 1, 2, 3$

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

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

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

Outline

- 1 The probability function
- 2 Conditional probability and independence
- 3 Total probability rule and Bayes's rule
- 4 Up next

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 $P(B_i) > 0 \forall i$. Then the “total” probability of event A is

$$P(A) = \sum_{i \in I} P(A | B_i) P(B_i).$$

“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 $P(B_i) > 0 \forall i \in I$. Then, provided $P(A) \neq 0$,

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

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: $P(A|B) = 0.98$.
- The test yields a ‘false positive’ result for 1 percent of the healthy subjects: $P(A|\bar{B}) = 0.01$.

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

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

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

Outline

- 1 The probability function
- 2 Conditional probability and independence
- 3 Total probability rule and Bayes's rule
- 4 Up next

Coming up

Random variables and their distribution

Random variables and their distributions

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

So why random variables?

We now have a working understanding of probability.

- This is based on a generic notion of events and (random) outcomes.
- In quantitative approaches we however work with **numeric** data!

So let's study probabilities of events built around numbers.
(And all the implications...)

Random variables and their distributions

- 1 (Univariate) Random variables
- 2 Probability density functions
- 3 Cumulative distribution functions
- 4 Up next

Outline

- 1 (Univariate) Random variables
- 2 Probability density functions
- 3 Cumulative distribution functions
- 4 Up next

We want numbers!

In many experiments it is easier to deal with a summary variable than with the original probability structure.

- Say you toss N coins but only care about the total number of heads/tails and not about which coin shows them,
- ... so let us call this number a variable X .
- This simplifies the sample space to the set $\{0, 1, 2, \dots, N\}$
- X depends on the outcomes of the experiment
- and is actually a function mapping from the probability space of the experiment to $\{0, 1, 2, \dots, N\}$.

Univariate random variable

Definition

Let $\{\mathcal{S}, \mathcal{Y}, P\}$ be a probability space. If $X : \mathcal{S} \rightarrow \mathbb{R}$ (or simply, X) is a real-valued function having as its domain the elements of \mathcal{S} , then $X : \mathcal{S} \rightarrow \mathbb{R}$ (or X) is called a random variable.

In some experiments random variables are implicitly used:

Experiment	Random variable
Toss two dice	$X = \text{sum of the numbers}$
Toss a coin 50 times	$X = \text{number of heads in 50 tosses}$
Toss a coin 50 times	$X = \text{squared number of heads in 50 tosses}$

Remark on notation

$X(\omega)$: denotes the image of $\omega \in \mathcal{S}$ generated by the random variable $X : \mathcal{S} \rightarrow \mathbb{R}$.

$x = X(\omega)$: (realized) value of the function X

Uppercase letters (X) will be used to denote random variables and corresponding lowercase letters (x) will denote the realized values.

Range of a random variable

Note that by defining a random variable, we have also defined a **new sample space**, namely, the **range of the random variable**.

This range, denoted by $R(X)$, is obtained as the set of all x -values which can be generated on the sample space \mathcal{S} using the function X :

$$R(X) = \{x : x = X(\omega), \omega \in \mathcal{S}\}.$$

This raises the following important questions:

How can we embed the new sample space $R(X)$ within a probability space that can be used for assigning probabilities to events in terms of random-variable outcomes?

Hence, what is the probability function on $R(X)$, say P_X ?

Induced probability function

- Suppose we have a discrete sample space

$$\mathcal{S} = \{\omega_1, \dots, \omega_n\} \quad \text{with a probability function } P(\cdot).$$

- Now define a random variable

$$X(\omega) \quad \text{with range } R(X) = \{x_1, \dots, x_m\}.$$

Assume that we observe $X = x_i$ iff the experiment's outcome is ω_j such that

$$x_i = X(\omega_j).$$

- Since the elementary event $\omega_j \in \mathcal{S}$ is equivalent to the event $x_i \in R(X)$, both events should have the same probability. Thus

$$P_X(X = x_i) = P(\{\omega_j : x_i = X(\omega_j), \omega_j \in \mathcal{S}\}).$$

Note that the function P_X on the left-hand side is **an induced probability set function on $R(X)$ defined in terms of the original function P** .

The eternal coins

Consider the experiment of tossing a fair coin two times.

- Define the random variable X to be the number of heads in the two tosses. Thus

Experiment's outcome $\omega \in \mathcal{S}$	(H,H)	(H,T)	(T,H)	(T,T)
Variable's Realization $x = X(\omega)$	2	1	1	0

- The random variable's range is $R(X) = \{0, 1, 2\}$
- Since, for example, $P_X(X = 1) = P(\{H, T\}) + P(\{T, H\})$, the induced probability function on $R(X)$ obtains as

x	0	1	2
$P_X(X = x)$	1/4	1/2	1/4

Outline

- 1 (Univariate) Random variables
- 2 Probability density functions
- 3 Cumulative distribution functions
- 4 Up next

Characterizing the probability

Tables are nice, but it is useful to have a representation of the induced probability set function, P_X , in a compact closed-form formula.

- This leads us to the definition of a so-called probability density function.
- They offer a convenient way of conveying the information contained in P_X .

Random variables can be either **discrete** or **continuous**. This dichotomy is inherited by the pdfs.

Discrete outcomes

Definition (Discrete random variable)

A random variable X is called discrete iff its range $R(X)$ is countable.

Definition (Discrete probability density function)

The discrete probability density function (pdf) of a discrete random variable X , denoted by f , is defined by

$$f : \mathbb{R} \rightarrow [0, 1] \quad \text{such that} \quad f(x) = \begin{cases} P_X(X = x) & \text{if } x \in R(X) \\ 0 & \text{else.} \end{cases}$$

- The discrete pdf is also called probability mass function (pmf).
- $R(X)$ may be countable, but the domain of the pmf is \mathbb{R} .
- This convention works for continuous rvs as well, but the discrete pdf is zero “almost everywhere”.

Working with the discrete pdf

The pdf allows us to obtain the probability for an event in $\mathcal{R}(X)$.

- Consider the event $A \subset \mathcal{R}(X)$, written as a union of elementary events $A = \cup_{x \in A} \{x\}$.
- Since elementary events are disjoint, we know from Axiom 1.3 that

$$P_X(A) = P_X(\cup_{x \in A} \{X = x\}) \stackrel{(Ax.3)}{=} \sum_{x \in A} P_X(x) = \sum_{x \in A} f(x).$$

- Thus, we can use the pdf to calculate probabilities for events on $\mathcal{R}(X)$ by summing the probabilities of the elementary events given by the pdf.

Example: Counting the dots I

Consider the experiment of tossing two fair dice and observing the number of dots facing up.

- The sample space is $\mathcal{S} = \{(i, j) : i = 1, \dots, 6; j = 1, \dots, 6\}$, where i, j are the number of dots. \mathcal{S} consists of 36 elementary events.
- Define a random variable X to be the sum of the dots, such that $x = X((i, j)) = i + j$.
- We can derive the pdf of X using elementary arguments.

Example: Counting the dots II

We obtain the following correspondence between outcomes of X and events in \mathcal{S} :

$x = X((i, j))$	$B_x = \{(i, j) : x = i + j, (i, j) \in \mathcal{S}\}$	$P_X(x) = f(x) = P(B_x)$
2	$\{(1, 1)\}$	$1/36$
3	$\{(1, 2), (2, 1)\}$	$2/36$
4	$\{(1, 3), (2, 2), (3, 1)\}$	$3/36$
	\vdots	
12	$\{(6, 6)\}$	$1/36$

- Consider the event $X \in \{3, 4\}$. The probability is given as $P_X(A) = \sum_{x \in A} f(x) = f(3) + f(4) = 5/36$.
- A compact algebraic form for the pdf f is $f(x) = \frac{6 - |x - 7|}{36} \mathbb{I}_{\{2, 3, \dots, 12\}}(x)$.

Continuous distributions

Definition (Continuous random variable)

A random variable X is called continuous iff its range $R(X)$ is not countable.

Problem:

- The range $R(X)$ is **continuous** with events A defined as intervals in $R(X) \subset \mathbb{R}$
- But can't use summation to add uncountably many probabilities!

(Heuristic) **Solution:** Substitute the summation operation $\sum_{x \in A}$ by integration $\int_{x \in A}$.

The “genuine” probability density function

Definition (Continuous probability density function)

A random variable X is called continuous iff

- its range $R(X)$ is uncountably infinite and
- there exists a function

$$f : \mathbb{R} \rightarrow [0, \infty) \quad \text{such that for any event } A, \quad P_X(A) = \int_{x \in A} f(x) dx$$

and

$$f(x) = 0 \quad \forall x \notin R(X).$$

The function f is called a continuous probability density function.

Cars (& Laplace)

Consider a Formula 1 circuit of 10 km. Suppose that accidents are equally likely to occur at each point of the circuit.

So define the continuous random variable X to be the point of a potential accident with range $R(X) = [0, 10]$.

In order to obtain the pdf for X ,

- consider the event A of an accident between two points a and b , such that $A = [a, b]$.
- Since all points are **equally likely**, $P_X(A) = \frac{\text{length of } A}{\text{length of } R(X)} = \frac{b-a}{10}$.

... and their accidents

According to the definition, the pdf f for X has to satisfy

$$\int_{x \in A} f(x)dx = \int_a^b f(x)dx \stackrel{!}{=} P_X(A) = \frac{b-a}{10}, \quad \forall \quad 0 \leq a \leq b \leq 10,$$

with

$$\frac{\partial [\int_a^b f(x)dx]}{\partial b} = f(b) \stackrel{!}{=} \frac{\partial [\frac{b-a}{10}]}{\partial b} = \frac{1}{10}, \quad \forall \quad b \in [0, 10].$$

Hence,

- the function $f(x) = \frac{1}{10}\mathbb{I}_{[0,10]}(x)$ can be used as a pdf for X ,
- and for any event A on $\mathcal{R}(X)$ we obtain $P_X(A) = \int_{x \in A} \frac{1}{10}dx$.
- E.g., the probability for $X \in A = [0, 5]$ is $P_X(A) = \int_0^5 \frac{1}{10}dx = 1/2$.

Singletons

The definition of the continuous pdf implies that the probability for an elementary event $A = \{a\}$ is zero, since

$$P_X(A) = \int_a^a f(x)dx = 0.$$

Still, some outcome *will* occur!

We may interpret this to mean that A is 'relatively impossible', relative to all other outcomes that can occur in $R(X) \setminus A$.

E.g., since $\{a\}$, $\{b\}$ and (a, b) are disjoint and $P_X(\{a\}) = P_X(\{b\}) = 0$,

$$P_X([a, b]) = P_X((a, b]) = P_X([a, b)) = P_X((a, b)) = \int_a^b f(x)dx.$$

Some comparisons

The interpretation of the function value of a continuous pdf $f(x)$ is fundamentally different from that of a discrete pdf:

- If f is discrete, $f(x) = P_X(x) = \text{probability of the outcome } x$.
- If f is continuous, $f(x)$ is not the probability of outcome x , which is $P_X(x) = 0$. (If $f(x)$ was a probability, we would have $f(x) = 0 \forall x$.)
- For a unified interpretation, imagine the discrete pdf as having point probability mass.¹

¹We'll discuss this later in the course but we need some additional motivation, so please be patient for now.

Requirements for pdfs

Pdfs should be such that the probabilities obtained from f adhere to the probability axioms.

Definition (Class of discrete pdfs)

The function $f : \mathbb{R} \rightarrow \mathbb{R}$ is a member of the class of discrete pdfs iff

- (i_a) the set $C = \{x : f(x) > 0, x \in \mathbb{R}\}$ is countable;
- (ii_a) $f(x) = 0 \forall x \in \bar{C}$;
- (iii_a) $\sum_{x \in C} f(x) = 1$.

Definition (Class of continuous pdfs)

The function $f : \mathbb{R} \rightarrow \mathbb{R}$ is a member of the class of continuous pdfs iff

- (i_b) $f(x) \geq 0 \forall x \in \mathbb{R}$;
- (ii_b) $\int_{x \in \mathbb{R}} f(x) dx = 1$.

Some checks

- 1) Consider the function $f(x) = (0.3)^x (0.7)^{1-x} \mathbb{I}_{\{0,1\}}(x)$. Can this f serve as pdf?

Since (i) $f(x) > 0$ on the countable set $\{0, 1\}$, and (ii) $\sum_{x=0}^1 f(x) = 1$, and (iii) $f(x) = 0 \forall x \notin \{0, 1\}$, the function f can serve as a pdf.

- 2) Consider the function $f(x) = (x^2 + 1)\mathbb{I}_{[-1,1]}(x)$. Can this f serve as pdf?

While $f(x) \geq 0 \forall x \in \mathbb{R}$, f does not integrate to 1:

$$\int_{\mathbb{R}} f(x)dx = \int_{-1}^1 (x^2 + 1)dx = \frac{8}{3} \neq 1.$$

Thus, f can not serve as a pdf. (Normalization gets us from f to a function which can serve as a pdf)

Outline

- 1 (Univariate) Random variables
- 2 Probability density functions
- 3 Cumulative distribution functions
- 4 Up next

Another description of probability

Definition (Cumulative distribution function)

The cumulative distribution function (cdf) of a random variable X , denoted by F , is defined by

$$F : \mathbb{R} \rightarrow [0, 1] \quad \text{such that} \quad F(b) = \Pr_X(X \leq b), \quad \forall b \in \mathbb{R}.$$

For a discrete random variable the cdf is obtained as

$$F(b) = \sum_{x \leq b} f(x), \quad \forall b \in \mathbb{R},$$

and for a continuous random variable as

$$F(b) = \int_{-\infty}^b f(x)dx, \quad \forall b \in \mathbb{R}.$$

A continuous example

Let the random variable X be the duration of a telephone call (in min), with range $R(X) = \{x : x > 0\}$.

- Let the pdf be: $f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \cdot \mathbb{I}_{(0,\infty)}(x)$, with $\lambda > 0$.
- The cdf is then $F(b) = \int_0^b \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx = (1 - e^{-\frac{b}{\lambda}}) \cdot \mathbb{I}_{(0,\infty)}(b)$
- Assume that $\lambda = 100$ (average duration). Then the probability that the duration is less than 50 min is: $F(50) = 1 - e^{-\frac{50}{100}} = 0.39$.

A discrete example

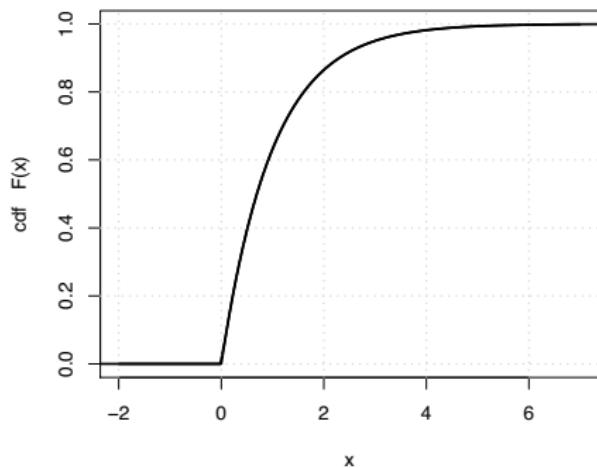
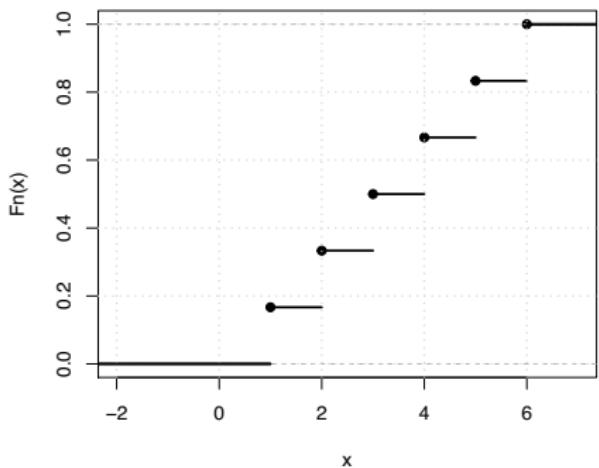
Let the random variable X be the number of dots observed rolling a die, with range $R(X) = \{1, 2, \dots, 6\}$.

- The pdf is: $f(x) = \frac{1}{6} \cdot \mathbb{I}_{\{1, \dots, 6\}}(x)$.
- The cdf is obtained as:

$$F(b) = \sum_{x \leq b} \frac{1}{6} \cdot \mathbb{I}_{\{1, \dots, 6\}}(x) = \frac{1}{6} \lfloor b \rfloor \cdot \mathbb{I}_{[1, \dots, 6]}(b) + \mathbb{I}_{(6, \infty)}(b)$$

($\lfloor b \rfloor$ denotes the integer part of the number b) – see following figure.

Cdfs of continuous vs. discrete RVs

Continuous distribution**Discrete distribution**

(And we may have mixtures of the two – nothing to be scared of.)

Properties

Theorem (2.1)

For any cdf F , we have that

- (i) $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$;
- (ii) $F(x)$ is a non decreasing function on x ; that is, $F(a) \leq F(b)$ for $a < b$;
- (iii) $F(x)$ is right-continuous; that is, $\lim_{h \downarrow 0} F(x + h) = F(x)$.

Relation to pdfs

Theorem (2.2)

Let $x_1 < x_2 < x_3 < \dots$ be the countable set of outcomes in the range of the discrete random variable X . Then the pdf for X obtains as

$$f(x_i) = \begin{cases} F(x_i), & i = 1 \\ F(x_i) - F(x_{i-1}), & i = 2, 3, \dots \\ 0, & x \notin R(X). \end{cases}$$

Theorem (2.3)

Let $f(x)$ and $F(x)$ denote the pdf and cdf of a continuous random variable X . Then the pdf for X obtains as

$$f(x) = \begin{cases} \frac{dF(x)}{dx}, & \text{wherever } f(x) \text{ is continuous} \\ 0, & \text{elsewhere.} \end{cases}$$

Jumps & co.

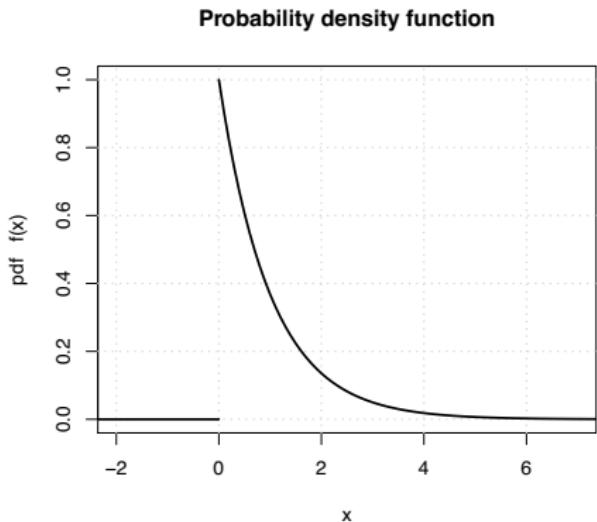
Recall X the duration of a telephone call, with cdf

$$F(x) = (1 - e^{-\frac{x}{\lambda}}) \cdot \mathbb{I}_{(0, \infty)}(x).$$

A pdf for X is given by

$$f(x) = \begin{cases} F'(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} & x > 0 \\ 0 & (say) \\ 0 & x = 0 \\ 0 & x < 0 \end{cases}$$

Note the (rather arbitrary)
choice at $x = 0$.



Outline

- 1 (Univariate) Random variables
- 2 Probability density functions
- 3 Cumulative distribution functions
- 4 Up next

Coming up

Multivariate random variables

Multivariate RVs and their distributions

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Increasing the numbers

Random variables allow us to include stochastic components in mathematical models.

The scalar case is often just the beginning,

... as we may need more than just one stochastic component.

So we have to discuss ways of dealing with **several** random variables

at the same time.

Multivariate RVs and their distributions

- 1 Multivariate random variables
- 2 Marginal distributions
- 3 Conditional distributions and independence
- 4 Up next

Outline

- 1 Multivariate random variables
- 2 Marginal distributions
- 3 Conditional distributions and independence
- 4 Up next

What if mapping to \mathbb{R}^K ?

Definition (Multivariate random variable)

Let $\{\mathcal{S}, \mathcal{Y}, P\}$ be a probability space. If $X : \mathcal{S} \rightarrow \mathbb{R}^n$ (or simply, X) is a real-valued vector function having as its domain the elements of \mathcal{S} , then $X : \mathcal{S} \rightarrow \mathbb{R}^n$ (or X) is called a **multivariate (n -variate) random variable**.

The realized value of the multivariate random variable is

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} X_1(\omega) \\ X_2(\omega) \\ \vdots \\ X_n(\omega) \end{pmatrix} = \mathbf{X}(\omega) \quad \text{for } \omega \in \mathcal{S},$$

and its range is

$$R(\mathbf{X}) = \{(x_1, \dots, x_n) : x_i = X_i(\omega), i = 1, \dots, n, \omega \in \mathcal{S}\}.$$

More complicated pdfs

Definition (Discrete multivariate pdf)

A multivariate random variable $\mathbf{X} = (X_1, \dots, X_n)$ is called discrete iff its range $R(\mathbf{X})$ is countable. The **discrete joint pdf** of a discrete random vector \mathbf{X} , denoted by f , is defined by

$f : \mathbb{R}^n \rightarrow [0, 1]$ such that

$$f(x_1, \dots, x_n) = \begin{cases} P_{\mathbf{X}}(X_1 = x_1, \dots, X_n = x_n) & \text{if } (x_1, \dots, x_n) \in R(\mathbf{X}) \\ 0 & \text{else.} \end{cases}$$

And the continuous case

Definition (Continuous multivariate pdf)

A multivariate random variable $\mathbf{X} = (X_1, \dots, X_n)$ is called continuous iff

- its range $R(\mathbf{X})$ is uncountably infinite and
- there exists a function

$f : \mathbb{R}^n \rightarrow [0, \infty)$ such that for any event A,

$$P_{\mathbf{X}}(A) = \int \cdots \int_{(x_1, \dots, x_n) \in A} f(x_1, \dots, x_n) dx_1 \cdots dx_n$$

and

$$f(x_1, \dots, x_n) = 0 \quad \forall (x_1, \dots, x_n) \notin R(\mathbf{X}).$$

The function f is called a **continuous joint pdf**.

Requirements

Definition (Class of discrete joint pdfs)

The function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a member of the class of discrete joint pdfs iff

- (i_a) the set $C = \{(x_1, \dots, x_n) : f(x_1, \dots, x_n) > 0, (x_1, \dots, x_n) \in \mathbb{R}^n\}$ is countable;
- (ii_a) $f(x_1, \dots, x_n) = 0 \quad \forall (x_1, \dots, x_n) \in \bar{C};$
- (iii_a) $\sum \cdots \sum_{(x_1, \dots, x_n) \in C} f(x_1, \dots, x_n) = 1.$

Definition (Class of continuous joint pdfs)

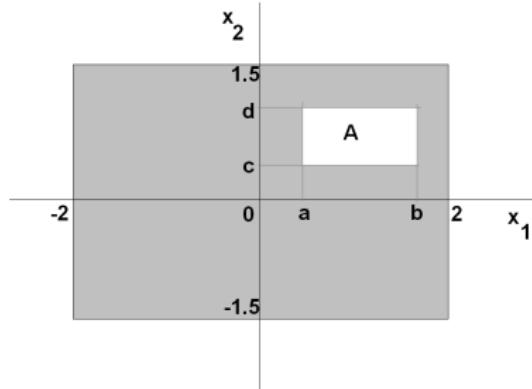
The function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a member of the class of continuous joint pdfs iff

- (i_b) $f(x_1, \dots, x_n) \geq 0 \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n;$
- (ii_b) $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_1 \cdots dx_n = 1.$

An example from outer space I

Consider that the NASA announces that a small meteorite will hit a rectangular area of $12\text{km}^2 = 4\text{km} \times 3\text{km}$.

- Define $\mathbf{X} = (X_1, X_2)$ to be the coordinates of the point of strike, with a range $R(\mathbf{X}) = \{(x_1, x_2) : x_1 \in [-2, 2], x_2 \in [-1.5, 1.5]\}$.
- Each point in that rectangle is equally likely to be struck.
- To obtain the continuous pdf of \mathbf{X} , consider a rectangle A in $R(\mathbf{X})$:



An example from outer space II

Since all points are equally likely, we obtain

$$P_{\mathbf{X}}(\mathbf{X} \in A) = \frac{\text{area of } A}{\text{area of } R(\mathbf{X})} = \frac{(b-a)(d-c)}{12}.$$

According to the definition, the pdf f for \mathbf{X} has to satisfy

$$\int_c^d \int_a^b f(x_1, x_2) dx_1 dx_2 \stackrel{!}{=} \frac{(b-a)(d-c)}{12},$$

$\forall -2 \leq a \leq b \leq 2; -1.5 \leq c \leq d \leq 1.5$, with

$$\frac{\partial^2 \left[\int_c^d \int_a^b f(x_1, x_2) dx_1 dx_2 \right]}{\partial d \partial b} = f(b, d) \stackrel{!}{=} \frac{\partial^2 [(b-a)(d-c)/12]}{\partial d \partial b} = \frac{1}{12},$$

$\forall b \in [-2, 2], d \in [-1.5, 1.5]$.

An example from outer space III

Hence, the function

$$f(x_1, x_2) = \frac{1}{12} \mathbb{I}_{[-2,2]}(x_1) \mathbb{I}_{[-1.5,1.5]}(x_2)$$

can be used as a joint pdf for \mathbf{X} , and for any event $A \in \mathcal{R}(\mathbf{X})$ we obtain
 $P_{\mathbf{X}}(A) = \int \int_{x \in A} \frac{1}{12} dx_1 dx_2.$

Multivariate cdfs

Definition (Joint cdf)

The joint cdf of an n -dimensional random variable \mathbf{X} , denoted by F , is defined by

$$F : \mathbb{R}^n \rightarrow [0, 1] \quad \text{with} \quad F(b_1, \dots, b_n) = P_{\mathbf{X}}(X_1 \leq b_1, \dots, X_n \leq b_n),$$

$$\forall (b_1, \dots, b_n) \in \mathbb{R}^n.$$

For a **discrete random variable** the joint cdf obtains as

$$F(b_1, \dots, b_n) = \sum_{x_1 \leq b_1} \cdots \sum_{x_n \leq b_n} f(x_1, \dots, x_n), \quad \forall (b_1, \dots, b_n) \in \mathbb{R}^n,$$

and for a **continuous random variable** as

$$F(b_1, \dots, b_n) = \int_{-\infty}^{b_n} \cdots \int_{-\infty}^{b_1} f(x_1, \dots, x_n) dx_1 \cdots dx_n, \quad \forall (b_1, \dots, b_n) \in \mathbb{R}^n.$$

Some properties

Theorem (2.4)

For any multivariate cdf F , we have that

- (i) $\lim_{b_i \rightarrow -\infty} F(b_1, \dots, b_n) = P_{\mathbf{X}}(\emptyset) = 0$, *for some* $i = 1, \dots, n$;
- (ii) $\lim_{b_i \rightarrow \infty, \forall i} F(b_1, \dots, b_n) = P_{\mathbf{X}}(R(\mathbf{X})) = 1$;
- (iii) F is a non decreasing function on (x_1, \dots, x_n) , that is, $F(\mathbf{a}) \leq F(\mathbf{b})$ for (the vector inequality)

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} < \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \mathbf{b};$$

- (iv) Discrete joint cdfs have a countable number of jump discontinuities and joint cdfs for continuous random variables are continuous without jump discontinuities.

Joint pdfs

Theorem (2.5)

Let (X, Y) be a discrete bivariate random variable with joint cdf $F(x, y)$ and range $R(X, Y) = \{x_1 < x_2 < x_3 < \dots, y_1 < y_2 < y_3 < \dots\}$. Then the joint pdf obtains as

$$f(x_1, y_1) = F(x_1, y_1),$$

$$f(x_1, y_j) = F(x_1, y_j) - F(x_1, y_{j-1}), \quad j \geq 2,$$

$$f(x_i, y_1) = F(x_i, y_1) - F(x_{i-1}, y_1), \quad i \geq 2.$$

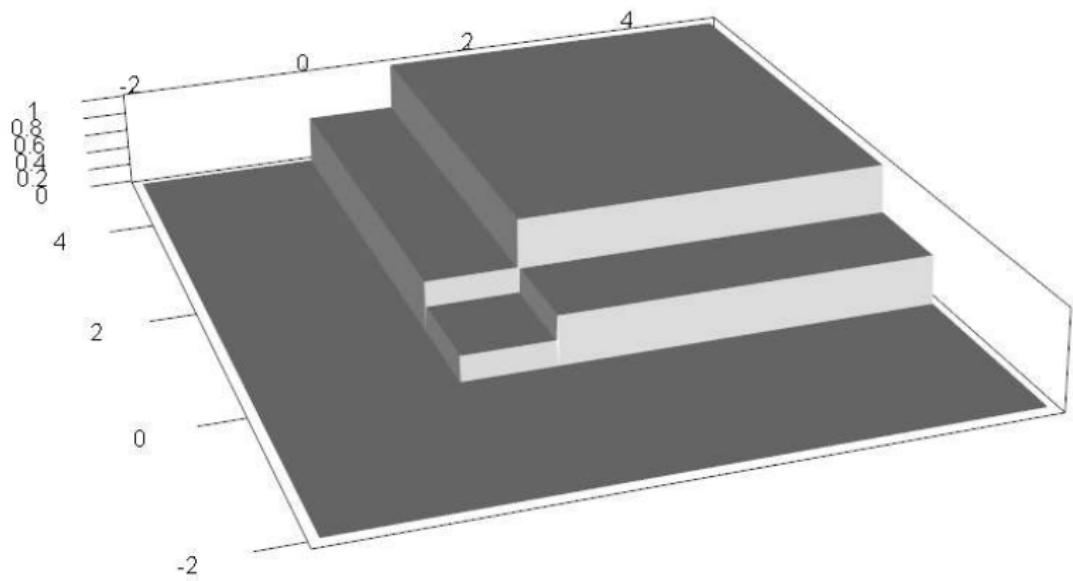
For $i, j \geq 2$,

$$f(x_i, y_j) = F(x_i, y_j) - F(x_i, y_{j-1}) - F(x_{i-1}, y_j) + F(x_{i-1}, y_{j-1}).$$

The result of the theorem for the **bivariate case** can be generalized to the ***n*-variate case**. (Cumbersome and omitted.)

Coin tosses

The cdf of two (0/1) fair coin tosses is as follows



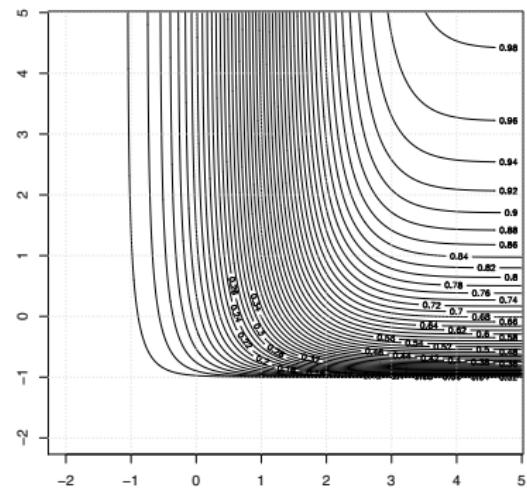
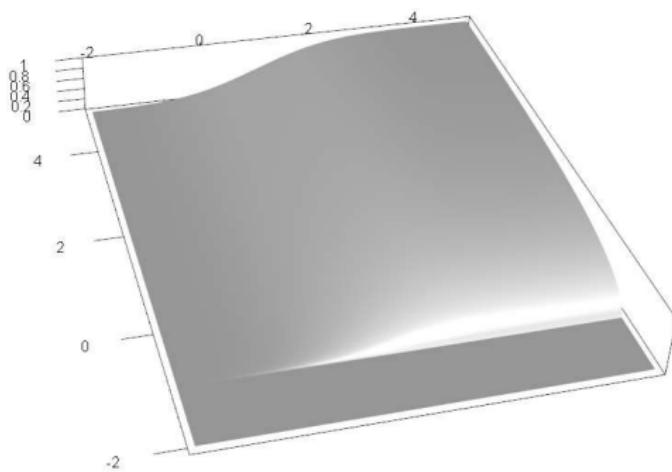
And the continuous case

Theorem (2.6)

Let $f(x_1, \dots, x_n)$ and $F(x_1, \dots, x_n)$ denote the joint pdf and cdf for a continuous multivariate random variable $\mathbf{X} = (X_1, \dots, X_n)$. Then the joint pdf for \mathbf{X} obtains as

$$f(x_1, \dots, x_n) = \begin{cases} \frac{\partial^n F(x_1, \dots, x_n)}{\partial x_1 \cdots \partial x_n}, & \text{wherever } f(\cdot) \text{ is continuous} \\ 0, & \text{elsewhere.} \end{cases}$$

An anonymous continuous example



Left: 3d plot; right: contour plot.

Outline

- 1 Multivariate random variables
- 2 Marginal distributions
- 3 Conditional distributions and independence
- 4 Up next

One component alone

Take for instance the joint distribution of $(X_1, X_2)'$,

$X_1 \setminus X_2$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$
$x_{1,1}$	0.2	0.2	0.05
$x_{1,2}$	0.1	0.2	0.25

Cells contain joint probabilities of $(X_1 = x_{1,i}, X_2 = x_{2,j})$.

For this distribution, what is the probability that $X_1 = x_{1,1}$ irrespective of X_2 ?

And how do we obtain the distribution of X_1 in general?

One component alone

Theorem (2.7)

Let $\mathbf{X} = (X_1, X_2)$ be a discrete random variable with joint pdf $f(x_1, x_2)$ and a range $R(\mathbf{X}) = R(X_1) \times R(X_2)$. The *marginal pdfs* are given by

$$f_1(x_1) = \sum_{x_2 \in R(X_2)} f(x_1, x_2), \quad \text{and} \quad f_2(x_2) = \sum_{x_1 \in R(X_1)} f(x_1, x_2).$$

To obtain the marginal pdf, we simply “*sum out*” the variables that are not of interest.

The continuous case

Theorem (2.8)

Let $\mathbf{X} = (X_1, X_2)$ be a continuous random variable with joint pdf $f(x_1, x_2)$. The corresponding **marginal pdfs** are given by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2, \quad \text{and} \quad f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1.$$

The concept of marginal pdfs can be straightforwardly generalized from the bivariate to the n -variate case as follows.

Many components

Definition (Marginal pdfs)

Let $f(x_1, \dots, x_n)$ be the joint pdf for the n -dimensional random variable (X_1, \dots, X_n) . Let $J = \{j_1, j_2, \dots, j_m\}$, $1 \leq m < n$, be a set of indices selected from the index set $I = \{1, 2, \dots, n\}$. Then the marginal density function for the m -dimensional random variable $(X_{j_1}, \dots, X_{j_m})$ is given by

$$f_{j_1 \dots j_m}(x_{j_1}, \dots, x_{j_m}) = \begin{cases} \sum_{(x_i \in R(X_i), i \in I - J)} \sum f(x_1, \dots, x_n) & (\text{discr.}) \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) \prod_{i \in I - J} dx_i & (\text{cont.}) \end{cases}$$

An abstract example

Consider the continuous random variable $\mathbf{X} = (X_1, X_2)$ with a **joint pdf**

$$f(x_1, x_2) = (x_1 + x_2)\mathbb{I}_{[0,1]}(x_1)\mathbb{I}_{[0,1]}(x_2).$$

The corresponding **marginal pdf** of X_1 obtains as

$$\begin{aligned} f_1(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 = \int_0^1 (x_1 + x_2)\mathbb{I}_{[0,1]}(x_1) dx_2 \\ &= \left[\left(x_1 x_2 + \frac{x_2^2}{2} \right) \mathbb{I}_{[0,1]}(x_1) \right]_{x_2=0}^{x_2=1} = \left(x_1 + \frac{1}{2} \right) \mathbb{I}_{[0,1]}(x_1). \end{aligned}$$

Outline

- 1 Multivariate random variables
- 2 Marginal distributions
- 3 Conditional distributions and independence
- 4 Up next

Assume that you know something...

Recall the joint distribution

	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$
$x_{1,1}$	0.2	0.2	0.05
$x_{1,2}$	0.1	0.2	0.25

E.g., the distribution of X_1 given $X_2 = x_2, 3$ is the 2-point distribution with

$$P(X_1 = x_{1,1} | X_2 = x_{1,3}) = 1/6$$

$$P(X_1 = x_{1,2} | X_2 = x_{1,3}) = 5/6.$$

So we can in principle derive the **conditional pdf of X_1 given X_2** , which can be used to assign the probability to the event $X_1 \in C$ given that (conditional on) $X_2 \in D$.

The discrete case

If (X_1, X_2) is a discrete random variable, the conditional pdf for X_1 given $X_2 \in D$ can be defined by

$$f(x_1|x_2 = d) = \frac{f(x_1, d)}{f_2(d)}$$

if D is a single point d , and, in general, by

$$f(x_1|x_2 \in D) = \frac{\sum_{x_2 \in D} f(x_1, x_2)}{\sum_{x_2 \in D} f_2(x_2)}.$$

From the conditional pdf we can straightforwardly derive the conditional cdf by using the conditional pdf in the general definition of a cdf. This holds in the continuous case as well, btw.

The continuous tentative

If (X_1, X_2) is a continuous random variable, we can substitute the summation operations by integrations, such that the conditional pdf for X_1 given $x_2 \in D$ is defined as

$$f(x_1|x_2 \in D) = \frac{\int_{x_2 \in D} f(x_1, x_2) dx_2}{\int_{x_2 \in D} f_2(x_2) dx_2}.$$

However, a problem arises when D is a single point d , such that

$$f(x_1|x_2 = d) = \frac{\int_d^d f(x_1, x_2) dx_2}{\int_d^d f_2(x_2) dx_2} = \frac{0}{0},$$

which is undefined!

This problem is circumvented by redefining the conditional probability in the continuous case in terms of a limit.

Concretely

Consider the definition

$$P(X_1 \in A | X_2 = d) \equiv \lim_{\epsilon \downarrow 0} P(X_1 \in A | d - \epsilon \leq X_2 \leq d + \epsilon).$$

After some algebra, we get

$$P(X_1 \in A | X_2 = d) = \int_{x_1 \in A} \frac{f(x_1, d)}{f_2(d)} dx_1.$$

Since pdfs are defined exactly this way, the desired conditional density must be the above integrand.

Hence the **conditional pdf of X_1 given $X_2 = d$** in the continuous case can be defined as

$$f(x_1 | x_2 = d) = \frac{f(x_1, d)}{f_2(d)}.$$

Note that it has exactly the same form as in the discrete case.

Example

Consider the continuous random variable $\mathbf{X} = (X_1, X_2)$ with joint pdf

$$f(x_1, x_2) = (x_1 + x_2)\mathbb{I}_{[0,1]}(x_1)\mathbb{I}_{[0,1]}(x_2),$$

and marginal pdf (see above):

$$f_2(x_2) = \left(x_2 + \frac{1}{2}\right)\mathbb{I}_{[0,1]}(x_2).$$

Then the conditional pdf of X_1 given $X_2 \leq .5$ obtains as

$$\begin{aligned} f(x_1|x_2 \leq .5) &\stackrel{(def.)}{=} \frac{\int_{-\infty}^{.5} f(x_1, x_2) dx_2}{\int_{-\infty}^{.5} f_2(x_2) dx_2} = \frac{\int_{-\infty}^{.5} (x_1 + x_2)\mathbb{I}_{[0,1]}(x_1)\mathbb{I}_{[0,1]}(x_2) dx_2}{\int_{-\infty}^{.5} \left(x_2 + \frac{1}{2}\right)\mathbb{I}_{[0,1]}(x_2) dx_2} \\ &= \left(\frac{4}{3}x_1 + \frac{1}{3}\right)\mathbb{I}_{[0,1]}(x_1). \end{aligned}$$

The conditional pdf of X_1 given $X_2 = .75$ is

$$f(x_1|x_2 = .75) \stackrel{(def.)}{=} \frac{f(x_1, .75)}{f_2(.75)} = \left(\frac{4}{5}x_1 + \frac{3}{5}\right)\mathbb{I}_{[0,1]}(x_1).$$

Multivariate case

Definition (Conditional pdfs)

Let $f(x_1, \dots, x_n)$ be the joint pdf for the n -dimensional random variable (X_1, \dots, X_n) . Let $J_1 = \{j_1, \dots, j_m\}$ and $J_2 = \{j_{m+1}, \dots, j_n\}$ be two mutually exclusive index sets whose union is equal to the index set $\{1, 2, \dots, n\}$. Then the conditional pdf for the m -dimensional random variable $(X_{j_1}, \dots, X_{j_m})$, given $(X_{j_{m+1}} = d_{m+1}, \dots, X_{j_n} = d_n)$ is given by

$$f(x_{j_1}, \dots, x_{j_m} \mid x_{j_i} = d_i, i = m + 1, \dots, n) = \frac{f(x_1, \dots, x_n)}{f_{j_{m+1} \dots j_n}(d_{m+1}, \dots, d_n)}$$

where $x_{j_i} = d_i$ if $j_i \in J_2$, when the marginal density in the denominator is positive valued.

From events to random variables

The independence of two events A and B means that
 $P(A \cap B) = P(A) \cdot P(B)$.

(How) Does this extend to random variables?

Definition (Independence of Random Variables)

The random variables X_1 and X_2 are said to be independent iff

$$P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1) \cdot P(X_2 \in A_2)$$

for all $A_1 \subset R(X_1)$ and $A_2 \subset R(X_2)$.

Factorization

The definition is not immediately operational since the factorization has to hold for all pairs of events.

Theorem (2.9)

The random variables X_1 and X_2 with joint pdf $f(x_1, x_2)$ and marginal pdfs $f_1(x_1)$ and $f_2(x_2)$ are independent, iff

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2) \quad \forall (x_1, x_2),$$

(except possibly at points of discontinuity for a joint continuous pdf f).

Conditional and marginal distributions

An important implication of the independence of X_1 and X_2 is that the conditional pdfs are identical to the corresponding marginal pdfs, that is,

$$f(x_1|x_2 = d) \stackrel{(def.)}{=} \frac{f(x_1, d)}{f_2(d)} = \frac{f_1(x_1)f_2(d)}{f_2(d)} = f_1(x_1).$$

Thus the probability of event $X_1 \in A$ is unaffected by the occurrence or nonoccurrence of event $X_2 = d$.

Meteorite again

Recall the meteorite example, where $\mathbf{X} = (X_1, X_2)$ is the point of strike with joint pdf

$$f(x_1, x_2) = \frac{1}{12} \mathbb{I}_{[-2,2]}(x_1) \mathbb{I}_{[-1.5,1.5]}(x_2),$$

Are X_1 and X_2 independent? The marginal pdfs are

$$f_1(x_1) = \frac{1}{12} \mathbb{I}_{[-2,2]}(x_1) \int_{-1.5}^{1.5} 1 dx_2 = \frac{1}{4} \mathbb{I}_{[-2,2]}(x_1)$$

$$f_2(x_2) = \frac{1}{12} \mathbb{I}_{[-1.5,1.5]}(x_2) \int_{-2}^2 1 dx_1 = \frac{1}{3} \mathbb{I}_{[-1.5,1.5]}(x_2)$$

Thus, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$, and X_1 and X_2 are independent.

Implications of independence

If X_1 and X_2 are independent, then knowing the marginal pdfs f_1 and f_2 is sufficient to determine the joint pdf: $f(x_1, x_2) = f_1(x_1)f_2(x_2)$.

This is not true in general: consider for some $\alpha \in [-1, 1]$ the joint pdf

$$f(x_1, x_2; \alpha) = [1 + \alpha(2x_1 - 1)(2x_2 - 1)]\mathbb{I}_{[0,1]}(x_1)\mathbb{I}_{[0,1]}(x_2).$$

For any choice of $\alpha \in [-1, 1]$, the marginal pdfs are

$$f_1(x_1) = \mathbb{I}_{[0,1]}(x_1) \quad \text{and} \quad f_2(x_2) = \mathbb{I}_{[0,1]}(x_2).$$

Hence, for all suitable values of α in the joint pdf f , we obtain the very same marginal pdfs f_1 and f_2 .

Thus, knowing f_1 and f_2 is insufficient to determine f and, in particular, the value of α .

The multivariate case

Definition (Independence in the n -variate case)

The random variables X_1, \dots, X_n are said to be independent iff

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n P(X_i \in A_i), \quad \text{for all } A_i \subset R(X_i).$$

Theorem (2.10)

The random variables X_1, \dots, X_n with joint pdf $f(x_1, \dots, x_n)$ and marginal pdfs $f_i(x_i)$, $i = 1, \dots, n$, are all independent of each other, iff

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i) \quad \forall (x_1, \dots, x_n),$$

(except possibly at points of discontinuity for a joint continuous pdf f).

Outline

- 1 Multivariate random variables
- 2 Marginal distributions
- 3 Conditional distributions and independence
- 4 Up next

Coming up

Transformations of random variables

Functions of Random Variables

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Random variables rule!

Focusing on distributions on the real line has its advantages,

... we may e.g. use calculus to characterize them.

Calculus can help with a further relevant question, namely

How to work out the distribution of some **transformation** of a random variable or vector.

Functions of Random Variables

- 1 Transformations of random variables
- 2 Distributions of functions of RVs
- 3 Numerical simulation
- 4 Up next

Outline

1 Transformations of random variables

2 Distributions of functions of RVs

3 Numerical simulation

4 Up next

Functions of random variables

A lot of models involving random components can be seen as signal processing units:

- Take deterministic inputs
- ... together with some random ones
- and (try to) analyze the outcome.

To do so, let's look at the most common case,

$$Y = g(X)$$

where the domain of g contains $\mathbb{R}(X)$.

(Models can be more complicated, say dynamic or high-dimensional, but we start simple.)

A (simple?) example

Say X is a standard normally distributed RV, i.e.

$$f_X = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

- This is a particular case of a **normal distribution $\mathcal{N}(\mu, \sigma^2)$** , with density

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}};$$

more about the normal family of distributions later.

Assume also that $g(x) = x + 5$.

What can we say about $Y = g(X)$?

What is Y ?

For starters, note that

- Y has random outcomes in general,
- since they depend on the outcomes of the *random* X .

So we want to find out,

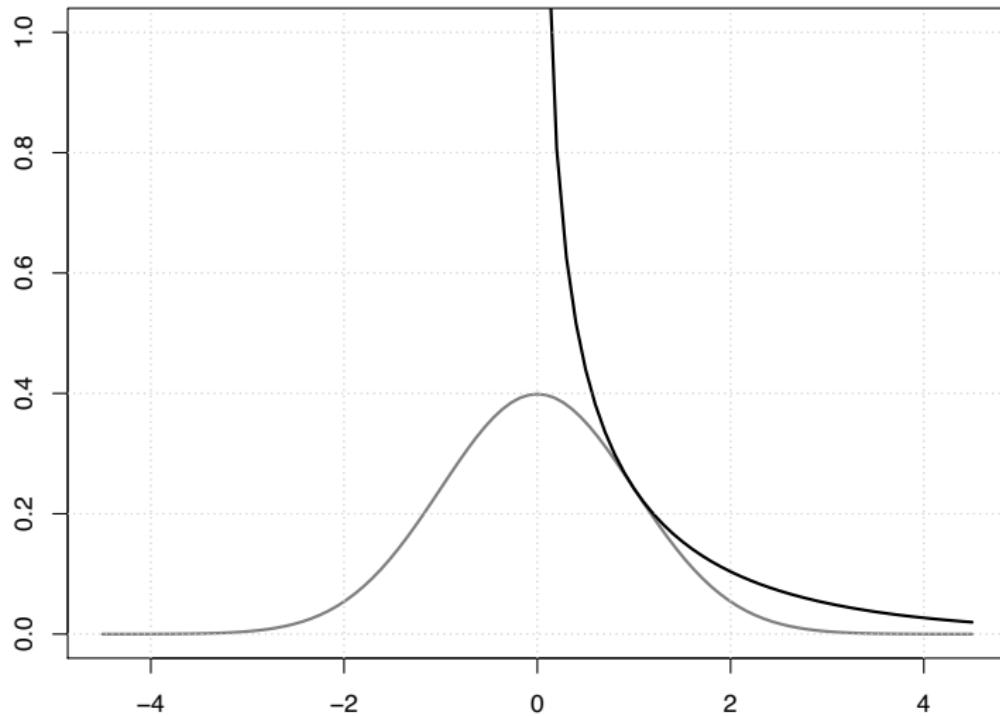
- (When) Is Y a random variable?
- What is the distribution of Y ?

Recall, X is a mapping $X(\omega) \mapsto \mathbb{R}$ (or \mathbb{R}^k).

- Therefore, $Y = g(X(\omega))$, and we do take Y to be an RV.¹
- How do we work out the distribution of Y ?

¹Technically, this is only the case if g is a so-called measurable function – but, since all functions we work with will be measurable, we'll just take for granted that Y is an RV. Measurability of g is only an issue with continuous RVs X , case in which it should simply be noted that (piecewise) continuity of g implies measurability.

Preview: standard normal vs. squared standard normal



Outline

1 Transformations of random variables

2 Distributions of functions of RVs

3 Numerical simulation

4 Up next

A discrete example

Let $\mathbf{X} = (X_1, X_2, X_3)'$ have a joint discrete pdf given by

x	(0,0,0)	(0,0,1)	(0,1,1)	(1,0,1)	(1,1,0)	(1,1,1)
$f_{\mathbf{X}}(\mathbf{x})$	1/8	3/8	1/8	1/8	1/8	1/8

What is the joint pdf of $\mathbf{Y} = (Y_1, Y_2)$ with $Y_1 = X_1 + X_2 + X_3$ and $Y_2 = |X_3 - X_2|$?

The mapping between the outcomes in the range of \mathbf{X} and that of \mathbf{Y} is

x	(0,0,0)	(0,0,1)	(0,1,1)	(1,0,1)	(1,1,0)	(1,1,1)
y	(0,0)	(1,1)	(2,0)	(2,1)	(2,1)	(3,0)

Hence the joint pdf of $\mathbf{Y} = (Y_1, Y_2)$ obtains as

y	(0,0)	(1,1)	(2,0)	(2,1)	(3,0)
$f_{\mathbf{Y}}(\mathbf{y})$	1/8	3/8	1/8	2/8	1/8

The equivalent events approach

This is the **equivalent-events approach**, applicable for discrete X . Concretely,

- Let $Y = g(X)$ be the function of interest, where X represents a discrete variable with pdf f_X ;
- Consider the set of elementary events x generating a particular elementary event y , i.e,

$$A_y = \{x : y = g(x), x \in \text{R}(X)\};$$

- Then the probability for the elementary event y can be written as

$$P_Y(y) = P_X(X \in A_y) = \sum_{\{x \in A_y\}} f_X(x) = f_Y(y),$$

which defines the discrete pdf (pmf) of Y .

(The extension to the case of multivariate variables is straightforward.)

Transformations of independent RVs

The equivalent events approach is in principle always valid,
... but is not immediately applicable for continuous RVs.

But here's a simple and intuitive, yet non-discriminating situation:

Theorem (2.11)

If X_1 and X_2 are independent random variables, and if Y_1 and Y_2 are defined as (measurable) functions thereof, $Y_1 = g_1(X_1)$ and $Y_2 = g_2(X_2)$, then Y_1 and Y_2 are independent.

Change of variables

In the continuous case, ...

Theorem (2.12)

Let X be a continuous random variable with a pdf $f(x)$ with support $\Xi = \{x : f(x) > 0\}$. Suppose that $y = g(x)$ is a continuously differentiable function with

- ① $\frac{dg(x)}{dx} \neq 0 \quad \forall x$ in some open interval Δ containing Ξ ,
- ② and an inverse $x = g^{-1}(y)$ defined $\forall y \in \Psi = \{y : y = g(x), x \in \Xi\}$.

Then the pdf of $Y = g(X)$ is given by

$$h(y) = f(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| \quad \text{for } y \in \Psi.$$

Example

Consider the Cobb-Douglas production function, $Q = \beta_0 \prod_{i=1}^k x_i^{\beta_i} e^W$, where Q : output, $x_i > 0$: deterministic quantities of input factors, β_i : corresponding partial production elasticities, $\beta_0 > 0$: efficiency parameter, and $W \sim \mathcal{N}(0, \sigma^2)$: stochastic error term.

What is the pdf of Q ? In order to answer this question rewrite Q as

$$Q = \exp \underbrace{\left\{ \ln \beta_0 + \sum_{i=1}^k \beta_i \ln x_i + W \right\}}_Z = \exp Z,$$

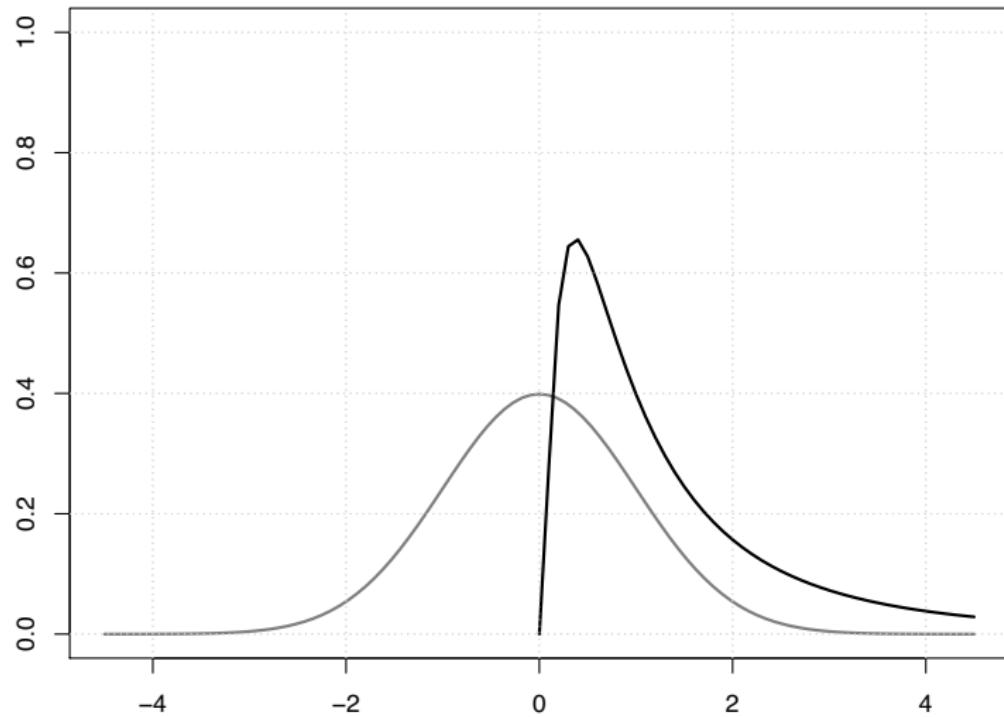
where $Z \sim \mathcal{N} \left(\underbrace{\ln \beta_0 + \sum_{i=1}^k \beta_i \ln x_i}_{\mu_Z}, \sigma^2 \right) = \mathcal{N}(\mu_Z, \sigma^2)$.

The function $q = \exp z$ is monotonic with $\frac{dq}{dz} = \exp z > 0 \forall z$. The inverse is $z = \ln q$ with $\frac{dz}{dq} = \frac{1}{q} > 0$. Thus Theorem 2.12 applies, and the pdf for Q is

$$h(q) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(\ln q - \mu_z)^2}{2\sigma^2} \right\}}_{f_Z(g^{-1}(q))} \cdot \underbrace{\left(\frac{1}{q} \right)}_{\frac{dg^{-1}(q)}{dq}}, \quad \text{for } q > 0.$$

This is the density of the **lognormal distribution**.

Standard normal vs. standard lognormal



Remarks

Theorem 2.12 **does not** apply to cases where the function g does not have an inverse – say $g(x) = x^2$ –, or is not smooth – say $g(x) = |x|$.

The trick is to apply Theorem 2.12 for each invertible piece of g , and then the law of total probability.

- E.g. for $X \sim \mathcal{N}(0, 1)$ and $g(x) = x^2$, Y follows a so-called chi-squared distribution with one degree of freedom.
- The χ^2 distribution with k degrees of freedom has density

$$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}\mathbb{I}(y \geq 0).$$

- If $X \sim \mathcal{N}(0, 1)$ and $g(x) = |x|$, Y follows the standard half-normal distribution with density $2/\sqrt{2\pi} \exp(-\frac{1}{2}y^2)\mathbb{I}(y \geq 0)$.

Furthermore, we may also go multivariate.

The multivariate case

Theorem (2.13)

Let \mathbf{X} be a continuous $(n \times 1)$ random vector with joint pdf $f(\mathbf{x})$ with support Ξ . Furthermore, let $\mathbf{g}(\mathbf{x})$ be a $(n \times 1)$ vector function which is

1. continuously differentiable $\forall \mathbf{x}$ in some open rectangle, $\Delta \supset \Xi$,
2. and with an inverse $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y}) \quad \forall \mathbf{y} \in \Psi = \{\mathbf{y} : \mathbf{y} = \mathbf{g}(\mathbf{x}), \mathbf{x} \in \Xi\}$.

Assume that the Jacobian matrix

$$\mathbf{J} = \left[\frac{\partial g_i^{-1}(\mathbf{y})}{\partial y_j} \right]_{i,j=1,\dots,n} \quad \text{satisfies} \quad \det(\mathbf{J}) \neq 0,$$

and assume that all partial derivatives in \mathbf{J} are continuous $\forall \mathbf{y} \in \Psi$. Then the joint pdf of $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is given by

$$h(\mathbf{y}) = f\left(g_1^{-1}(\mathbf{y}), \dots, g_n^{-1}(\mathbf{y})\right) |\det(\mathbf{J})| \quad \text{for } \mathbf{y} \in \Psi.$$

Further remarks

In the multivariate change-of-variable case, there are as many coordinates in \mathbf{y} as there are elements in the argument \mathbf{x} , i.e., n .

In cases where $\dim(\mathbf{y}) < \dim(\mathbf{x}) = n$, we need to

- introduce *auxiliary variables* to obtain an n -to- n function, and
- integrate out the auxiliary variables from the derived joint pdf.

This is e.g. how the so-called t distribution is derived.

The t density

Let $Z \sim \mathcal{N}(0, 1)$, let $Y \sim \chi_k^2$, and let Z and Y be independent. Then

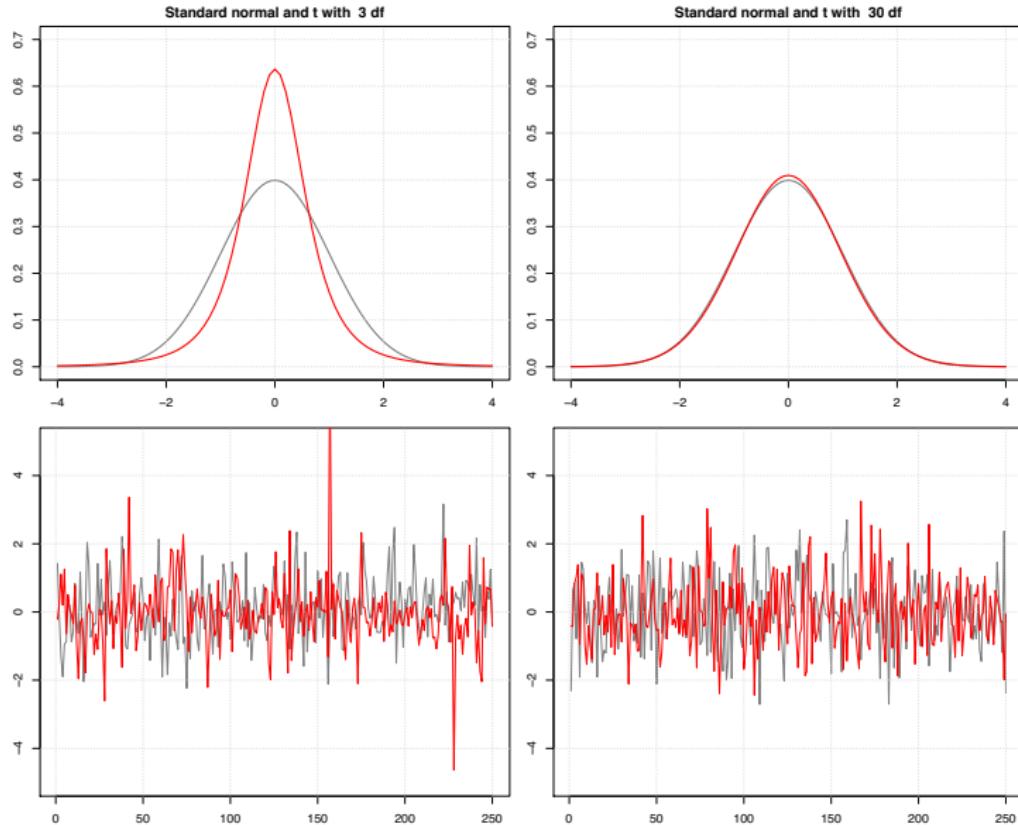
$T = \frac{Z}{\sqrt{Y/k}}$ is t -distributed with k degrees of freedom,

with density

$$f(t; k) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(k/2)\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-\left(\frac{k+1}{2}\right)}.$$

- The t -density is symmetric about 0 and has **fatter tails** than a normal,
- ... but its density converges to that of a normal for $v \rightarrow \infty$.

Comparison for standardized densities



Outline

- 1 Transformations of random variables
- 2 Distributions of functions of RVs
- 3 Numerical simulation
- 4 Up next

Closed form results are not that common

It is not always possible to derive the distribution of $g(X)$:

- The input density f_X may have a complicated expression
- The function g (or rather the inverse) may be complicated too
- In the multivariate case there may be “many” random inputs making analytical derivation of the Jacobian difficult.

One idea would be to resort to numerical approximations of (partial) derivatives and use computers to compute F_Y at any desired point.

Another, much more common in statistics, would be to *simulate* the random behavior of $g(X)$.

(See also the course on Statistical Computing.)

Introducing Monte Carlo simulation

The idea is to

- simulate a long sequence of realizations y_1, \dots, y_n from the distribution F_Y of interest,
- and use them rather than working with F_Y or so.

This is easily done here by generating x_1, \dots, x_n from the distribution F_X , and computing $y_i = g(x_i)$.

There is of course the question of whether approximating F_Y by the so-called *empirical* cdf of y_1, \dots, y_n is a good idea – as we'll see, the LLN (see chapter on asymptotics) guarantees that it is.

But what does “simulating a random sequence” involve?

Pseudo-random numbers

Any Monte Carlo (MC) simulation actually produces a *deterministic* sequence of numbers, which we call *pseudo-random*.

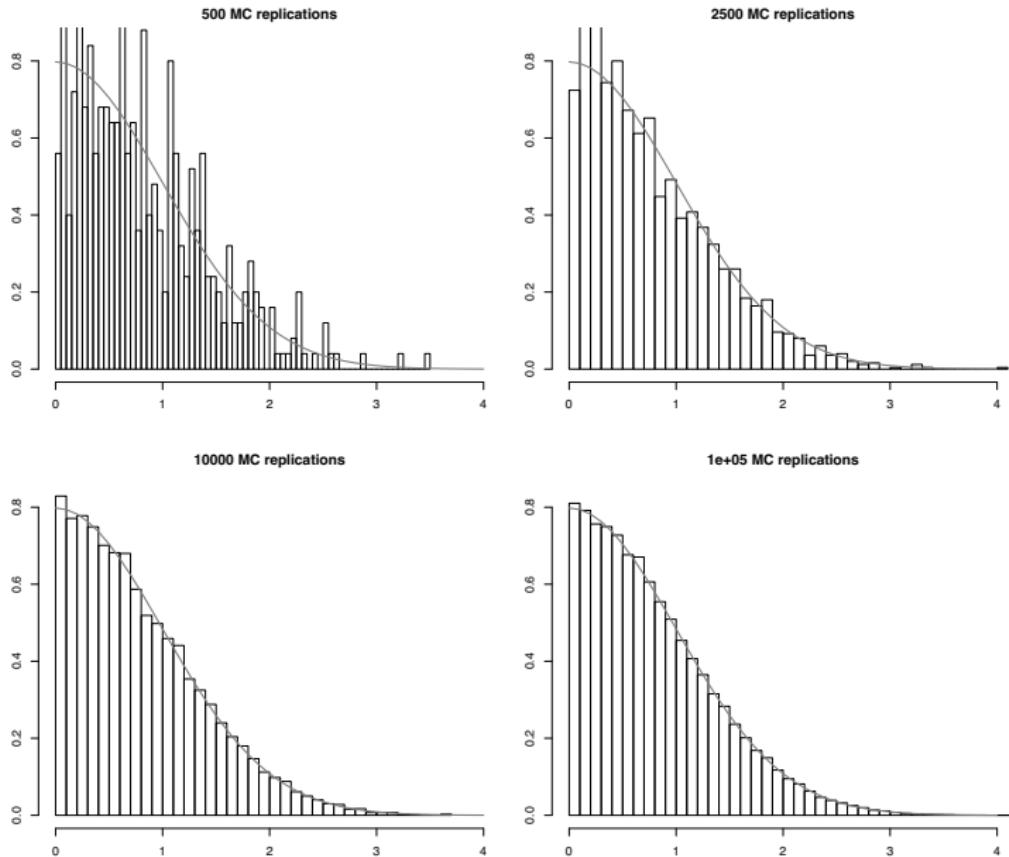
The trick is to generate them such that they look, for all purposes, as being random:

- They should follow the distribution of interest
- They should not be (statistically) distinguishable from random noise

Such numerical generators do exist, and we use them widely, but we don't go into details here.²

²See, again, Statistical Computing.

Example: simulating the standard half-normal distribution



Outline

- 1 Transformations of random variables
- 2 Distributions of functions of RVs
- 3 Numerical simulation
- 4 Up next

Coming up

Expectations of random variables

Expectations

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

The cdf is very informative

The cdf (or pdf) describes *an entire distribution*.

This may be overkill at times,
and we may want to focus on **specific characteristics**.

We often resort to **average** characteristics of possible outcomes,
... so today we formalize this notion (and also give it a new name).

Expectations

- 1 Expectation of a random variable
- 2 Properties of the expectation operator
- 3 Representing discrete pdfs via expectations
- 4 Up next

Outline

- 1 Expectation of a random variable
- 2 Properties of the expectation operator
- 3 Representing discrete pdfs via expectations
- 4 Up next

The nexus

The **expected value**, or expectation, of a random variable represents its probability-weighted average value;

It gives a **measure of the location** of X (the center of gravity of its pdf).

Definition (Expectation; discrete case)

The expected value of a discrete random variable exists, and is defined by

$$E(X) = \sum_{x \in R(X)} x \cdot f(x), \quad \text{iff} \quad \sum_{x \in R(X)} |x \cdot f(x)| = \sum_{x \in R(X)} |x| \cdot f(x) < \infty.$$

- The existence condition ensures that the sum $\sum_{x \in R(X)} x f(x)$ defining the expectation is **absolutely convergent**.
- The condition is sometimes called integrability of X .

Worth thinking about

- Thanks to the triangle inequality, absolute convergence implies standard convergence:

$$\sum_{x \in R(X)} |x| \cdot f(x) < \infty \quad \Rightarrow \quad \left| \sum_{x \in R(X)} x \cdot f(x) \right| < \infty.$$

such that the sum defining the expectation is finite and exists.

- Also, if $R(X)$ is finite and $|x| < \infty \forall x \in R(X)$, then $\sum_{x \in R(X)} |x| \cdot f(x) < \infty$ automatically.
- But if $R(X)$ is countably infinite there is no guarantee that $\sum_{x \in R(X)} |x| \cdot f(x) < \infty$.

Example

Consider a discrete random variable with pdf

$$f(x_k) = \frac{1}{2^k} \quad \text{with} \quad R(X) = \left\{ x_k = (-1)^k \frac{2^k}{k}, k = 1, 2, \dots \right\}.$$

The sum defining the expectation is

$$\begin{aligned} \sum_{k=1}^{\infty} x_k f(x_k) &= \sum_{k=1}^{\infty} \frac{(-1)^k}{k} = - \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} 1^k \\ &= -\ln(1+1). \quad \left[\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} x^k \Big|_{x \in (-1,1]} = \ln(1+x) \right] \end{aligned}$$

Thus, the sum is convergent, but not absolutely convergent since

$$\sum_{k=1}^{\infty} |x_k| f(x_k) = \sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>1/2} + \underbrace{\frac{1}{5} + \cdots + \frac{1}{8}}_{>1/2} + \cdots = \infty.$$

Moving on to integrals

Definition (Expectation; continuous case)

The expected value of a continuous random variable exists, and is defined by

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx, \quad \text{iff} \quad \int_{-\infty}^{\infty} |x| \cdot f(x) dx < \infty.$$

The existence condition is necessary to ensure that the improper Riemann integral $\int_{-\infty}^{\infty} x \cdot f(x) dx$ (and hence the expectation) exists.

Theorem (3.1)

If $|x| < c \forall x \in R(X)$, for some choice of $c \in (0, \infty)$. Then $E(X)$ exists.

The Cauchy (counter) example

Consider a random variable with pdf

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad -\infty < x < \infty \quad (\text{Cauchy distribution}).$$

Write now

$$\int_{-\infty}^{\infty} |x| f(x) dx = \int_{-\infty}^{\infty} \frac{|x|}{\pi} \frac{1}{1+x^2} dx = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2} dx,$$

For any positive number a we obtain

$$\int_0^a \frac{x}{1+x^2} dx = \left[\frac{\ln(1+x^2)}{2} \right]_{x=0}^{x=a} = \frac{\ln(1+a^2)}{2}.$$

Thus,

$$\int_{-\infty}^{\infty} |x| f(x) dx = \lim_{a \rightarrow \infty} \frac{2}{\pi} \int_0^a \frac{x}{1+x^2} dx = \frac{1}{\pi} \lim_{a \rightarrow \infty} \ln(1+a^2) = \infty.$$

Outline

- 1 Expectation of a random variable
- 2 Properties of the expectation operator
- 3 Representing discrete pdfs via expectations
- 4 Up next

Expectation of a function of random variables

We sometimes need to work with transformations of RVs, $Y = g(X)$.

If we only need $E(Y)$, we don't have to derive the pdf of Y .¹

Theorem (3.2)

Let X be a random variable with pdf $f(x)$. Then the expectation of random variable $Y = g(X)$ is given by

$$E(g(X)) = \begin{cases} \sum_{x \in R(X)} g(x) f(x) & (\text{discrete}) \\ \int_{-\infty}^{\infty} g(x) f(x) dx & (\text{continuous}). \end{cases}$$

¹Fortunately.

Probability and expectation

An application of the theorem is that the expectation of an indicator function equals the probability of the set being indicated:

Example

Let X be a variable with pdf f and recall that $\mathbb{I}_A(x) = \begin{cases} 1 & x \in A \\ 0 & \text{else} \end{cases}$.

Then,

$$\mathbb{E}(\mathbb{I}_A(X)) = \begin{cases} \sum_{x \in R(X)} \mathbb{I}_A(x) \cdot f(x) = \sum_{x \in A} f(x) = P(X \in A) & \text{(discrete)} \\ \int_{x \in R(X)} \mathbb{I}_A(x) \cdot f(x) dx = \int_{x \in A} f(x) dx = P(X \in A) & \text{(cont).} \end{cases}$$

Markov's inequality

Theorem (3.11 (Markov's inequality))

Let X be a random variable with pdf f , and let g be a nonnegative function of X . Then

$$\mathrm{P}(g(X) \geq a) \leq \frac{\mathrm{E}(g(X))}{a} \quad \text{for any } a > 0.$$

Note that a should be large to have nontrivial bounds, though.

Can we say anything about $E(g(X))$ in relation to $E(X)$?

Theorem (3.3 (Jensen's Inequality))

Let X be a non-degenerate random variable with expectation $E(X)$, and let g be a continuous function on an open interval I containing $R(X)$ (that is $R(X) \subseteq I$).

If g is convex on I , then $E(g(X)) \geq g(E(X))$;
if g is strictly convex on I , then $E(g(X)) > g(E(X))$.

Jensen's Inequality also applies to concave functions...

One immediate application of Jensen's Inequality shows that

$$E(X^2) \geq (E(X))^2, \quad \text{since } g(x) = x^2 \text{ is convex.}$$

Note that this implies that $\text{Var}(X) = E(X^2) - (E(X))^2 \geq 0$.²

²In case you have not met the variance before: we introduce it next week formally.

Some more properties of the expectation

Theorem (3.4)

If c is a constant, then $E(c) = c$.

Theorem (3.5)

If c is a constant, then $E(cX) = cE(X)$.

Theorem (3.6)

$$E\left(\sum_{i=1}^k g_i(X)\right) = \sum_{i=1}^k E(g_i(X)).$$

Corollary (3.1)

$$E(a + bX) = a + bE(X).$$

Outline

- 1 Expectation of a random variable
- 2 Properties of the expectation operator
- 3 Representing discrete pdfs via expectations
- 4 Up next

Recall: the Riemann integral

Definition

A function g is said to be Riemann-integrable on an interval $[a, b]$ if, for any partition $a = x_0, \dots, x_n = b$ of the interval $[a, b]$, the limit

$$\lim_{\max|x_i - x_{i-1}| \rightarrow 0} \sum_{i=1}^n g(\xi_i) (x_i - x_{i-1})$$

exists and is finite for any $\xi_i \in [x_i, x_{i-1}]$.

- The limit is then denoted by $\int_a^b g(x)dx$
- Piecewise continuity is sufficient for (Riemann) integrability
- The (Riemann) integral is “the area under the curve”

A generalization of Riemann integration

Definition

A function g is said to be Stieltjes-integrable on an interval $[a, b]$ with integrator F if, for any partition $a = x_0, \dots, x_n = b$ of the interval $[a, b]$, the limit

$$\lim_{\max|x_i-x_{i-1}| \rightarrow 0} \sum_{i=1}^n g(\xi_i) (F(x_i) - F(x_{i-1}))$$

exists and is finite for any $\xi_i \in [x_i, x_{i-1}]$.

- The limit is then denoted by $\int_a^b g(x)dF(x)$
- Piecewise continuity of g and monotonicity of F are sufficient for Stieltjes integrability, provided that discontinuities of g and F are not common.
- Improper integrals and integrals over unions of intervals are treated the usual (Riemann) way

Equivalence

The Stieltjes and Riemann integrals are closely related

- In fact, if $F(x)$ is linear, they are (more or less) the same.
- Moreover, if F is smooth, $\int_A g(x)dF(x) = \int_A g(x)F'(x)dx$
- This is relevant for distributions:

Example

Let f be the pdf of a continuous random variable X and $F(f)$ the associated cdf (pdf). Then,

$$E(g(X)) = \int_{R(X)} g(x)f(x)dx = \int_{R(X)} g(x)dF(x).$$

Properties at a glance

Theorem

Let $g : [a, b] \rightarrow \mathbb{R}$ be Stieltjes integrable w.r.t. right-continuous F . Then,

- ① **Linearity:** for $A, B \in \mathbb{R}$,

$$\int_a^b (Ag_1(x) + Bg_2(x)) dF(x) = A \int_a^b g_1(x) dF(x) + B \int_a^b g_2(x) dF(x)$$

$$\int_a^b g(x) d(AF_1(x) + BF_2(x)) = A \int_a^b g(x) dF_1(x) + B \int_a^b g(x) dF_2(x)$$

$$\int_a^b g(x) dF(x) = \int_a^c g(x) dF(x) + \int_c^b g(x) dF(x) \quad \text{where } c \in (a, b).$$

- ② **Integration by parts:** $\int_a^b g(x) dF(x) = g(x)F(x)|_a^b - \int_a^b F(x) dg(x)$.
- ③ **Equivalence with Riemann integral when F is smooth.**
- ④ **Change of variables:** $\int_c^d g(h(y)) dF(h(y)) = \int_{h(c)}^{h(d)} g(x) dF(x)$.

The Stieltjes integral is more flexible

The integrator F does not have to be continuous!

Lemma

Let F be piecewise smooth, right-continuous with a jump discontinuity at $x = x_0 \in [a, b]$, and g piecewise continuous, continuous at x_0 . Then,

$$\int_a^b g(x) dF(x) = \lim_{c \nearrow x_0} \int_a^c g(x) dF(x) + \int_{x_0}^b g(x) dF(x) \\ + g(x_0) (F(x_0+) - F(x_0-))$$

where $F(x_0+)$ ($F(x_0-)$) stands for the limit of F at x_0 to the right (to the left).

Unifying discrete and continuous distributions

Example

Take the two-point distribution given by

$$P(X = 0) = 1 - p \quad \text{and} \quad P(X = 1) = p$$

(the Bernoulli distribution with success probability p), with expectation

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Its cdf is

$$F(x) = (1 - p)\mathbb{I}(x \geq 0) + p\mathbb{I}(x \geq 1),$$

and the above lemma indicates that

$$\int_{-\infty}^{\infty} x dF(x) = p = E(X).$$

A unified notation

Recall: if F is smooth,

$$\int_a^b g(x)dF(x) = \int_a^b g(x)F'(x)dx.$$

Then...

- The Stieltjes integral on the l.h.s. exists for discontinuous F as well.
- May exploit the equality to define a “derivative” of F at its jumps.
- Focus to this end on piecewise smooth F with one finite jump at $x_0 \in (a, b)$.

Jumps I

Split F in smooth and nonsmooth components,

$$F(x) = \tilde{F}(x) + (F(x_0+) - F(x_0-)) H(x - x_0)$$

where \tilde{F} is smooth and $H(x)$ is a jump function at 0,

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}.$$

With $C = F(x_0+) - F(x_0-)$, the previous Lemma delivers

$$\int_a^b g(x) dF(x) = \int_a^b g(x) d\tilde{F}(x) + C \int_a^b g(x) dH(x - x_0).$$

Jumps II

Would we be able to differentiate H ,

$$\int_a^b g(x) dF(x) = \int_a^b g(x) \tilde{F}'(x) dx + C \int_a^b g(x) H'(x - x_0) dx.$$

At the same time (see previous Lemma),

$$\begin{aligned} \int_a^b g(x) dF(x) &= \int_a^{x_0} g(x) dF(x) + \int_{x_0}^b g(x) dF(x) + Cg(x_0) \\ &= \int_a^b g(x) \tilde{F}'(x) dx + Cg(x_0) \end{aligned}$$

since $d\tilde{F} = dF$ for $x < x_0$ and $x > x_0$. Therefore, we should have

$$\int_a^b g(x) H'(x - x_0) dx = g(x_0).$$

At the same time, $H' = 0$ for all $x \neq 0$, but H' is undefined at $x = 0$.

The δ function

We now have all the ingredients we need:

Definition (Dirac's δ)

The generalized function $\delta(x)$ satisfying

- ① $\delta(x) = 0$ for all $x \neq 0$ and
- ② $\int_{-\infty}^{\infty} g(x)\delta(x)dx = g(0)$

is called the Dirac's δ , and we write $\delta(x) = H'(x)$.

Dirac's δ can also be seen as the pdf of the limit of convergence in probability to a constant.

Discrete pdfs

We may now write for any discrete pdf

$$f(x) = f(x)\mathbb{I}_{R(X)}(x)$$

with the interpretation that $f(x) = \text{P}(X = x)$ ³ in a more intuitive way,

$$f(x) = \sum_{x_0 \in R(X)} \text{P}(X = x_0) \delta(x - x_0),$$

one that also allows for nice integration.

The (generalized) pdf of a point mass distribution (i.e. cdf $H(x - x_0)$) is thus $\delta(x - x_0)$.

³But not with the properties of a “proper” density function, since, in the Riemann world, $\int_{R(X)} f(x)\mathbb{I}_A(x)dx = 0$ for discrete sets A .

Outline

- 1 Expectation of a random variable
- 2 Properties of the expectation operator
- 3 Representing discrete pdfs via expectations
- 4 Up next

Coming up

Moments and other functionals

Moments of Random Variables

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Summarizing a distribution?

Expectations allow one to discuss “average”, or “typical”, outcomes in a rigorous manner.

We motivated this by the need to analyze distributions using selected average characteristics rather than the pdf or cdf.

- We note that such averages do not replace the cdf/pdf,
- ... they rather help us grasp various aspects of cdfs/pdfs more intuitively.

Today, we'll get more specific and target particular aspects of probability distributions by focusing on specific expectations.

Moments of Random Variables

- 1 Moments of a random variable
- 2 Special moments and the MGF
- 3 Other statistical functionals
- 4 Up next

Outline

1 Moments of a random variable

2 Special moments and the MGF

3 Other statistical functionals

4 Up next

Special expectations

Moments are **expectations of power functions of random variables**. They can be used to measure certain characteristics of given distributions.

We distinguish between non-central and central moments.

Definition (*r*th non-central moment)

Let X be a random variable with pdf $f(x)$. Then the *r*th non-central moment of X , denoted by μ'_r , is defined as

$$\mu'_r = E(X^r) = \begin{cases} \sum_{x \in R(X)} x^r f(x) & (\text{discrete}) \\ \int_{-\infty}^{\infty} x^r f(x) dx & (\text{continuous}). \end{cases}$$

Remarks

- Note that $\mu'_0 = E(X^0) = 1$.
- The first non-central moment is simply the expectation (also called the mean) of the random variable, that is $\mu'_1 = E(X)$; it is often denoted by μ .
- The mean is informative about the location of a distribution.
- Central moments aim at a location-free characterization of a distribution.

... and even more special ones

Definition (r th central moment)

Let X be a random variable with pdf $f(x)$. Then the r th central moment of X , denoted by μ_r , is defined as

$$\mu_r = E((X - \mu)^r) = \begin{cases} \sum_{x \in R(X)} (x - \mu)^r f(x) & (\text{discrete}) \\ \int_{-\infty}^{\infty} (x - \mu)^r f(x) dx & (\text{continuous}). \end{cases}$$

- Note that $\mu_0 = E((X - \mu)^0) = 1$, and $\mu_1 = E(X - \mu) = 0$.
- The second central moment is commonly known as the variance.

Existence of Moments

Theorem (3.12)

If $E(|X|^r)$ exists for an $r > 0$, then $E(|X|^s)$ exists $\forall s \in [0, r]$.

The theorem implies, that if $E(|X|^r)$ does not exist, then necessarily $E(|X|^s)$ cannot exist for $s > r$.

Also, it noncentral moments of order $r \geq 1$ exist, so do noncentral ones.

Also,

Theorem (3.13)

If $E(|Y - \mu|^r)$ exists for an $r > 0$, then $E(|Y - \mu|^s)$ exists $\forall s \in [0, r]$.

Example

Consider the pdf

$$f(x) = \frac{2}{(x+1)^3} \mathbb{I}_{[0,\infty)}(x).$$

Examine $E(X^\alpha)$, i.e.

$$E X^\alpha = \int_0^\infty \frac{x^\alpha 2}{(x+1)^3} dx = 2 \int_1^\infty (y-1)^\alpha y^{-3} dy,$$

(obtained by substituting $y = x + 1$, so that $y - 1 = x$ and $dy = dx$). If $\alpha = 2$, we get

$$E(X^2) = 2 \lim_{b \rightarrow \infty} \left[\ln(y) + 2y^{-1} - \frac{1}{2}y^{-2} \right]_{y=1}^{y=b} = \infty.$$

Thus, $E(X^2)$ does not exist. By the theorem, moments of order larger than 2 also do not exist.

Outline

1 Moments of a random variable

2 Special moments and the MGF

3 Other statistical functionals

4 Up next

Variance and standard deviation

Definition

The **variance** of a random variable X is the 2nd central moment, $\text{Var}(X) = E((X - \mu)^2)$, and will be denoted by the symbol σ^2 .

The non-negative square root of $\text{Var}(X)$ is the **standard deviation** of X and will be denoted by the symbol σ .

- The variance and standard deviation are measures of the dispersion of a distribution around the mean.
- We have a simple connection to noncentral moments,
$$E((X - \mu)^2) = E(X^2) - \mu^2.$$

A special case

Corollary (3.3 (Chebyshev's inequality))

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \text{for } k > 0.$$

Equivalently, $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$.

If $\sigma^2 \rightarrow 0$, then the distribution of X “collapses” to that of μ .

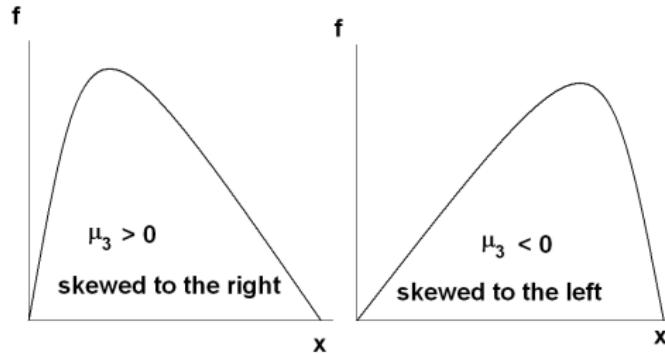
Further moments

Definition (Symmetry of a pdf)

The pdf f is said to be symmetric around μ iff

$f(\mu + \delta) = f(\mu - \delta)$ for any $\delta > 0$. Otherwise f is said to be skewed.

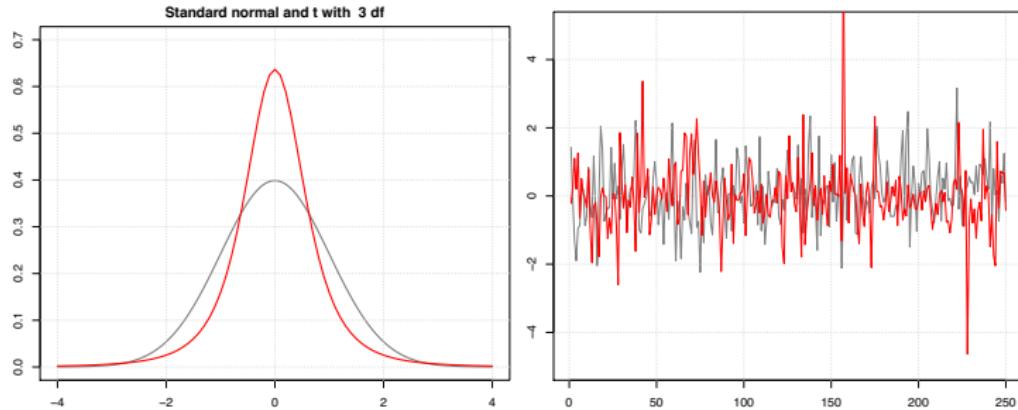
- A symmetric pdf has necessarily $\mu_3 = E((X - \mu)^3) = 0$.
(In fact, $\mu_{2K+1} = 0$ under symmetry.)
- For $\mu_3 > 0$ ($\mu_3 < 0$) the pdf said to be **skewed to the right (left)**.



The kurtosis

A distribution is said to have fat tails if it tends to generate too many large outcomes away from the bulk of the distribution.

To put things in perspective, recall the (standardized) t distribution, compared to the standard normal.



Relation between central and noncentral moments

Given noncentral moments we may derive the central ones, since

$$\mu_r = E((X - \mu)^r) = \sum_{i=0}^r \binom{r}{i} (-1)^i E(X^{r-i}) \mu^i.$$

E.g. for the central 4th order moment we have

$$\mu_4 = E(X^4) - 4\mu E(X^3) + 6\mu^2 E(X^2) - 3\mu^4.$$

(with μ the somewhat misleading but very common notation for $\mu'_1 = E(X)$.)

Mnemotechnics

The Moment-Generating Function (MGF) can be used to determine moments of a random variable.

Definition (Moment-Generating Function)

The MGF of a random variable X , denoted by $M_X(t)$, is

$$M_X(t) = \mathbb{E}(e^{tX}) = \begin{cases} \sum_{x \in R(X)} e^{tx} f(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx & \text{(continuous)}, \end{cases}$$

provided that the expectation exists for t in some neighborhood of 0. That is, there exists an $h > 0$ such that $\mathbb{E}(e^{tX})$ exists $\forall t \in (-h, h)$.

The main practical use of the MGF is not to generate moments, but to help in characterizing a distribution.

The first use

The condition that $M_X(t)$ be defined $\forall t \in (-h, h)$ is a technical condition ensuring that $M_X(t)$ is differentiable at the point $t = 0$.

Differentiability is useful.

Theorem (3.14)

Let X be a random variable for which the MGF $M_X(t)$ exists. Then

$$\mu'_r = E(X^r) = \frac{d^r M_X(t)}{dt^r} \Big|_{t=0}$$

Also useful: one may “invert” the MGF, $f(x) = \frac{1}{2\pi} \int_0^{2\pi} M_X(i\theta) e^{-ix\theta} d\theta$.

Example

Consider the pdf

$$f(x) = e^{-x} \mathbb{I}_{(0, \infty)}(x) \quad (\text{pdf of an exponential distribution}).$$

The MGF is given by

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} e^{-x} \mathbb{I}_{(0, \infty)}(x) dx = \int_0^{\infty} e^{x(t-1)} dx \\ &= \left[\frac{e^{x(t-1)}}{t-1} \right]_{x=0}^{x=\infty} \Bigg|_{t<1} = 0 - \frac{1}{t-1} = \frac{1}{1-t}. \end{aligned}$$

The mean and the 2nd non-central moment are given by

$$\mu = \frac{dM_X(t)}{dt} \Bigg|_{t=0} = \frac{1}{(1-t)^2} \Bigg|_{t=0} = 1, \quad \mu'_2 = \frac{d^2M_X(t)}{dt^2} \Bigg|_{t=0} = \frac{2}{(1-t)^3} \Bigg|_{t=0} = 2.$$

Taylor around $t = 0$

The MGF $M_X(t) = \text{E}(e^{tX})$ can be written as a series expansion in terms of the moments of the pdf of X :

$$\begin{aligned} M_X(t) = \text{E}(e^{tX}) &= \text{E}\left(e^{0X} + \frac{1}{1!}[Xe^{0x}]t + \frac{1}{2!}[X^2 e^{0x}]t^2 + \frac{1}{3!}[X^3 e^{0x}]t^3 + \dots\right) \\ &= 1 + \mu'_1 t + \frac{1}{2!} \mu'_2 t^2 + \frac{1}{3!} \mu'_3 t^3 + \dots \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} t^i \mu'_i. \end{aligned}$$

Therefore: if the MGF exists, it characterizes an infinite set of moments.

The existence of all moments does not imply the existence of the MGF, though.

Useful properties

Let X_1, \dots, X_n be independent random variables having MGFs $M_{X_i}(t)$, $i = 1, \dots, n$. Then we get

- for $Y = aX_i + b$ the MGF

$$M_Y(t) = E\left(e^{(aX_i+b)t}\right) = e^{bt} M_{X_i}(at);$$

- for $Y = \sum_{i=1}^n X_i$ the MGF

$$M_Y(t) = E\left(e^{(\sum_{i=1}^n X_i)t}\right) = \underbrace{E\left(\prod_{i=1}^n e^{X_i t}\right)}_{\text{by independence}} = \prod_{i=1}^n E e^{X_i t} = \prod_{i=1}^n M_{X_i}(t);$$

- for $Y = \sum_{i=1}^n a_i X_i + b$ the MGF

$$M_Y(t) = e^{bt} \prod_{i=1}^n M_{X_i}(a_i t).$$

The second use

Theorem (3.15 (MGF Uniqueness Theorem; scalar case))

If an MGF exists for a random variable X having pdf $f(x)$, then

- the MGF is unique;
- and, conversely, the MGF determines the pdf of X uniquely, at least up to a set of points having probability 0.

This ...

- allows us e.g. to identify distributions:
 - if a distribution has the MGF $e^{t^2/2}$
 - (which, as we'll see in future classes, is the MGF of the standard normal)
 - then it must be the standard normal distribution, and
- holds for random vectors as well.

Example

Suppose Z has an MGF defined by $M_Z(t) = \frac{1}{1-t}$ for $|t| < 1$.

Now, consider the pdf

$$f(x) = e^{-x} \mathbb{I}_{(0,\infty)}(x), \quad \text{which has an MGF } M_X(t) = \frac{1}{1-t}$$

(see the previous Example). Then, by the uniqueness theorem, the pdf of Z must be

$$Z \sim f(z) = e^{-z} \mathbb{I}_{(0,\infty)}(z).$$

(Almost everywhere.)

Outline

1 Moments of a random variable

2 Special moments and the MGF

3 Other statistical functionals

4 Up next

Functionals

Moments are functions of the cdf F , say $\tau(F)$. Since the argument of τ is a function, we call τ a functional.

Moments are so-called linear functionals (of the form $\int g(x)dF(x)$); and they are not the only functionals of interest ...

Definition (Median)

Any number, b , satisfying

$$\mathrm{P}(X \leq b) \geq 1/2 \quad \text{and} \quad \mathrm{P}(X \geq b) \geq 1/2$$

is called a **median** of X and is denoted by $\mathrm{med}(X)$.

The median is an alternative location measure for the distribution.¹

¹ And statistical folklore says it's more robust to so-called outliers.

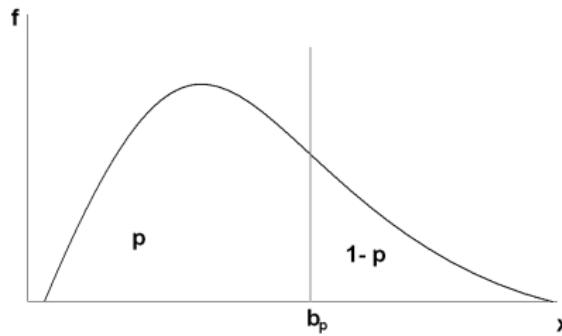
The median is a special quantile

Definition (Quantile)

A number q_p is a **quantile** of X of order p (or the $(100p)$ th percentile of X) if

$$P(X \leq q_p) \geq p \quad \text{and} \quad P(X \geq q_p) \geq 1 - p.$$

In case of non-uniqueness, a convenient choice is $q_p = \inf\{x : F(x) \geq p\}$.



And unique quantiles have special properties

- Quantiles at any p are unique if the cdf F is continuous; $q_p = F^{-1}(p)$.
- Quantiles are (location) equivariant, i.e. if X has quantile q_p , then $X + \mu$ has quantile $q_p + \mu$
- Quantiles are linear in scale, i.e. if X has quantile q_p , then σX has quantile σq_p
- Unique quantiles minimize a certain expectation,

$$q_p = \arg \min_{q^*} \mathbb{E} (\rho_p (X - q^*))$$

where ρ_p is the so-called quantile check function,

$$\rho_p(u) = \begin{cases} -(1-p)u, & u < 0; \\ pu, & u \geq 0. \end{cases}$$

Outline

- 1 Moments of a random variable
- 2 Special moments and the MGF
- 3 Other statistical functionals
- 4 Up next

Coming up

Joint and conditional moments

Multivariate Expectations and Moments

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Getting multivariate

We defined moments and related quantities (like the MGF) for **scalar** random variables.

We may wonder what happens with random **vectors**.

Obviously, we may work with the moments of the marginal distributions of each element of the random vector.

- But, just as the joint distribution was more than just the set of marginal distributions,
- ... there is something to learn from **joint** moments.

Multivariate Expectations and Moments

- 1 Multivariate expectations and moments
- 2 Covariance and correlation
- 3 Conditional expectations
- 4 Up next

Outline

1 Multivariate expectations and moments

2 Covariance and correlation

3 Conditional expectations

4 Up next

The scalar case is not enough...

So far, we considered the expectation of a **function of a univariate random variable**. But...

Theorem (3.7)

Let (X_1, \dots, X_n) be a multivariate random variable with joint pdf $f(x_1, \dots, x_n)$. Then the expectation of random variable $Y = g(X_1, \dots, X_n)$ is given by

$$E(Y) = \begin{cases} \sum_{(x_1, \dots, x_n) \in R(X)} g(x_1, \dots, x_n) f(x_1, \dots, x_n) & (\text{discrete}) \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \cdots dx_n & (\text{continuous}). \end{cases}$$

Multivariate results

Theorem (3.8)

$$\mathrm{E}\left(\sum_{i=1}^k g_i(X_1, \dots, X_n)\right) = \sum_{i=1}^k \mathrm{E}(g_i(X_1, \dots, X_n)).$$

Corollary (3.2)

$$\mathrm{E}\left(\sum_{i=1}^k X_i\right) = \sum_{i=1}^k \mathrm{E}(X_i).$$

And the much more interesting

Theorem (3.9)

Let X_1, \dots, X_n be independent random variables. Then

$$\mathrm{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathrm{E}(X_i).$$

Joint distributions...

In the case of multivariate random variables, *joint moments* characterize the relationship between the individual variables.

Definition (Joint non-central moment)

Let X and Y be two random variables with joint pdf $f(x, y)$. Then the joint non-central moment of (X, Y) of order (r, s) is defined as

$$\mu'_{r,s} = \mathbb{E}(X^r Y^s) = \begin{cases} \sum_{x \in R(X)} \sum_{y \in R(Y)} x^r y^s f(x, y) & \text{(discrete)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s f(x, y) dx dy & \text{(continuous).} \end{cases}$$

Multivariate MGFs

Definition (Moment-Generating Function; multivariate)

The MGF of a multivariate random variable $\mathbf{X} = (X_1, \dots, X_N)'$ is

$$M_{\mathbf{X}}(\mathbf{t}) = E\left(e^{\mathbf{t}' \mathbf{X}}\right) = E\left(e^{\sum_{i=1}^n t_i X_i}\right), \quad \text{where } \mathbf{t} = (t_1, \dots, t_n)',$$

if the expectation exists for all t_i in some neighborhood of 0, $i = 1, \dots, n$.

I.e. $\exists h > 0$ such that $E\left(e^{\mathbf{t}' \mathbf{X}}\right)$ exists $\forall t_i \in (-h, h)$, $i = 1, \dots, n$.

The r th order non-central moment of X_i obtains from the r th order partial derivative w.r.t. t_i ,

$$\mu'_r(X_i) = E(X_i^r) = \frac{\partial^r M_{\mathbf{X}}(\mathbf{t})}{\partial t_i^r} \Big|_{\mathbf{t}=0}.$$

Cross partial derivatives deliver *joint non-central moments*,

$$E(X_i^r X_j^s) = \frac{\partial^{r+s} M_{\mathbf{X}}(\mathbf{t})}{\partial t_i^r \partial t_j^s} \Big|_{\mathbf{t}=0}.$$

The central version

Definition (Joint central moment)

Let X and Y be two random variables with joint pdf $f(x, y)$. Then the joint central moment of (X, Y) of order (r, s) is defined as

$$\mu_{r,s} = \begin{cases} \sum_{x \in R(X)} \sum_{y \in R(Y)} (x - E(X))^r (y - E(Y))^s f(x, y) & (\text{discrete}) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))^r (y - E(Y))^s f(x, y) dx dy & (\text{continuous}). \end{cases}$$

The joint moment of order $(1, 1)$, namely $\mu_{1,1}$, is commonly known as the **covariance**, which measures the ‘linear association’ between X and Y .

We use it so often that it pays to discuss it in more detail.

Outline

- 1 Multivariate expectations and moments
- 2 Covariance and correlation
- 3 Conditional expectations
- 4 Up next

The queen of joint moments

Definition (Covariance)

The **covariance between the random variables X and Y** is the joint central moment of the order $(1, 1)$,

$$\sigma_{XY} = \text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))).$$

The covariance can be represented in terms of non-central moments:

$$\begin{aligned}\sigma_{XY} &= E((X - E(X))(Y - E(Y))) \\ &= E(XY - E(X)Y - E(Y)X + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y).\end{aligned}$$

From this relationship we obtain e.g. that

$$E(XY) = E(X)E(Y) \quad \text{iff} \quad \sigma_{XY} = 0.$$

Some results

Theorem (3.16 (Cauchy-Schwarz Inequality))

$$(E(WZ))^2 \leq E(W^2) E(Z^2).$$

Theorem (3.17 (Covariance bound))

$$|\sigma_{XY}| \leq \sigma_X \sigma_Y.$$

Using this upper bound, we can define a normalized version of the covariance, the so-called **correlation**.

Definition (Correlation)

The correlation between the random variables X and Y is defined by

$$\text{corr}(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

More on correlation

Theorem (3.18 (Correlation bound))

$$-1 \leq \rho_{XY} \leq 1.$$

A fundamental relationship between the covariance and the stochastic (in)dependence is indicated in the next theorem.

Theorem (3.19)

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$.

The converse of the theorem is not true: The fact that $\sigma_{XY} = 0$ does not necessarily imply that X and Y are independent:

Let X and Y have the joint pdf $f(x, y) = 1.5\mathbb{I}_{[-1,1]}(x)\mathbb{I}_{[0,x^2]}(y)$. The correlation is zero but the variables are not independent...

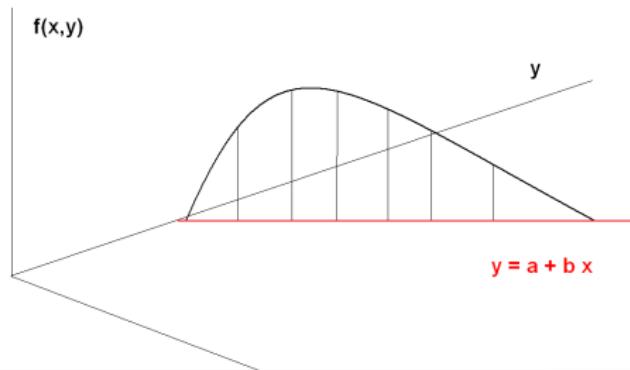
At the other end of the scale

Theorem (3.20)

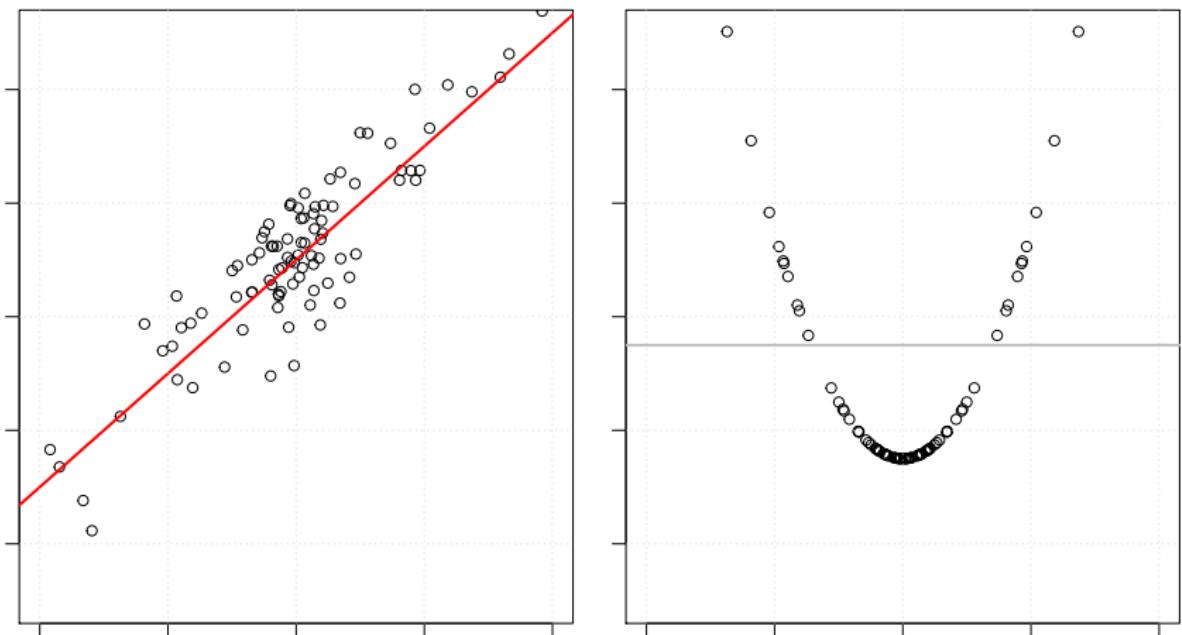
If $\rho_{XY} = 1$ or -1 , then $P(Y = a + bX) = 1$, where $b \neq 0$.

If $\rho_{XY} = 1$ or -1 such that $P(Y = a + bx) = 1$, then the joint pdf $f(x, y)$ is **degenerate**. All the probability mass of $f(x, y)$ is concentrated above the line $y = a + bx$.

This generates a perfect linear relationship between X and Y .



Don't overinterpret!



Left: correlation, imperfect relation; Right: no correlation, perfect dependence.

Mean and variance of linear combinations

Theorem (3.21)

Let $Y = \sum_{i=1}^n a_i X_i$, where a_i are constant. Then $E(Y) = \sum_{i=1}^n a_i E(X_i)$.

The matrix representation of this result obtains as follows. Let

$$\mathbf{a} = (a_1, \dots, a_n)' \quad \text{and} \quad \mathbf{X} = (X_1, \dots, X_n)'.$$

Then $Y = \mathbf{a}' \mathbf{X}$ such that $E(Y) = \mathbf{a}' E(\mathbf{X})$.

Theorem (3.22)

Let $Y = \sum_{i=1}^n a_i X_i$, where the a_i s are constants. Then

$$\sigma_Y^2 = \sum_{i=1}^n a_i^2 \sigma_{X_i}^2 + 2 \sum_{i < j} a_i a_j \sigma_{X_i X_j}.$$

In order to rewrite this result in matrix notation we shall define the **covariance matrix** of a multivariate random variable.

Covariance matrix

Definition

The covariance matrix of the n -dimensional random vector $\mathbf{X} = (X_1, \dots, X_n)'$ is the $n \times n$ symmetric matrix

$$\text{Cov}(\mathbf{X}) = E((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))') = \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma_{X_n}^2 \end{pmatrix}$$

- The variance of the i th variable in \mathbf{X} is given by the (i, i) th diagonal entry in the covariance matrix.
- A covariance matrix is symmetric, that is $\text{Cov}(\mathbf{X}) = \text{Cov}(\mathbf{X})'$.

The matrix expressions

Let $\mathbf{a} = (a_1, \dots, a_n)'$ and $\mathbf{X} = (X_1, \dots, X_n)'$. Then the variance of $Y = \mathbf{a}' \mathbf{X}$ given in the theorem can obviously be represented as

$$\sigma_Y^2 = \mathbf{a}' \text{Cov}(\mathbf{X}) \mathbf{a}.$$

Note that since a variance is non-negative ($\sigma_Y^2 \geq 0$) the expression $\mathbf{a}' \text{Cov}(\mathbf{X}) \mathbf{a}$ is also non-negative for any a .

This implies that a covariance matrix is necessarily positive semidefinite!

More matrices

Theorem (3.23)

Let $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A} = (a_{hm})$ is a $k \times n$ matrix of real constants, and $\mathbf{X} = (X_i)$ is an $n \times 1$ vector of random variables. Then $E(\mathbf{Y}) = \mathbf{A}E(\mathbf{X})$.

Theorem (3.24)

Let $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A} = (a_{hm})$ is a $k \times n$ matrix of real constants, and $\mathbf{X} = (X_i)$ is a $n \times 1$ vector of random variables. Then
 $\text{Cov}(\mathbf{Y}) = \mathbf{A} \text{Cov}(\mathbf{X}) \mathbf{A}'$.

Outline

- 1 Multivariate expectations and moments
- 2 Covariance and correlation
- 3 Conditional expectations
- 4 Up next

The same thing?

- So far, we have considered unconditional expectations, this means the expectations of unconditional/marginal distributions.
- If we take the expectation w.r.t. a **conditional** distribution, we have the **conditional** expectation.
- The conditional expectation is one of the most important concepts used in econometrics and empirical economics.
- It is for instance the key element of regression analysis, telling us how a variable reacts on the average to changes in other variables.¹

¹This is one interpretation, don't expect uniqueness thereof.

A conditional distribution is just a distribution

Definition (Conditional expectation)

Let (X_1, \dots, X_n) and (Y_1, \dots, Y_m) be random vectors with joint pdf $f(x_1, \dots, x_n, y_1, \dots, y_m)$. The conditional expectation of $g(Y_1, \dots, Y_m)$, given $(X_1, \dots, X_n) \in B$, is defined as

$$\text{(discrete)} \quad E(g(Y_1, \dots, Y_m) | (X_1, \dots, X_n) \in B)$$

$$= \sum_{(y_1, \dots, y_m) \in R(Y)} \cdots \sum g(y_1, \dots, y_m) f(y_1, \dots, y_m | (x_1, \dots, x_n) \in B)$$

$$\text{(continuous)} \quad E(g(Y_1, \dots, Y_m) | (X_1, \dots, X_n) \in B)$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, \dots, y_m) f(y_1, \dots, y_m | (x_1, \dots, x_n) \in B) dy_1 \cdots dy_m.$$

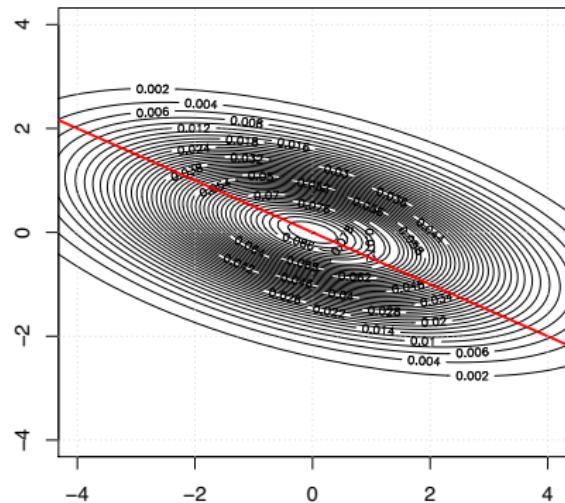
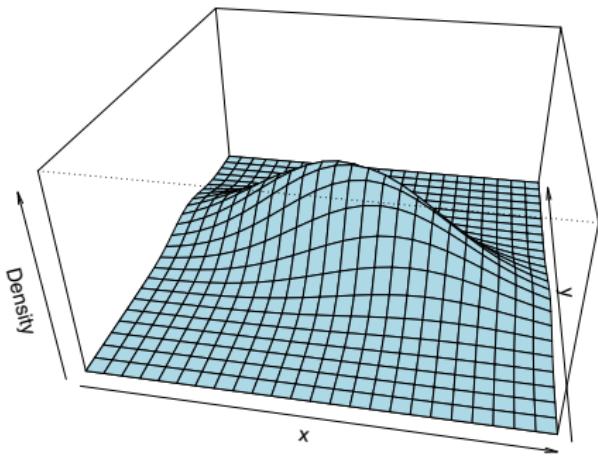
The regression function

An important special case of the definition given above obtains by setting $g(Y_1, \dots, Y_n) = Y$, where Y is a univariate random variable, and B is an elementary event.

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = \begin{cases} \sum_{y \in \mathrm{R}(Y)} y \cdot f(y \mid \mathbf{X} = \mathbf{x}) & \text{(discrete)} \\ \int_{-\infty}^{\infty} y \cdot f(y \mid \mathbf{X} = \mathbf{x}) \, dy & \text{(continuous).} \end{cases}$$

This is a function of x ; we call it **the regression curve** of Y on \mathbf{X} .

An example



Left: bivariate pdf, correlation; Right: level curves and regression line

A nonlinear example

Take the bivariate random variable with joint pdf

$$f(x, y) = \frac{1}{96}(x^2 + 2xy + 2y^2)\mathbb{I}_{[0,4]}(x)\mathbb{I}_{[0,2]}(y).$$

The regression function of Y on X is obtained as

$$\begin{aligned} E(Y|X = x) &= \int_{-\infty}^{\infty} y \cdot \frac{f(x, y)}{f_X(x)} dy = \int_0^2 \frac{y \cdot (x^2 + 2xy + 2y^2)\mathbb{I}_{[0,4]}(x)}{(2x^2 + 4x + \frac{16}{3})\mathbb{I}_{[0,4]}(x)} dy \\ &= \frac{2x^2 + \frac{16}{3}x + 8}{2x^2 + 4x + \frac{16}{3}} \quad \text{for } x \in [0, 4]. \end{aligned}$$

For $x \notin [0, 4]$, the regression function is not defined.

Getting more random

- The conditional expectation $E(Y|(X_1, \dots, X_n) \in B)$ was introduced as being conditional on a *particular event* B ,
e.g. $B = ((X_1, \dots, X_n) = (x_1, \dots, x_n))$.
- Rather than specifying a particular event, we might conceptualize leaving the event for (X_1, \dots, X_n) *unspecified* and interpret the conditional expectation of Y as a function of (X_1, \dots, X_n) denoted by $E(Y|X_1, \dots, X_n)$.
- Note that $E(Y|X_1, \dots, X_n)$ is then a function of random variables and, therefore, itself a random variable.
- $E(Y|(X_1, \dots, X_n) = (x_1, \dots, x_n)) = E(Y|x_1, \dots, x_n)$ is referred to as the **regression function of a regression of Y on the X_i s**.

Recovering the unconditional

One might ask whether there's any relation between unconditional and conditional expectations. And...

Theorem (3.10)

$$\mathrm{E}(\mathrm{E}(g(Y)|\boldsymbol{X})) = \mathrm{E}(g(Y)).$$

For random vectors, we get

$$\mathrm{E}(\mathrm{E}(g(Y_1, \dots, Y_n)|X_1, \dots, X_n)) = \mathrm{E}(g(Y_1, \dots, Y_n)).$$

Final remark: All properties of expectations discussed above also apply analogously to conditional expectations.

Outline

- 1 Multivariate expectations and moments
- 2 Covariance and correlation
- 3 Conditional expectations
- 4 Up next

Coming up

Parametric families of distributions

Parametric families of (univariate) distributions

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Overview

There are many types of data for which we would like to have a suitable distributional model:

- **Discrete**: categorical, ordinal, counts
- **Continuous**: durations, generic errors

In practice, one usually works with a suitable **parametric family** of densities, and we do the same here.¹

¹In Advanced Statistics III, we will discuss **nonparametric** approaches.

Parametric models

- We will use the generic notation $f(x; \theta)$ (and $F(x; \theta)$ for the cdf):
 - This denotes a family of densities for random variable X .
 - A given value for θ pins down a specific member of the family.
- The admissible values θ of the parameters are called the **parameter space** and will be denoted by Ω . (Vector θ allowed for as well.)
- Each family is more suitable for certain tasks and comes with specific parameter interpretations/symbols.² So use with care.

²In Advanced Statistics II, we shall discuss **estimation** of θ given sample data.

Parametric families of (univariate) distributions

- 1 Models for categorical data
- 2 Models for counts
- 3 Models for durations
- 4 Up next

Outline

1 Models for categorical data

2 Models for counts

3 Models for durations

4 Up next

The discrete uniform distribution

Family Name: Discrete Uniform

Parameterization $N \in \Omega = \{N : N \text{ is a positive integer}\}$

Density Definition $f(x; N) = \frac{1}{N} \mathbb{I}_{\{1,2,\dots,N\}}(x)$

Moments $\mu = (N + 1)/2, \sigma^2 = (N^2 - 1)/12, \mu_3 = 0$

MGF $M_X(t) = \sum_{j=1}^N e^{jt}/N$

This is (only) suitable if your outcomes are equally likely.

Example

Consider the experiment of rolling a die. The pdf of the number of dots facing up is $f(x; N = 6) = \frac{1}{6} \mathbb{I}_{\{1,2,\dots,6\}}(x)$, and belongs to the family of discrete uniforms.

The Bernoulli distribution

Family Name: Bernoulli

Parameterization $p \in \Omega = \{p : 0 \leq p \leq 1\}$

Density Definition $f(x; p) = p^x(1 - p)^{1-x}\mathbb{I}_{\{0,1\}}(x)$

Moments $\mu = p, \sigma^2 = p(1 - p), \mu_3 = 2p^3 - 3p^2 + p$

MGF $M_X(t) = pe^t + (1 - p)$

This works perfectly if you only have two possible outcomes – not necessarily equally likely.

Usually 0 stands for one out of two categories and 1 for the other.

The binomial distribution

Family Name: Binomial

Parameterization $(n, p) \in \Omega = \{(n, p) : n \text{ is a positive integer}, 0 \leq p \leq 1\}$

Density Definition $f(x; n, p) = \begin{cases} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, & \text{for } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$

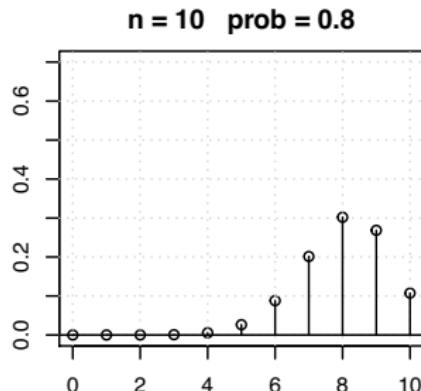
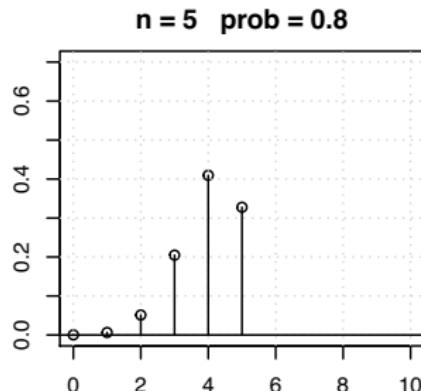
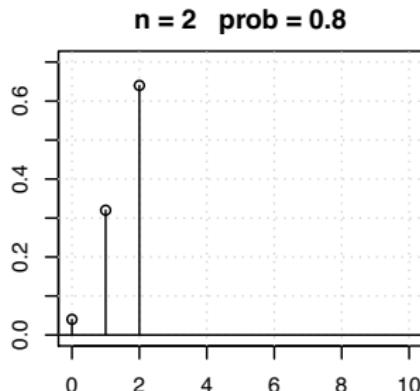
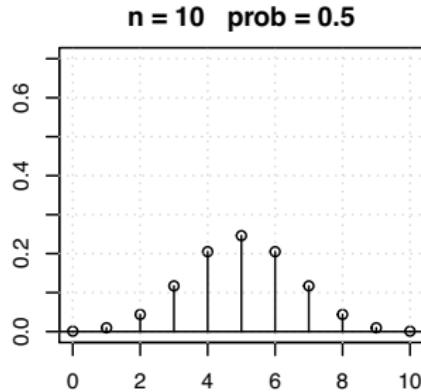
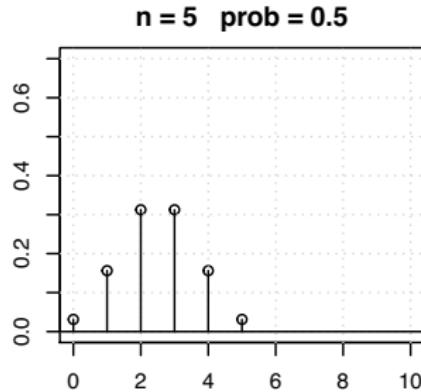
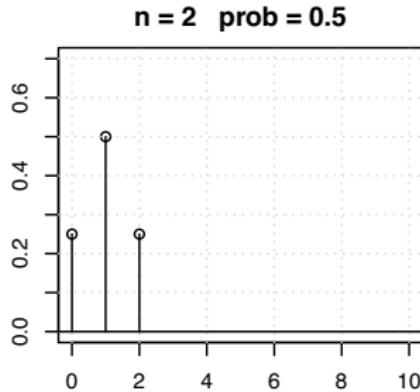
Moments $\mu = np, \sigma^2 = np(1-p), \mu_3 = np(1-p)(1-2p)$

MGF $M_X(t) = (1 - p + pe^t)^n$

The binomial density is used to model an experiment that consists of n independent repetitions of a Bernoulli-type experiment with a success probability p .

The quantity of interest x is the total number of successes in n of such Bernoulli trials. (Compare the MGFs)

Some specific binomial pmfs (or discrete pdfs)



Example

What is the probability of obtaining at least one '6' in four rolls of a fair die?

- This experiment can be modeled as a sequence of $n = 4$ $\text{Ber}(p)$ trials with success probability $p = 1/6 = P(6 \text{ dots face up})$.
- Define the random variable $X = \text{total number of 6s in four rolls}$.
- Then $X \sim \text{Binom}(n = 4, p = 1/6)$ and

$$\begin{aligned} P(\text{at least one '6'}) &= P(X > 0) = 1 - P(X = 0) \\ &= 1 - \binom{4}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^4 = .518. \end{aligned}$$

The multinomial distribution

Family Name: Multinomial

Parameterization $(n, p_1, \dots, p_m) \in \Omega = \{(n, p_1, \dots, p_m) : n \text{ is a positive integer, } 0 \leq p_i \leq 1, \forall i, \sum_{i=1}^m p_i = 1\}$

Density Definition $f(x_1, \dots, x_m; n, p_1, \dots, p_m)$

$$= \begin{cases} \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i} & \text{for } x_i = 0, 1, 2, \dots, n \forall i, \quad \sum_{i=1}^m x_i = n \\ 0 & \text{otherwise} \end{cases}$$

Moments $\mu_i = np_i, \sigma_i^2 = np_i(1-p_i), \mu_{3,i} = np_i(1-p_i)(1-2p_i),$

$$\text{Cov}(X_i, X_j) = -np_i p_j$$

MGF $M_X(t) = (\sum_{i=1}^m p_i e^{t_i})^n$

The quantities of interest x_1, \dots, x_m are the total numbers of each of the m different possible outcomes in n independent repetitions of the experiment.

Note that the range of the random vector (X_1, \dots, X_n) is given by

$$R(\mathbf{X}) = \{(x_1, \dots, x_n) : x_i \in \{0, 1, \dots, n\} \forall i, \quad \sum_{i=1}^m x_i = n\}.$$

Outline

1 Models for categorical data

2 Models for counts

3 Models for durations

4 Up next

The negative binomial distribution

Family Name: Negative Binomial (Pascal)

Parameterization $(r, p) \in \Omega \{ (r, p) : r \text{ is a positive integer}, 0 < p < 1 \}$

Density Definition $f(x; r, p) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & \text{for } x = r, r+1, r+2, \dots \\ 0 & \text{otherwise} \end{cases}$

Moments $\mu = \frac{r}{p}, \sigma^2 = \frac{r}{p^2}(1-p), \mu_3 = \frac{r}{p^3} \left((1-p) + (1-p)^2 \right)$

MGF $M_X(t) = e^{rt} p^r (1 - (1-p)e^t)^{-r} \text{ for } t < -\ln(1-p)$

This models (randomly many) independent $\text{Ber}(p)$ experiments/trials.

- The quantity of interest x is the **number of Bernoulli trials** which are necessary to obtain r successes.
- Compared to the binomial, the **number of trials** and the **number of successes** are reversed w.r.t. what is **random** and what is a **parameter**.

A special case: the geometric distribution

Set $r = 1$ such that

$$f(x; p) = p(1 - p)^{x-1} \quad \text{for } x = 1, 2, \dots$$

The quantity of interest x is the Bernoulli trial at which the first success occurs.³

The **geometric distribution** has a property known as the **memoryless property**. It means that for some positive integers i and j we obtain

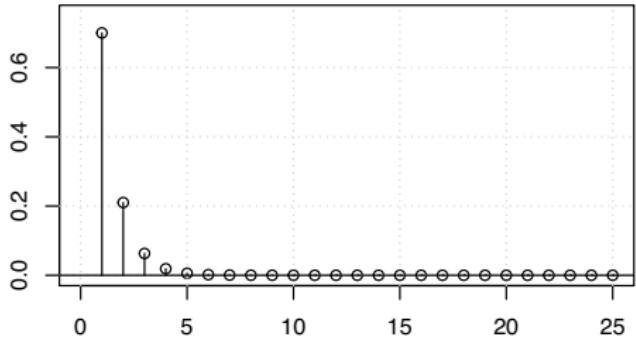
$$\mathrm{P}(X > i + j | X > i) = \mathrm{P}(X > j).$$

The memoryless property can be interpreted as a *lack-of-aging* property.

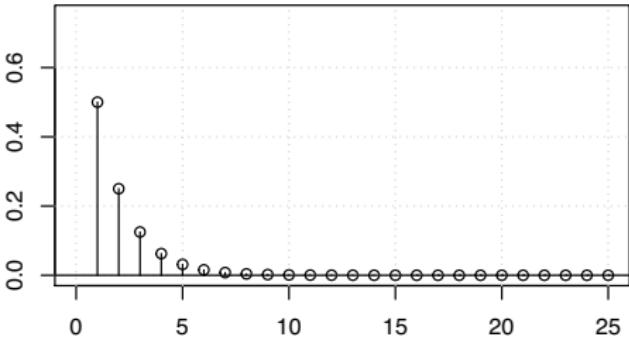
³Some (e.g. in R) take x to be the number of failures before the first success.

Some specific geometric pmfs

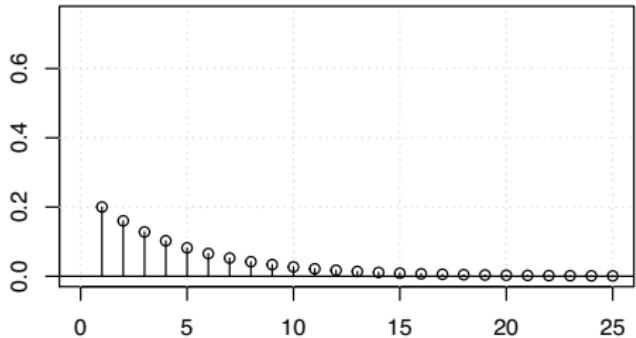
prob = 0.7



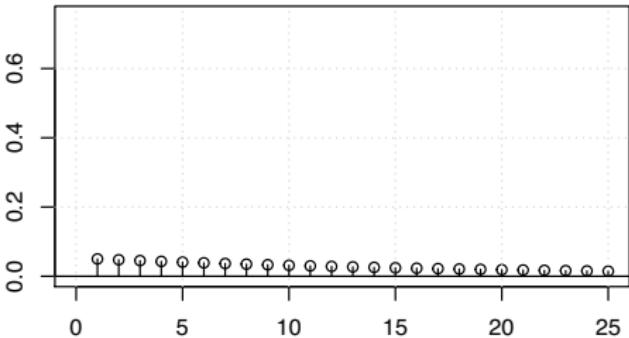
prob = 0.5



prob = 0.2



prob = 0.05



And the last one

Family Name: Poisson

Parameterization $\lambda \in \Omega = \{\lambda : \lambda > 0\}$

Density Definition $f(x; \lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!}, & \text{for } x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$

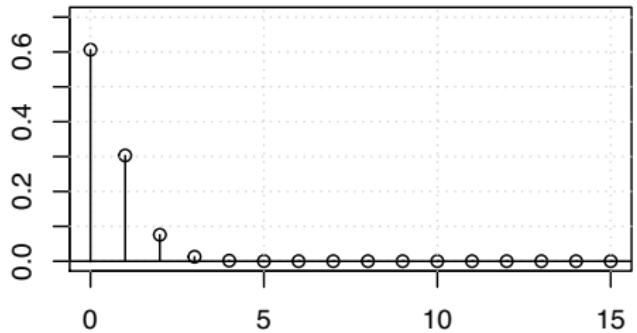
Moments $\mu = \lambda, \sigma^2 = \lambda, \mu_3 = \lambda$

MGF $M_X(t) = e^{\lambda(e^t - 1)}$

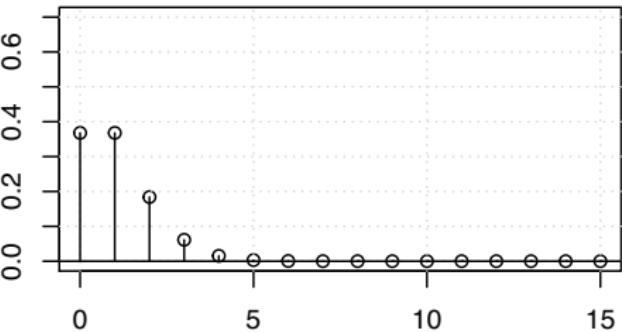
The Poisson distribution is also called the **law of rare events** (see below why).

Some Poisson pmfs

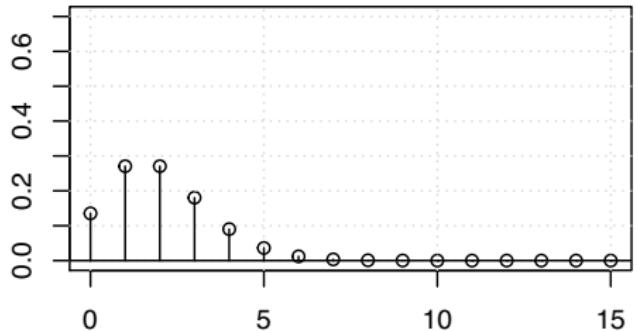
lambda = 0.5



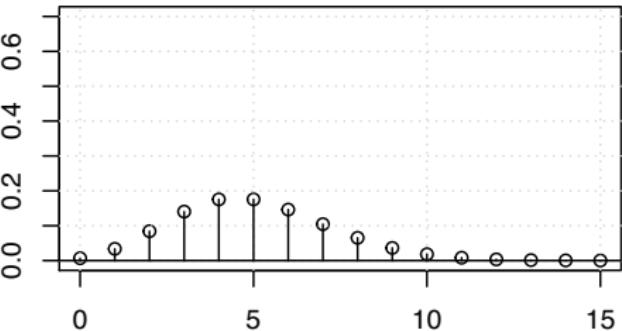
lambda = 1



lambda = 2



lambda = 5



The Poisson process

The Poisson distribution models experiments whose outcomes are governed by the so-called **Poisson process**:

Definition (Poisson process)

Let an experiment consist of observing the occurrence of a certain event over a time interval $[0, t]$. The experiment follows a Poisson process if:

- 1) the probability that the event occurs **once** over a small time interval Δt is approximately proportional to Δt as^a $\gamma \cdot (\Delta t) + o(\Delta t)$, where $\gamma > 0$,
- 2) the probability that the event occurs **twice or more often** over a small time interval Δt is negligible being of order of magnitude $o(\Delta t)$,
- 3) the numbers of occurrences of the event that are observed in non-overlapping intervals are independent events.

^a $o(\Delta t)$ stands for *of smaller order than* Δt and means that the values of $o(\Delta t)$ approach zero at a rate faster than Δt . That is $\lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0$.

... and the formal connection

Theorem (4.1)

Let X be the number of times a certain event occurs in the interval $[0, t]$. If the experiment underlying X follows a Poisson process, then $X \sim \text{Po}(\lambda)$.

- The parameter γ is interpreted as the mean rate of occurrence of the event per unit of time or the intensity of the Poisson process;
- This follows from the fact that for a Poisson variable $E(X) = \lambda = \gamma t$ such that $E(X/t) = \gamma$.

A shortcut to the binomial

The Poisson distribution provides an approximation to the probabilities generated by the binomial distribution.

- In fact, the limit of the binomial density as the number of Bernoulli trials $n \rightarrow \infty$ is the Poisson density if $np \rightarrow \lambda > 0$.
- For a large number of trials n and thus for a small success probability $p = \lambda/n$, we can use

$$\frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \approx \frac{(np)^x e^{-np}}{x!}.$$

- The Poisson density is relatively easy to evaluate, whereas, for large n , the calculation of the factorial expressions is not.

Outline

1 Models for categorical data

2 Models for counts

3 Models for durations

4 Up next

The Gamma distribution

Family Name: Gamma

Parameterization $(\alpha, \beta) \in \Omega = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$

Density Definition $f(x; \alpha, \beta) = \frac{1}{(\beta^\alpha \Gamma(\alpha))} x^{\alpha-1} e^{-x/\beta} \mathbb{I}_{(0, \infty)}(x),$
 where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$

Moments $\mu = \alpha\beta, \sigma^2 = \alpha\beta^2, \mu_3 = 2\alpha\beta^3$

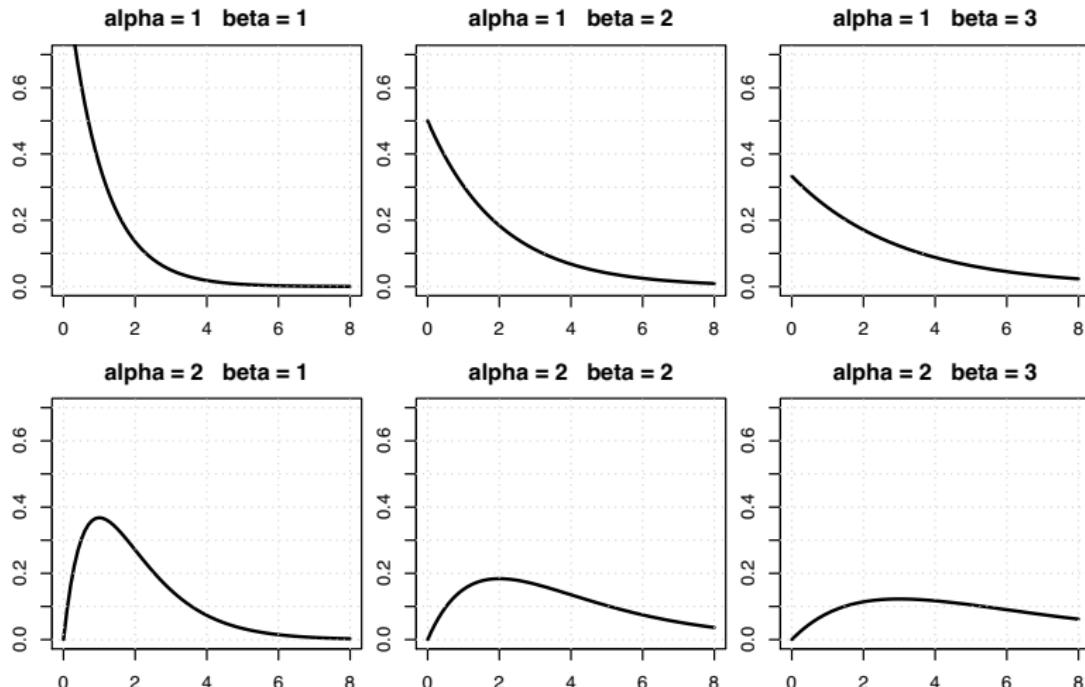
MGF $M_X(t) = (1 - \beta t)^{-\alpha} \text{ for } t < \beta^{-1}$

The gamma function has the following properties.

- For any real $\alpha > 0$, the gamma function satisfies the recursion $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. This can be verified through integration by parts.
- $\Gamma(1) = \int_0^\infty e^{-y} dy = 1$ and $\Gamma(1/2) = \pi^{1/2}$.
- If $\alpha > 0$ is an integer, then $\Gamma(\alpha) = (\alpha - 1)!$.

Useful

Since the range of the Gamma distribution is \mathbb{R}_+ , it's a natural choice for modelling durations. Must accept skewness to the right though.



Some properties

- The Gamma distribution models the **waiting time (duration)** between occurrences of events under a Poisson process.
- The gamma distribution has an **additivity property**, see below.
- A rescaled gamma distribution is also gamma, see below.

Theorem (4.2)

Let X_1, \dots, X_n be independent RVs with $X_i \sim \text{Gamma}(\alpha_i, \beta)$, $i = 1, \dots, n$.
Then $Y = \sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.

Theorem (4.3)

Let $X \sim \text{Gamma}(\alpha, \beta)$. Then, for any $c > 0$, $Y = cX \sim \text{Gamma}(\alpha, \beta c)$.

(Like before, compare MGFs)

The exponential special case

Gamma Subfamily Name: Exponential

Parameterization $\theta \in \Omega = \{\theta : \theta > 0\}$

Density Definition $f(x; \theta) = \frac{1}{\theta} e^{-x/\theta} \mathbb{I}_{(0, \infty)}(x)$

$$f(x; \lambda) = \lambda e^{-\lambda x} \mathbb{I}_{(0, \infty)}(x)$$

Moments $\mu = \theta, \sigma^2 = \theta^2, \mu_3 = 2\theta^3$

MGF $M_X(t) = (1 - \theta t)^{-1}$ for $t < \theta^{-1}$

A specific application of the exponential distribution is the modeling of the time that passes until a Poisson process produces the first success.

No memory

The exponential distribution has the **memoryless property** (too):

Theorem (4.4)

If $X \sim \text{Exp}(\theta)$, then $P(X > s + t | X > s) = P(X > t) \forall (t, s) > 0$.

This indicates that the exponential distribution is not appropriate to model lifetimes for which the failure probability is expected to increase with time.

The χ^2 special case

A further important special case of the gamma distribution, obtained by setting $\alpha = v/2$ and $\beta = 2$, is the **chi-square distribution**.

Gamma Subfamily Name: Chi-Square

Parameterization $v \in \Omega = \{v : v \text{ is a positive integer}\}$
 v is called the **degrees of freedom**

Density Definition $f(x; v) = \frac{1}{2^{v/2}\Gamma(v/2)}x^{(v/2)-1}e^{-x/2}\mathbb{I}_{(0,\infty)}(x)$

Moments $\mu = v, \sigma^2 = 2v, \mu_3 = 8v$

MGF $M_X(t) = (1 - 2t)^{-v/2}$ for $t < \frac{1}{2}$

The chi-square distribution plays an important role in **statistical inference**.
 In particular, (as we will show later) the **sum of the squares of v independent standard normal random variables** has a χ_v^2 -distribution.

Related, though not a duration

Family Name: Beta

Parameterization $(\alpha, \beta) \in \Omega = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$

Density Definition $f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbb{I}_{(0,1)}(x),$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$ is the beta function⁴.

Moments $\mu = \alpha / (\alpha + \beta), \quad \sigma^2 = \alpha\beta / [(\alpha + \beta + 1)(\alpha + \beta)^2],$

$\mu_3 = 2(\beta - \alpha)(\alpha\beta) / [(\alpha + \beta + 2)(\alpha + \beta + 1)(\alpha + \beta)^3]$

MGF $M_X(t) = \sum_{r=1}^{\infty} (B(r + \alpha, \beta) / B(\alpha, \beta)) (t^r / r!)$

The beta density can be used to model experiments whose outcomes are coded as **real numbers on the interval $[0, 1]$** . It has obvious applications in modeling random variables representing **proportions**.

⁴Some useful properties of the beta function include the fact that $B(\alpha, \beta) = B(\beta, \alpha)$ and $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$.

A particular case: the uniform distribution

Family Name: Continuous Uniform

Parameterization $(a, b) \in \Omega = \{(a, b) : -\infty < a < b < \infty\}$

Density Definition $f(x; a, b) = \frac{1}{b-a} \mathbb{I}_{[a,b]}(x)$

Moments $\mu = (a + b) / 2, \sigma^2 = (b - a)^2 / 12, \mu_3 = 0$

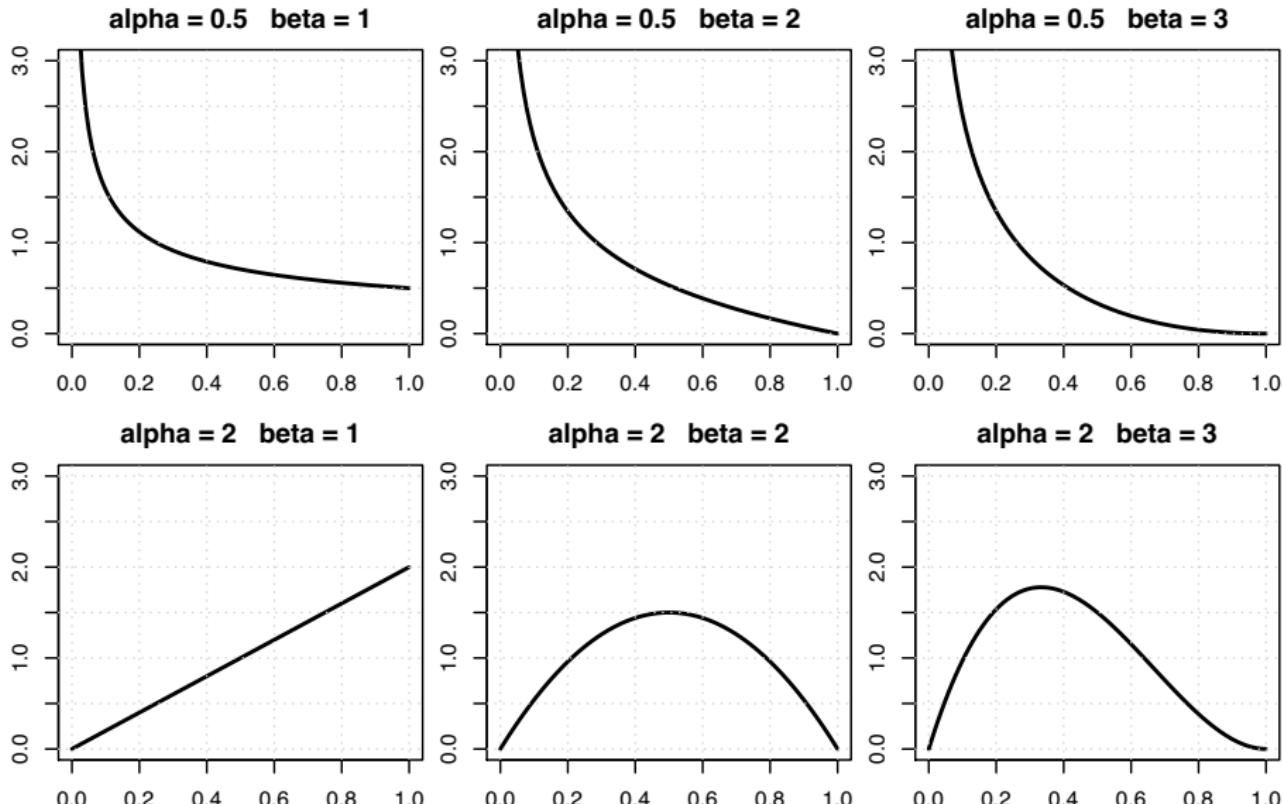
MGF
$$M_X(t) = \begin{cases} \frac{e^{bt} - e^{at}}{(b-a)t} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$$

Fun fact: let $X \sim F$ be a continuous RV; then, $F(X) \sim \text{Unif}(0, 1)$.
 Anyway, the Beta(1, 1) distribution is the same as Unif(0, 1).

Example

Spin a wheel of fortune with radius r . The point X at which the wheel stops is uniformly distributed with $a = 0$ and $b = 2\pi r$.

A lot of flexibility



Outline

- 1 Models for categorical data
- 2 Models for counts
- 3 Models for durations
- 4 Up next

Coming up

The normal family of distributions

The normal family

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

The normal family

- 1 The univariate normal
- 2 Some generalizations
- 3 The multivariate normal
- 4 Up next

Outline

- 1 The univariate normal
- 2 Some generalizations
- 3 The multivariate normal
- 4 Up next

One of Gauss' many ideas

The **normal (Gaussian) family** of distributions is the most extensively used distribution in statistics and econometrics. There are three main reasons for this.

- 1) The normal distribution is very **tractable analytically**.
- 2) The normal density has a **bell shape**, whose symmetry makes it an appealing candidate to model the probability space of many experiments.
- 3) There is the **Central Limit Theorem** (which we will discuss in Chapter 5), which indicates that under mild conditions, the normal distribution can be used to approximate a large variety of distributions in large samples.

The normal distribution

Family Name: Univariate Normal

Parameterization $(\mu, \sigma) \in \Omega = \{(\mu, \sigma) : \mu \in (-\infty, \infty), \sigma > 0\}$

Density Definition $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$

Moments $E(X) = \mu, \quad \text{Var}(X) = \sigma^2, \quad \mu_3 = 0$

MGF $M_X(t) = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}$

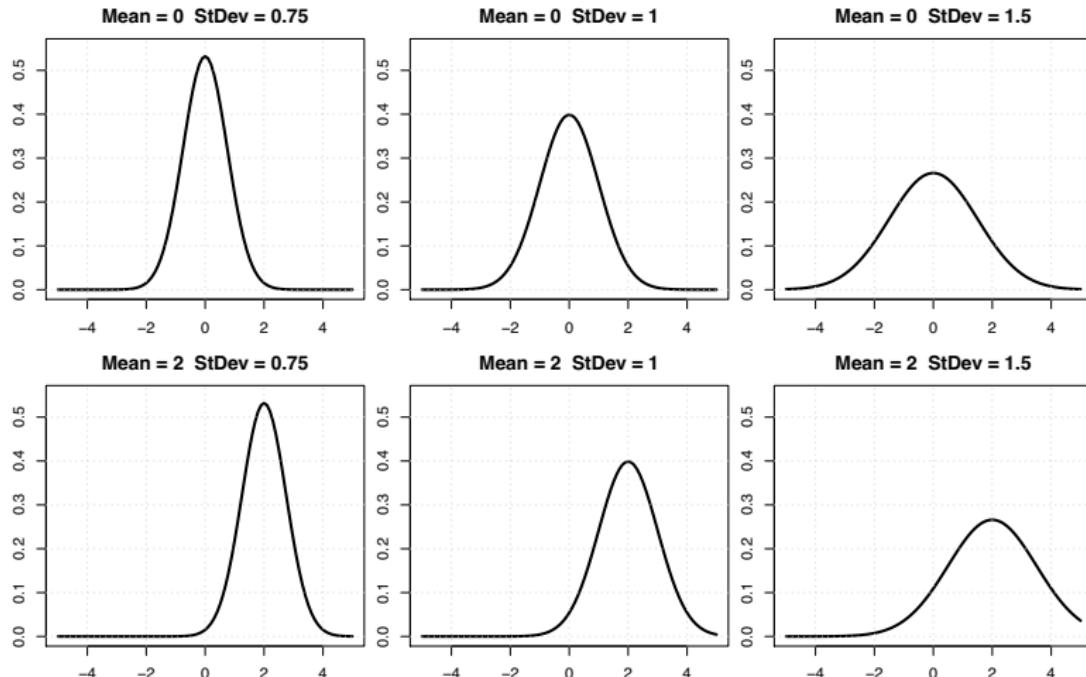
The normal family of densities is indexed by the two parameters μ and σ which correspond to the mean and the standard deviation, respectively.

In order to denote a normally distributed random variable with mean μ and variance σ^2 , we will use the usual notation $X \sim \mathcal{N}(\mu, \sigma^2)$.

A normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called **standard normal distribution**, and is abbreviated by $\mathcal{N}(0, 1)$. It has density φ and cdf Φ .

The bell shape

The normal density is symmetric about its mean μ , has its maximum at $x = \mu$ and inflection points at $x = \mu \pm \sigma$:



Some properties

Theorem (4.5)

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.

Hence, the standard normal distribution is sufficient to assign probabilities to **all** events involving Gaussian random variables.

Let $X \sim \mathcal{N}(17, 1/4)$. The probability of the event $X \in [16, 18]$ can be computed as

$$\begin{aligned} P(16 \leq X \leq 18) &= P\left(\frac{16 - 17}{(1/2)} \leq \frac{X - 17}{(1/2)} \leq \frac{18 - 17}{(1/2)}\right) \\ &= P(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) = 0.9544, \end{aligned}$$

where $\Phi(\cdot)$ denotes the cdf of a standard normal distribution.

Relation to gamma

Normal and chi-square distribution: There is relationship between standard normal random variables and the χ^2 distribution which is subject of the following two theorems:

Theorem (4.6)

If $X \sim \mathcal{N}(0, 1)$, then $Y = X^2 \sim \chi_1^2$.

Theorem (4.7)

Let (X_1, \dots, X_n) independent $\mathcal{N}(0, 1)$ -distributed random variables. Then $Y = \sum_{i=1}^n X_i^2 \sim \chi_n^2$.

(The MGF works miracles...)

Outline

- 1 The univariate normal
- 2 Some generalizations
- 3 The multivariate normal
- 4 Up next

The generalized normal distribution

Family Name: Generalized normal

Parameterization $\mu \in \mathbb{R}, \alpha, \beta \in (0, \infty)$

Density Definition $f(x; \mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-|\frac{x-\mu}{\alpha}|^\beta}$

CDF $F(x) = \frac{1}{2} + \text{sgn}(x - \mu) \frac{\gamma(\frac{1}{\beta}, |\frac{x-\mu}{\alpha}|^\beta)}{2\Gamma(\frac{1}{\beta})}$
 where $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} ds$

Moments $E(X) = \mu, \text{Var}(X) = \frac{\alpha^2 \Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})}$

Other names: generalized error distribution, exponential power distribution, generalized Gaussian distribution

- Setting $\beta = 1$ leads to the Laplace (double exponential) distribution
- Setting $\beta = 2$ leads to the normal (note the missing 1/2 in the exp.)

Skewed distributions

The already discussed skewed distributions are sometimes not flexible enough...

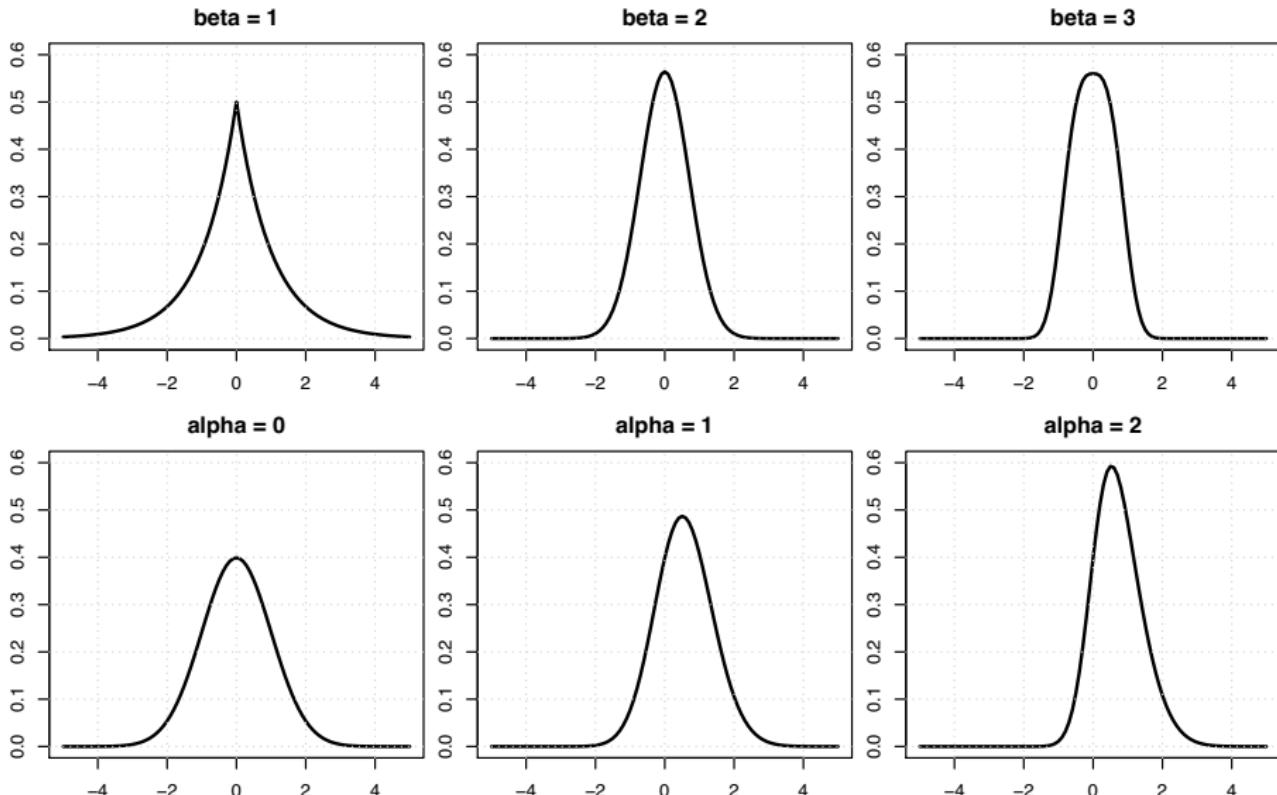
Family Name: Skew normal distribution

Parameterization $\alpha \in \mathbb{R}$,

Density Definition $f(x) = \frac{2}{\omega} \phi\left(\frac{x-\xi}{\omega}\right) \Phi\left(\alpha \frac{x-\xi}{\omega}\right)$

- For $\alpha = 0$, symmetry is recovered.
- For $\alpha \rightarrow \pm\infty$, $f(x)$ converges to the positive (negative) half-normal distribution given by $f(x) = \frac{2}{\omega} \phi\left(\frac{x-\xi}{\omega}\right) \mathbb{I}_{(\xi, \infty)}(x)$
- This can be generalized to $f(x) = \frac{2}{\omega} h\left(\frac{x-\xi}{\omega}\right) G\left(\alpha \frac{x-\xi}{\omega}\right)$ with h, g continuous densities, symmetric about 0 (and G the associated cdf).

Generalized (top) and skew normal (bottom) pdfs



Location-scale families

What if the shape is of secondary interest?

Family Name: Location-scale (univariate)

Parameterization $\mu \in \mathbb{R}, \sigma \in (0, \infty), g$ a pdf

Density Definition $f(x; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)$

CDF $F(x) = G\left(\frac{x-\mu}{\sigma}\right)$, G the corresponding cdf

Moments μ, σ^2 (if g is standardized with finite variance)

MGF $M_X(t) = e^{\mu t} M_Z(\sigma t)$

Note that the family can actually be defined for base densities that do not have finite variance (or even expectation).

If X has a location-scale distribution (with a given base g), then so does $Y = a + bX$ for any $a, b \neq 0$.

And finally: Gaussian mixtures

Another approach considers building densities from adding simpler basic elements:

Family Name: Gaussian mixture distributions (countable)

Parameterization $w_i \geq 0, \sum_{i \geq 1} w_i = 1, \mu_i, \sigma_i^2$

Density definition $f(x) = \sum_i w_i \frac{1}{\sigma_i} \phi\left(\frac{x-\mu_i}{\sigma_i}\right),$

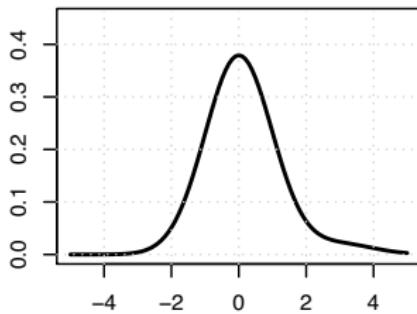
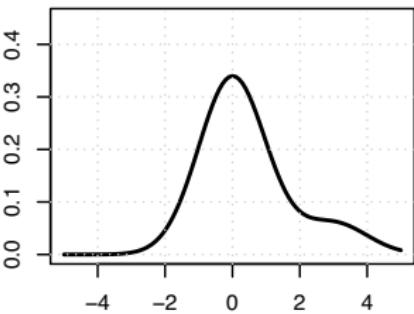
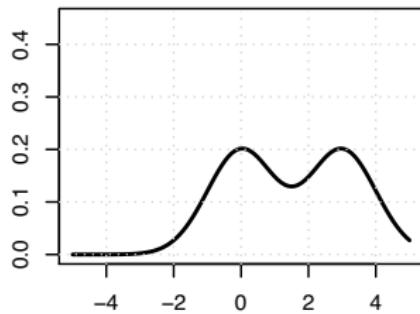
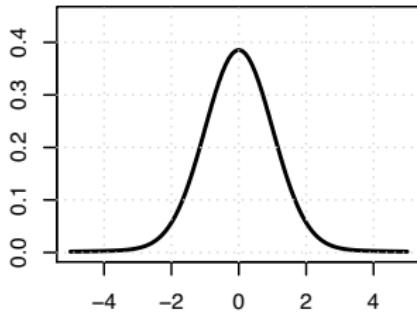
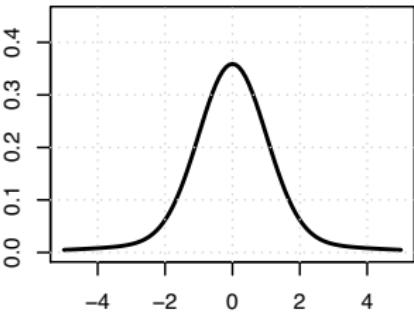
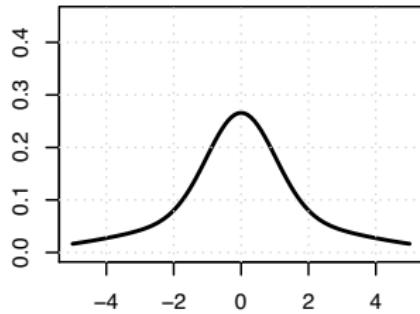
Moments $\mu = \sum_i w_i \mu_i,$

$$\sigma^2 = \sum_i w_i \left((\mu_i - \bar{\mu})^2 + \sigma_i^2 \right)$$

Clearly, this can be extended to other base distributions (which may also be multivariate).

One may consider uncountable versions thereof with
 $f(x) = \int_{\theta} f(x; \theta) w(\theta) d\theta, w$ some pdf.

Some mixtures

w = 0.05**w = 0.15****w = 0.5****w = 0.05****w = 0.15****w = 0.5**

Top: $w \cdot \mathcal{N}(0, 1) + (1 - w)\mathcal{N}(3, 1)$; Bottom: $w \cdot \mathcal{N}(0, 1) + (1 - w)\mathcal{N}(0, 3)$

Outline

- 1 The univariate normal
- 2 Some generalizations
- 3 The multivariate normal
- 4 Up next

Adding some variates

The univariate normal distribution discussed so far has a straightforward multivariate generalization.

Family Name: Multivariate Normal

Parameterization $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{pmatrix}$

$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Omega = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbf{R}^n,$$

$\boldsymbol{\Sigma}$ is a $(n \times n)$ p.d. symmetric matrix}

Density Definition $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$

Moments $E(\mathbf{X}) = \boldsymbol{\mu}, \quad \text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}, \quad \boldsymbol{\mu}_{3(n \times 1)} = [\mathbf{0}]$

MGF $M_{\mathbf{X}}(\mathbf{t}) = \exp\{\boldsymbol{\mu}' \mathbf{t} + (1/2) \mathbf{t}' \boldsymbol{\Sigma} \mathbf{t}\}, \text{ where } \mathbf{t} = (t_1, \dots, t_n)'.$

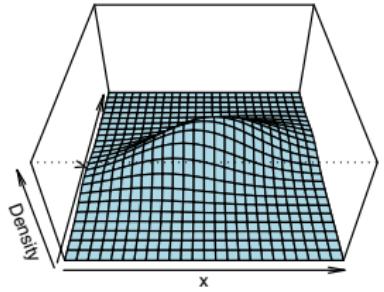
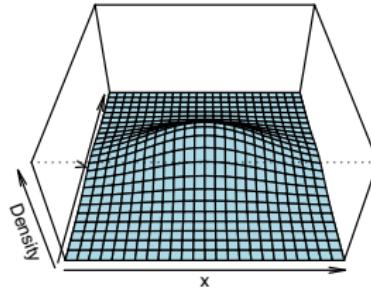
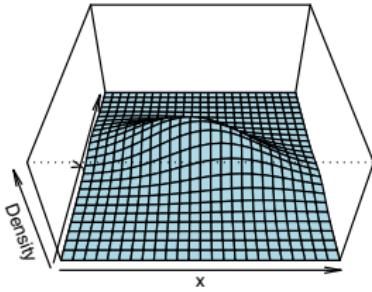
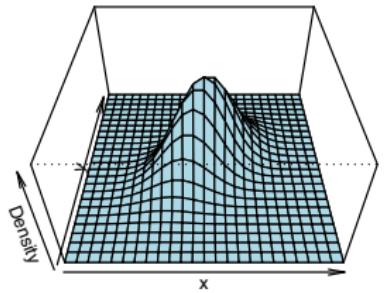
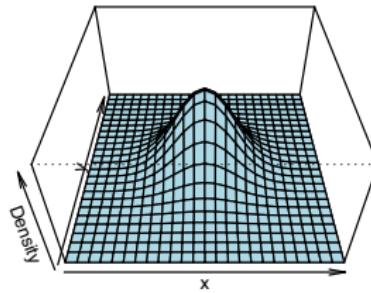
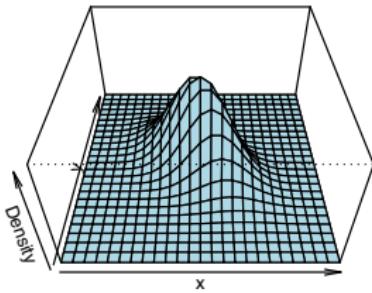
More details

The n -variate normal family of distribution is indexed by $n + n(n + 1)/2$ parameters: In the mean vector (μ) n parameters and in the covariance matrix (Σ) $n + (n^2 - n)/2$ parameters.

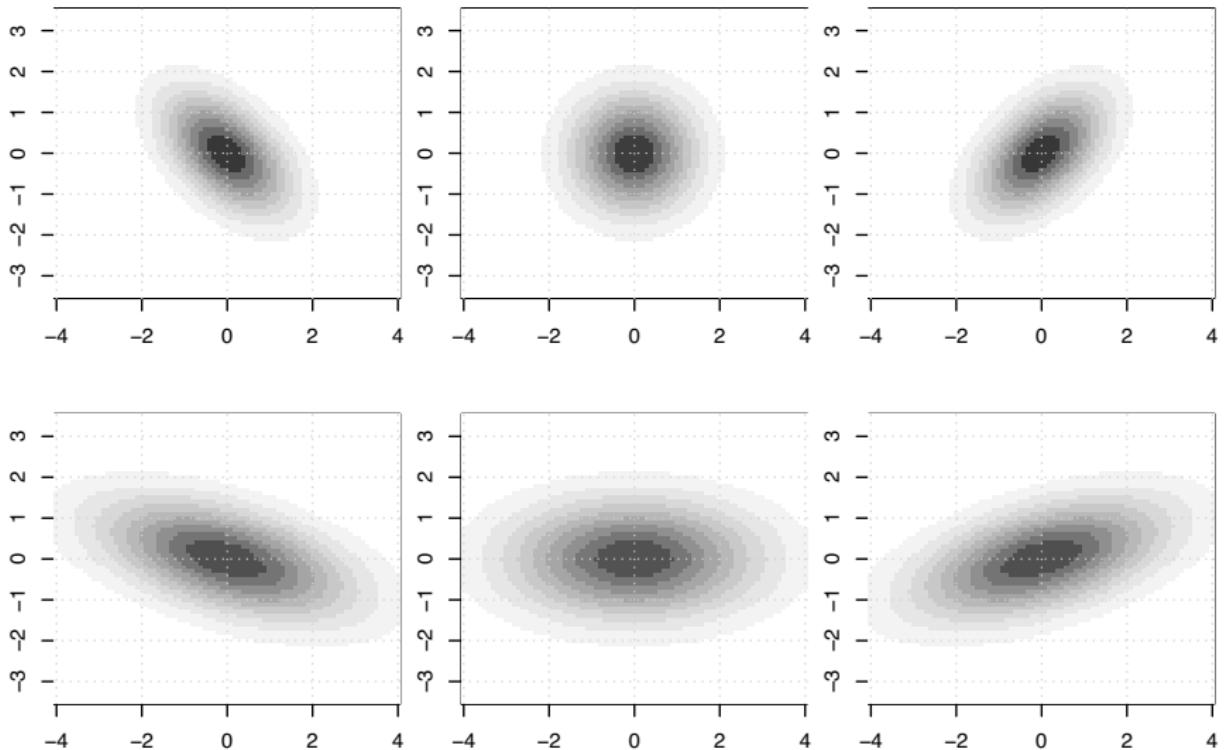
In order to illustrate graphically some of the characteristics of a multivariate Gaussian density, we consider the bivariate case with $n = 2$.

- The multivariate Gaussian density is bell-shaped and has its maximum at $\mathbf{x} = (x_1, x_2) = \boldsymbol{\mu} = (\mu_1, \mu_2)$.
- The iso-density contours, given by the set of points $(x_1, x_2) \in \{(x_1, x_2) : f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c\}$, have the form of an ellipse. Its origin is given by $\boldsymbol{\mu}$ and its direction depends on $\boldsymbol{\Sigma}$.

Various bivariate normal pdfs



... and the contour plots



Properties of Multivariate Normal Distributions

A useful property is that **linear combinations** of a vector of multivariate normally distributed random variables are also normally distributed as stated in the following theorem.

Theorem (4.8)

Let \mathbf{X} be an n -dimensional $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ -distributed random variable. Let \mathbf{A} be any $(k \times n)$ matrix of constants with $\text{rank}(\mathbf{A}) = k$, and let \mathbf{b} be any $(k \times 1)$ vector of constants. Then the $(k \times 1)$ random vector $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ is $\mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'')$ distributed.

This theorem can be used to **standardize** a normally distributed random vector.

Standardizing in the multivariate case

- Let Z be a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distributed $(n \times 1)$ random vector, that is a vector of n uncorrelated $\mathcal{N}(0, 1)$ distributed random variables.
- Then the $(n \times 1)$ random vector Y with a $\mathcal{N}(\mu, \Sigma)$ distribution can be represented in terms of Z as

$$Y = \mu + AZ, \quad \text{where } A \text{ is selected such that}^1 \quad AA' = \Sigma.$$

This is because $Y \sim \mathcal{N}(A\mathbf{0} + \mu, A\mathbf{I}A') = \mathcal{N}(\mu, \Sigma)$.

- Furthermore, the inversion of the function $Y = \mu + AZ$ **standardizes the normally distributed vector Y**

$$A^{-1}(Y - \mu) = Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

¹If A is a lower triangular matrix, we call it Cholesky factor, and $AA' = \Sigma$ denotes the so-called Cholesky decomposition.

Margins are also normal

Theorem (4.9)

Let Z be an n -dimensional $\mathcal{N}(\mu, \Sigma)$ -distributed random variable, where

$$Z = \begin{bmatrix} Z_{(1)} \\ \vdots \\ Z_{(n-m)} \end{bmatrix}_{(m \times 1)}, \quad \mu = \begin{bmatrix} \mu_{(1)} \\ \vdots \\ \mu_{(n-m)} \end{bmatrix}_{(n-m) \times 1}, \quad \Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right]_{(n-m) \times m | (n-m) \times (n-m)}.$$

Then the marginal pdf of $Z_{(1)}$ is $\mathcal{N}(\mu_1, \Sigma_{11})$, and the marginal PDF of $Z_{(2)}$ is $N(\mu_2, \Sigma_{22})$.

Note that Theorem 4.9 can be applied to obtain the marginal pdf of **any subset** of the normal random variable (Z_1, \dots, Z_n) by simply ordering them appropriately in the definition of Z in the theorem.

... as are the conditional distributions

Theorem (4.10)

Let Z be an n -dimensional $\mathcal{N}(\mu, \Sigma)$ -distributed random variable, where

$$Z = \begin{bmatrix} Z_{(1)} \\ Z_{(2)} \end{bmatrix}_{\begin{smallmatrix} (m \times 1) \\ (n-m) \times 1 \end{smallmatrix}}, \quad \mu = \begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}_{\begin{smallmatrix} (m \times 1) \\ (n-m) \times 1 \end{smallmatrix}}, \quad \Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline (\text{---})^{(m \times m)} & m \times (n-m) \\ \Sigma_{21} & \Sigma_{22} \\ (\text{---})^{(n-m) \times m} & (n-m) \times (n-m) \end{array} \right];$$

and let z^0 be an n -dimensional vector of constants partitioned conformably with the partition Z into $z_{(1)}^0$ and $z_{(2)}^0$. Then,

$$Z_{(1)} | (Z_{(2)} = z_{(2)}^0) \sim \mathcal{N} \left(\mu_{(1)} + \Sigma_{12} \Sigma_{22}^{-1} [z_{(2)}^0 - \mu_{(2)}], \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right)$$

$$Z_{(2)} | (Z_{(1)} = z_{(1)}^0) \sim \mathcal{N} \left(\mu_{(2)} + \Sigma_{21} \Sigma_{11}^{-1} [z_{(1)}^0 - \mu_{(1)}], \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right).$$

The conditional expectation

Note that the mean of the conditional distribution given by

$$E(\mathbf{Z}_{(1)} | \mathbf{Z}_{(2)} = \mathbf{z}_{(2)}) = \boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{z}_{(2)} - \boldsymbol{\mu}_{(2)})$$

is a linear function in the 'conditioning variable' $\mathbf{z}_{(2)}$.

This linearity of the conditional mean is a specific feature of the multivariate normal distribution as a member of the family of *elliptically contoured distributions*.

Consider the special case where $Z_{(1)}$ is a scalar and $\mathbf{Z}_{(2)}$ is a $(k \times 1)$ vector. Then the conditional mean of $Z_{(1)}$ given $\mathbf{z}_{(2)}$, that is, the regression function of $Z_{(1)}$ on $\mathbf{Z}_{(2)}$ has the form

$$E(Z_{(1)} | \mathbf{z}_{(2)}) = {}_{(1 \times 1)} a + {}_{(1 \times k)} b \mathbf{z}_{(2)},$$

where $a = \mu_{(1)} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\mu}_{(2)}$, $b = \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1}$.

Correlation and independence

The following theorem states that in the case of a normal distribution, zero covariance implies independence of the random variables, which in general is not true for other distributions.

Theorem (4.11)

Let $\mathbf{X} = (X_1, \dots, X_n)'$ be a $\mathcal{N}(\mu, \Sigma)$ -distributed random variable. Then (X_1, \dots, X_n) are independent iff Σ is a diagonal matrix with all covariances being zero.

Quadratic forms

Sometimes, one may be interested in the behavior of so-called quadratic forms in \mathbf{X} ,

$$Q = \mathbf{X}' \mathbf{A} \mathbf{X}$$

with \mathbf{A} some conformable matrix.

- Means and variances of Q may be derived for Gaussian \mathbf{X}
- Quite useful: if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbf{X}' \mathbf{\Sigma}^{-1} \mathbf{X} \sim \chi^2(\dim(\mathbf{X}), \boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu})$$

with $\chi^2(r, \lambda)$ a so-called non-central chi-squared distribution with r degrees of freedom and non-centrality parameter λ

- If $\lambda = 0$, the usual χ^2 with r degrees of freedom is recovered.

Outline

- 1 The univariate normal
- 2 Some generalizations
- 3 The multivariate normal
- 4 Up next

Coming up

More on modelling joint distributions

More on multivariate distributions

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

More on multivariate distributions

- 1 Relaxing the multivariate normal
- 2 Conditional distributions and functionals
- 3 The generalized linear model
- 4 Up next

Outline

- 1 Relaxing the multivariate normal
- 2 Conditional distributions and functionals
- 3 The generalized linear model
- 4 Up next

Location-scale families

Recall that the n -variate normal $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be decomposed as

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{H}\mathbf{Z}, \quad \text{with } \mathbf{H} \text{ a } n \times n \text{ matrix s.t. } \mathbf{H}\mathbf{H}' = \boldsymbol{\Sigma},$$

where \mathbf{Z} is a vector of n independent standard normals.

If \mathbf{Z} is not normal, but has zero mean and uncorrelated elements, we still have $E(\mathbf{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{X}) = \mathbf{H}\mathbf{H}' = \boldsymbol{\Sigma}$, leading to

Family Name: Location-scale (multivariate)

Parameterization $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ pos. def., g multivariate pdf

Density Definition $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} g(\boldsymbol{\Sigma}^{-0.5}(\mathbf{x} - \boldsymbol{\mu}))$

Moments $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ (g is standardized with finite variance)

It is often convenient to pick g such that it is the density of a vector of independent standardized random variables.

Elliptical distributions

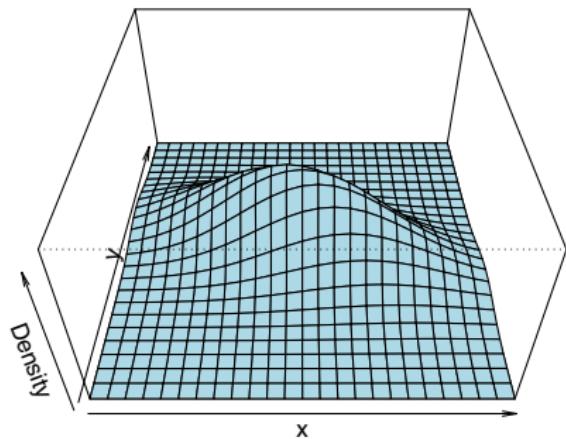
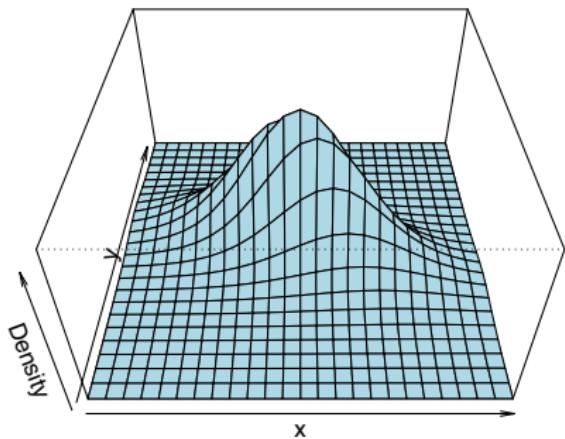
One interesting subclass of multivariate location-scale distributions of the class of the elliptical distributions, defined as

Family Name: Elliptical distributions

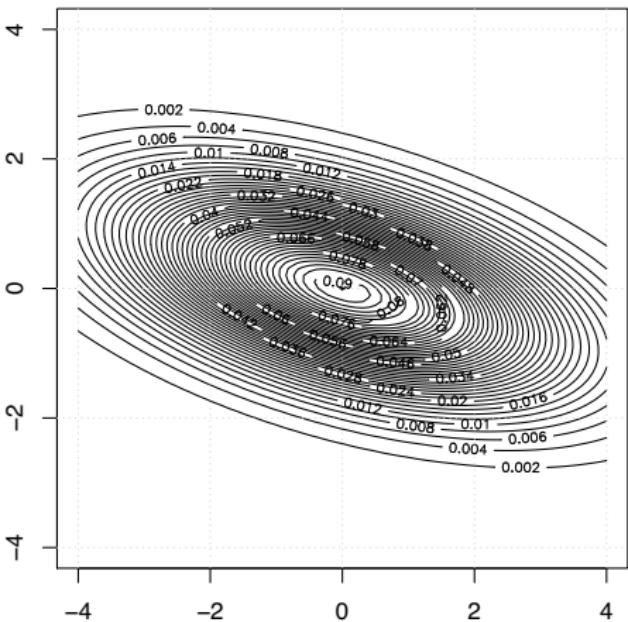
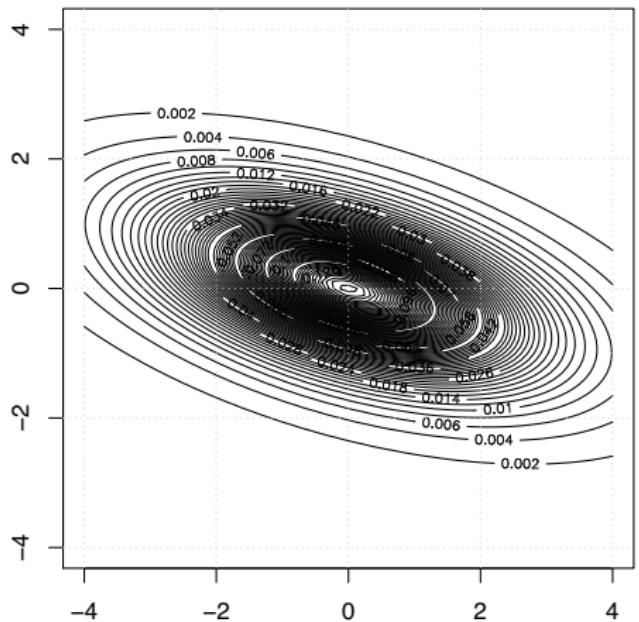
Parameterization $\mu \in \mathbb{R}^n$, Σ pos. def., $g : g(x^2)$ integrable
Density kernel $f(x; \mu, \Sigma) \propto g((x - \mu)' \Sigma^{-1} (x - \mu))$

- The name comes from the fact that the level curves of the density function are ellipses, like for the multivariate normal,
- ... which is a particular case with $g(u) = e^{-u/2}$.
- The covariance matrix (if finite) is proportional to Σ (for this reason, the correlations are $\sigma_{i,j}/(\sigma_{i,i}\sigma_{j,j})$ with $\sigma_{i,j}$ the elements of Σ).

Elliptical bivariate $t(5)$ & normal (same covariance matrix)



... and the level curves



Factor models

Consider $\mathbf{X} = \boldsymbol{\mu} + \mathbf{HZ}$ beyond $\dim \mathbf{X} = \dim \mathbf{Z} = n$.

In particular, the case $\dim \mathbf{Z} = r < n$ may be interesting:

- Of course, \mathbf{H} (and thus Σ) is of rank $r < n$.
- To alleviate this, add some randomness, say

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{HZ} + \mathbf{E}$$

where \mathbf{E} is a vector of uncorrelated RVs, also uncorrelated with \mathbf{Z} .

- This implies

$$\text{Cov}(\mathbf{X}) = \mathbf{H} \text{Cov}(\mathbf{Z}) \mathbf{H}' + \text{Cov}(\mathbf{E})$$

where $\text{Cov}(\mathbf{E})$ is diagonal, making $\text{Cov}(\mathbf{X})$ full-rank again.

Such models are flexible (imagine e.g. a large set of variables depending on a small number of “**common factors**”) – and work for discrete RVs too.

As opposed to multivariate mixtures, here we combine random variables linearly and not densities.

Towards an encompassing class

Take the univariate normal density, $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ and rewrite it as

$$f(x) = \exp\left(\frac{\mu}{\sigma^2} \cdot x - \frac{1}{2\sigma^2} \cdot x^2 - \frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \frac{\mu^2}{\sigma^2}\right)$$

We may do the same for other distributions:

- Bernoulli: $f(x) = \exp\left(x \log \frac{p}{1-p} + \log p + \log \mathbb{I}_{\{0,1\}}(x)\right)$
- Poisson: $f(x) = \exp(x \log \lambda - \lambda - \log x! + \log \mathbb{I}_{\mathbb{N}}(x))$
- Exponential:¹ $f(x) = \exp\left(-x\lambda + \log \lambda + \log \mathbb{I}_{\mathbb{R}_+}(x)\right)$

... and note how parameters and density arguments interact.

¹In the λ and not the θ parameterization, $f(x) = \lambda \exp(-\lambda x)$.

The exponential class

Definition

The pdf $f(\mathbf{x}; \boldsymbol{\theta})$ is a member of the exponential class of pdfs iff it has the form

$$f(\mathbf{x}; \boldsymbol{\theta}) = \begin{cases} \exp \left\{ \sum_{i=1}^k c_i(\boldsymbol{\theta}) g_i(\mathbf{x}) + d(\boldsymbol{\theta}) + z(\mathbf{x}) \right\} & \text{for } \mathbf{x} \in A \\ 0 & \text{otherwise} \end{cases},$$

where

$$\mathbf{x} = (x_1, \dots, x_n)';$$

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)';$$

$c_i(\boldsymbol{\theta}), d(\boldsymbol{\theta})$: real-valued functions of $\boldsymbol{\theta}$ that do not depend on \mathbf{x} ;

$g_i(\mathbf{x}), z(\mathbf{x})$: real-valued functions of \mathbf{x} that do not depend on $\boldsymbol{\theta}$;

$A \subset \mathbb{R}^n$: a range/support which **does not depend** on $\boldsymbol{\theta}$.

Members of the exponential class

For $\mathcal{N}(\mu, \sigma^2)$ with $n = 1$ and $k = 2$ (# of parameters),

$$c_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}, \quad c_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}, \quad g_1(x) = x, \quad g_2(x) = x^2;$$

$$d(\boldsymbol{\theta}) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \frac{\mu^2}{\sigma^2}, \quad z(x) = 0, \quad A = \mathbb{R}.$$

The multivariate normal also fits the exponential class, btw. And so do *Bernoulli*, *binomial*, *multinomial*, *negative binomial*, *Poisson*, *geometric*, *gamma*, *chi-square*, *exponential*, *beta*, etc.

Distributions that do not belong to the exponential class are, e.g.: *discrete uniform*, *continuous uniform*, *hypergeometric*.

The exponential class of densities is a very popular model (and also has nice inferential properties; see Advanced Statistics II).

Outline

- 1 Relaxing the multivariate normal
- 2 Conditional distributions and functionals
- 3 The generalized linear model
- 4 Up next

Modelling dependence

Take the random vector $(Y, \mathbf{X}')'$ with joint density $f_{Y,\mathbf{X}}$.

- If Y and \mathbf{X} are statistically independent,

$$f_{Y,\mathbf{X}} = f_Y f_{\mathbf{X}}$$

where $f_Y, f_{\mathbf{X}}$ are the respective marginal distributions.

- If not, we may write

$$f_{Y,\mathbf{X}} = f_{Y|\mathbf{X}} f_{\mathbf{X}},$$

where $f_{Y|\mathbf{X}}$ is the conditional distribution of Y given \mathbf{X} .

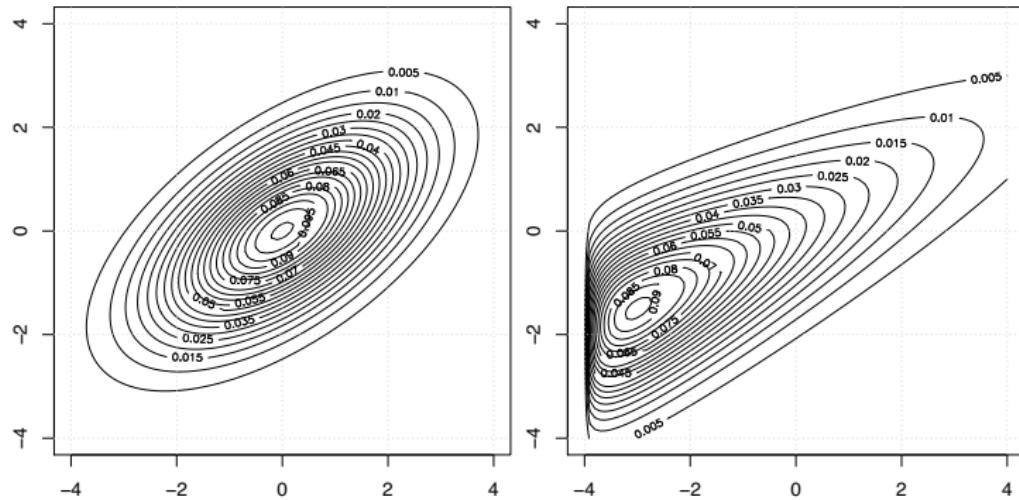
- The conditional distribution changes, in general, with \mathbf{X} !

This helps understand how (features of) Y depend(s) on \mathbf{X} , and use such knowledge to set up models suitable for dependent data.²

²Beware of causal interpretations, though!

Joint distributions

Different marginal distributions imply different joint ones.



But: If only dependence is of interest, f_X need not receive any attention.³

³Copulas offer another way of decomposing joint distributions; see the lecture notes.

The linear conditional normal model

For the multivariate normal,

- the conditional distribution is normal, and
- the conditional expectation is linear.

We may write

$$Y = \beta_0 + \boldsymbol{\beta}' \mathbf{X} + E$$

where $E \sim \mathcal{N}(0, \sigma^2)$, independent of \mathbf{X} .

We then obtain the linear conditionally normal model

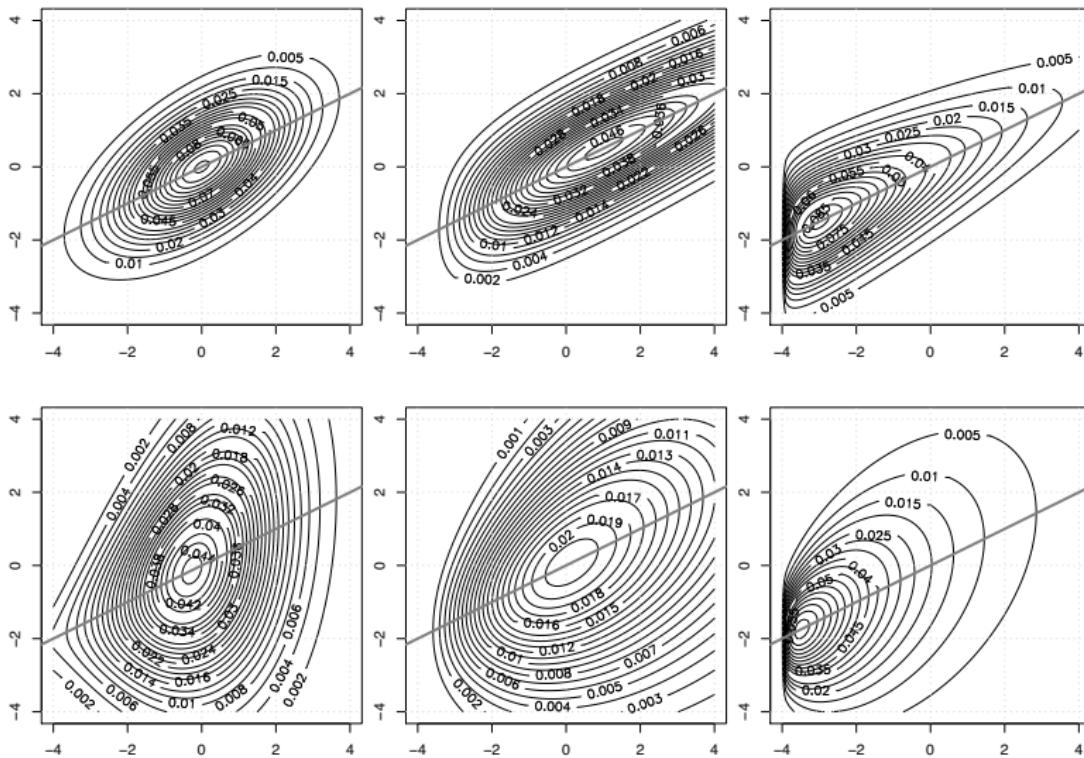
$$Y | \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\beta_0 + \boldsymbol{\beta}' \mathbf{x}, \sigma^2).$$

(Y is only marginally normal if \mathbf{X} is normal, independent of the errors.)

For non-Gaussian E or varying σ^2 , we in fact *only* model $E(Y | \mathbf{X})$ ⁴

⁴This is what people usually understand under a **regression model**.

Same linear regression curve, different cond. distributions



Possible extensions

Such models for the conditional mean can be used

- for forecasting when outcomes for X are observed, and
- also for causal analyses (with some care of course; see any econometrics course).

This makes them quite useful...

We would also like to have such models for

- other conditional functionals, say quantiles,
- other distributions than the normal, or
- other functional forms than linearity.

Each relaxation (not to mention all at the same time) brings up interesting models...

Conditional quantiles

Let Y and \mathbf{X} have a continuous joint distribution.

- Then, q_p such that $P(Y \leq q_p) = p$ is the marginal p -quantile of Y .
- When focussing on conditional distributions, we naturally obtain conditional quantiles...

So we call $q_p(\mathbf{x})$ the conditional p -quantile of Y given $\mathbf{X} = \mathbf{x}$ if

$$P(Y \leq q_p(\mathbf{x}) | \mathbf{X} = \mathbf{x}) = p.$$

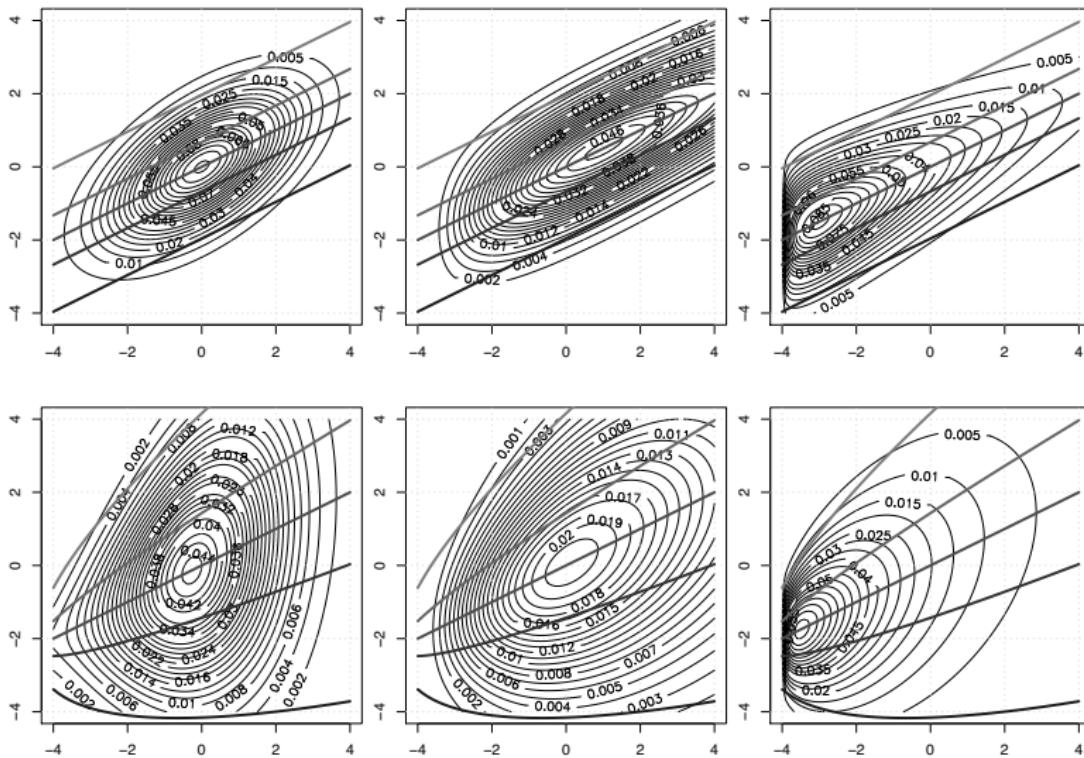
If q_p is linear in \mathbf{x} , we obtain a linear quantile regression model.

Note that q_p will be different for different levels p !

We may rewrite this as

$$Y = q_p(\mathbf{X}) + E_p$$

where E_p has zero conditional quantile given \mathbf{X} .

Quantile functions for $p \in \{0.025, 0.25, 0.5, 0.75, 0.975\}$ 

Outline

- 1 Relaxing the multivariate normal
- 2 Conditional distributions and functionals
- 3 The generalized linear model
- 4 Up next

Other types of data

A linear regression model is clearly a bad model for binary or count data (or for durations for that matter).

In the univariate case, we used e.g. the Bernoulli or Gamma distributions.

- So we would have to set up models where the conditional distribution is Bernoulli etc.
- Like for the normal regression, we make parameters of these distributions depend on X !

We'll do this in the framework of the exponential class of densities, which leads us to **generalized linear models**.

The GLM

To keep things nicely interpretable, we'll resort to a special version of the univariate exponential class. Start with

$$f(y, \theta) = \exp(\theta \cdot y - b(\theta) + z(y)).$$

(We say the density is in **canonical form** iff $c(\theta) = \theta$.)

You can represent the Bernoulli and Poisson distributions this way, but not the normal. So add a second parameter to obtain a so-called overdispersed version, namely

$$f(y, \theta, \psi) = \exp\left(\frac{\theta \cdot y - b(\theta)}{a(\psi)} + d(y, \psi) + z(y)\right).$$

The Bernoulli case

If Y is Bernoulli distributed,

$$f(y) = \exp \left(y \log \left(\frac{p}{1-p} \right) + \log(1-p) + \log \mathbb{I}_{\{0,1\}}(y) \right).$$

With $\theta := \log \left(\frac{p}{1-p} \right)$, we have an exponential family with no extra parameter ψ ,

$$a(\psi) = 1, \quad b(\theta) = -\log(1-p) = \log(1+e^\theta), \quad d(y, \psi) = 0.$$

(We call $\log \left(\frac{p}{1-p} \right)$ the logit or log-odds transform, and its inverse $e^\theta / (1 + e^\theta)$ the logistic transform.)

... and the rest

Thus,

- Poisson:

$$\theta = \log \lambda, \quad b(\theta) = \lambda = e^\theta, \quad b'(\theta) = e^\theta, \quad a(\psi) = 1;$$

- Exponential:⁵

$$\theta = \lambda, \quad b(\theta) = \log \lambda = \log \theta, \quad b'(\theta) = \frac{1}{\theta}, \quad a(\psi) = -1 :$$

- Gaussian:

$$\theta = \mu, \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}, \quad b'(\theta) = \theta, \quad a(\psi) = \sigma^2.$$

⁵Again, in λ parameterization to avoid confusions.

Properties of the canonical form

Lemma

Regularity conditions assumed, we have for the above representation

$$\begin{aligned}\mathrm{E}(Y) &= b'(\theta), \\ \mathrm{Var}(Y) &= b''(\theta)a(\psi).\end{aligned}$$

For linear regression we had

$$\mathrm{E}(Y|\boldsymbol{X}) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}.$$

Putting this in GLM language, we have equivalently

$$b'(\theta) := \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} = \mu(\boldsymbol{X})$$

This way obtain a conditional model for Y given $\boldsymbol{X} = \boldsymbol{x}$.

The link function

What we did for the Gaussian was to choose θ to be a linear function of \mathbf{X} .

Generally we set $E(Y|\mathbf{X}) = b'(\theta) = G(\beta_0 + \boldsymbol{\beta}'\mathbf{X})$ where G is called **link function**.

Then, $b'(\theta)$ gives **the canonical** link:

- Gaussian regression: G is the identity function;
- (Exponential) hazard model: G is the reciprocal function;
- Poisson regression: G is the exponential link;
- (Bernoulli) Logit regression: G is the logistic transform $e^\theta/(1 + e^\theta)$

One may even use other link functions instead of the above canonical ones.

And of course linearity of b' in \mathbf{X} may be relaxed.

More nonlinearity

Sofar, the GLM used a transformation of a linear combination of the \mathbf{X} s⁶ to model conditional means.

But there is nothing that keeps us from making more generalizations:

- We may take nonlinear functions of \mathbf{X}
 - E.g. for the Gaussian GLM, this leads to nonlinear regression models
- We may also let ψ also depend on \mathbf{X}
 - This usually amounts modeling the conditional mean *and* the conditional variance
 - ... leading for the Gaussian GLM to location-scale models

Conditional location-scale models can also be quite useful in practice...

⁶Such models are called single-index models, btw.

Conditional location-scale models

Write

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})E$$

where E has zero conditional mean and unit conditional variance.⁷ Then,

$$\text{E}(Y|\mathbf{X}) = m(\mathbf{X}), \quad \text{Var}(Y|\mathbf{X}) = \sigma^2(\mathbf{X}).$$

If E and \mathbf{X} are independent, there's a direct implication for quantiles:

- Let $q_{p,E}$ be the p -quantile of E : under independence of E and \mathbf{X} ,

$$\text{P}(E - q_{p,E} < 0) = p = \text{P}(E - q_{p,E} < 0|\mathbf{X})$$

- This implies $\text{P}(Y < \mu(\mathbf{x}) + q_{p,E}\sigma(\mathbf{x})|\mathbf{X} = \mathbf{x}) = p$, or

$$Y = \mu(\mathbf{x}) + q_{p,E}\sigma(\mathbf{x}) + U_p$$

where $U_p = \sigma(\mathbf{x})(E - q_{p,E})$ has zero conditional p -quantile.

⁷This is satisfied if E is standardized and independent of \mathbf{X} .

Outline

- 1 Relaxing the multivariate normal
- 2 Conditional distributions and functionals
- 3 The generalized linear model
- 4 Up next

Coming up

Basic asymptotics

Asymptotics: Stochastic convergence

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Why bother?

We often encounter **random quantities** of the form

$$Y_{\textcolor{blue}{n}} = g(X_1, \dots, X_{\textcolor{blue}{n}}), \quad \text{where } n = 1, 2, 3, \dots.$$

A simple – but ubiquitous – case is the average of n random variables

$$Y_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

The objective of asymptotic theory is to establish results relating to the stochastic behavior of such sequences $Y_{\textcolor{blue}{n}}$ when $\textcolor{blue}{n} \rightarrow \infty$.

- $Y_{\textcolor{blue}{n}}$ may converge to a constant in various ways,
- or the distribution of $Y_{\textcolor{blue}{n}}$ may converge to some ‘limit distribution’.

We do this to obtain an approximation of the behavior of Y_n !

Asymptotics: Stochastic convergence

- 1 Convergence of number and function sequences
- 2 Shorthand notation: Landau symbols
- 3 Convergence of sequences of random variables
- 4 Up next

Outline

- 1 Convergence of number and function sequences
- 2 Shorthand notation: Landau symbols
- 3 Convergence of sequences of random variables
- 4 Up next

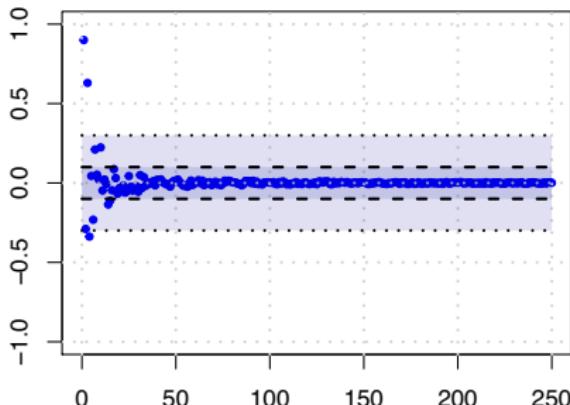
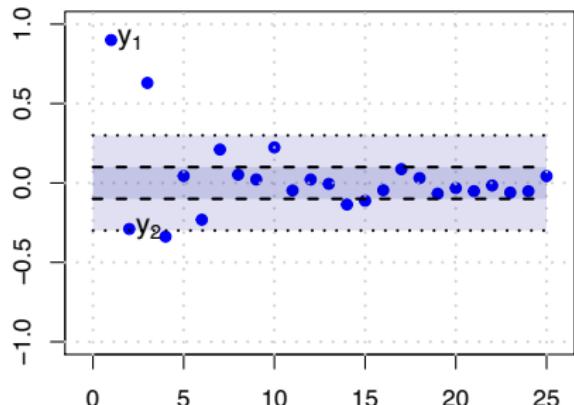
Real sequences and limits

Definition (Convergence of real number sequences)

A sequence of real numbers $\{y_n\}$ converges to $y \in \mathbb{R}^1$ iff for every real $\epsilon > 0$ there exists an integer $N(\epsilon)$ such that

$$|y_n - y| < \epsilon \quad \forall n \geq N(\epsilon).$$

The existence of the limit is denoted by $y_n \rightarrow y$ or $\lim_{n \rightarrow \infty} y_n = y$.



Relation to boundedness

For the limit of a sequence of numbers to exist, it is necessary (but not sufficient) that the sequence be **bounded**.

Example

The sequence $y_n = 3 + n^{-2}$, $n \in \mathbb{N}$

- is bounded, since $|y_n| \leq 4 \quad \forall n \in \mathbb{N}$,
- and has a limit $y_n \rightarrow 3$.

Example

The sequence $y_n = \sin n$, $n \in \mathbb{N}$

- is bounded, since $|\sin x| \leq 1 \quad \forall x$,
- but does not have a limit, since $\sin x$ cycles between +1 and -1.

Extensions

Convergence of sequences of vectors and matrices is defined **elementwise**.

Definition (Convergence of function sequences)

Let $\{f_n(x)\}$, $n \in \mathbb{N}$, be a sequence of functions having a common domain $D \subset \mathbb{R}^m$. The function sequence $\{f_n(x)\}$ converges to a function $f(x)$ with domain $D_0 \subset D$ iff for $n \rightarrow \infty$

$$f_n(x) \rightarrow f(x) \quad \forall x \in D_0.$$

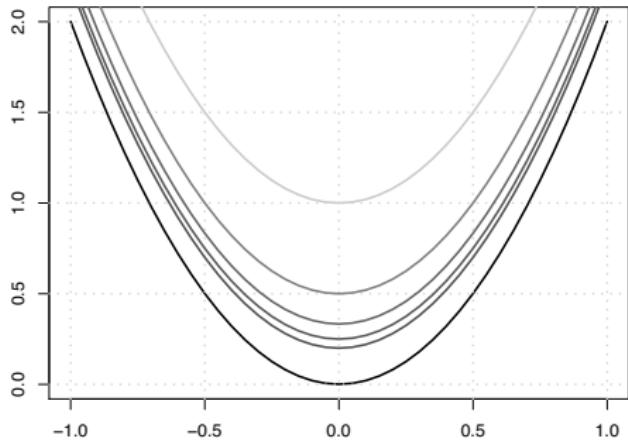
f is called the **limiting function** of $\{f_n\}$.

The definition implies that the values of the functions $f_n(x)$, $n = 1, 2, 3, \dots$ converge to $f(x)$ **pointwise** for each single $x \in D_0$.

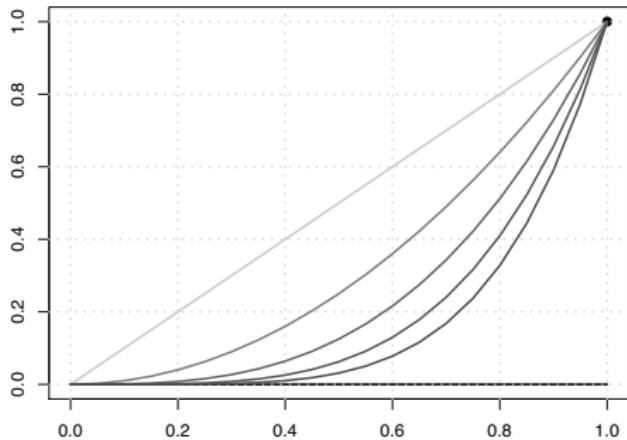
Hence, $f(x)$ can be viewed as an approximation of $f_n(x)$ when n is large **for a given x** .

Two examples

$$f_n(x) = n^{-1} + 2x^2, x \in [-1, 1]$$



$$f_n(x) = x^n, x \in [0, 1]$$



In fact, the latter is an example of **non-uniform** convergence.

Uniform convergence

Uniformity of the convergence requires $\sup_x |f_n(x) - f(x)| \rightarrow 0$ as $n \rightarrow \infty$ and is (much) stricter.

- In fact, $\sup_x |f_n(x) - f(x)|$ may be seen as a distance between two functions.
- Distances (between elements of a vector space) may be characterized formally:
 - Should be zero iff the two elements are the same
 - Should obey the triangle inequality.
 - Should be symmetric and nonnegative.
- The sup distance does fulfil these, but there are other metrics, e.g. the so-called L_2 distance, $\sqrt{\int (f_n(x) - f(x))^2 dx}$.

Outline

- 1 Convergence of number and function sequences
- 2 Shorthand notation: Landau symbols
- 3 Convergence of sequences of random variables
- 4 Up next

Big-Oh, Small-oh (Landau) notations

Definition (Order of magnitude of a sequence)

Let $\{y_n\}$ be a real number sequence.

- $\{y_n\}$ is said to be **at most of order n^k** , denoted by $y_n = O(n^k)$, if there exists a finite constant c such that

$$\left| \frac{y_n}{n^k} \right| \leq c \quad \forall n \in \mathbb{N}.$$

- $\{y_n\}$ is said to be **of order smaller than n^k** , denoted by $y_n = o(n^k)$, if

$$\frac{y_n}{n^k} \rightarrow 0.$$

More generally, one may replace n^k by some sequence \tilde{y}_n .

Revisiting convergence and boundedness

- Write $O(n^0)$ and $o(n^0)$ as $O(1)$ and $o(1)$.
- They have a special interpretation...

Example

The sequence $y_n = 3 + n^{-2}$, $n \in \mathbb{N}$

- is $O(1)$, since $|y_n| \leq 4 \quad \forall n \in \mathbb{N}$,
- and has a limit, $y_n - 3 = o(1)$.

Example

The sequence $y_n = \sin n$, $n \in \mathbb{N}$

- is just $O(1)$, since $|\sin x| \leq 1 \quad \forall x$.

Typical relations

Note e.g. that

- | | | |
|----------------------------|------|---|
| if $\{y_n\}$ is $O(n^k)$, | then | $\{y_n\}$ is $o(n^{k+\epsilon}) \forall \epsilon > 0$; |
| if $\{y_n\}$ is $o(n^k)$, | then | $\{y_n\}$ is $O(n^k)$. |

Example

Let $\{y_n\}$ be defined by $y_n = 3n^3 - n^2 + 2, n \in \mathbb{N}$. Since

$$\frac{y_n}{n^3} = 3 - \frac{1}{n} + \frac{2}{n^3} \rightarrow 3 < \infty, \quad \text{we have} \quad y_n = O(n^3);$$

Since for a positive ϵ

$$\frac{y_n}{n^{3+\epsilon}} = \frac{3}{n^\epsilon} - \frac{1}{n^{1+\epsilon}} + \frac{2}{n^{3+\epsilon}} \rightarrow 0, \quad \text{we have} \quad y_n = o(n^{3+\epsilon}).$$

But note that $y_n = O(n^\alpha)$ **does not imply** that $\frac{1}{y_n} = O(n^{-\alpha})!$

And some rules

If $x_n = O(\tilde{x}_n)$ and $y_n = O(\tilde{y}_n)$, then

$$\begin{aligned}x_n + y_n &= O(\max\{\tilde{x}_n; \tilde{y}_n\}), \\x_n - y_n &= O(\max\{\tilde{x}_n; \tilde{y}_n\}), \\x_n y_n &= O(\tilde{x}_n \tilde{y}_n).\end{aligned}$$

If $x_n = O(\tilde{x}_n)$ and $y_n = o(\tilde{y}_n)$, then

$$\begin{aligned}x_n + y_n &= O(\max\{\tilde{x}_n; \tilde{y}_n\}), \\x_n - y_n &= O(\max\{\tilde{x}_n; \tilde{y}_n\}), \\x_n y_n &= o(\tilde{x}_n \tilde{y}_n).\end{aligned}$$

Outline

- 1 Convergence of number and function sequences
- 2 Shorthand notation: Landau symbols
- 3 Convergence of sequences of random variables
- 4 Up next

Convergence depends on distance

In this section we extend the converge concepts for real number sequences to sequences of random variables.

For sequences of random variables, we distinguish among the following types/modes of convergence:

- 1.) *convergence in distribution*;
- 2.) *convergence in probability*;
- 3.) *convergence in mean square*;
- 4.) *almost-sure convergence*.

They express different ways in which the sequence may be close to its limit.

Convergence in Distribution

Definition (Convergence in Distribution)

Let $\{Y_n\}$ be a sequence of random variables with an associated sequence of cdfs $\{F_n\}$. If there exists a cdf F such that as $n \rightarrow \infty$

$$F_n(y) \rightarrow F(y) \quad \forall y \quad \text{at which } F \text{ is continuous,}$$

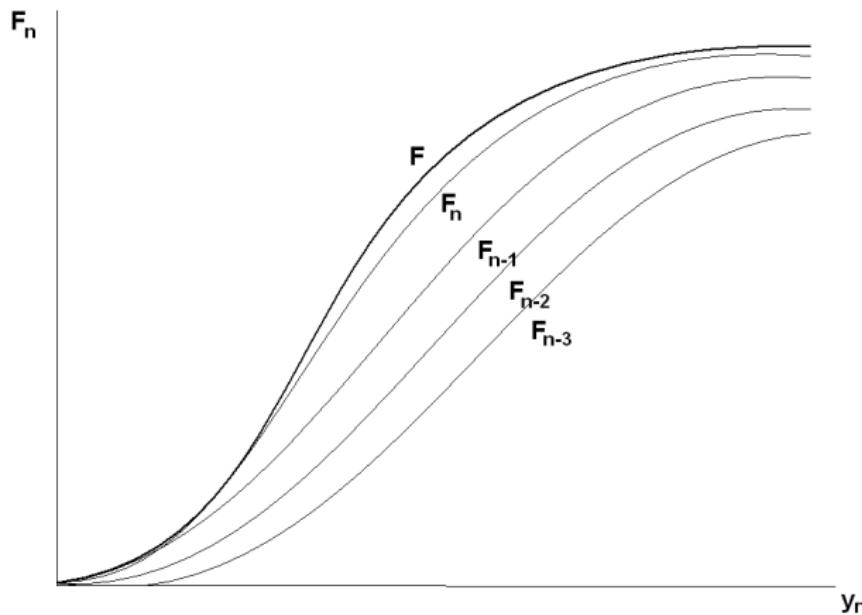
then Y_n converges in distribution to the random variable Y with cdf F .

We denote this by $Y_n \xrightarrow{d} Y$ or $Y_n \xrightarrow{d} F$. The function F is called the **limiting cdf/limiting distribution** of $\{Y_n\}$.

- The limiting cdf F can be the cdf of a *degenerate random variable* with $Y = c$, where c is a constant.
- In this case, we say that Y_n **converges in distribution to a constant**, and we denote this by $Y_n \xrightarrow{d} c$.

Convergence of distribution functions

- If the limit is continuous, convergence is also uniform.
- If $Y_n \xrightarrow{d} Y$, then as n becomes large, the actual cdf of Y_n can be approximated by the cdf F of the random variable Y .



... more concretely

Example

Let $\{Y_n\}$ be a sequence of random variables with an associated sequence of cdfs $\{F_n\}$ given by

$$F_n(y) = \begin{cases} 0 & \text{for } y < 0 \\ (\frac{y}{\theta})^n & \text{for } 0 \leq y < \theta \\ 1 & \text{for } y \geq \theta \end{cases} .$$

We see that as $n \rightarrow \infty$,

$$F_n(y) \rightarrow F(x) = \begin{cases} 0 & \text{for } y < \theta \\ 1 & \text{for } y \geq \theta \end{cases} ,$$

which is the cdf a *degenerate random variable*, and we have $Y_n \xrightarrow{d} \theta$.

Establishing convergence in distribution

The following theorem is based upon the uniqueness of MGFs, and is very useful for identifying limiting distributions.

Theorem (5.1)

Let $\{Y_n\}$ be a sequence of random variables having an associated sequence of MGFs $\{M_{Y_n}(t)\}$. Let $M_Y(t)$ be the MGF of Y . Then

$$Y_n \xrightarrow{d} Y \quad \text{iff} \quad M_{Y_n}(t) \rightarrow M_Y(t) \quad \forall t \in (-h, h), \text{ for some } h > 0.$$

What about densities?

- Convergence of cdfs does not imply convergence of pdfs
- ... but the converse holds; see Mittelhammer (1996, Theorem 5.1).

The χ^2 case

Example

Let $X_n \sim \chi_{(n)}^2$ with MGF $M_{X_n}(t) = (1 - 2t)^{-\frac{n}{2}}$, $t < 1/2$ and

$$Z_n = \frac{X_n - n}{\sqrt{2n}} = -\sqrt{\frac{n}{2}} + \frac{1}{\sqrt{2n}} X_n.$$

Therefore

$$M_{Z_n}(t) = e^{-\sqrt{\frac{n}{2}}t} \cdot M_{X_n}\left(\frac{1}{\sqrt{2n}} \cdot t\right) = e^{-\sqrt{\frac{n}{2}}t - \frac{n}{2} \ln\left(1 - \sqrt{\frac{2}{n}}t\right)}.$$

But $\ln(1 - x) = -x + x^2/2 + o(x^2)$ so

$$M_{Z_n}(t) = e^{\frac{t^2}{2} + o(1)} = e^{\frac{t^2}{2}} + o(1)$$

which is the same as $M_{Z_n}(t) \rightarrow e^{\frac{t^2}{2}}$, the MGF of the standard normal.

Asymptotic Distributions

The **asymptotic distribution** for a random variable Z_n is any distribution that provides an **approximation to the true distribution** of Z_n for large n .

If $\{Z_n\}$ has a **limiting distribution**,

- this may be considered as an **asymptotic distribution**, since
- it provides an approximation to the distribution of Z_n for large n .

The following definition of the asymptotic distribution

- generalizes the concept of approximating distributions for large n and
- includes cases where Z_n has no limiting distribution or a degenerate limiting distribution.

A bit of a misnomer

Definition (Asymptotic Distribution)

Let $\{Z_n\}$ be a sequence of random variables defined by

$$Z_n = h(X_n, a_n), \quad \text{where} \quad X_n \xrightarrow{d} X \text{ (nondegenerate),}$$

a_n : sequence of numbers/parameters.

An *asymptotic distribution* for Z_n is the distribution of $h(X, a_n)$,

$$Z_n \xrightarrow{a} h(X, a_n) \quad \text{"}Z_n \text{ is asymptotically distributed as } h(X, a_n)\text{"}.$$

Example

We know e.g. that $W_n = \frac{X_n - n}{\sqrt{2n}} \xrightarrow{d} W \sim \mathcal{N}(0, 1)$ if $X_n \sim \chi^2_{(n)}$.

Consider now $Y_n = h(W_n, n) = \sqrt{2n} \cdot W_n + n (= X_n)$.

Then $Y_n = h(\textcolor{blue}{W}_n, n) \xrightarrow{a} h(\textcolor{blue}{W}, n) = \sqrt{2n} \cdot \textcolor{blue}{W} + n \sim \mathcal{N}(n, 2n)$.

The continuous mapping theorem

Theorem (5.2 (CMT))

Let $X_n \xrightarrow{d} X$, and let $Y_n = g(X_n)$ with g a continuous function which depends on n only via X_n . Then $g(X_n) \xrightarrow{d} g(X)$.

This helps working with limiting/asymptotic distributions

Example

Consider e.g. $Z_n \xrightarrow{d} Z \sim \mathcal{N}(0, 1)$. Then

- $g(Z_n) = 2Z_n + 5 \xrightarrow{d} 2Z + 5 \sim \mathcal{N}(5, 4)$;
- $g(Z_n) = Z_n^2 \xrightarrow{d} Z^2 \sim \chi_{(1)}^2$.

Convergence in Probability

Focus on more than just distributions converging...

Definition (Convergence in probability)

The sequence of random variables $\{Y_n\}$ *converges in probability* to the random variable Y iff

$$\lim_{n \rightarrow \infty} P(|Y_n - Y| < \epsilon) = 1 \quad \forall \epsilon > 0.$$

We denote this by $Y_n \xrightarrow{p} Y$, or $\text{plim } Y_n = Y$, where Y is called the **probability limit** of Y_n .

The definition implies that if n is large enough, observing outcomes of Y_n is essentially equivalent to observing outcomes of Y .

Also note that the probability limit Y can be a *degenerate random variable* with $Y = c$, where c is a constant. We denote this by $Y_n \xrightarrow{p} c$.

Playing with normals

Example

Consider the random variable Y_n with pdf

$$f_n(y) = \frac{1}{n} \mathbb{I}_{\{0\}}(y) + \left(1 - \frac{1}{n}\right) \mathbb{I}_{\{1\}}(y) \quad \rightarrow_{n \rightarrow \infty} \mathbb{I}_{\{1\}}(y).$$

Hence we have $P(|Y_n - 1| = 0) \rightarrow 1$ as $n \rightarrow \infty$, so that

$$\lim_{n \rightarrow \infty} P(|Y_n - 1| < \epsilon) = 1 \quad \forall \epsilon > 0, \quad \text{and} \quad \text{plim } Y_n = 1.$$

Example

Let $Y \sim \mathcal{N}(0, 1)$ and $Z_n \sim \mathcal{N}(0, \frac{1}{n})$, independent. Let $Y_n = Z_n + Y$ such that $Y_n \sim \mathcal{N}\left(0, [1 + \frac{1}{n}]\right)$. Then, $\text{plim } Y_n = Y$ since

$$\lim_{n \rightarrow \infty} P(|Y_n - Y| < \epsilon) = \underbrace{\lim_{n \rightarrow \infty} P(|Z_n| < \epsilon)}_{\text{Chebyshev's Ineq.}} \geq \lim_{n \rightarrow \infty} \left(1 - \frac{\text{Var}(Z_n)}{\epsilon^2}\right) = 1.$$

Probabilistic Landau symbols

Definition

A (stochastic) sequence Y_n is said to be $o_p(n^k)$ if $\frac{Y_n}{n^k} \xrightarrow{p} 0$.

Things are slightly different for the big-Oh case, though:

Definition

- ① A stochastic sequence Y_n is said to be **uniformly bounded in probability** if $\forall \epsilon > 0 \exists C(\epsilon)$ such that $\sup_n P(|Y_n| > C(\epsilon)) < \epsilon$.
- ② A stochastic sequence Y_n is said to be $O_p(n^k)$ if $\frac{Y_n}{n^k}$ is uniformly bounded in probability.

The same rules as for deterministic Landau symbols apply for finite sums, products, etc.

The continuous mapping theorem (again?)

Theorem (5.3)

Let $X_n \xrightarrow{p} X$, and let $Y_n = g(X_n)$ with g a continuous function which depends on n only via X_n . Then,

$$\text{plim } Y_n = \text{plim } g(X_n) = g(\text{plim } X_n) = g(X).$$

The theorem implies that the plim operator acts analogously to the standard lim operator of real analysis. You may in fact write

$$g(X + o_p(1)) = g(X) + o_p(1).$$

Example

Let $X_n \xrightarrow{p} 3$. Then the probability limit of $Y_n = \ln(X_n) + \sqrt{X_n}$ is

$$\text{plim } Y_n = \ln(\text{plim } X_n) + \sqrt{\text{plim } X_n} = \ln(3) + \sqrt{3}.$$

Special cases

Theorem (5.4)

For the sequences of random variables X_n , Y_n , and the constant a .

- a. $\text{plim}(aX_n) = a(\text{plim } X_n);$
- b. $\text{plim}(X_n + Y_n) = \text{plim } X_n + \text{plim } Y_n$ (*the plim of a sum = the sum of the plims*);
- c. $\text{plim}(X_n Y_n) = \text{plim } X_n \text{ plim } Y_n$ (*the plim of a product = the product of the plims*);
- d. $\text{plim}(X_n/Y_n) = (\text{plim } X_n)/(\text{plim } Y_n)$ (*if denominators are nonzero*).

The results of Theorem 5.4 extend to matrices by applying them to matrices element-by-element – see Mittelhammer (1996, p. 244-245).

Relation to convergence in distribution

Theorem (5.5)

$$Y_n \xrightarrow{p} Y \Rightarrow Y_n \xrightarrow{d} Y.$$

Example

Let e.g. $Y_n = (2 + \frac{1}{n})X + 3$, where $X \sim \mathcal{N}(1, 2)$. Then the plim of Y_n is

$$\text{plim } Y_n = \text{plim} \left(2 + \frac{1}{n} \right) \text{plim } X + \text{plim } 3 = 2X + 3 = Y \sim \mathcal{N}(5, 8).$$

Hence $Y_n \xrightarrow{d} Y \sim \mathcal{N}(5, 8)$.

More limits

Theorem (5.6)

$$Y_n \xrightarrow{d} c \Rightarrow Y_n \xrightarrow{p} c.$$

Theorem (5.7 (Slutsky's Theorem(s)))

Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$. Then,

- a. $X_n + Y_n \xrightarrow{d} X + c$;
- b. $X_n \cdot Y_n \xrightarrow{d} X \cdot c$;
- c. $X_n / Y_n \xrightarrow{d} X/c$.

Convergence in mean square

Definition (Convergence in mean square)

The sequence of random variables $\{Y_n\}$ **converges in mean square** to the random variable Y , iff

$$\lim_{n \rightarrow \infty} E((Y_n - Y)^2) = 0.$$

We denote this by $Y_n \xrightarrow{m} Y$.

First- and second order moments of Y_n and Y converge to one another; see below.

The basic tool

Theorem (5.8)

$Y_n \xrightarrow{m} Y$ iff

- a. $E(Y_n) \rightarrow E(Y),$
- b. $\text{Var}(Y_n) \rightarrow \text{Var}(Y),$
- c. $\text{Cov}(Y_n, Y) \rightarrow \text{Var}(Y).$

The necessary and sufficient conditions in Theorem 5.8 simplify, when Y is a constant, as stated in the following corollary.

Corollary (5.1)

$Y_n \xrightarrow{m} c$ iff $E(Y_n) \rightarrow c$ and $\text{Var}(Y_n) \rightarrow 0.$

Thank Markov/Chebychev for this shortcut

Theorem (5.9)

$$Y_n \xrightarrow{m} Y \Rightarrow Y_n \xrightarrow{p} Y.$$

- Often convergence in mean square is relatively easy to demonstrate.
- The sample average of $iid(0, \sigma^2)$ variables is the leading example.
- But the converse is not true!

Example

Let e.g. the pdf of Y_n be given by $f_n(y) = \begin{cases} 1 - \frac{1}{n^2} & \text{for } y_n = 0 \\ \frac{1}{n^2} & \text{for } y_n = n \end{cases}$.

It immediately follows that $P(y_n = 0) \rightarrow 1$ so that $\text{plim } Y_n = 0$. However,

$$E((Y_n - 0)^2) = E(Y_n^2) = 0^2 \cdot (1 - \frac{1}{n^2}) + n^2 \cdot \frac{1}{n^2} = 1 \quad \forall n,$$

so that $Y_n \xrightarrow{m} 0$.

Outline

- 1 Convergence of number and function sequences
- 2 Shorthand notation: Landau symbols
- 3 Convergence of sequences of random variables
- 4 Up next

Coming up

Limiting theorems

Asymptotics: Limiting results

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Asymptotics: Limiting results

- 1 Weak Laws of Large Numbers
- 2 Central Limit Theorems
- 3 The delta method
- 4 Up next

Outline

1 Weak Laws of Large Numbers

2 Central Limit Theorems

3 The delta method

4 Up next

Focus on the sample average

Definition (Weak Law of Large Numbers)

Let $\{X_n\}$ be a sequence of random variables with finite expected values $E(X_n) = \mu_n$. We say that $\{X_n\}$ obeys the weak law of large numbers (WLLN), if

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) = \bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0.$$

(May generalize with a more generic sequence b_n instead of n .)

For $E(X_i) = \mu_i = \mu \forall i$ the WLLN would imply that

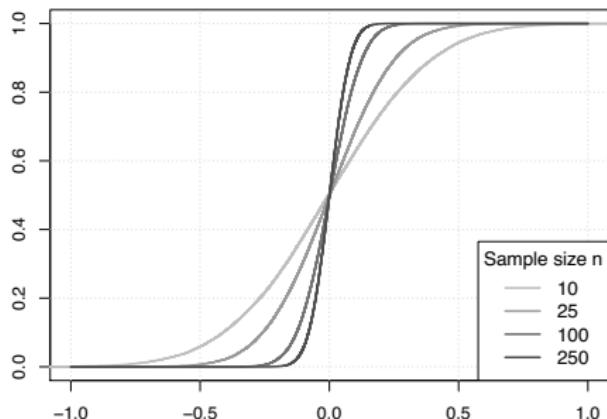
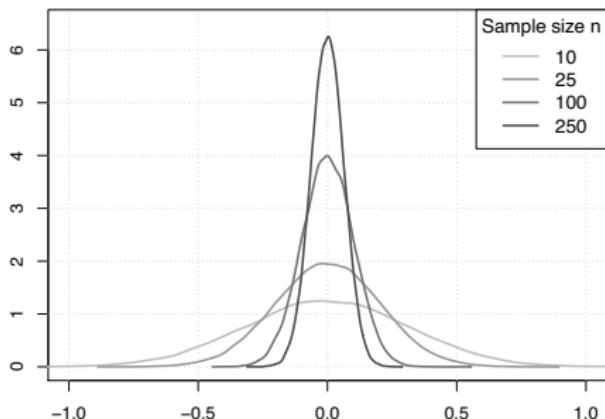
$$\frac{1}{n} \sum_{i=1}^n X_i - \mu \xrightarrow{p} 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mu,$$

such that the sample average converges in probability to the expectation.

The basic flavour

Theorem (5.10 (Khinchin's WLLN))

Let $\{X_n\}$ be a sequence of iid random variables with finite expectations $E(X_n) = \mu \forall n$. Then $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mu$.



$X_n \sim iid \mathcal{N}(0, 1)$. Left: pdf of \bar{X}_n ; Right: cdf of \bar{X}_n

Gamma distributions

Example

Let $\{X_n\}$ be a sequence of iid random variables, with $X_n \sim \text{Gamma}(\alpha, \beta)$ such that $E(X_n) = \alpha\beta$. Khinchin's WLLN implies that

$$\bar{X}_n \xrightarrow{P} E(X_n) = \alpha\beta.$$

Hence, for large enough n , the outcome of the random variable \bar{X}_n can be taken as a close approximation of $\alpha\beta$.

This is *the* property of a *consistent estimator* for $\alpha\beta$ as we shall discuss in the course *Advanced Statistics II*.

Relax some, strengthen other assumptions...

Theorem (5.11)

Let $\{X_n\}$ be a sequence of random variables with finite variances, and let $\{\mu_n\}$ be the corresponding sequence of their expectations, Then

$$\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0 \quad \text{iff} \quad E \left[\frac{(\bar{X}_n - \bar{\mu}_n)^2}{1 + (\bar{X}_n - \bar{\mu}_n)^2} \right] \rightarrow 0.$$

Theorem (5.12)

Let $\{X_n\}$ be a sequence of random variables with respective expectations given by $\{\mu_n\}$. If

$$\text{Var}(\bar{X}_n) \rightarrow 0, \quad \text{then} \quad \bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0.$$

Variances

Example

Let $\{X_n\}$ be a sequence random variables, with

$$\mathbb{E}(X_i) = \mu_i = \frac{1}{2^i}, \quad \text{Var}(X_i) = 4, \quad \text{and} \quad \text{Cov}(X_i, X_j) = 0 \quad \forall i \neq j.$$

The mean and variance for the average \bar{X}_n are

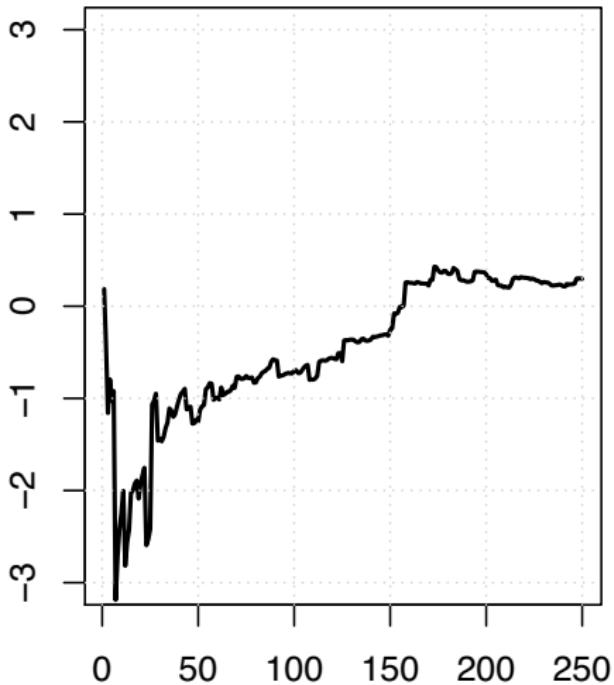
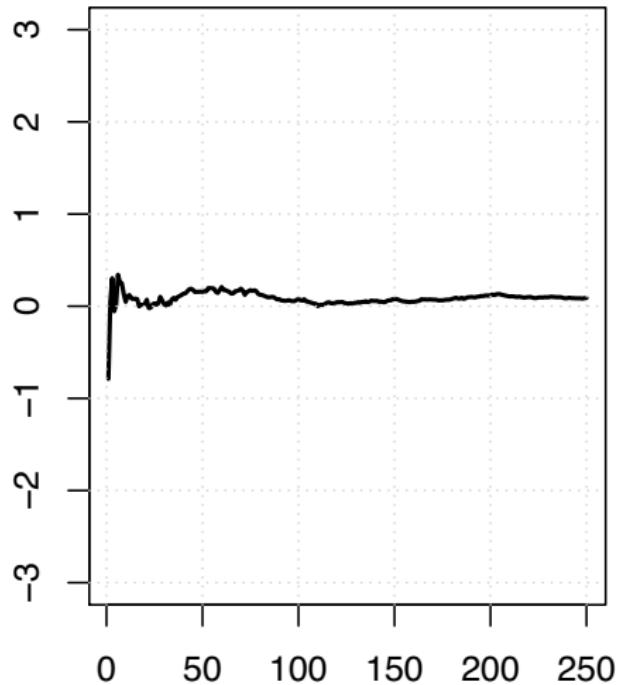
$$\mathbb{E}(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n \mu_i = \frac{1}{n} \sum_{i=1}^n \frac{1}{2^i} = \frac{1 - (\frac{1}{2})^n}{n}, \quad \text{Var}(\bar{X}_n) = \frac{4}{n}.$$

Since $\text{Var}(\bar{X}_n) \rightarrow 0$, it follows by Theorem 5.12, that

$$\bar{X}_n - \bar{\mu}_n = \bar{X}_n - \frac{1 - (\frac{1}{2})^n}{n} \xrightarrow{p} 0.$$

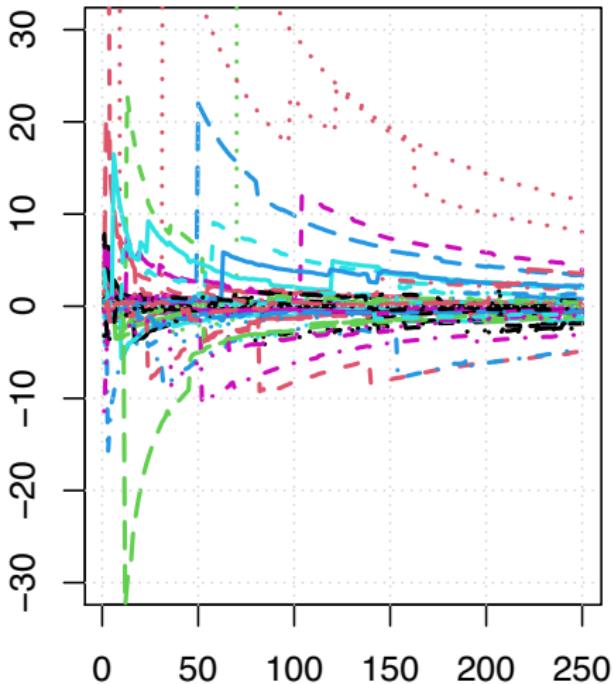
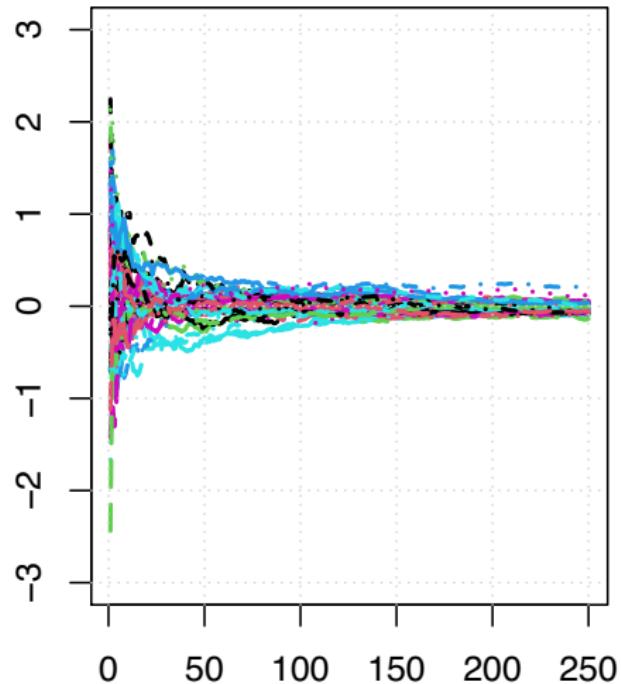
Also note that $\bar{\mu}_n = \frac{1 - (\frac{1}{2})^n}{n} \rightarrow 0$, such that $\bar{X}_n \xrightarrow{p} 0$.

Finite vs. infinite mean: single outcome paths



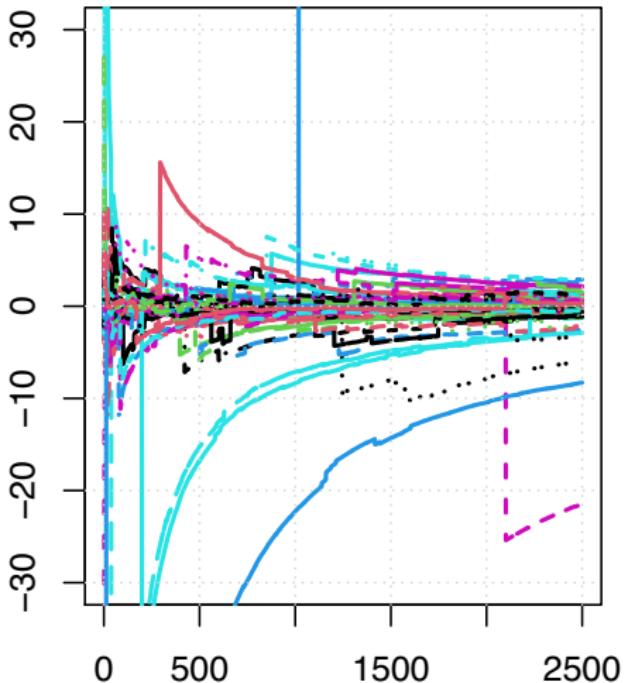
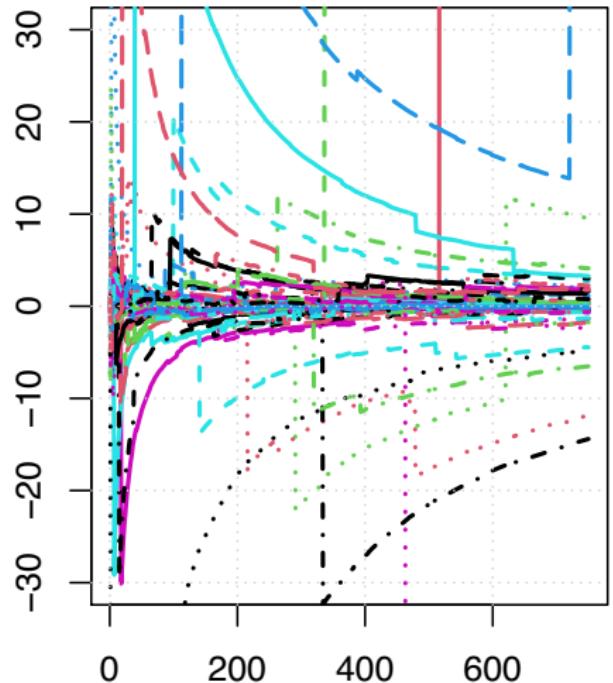
Left: path of \bar{X}_n for $X_n \sim iid \mathcal{N}(0, 1)$; Right: path of \bar{X}_n for $X_n \sim iid t(1)$

Finite vs. infinite mean: the bigger picture



50 paths for $\mathcal{N}(0, 1)$ and $t(1)$ distributed summands.

Finite vs. infinite mean: the scary picture



No apparent convergence for larger n either.

Outline

1 Weak Laws of Large Numbers

2 Central Limit Theorems

3 The delta method

4 Up next

Terminology

Central limit theorems (CLTs) are concerned with the conditions under which sequences of random variables **converge in distribution** to known families of distribution.

Definition

Let $\{X_n\}$ be a sequence of random variables, and let $S_n = \sum_{i=1}^n X_i$, $n = 1, 2, \dots$. Here we focus on the convergence in distribution of sequences of random variables of the following form

$$b_n^{-1}(S_n - a_n) \xrightarrow{d} Y \sim \mathcal{N}(0, \Sigma),$$

where $\{a_n\}$ and $\{b_n\}$ are sequences of appropriately chosen real constants.

A statement of conditions on $\{X_n\}$, $\{a_n\}$, and $\{b_n\}$ that ensure the convergence in distribution result constitutes a particular CLT. (May have other distributions than the normal as limits, actually.)

Why bother?

If a CLT applies, then

$$S_n \xrightarrow{asy} \mathcal{N}(a_n, b_n^2 \Sigma).$$

In fact:

- As we shall see in *Advanced Statistics II*, many procedures for parameter estimation and hypothesis testing are specified as functions of sums of random variables such as $S_n = \sum_{i=1}^n X_i$.
- CLTs are then often useful for establishing the asymptotic distributions for those procedures.

The basic flavour

Similar to the case of the WLLN, there are a variety of conditions that can be placed on the variables in the sum $S_n = \sum_{i=1}^n X_i$.

Theorem (5.13 (Lindeberg-Lévy))

Let $\{X_n\}$ be a sequence of iid random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 \in (0, \infty) \forall i$. Then

$$Y_n = \frac{1}{\sqrt{n}\sigma} \left(\sum_{i=1}^n X_i - n\mu \right) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

Obtaining asymptotic approximations

For the variable $S_n = \sum_{i=1}^n X_i$, for example, we obtain

$$S_n = \sqrt{n}\sigma Y_n + n\mu \underset{\text{def.}}{\left| \begin{array}{c} \\ \end{array} \right.} \stackrel{a}{\sim} \sqrt{n}\sigma Y + n\mu \quad \Rightarrow \quad S_n \stackrel{a}{\sim} \mathcal{N}(n\mu, n\sigma^2).$$

For the average $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ we have

$$\bar{X}_n = \frac{\sigma}{\sqrt{n}} Y_n + \mu \underset{\text{def.}}{\left| \begin{array}{c} \\ \end{array} \right.} \stackrel{a}{\sim} \frac{\sigma}{\sqrt{n}} Y + \mu \quad \Rightarrow \quad \bar{X}_n \stackrel{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Don't forget that, actually, $|S_n| \xrightarrow{p} \infty$ and $\bar{X}_n \xrightarrow{p} \mu$.

Average durations

Example

Let X_n be a sequence of independent exponentially distributed durations,

$$f(x) = \lambda \exp(-\lambda x) \mathbb{I}(x \geq 0).$$

We are interested in the average duration $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ and its distribution!

- ① The sum $S_n = \sum_{i=1}^n X_i$ is Gamma-distributed with parameters $\alpha = n$ and $\beta = 1/\lambda$,

$$f_S(x) = \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x}.$$

- ② So $\bar{X}_n = \frac{1}{n} S$ is Gamma-distributed with parameters $\alpha = n$ and $\beta = 1/(n\lambda)$,

$$f_{\bar{X}_n}(x) = \frac{(n\lambda)^n}{\Gamma(n)} x^{n-1} e^{-n\lambda x}.$$

... and their approximate distribution

Example (cont'd)

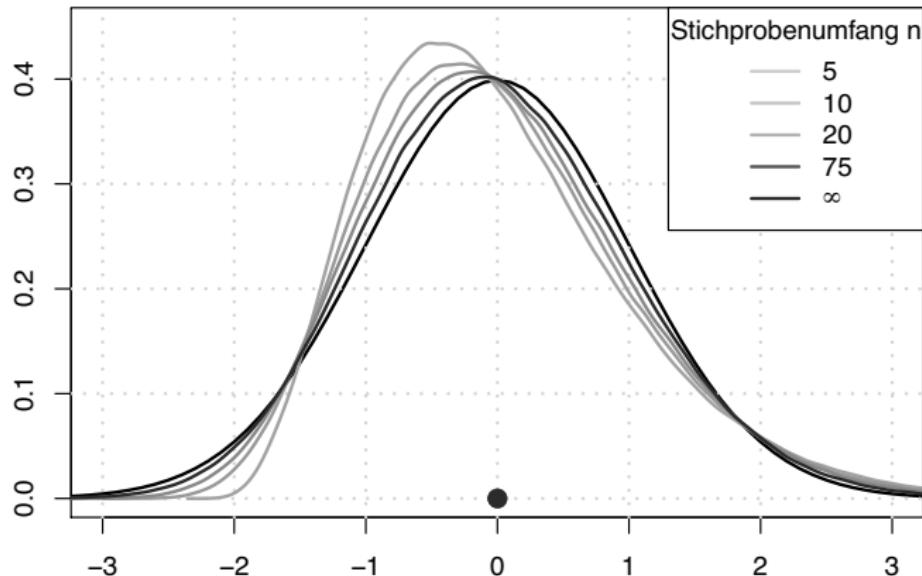
- Since $X_n \sim iid$ with $E(X_n) = \frac{1}{\lambda}$ & $\text{Var}(X_n) = \frac{1}{\lambda^2} < \infty$,

$$\sqrt{n} \left(\bar{X}_n - \frac{1}{\lambda} \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{1}{\lambda^2} \right),$$

- ... leadit to the approximation

$$\bar{X}_n \stackrel{\text{approx}}{\sim} \mathcal{N} \left(\frac{1}{\lambda}, \frac{1}{n\lambda^2} \right).$$

How good is the normal approximation?



Standard normal and exact distribution of $\sqrt{n}\lambda^2(\bar{X}_n - 1/\lambda)$,
 $\lambda = 1, n \in \{5, 10, 20, 75\}$.

The iid assumption

The CLT of Lindeberg-Levy requires that the **random variables are iid**.

- However, in many applications the assumption that the variables are iid is violated since we have variables which are correlated and/or have different distributions.
- Fortunately, there are various other CLTs, which do not need the iid condition. Instead, they place alternative conditions on the stochastic behavior of the random variables in the sequence $\{X_n\}$.

In Time Series Analysis, we allow for serial dependence. Below, we allow for heterogeneity.

Lindeberg's CLT

Theorem (5.14 (Lindeberg's CLT))

Let $\{X_n\}$ be a sequence of **independent** random variables with $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2 < \infty \forall i$. Define $b_n^2 = \sum_{i=1}^n \sigma_i^2$, $\bar{\sigma}_n^2 = n^{-1} \sum_{i=1}^n \sigma_i^2$, $\bar{\mu}_n = n^{-1} \sum_{i=1}^n \mu_i$, and let f_i be the pdf of X_i . If $\forall \varepsilon > 0$,

$$(\text{continuous case:}) \quad \lim_{n \rightarrow \infty} \frac{1}{b_n^2} \sum_{i=1}^n \int_{(x_i - \mu_i)^2 \geq \varepsilon b_n^2} (x_i - \mu_i)^2 f_i(x_i) dx_i = 0,$$

$$(\text{discrete case:}) \quad \lim_{n \rightarrow \infty} \frac{1}{b_n^2} \sum_{i=1}^n \sum_{\substack{(x_i - \mu_i)^2 \geq \varepsilon b_n^2 \\ f_i(x_i) > 0}} (x_i - \mu_i)^2 f_i(x_i) = 0,$$

then

$$\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}} = \frac{n^{1/2} (\bar{X}_n - \bar{\mu}_n)}{\bar{\sigma}_n} \xrightarrow{d} \mathcal{N}(0, 1).$$

Bounded random variables satisfy this (see Theorem 5.15 in lecture notes).

Multivariate Central Limit Theorems

The CLTs presented so far are applicable to sequences of **random scalars**.

In order to discuss CLTs for a sequence of **random vectors** a result of Cramér and Wold, termed the **Cramér-Wold device** is very useful.

The Cramér-Wold device allows to reduce the question of convergence in distribution for multivariate random vectors to the question of convergence in distribution for random scalars.

Thus it facilitates the use of CLTs for random scalars in order to obtain multivariate CLTs.

The CW device

Theorem (5.16 (Cramér-Wold Device))

The sequence of $(k \times 1)$ -dim. random vectors $\{\mathbf{X}_n\}$ converges in distribution to the $(k \times 1)$ -dim. random vector \mathbf{X} iff

$$\ell' \mathbf{X}_n \xrightarrow{d} \ell' \mathbf{X} \quad \forall \ell \in \mathbb{R}^k.$$

This really means “any linear combination” ...

Corollary (5.2)

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{iff} \quad \ell' \mathbf{X}_n \xrightarrow{d} \ell' \mathbf{X} \sim \mathcal{N}(\ell' \boldsymbol{\mu}, \ell' \boldsymbol{\Sigma} \ell).$$

... allows for multivariate CLTs

Theorem (5.17 (Multivariate Lindeberg-Lévy))

Let $\{\mathbf{X}_n\}$ be a sequence of iid $(k \times 1)$ random vectors with $E(\mathbf{X}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma} \forall i$, where $\boldsymbol{\Sigma}$ is a $(k \times k)$ positive definite matrix. Then

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i - \boldsymbol{\mu} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

It follows from the multivariate Lindeberg-Lévy CLT that
 $\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \xrightarrow{a} \mathcal{N}(\boldsymbol{\mu}, \frac{1}{n} \boldsymbol{\Sigma})$.

Outline

1 Weak Laws of Large Numbers

2 Central Limit Theorems

3 The delta method

4 Up next

A transformation

Take the exponential durations again, where we are interested in $\frac{1}{\bar{X}_n}$:

- Since $\bar{X}_n \xrightarrow{p} \frac{1}{\lambda}$, the CMT implies $\frac{1}{\bar{X}_n} \xrightarrow{p} \lambda$
- ... and we may *estimate* an unknown λ by means of $\frac{1}{\bar{X}_n}$.

Since \bar{X}_n is Gamma-distributed, $1/\bar{X}_n$ follows a so-called inverse Gamma distribution which is not just as tractable.

Recall however that $\bar{X}_n \stackrel{asy}{\sim} \mathcal{N}\left(\frac{1}{\lambda}, \frac{1}{n\lambda^2}\right)$.

- We could use that by working out the distribution of $1/\mathcal{N}(\mu, \sigma^2)$.
Again not very tractable.
- But if we linearized $1/x$ in a neighbourhood of $\frac{1}{\lambda}$...

Functions of asymptotically normally distributed RVs

Theorem (5.18 (The delta method))

Take $\{\mathbf{X}_n\} \in \mathbb{R}^k$ where $\sqrt{n}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Let $g(\mathbf{x})$ be a function that has first-order partial derivatives in a neighborhood of the point $\mathbf{x} = \boldsymbol{\mu}$ that are continuous at $\boldsymbol{\mu}$, and suppose the gradient vector of $g(\mathbf{x})$ evaluated at $\mathbf{x} = \boldsymbol{\mu}$,

$$\mathbf{g}_{(1 \times k)} = [\partial g(\boldsymbol{\mu})/\partial x_1 \dots \partial g(\boldsymbol{\mu})/\partial x_k],$$

is not the zero vector.^a Then

$$\sqrt{n}(g(\mathbf{X}_n) - g(\boldsymbol{\mu})) \xrightarrow{d} \mathcal{N}(0, \mathbf{g}\boldsymbol{\Sigma}\mathbf{g}')$$

and correspondingly $g(\mathbf{X}_n) \stackrel{asy}{\sim} \mathcal{N}(g(\boldsymbol{\mu}), n^{-1}\mathbf{g}\boldsymbol{\Sigma}\mathbf{g}')$.

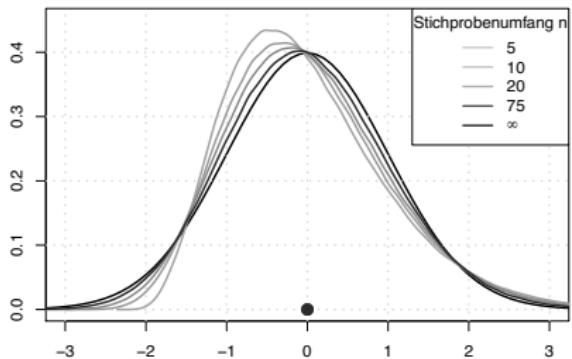
^aThis is the row version of the gradient.

Exponential durations take three

We have with $\mu = 1/\lambda$ that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \frac{1}{\lambda^2})$, and $g' = -1/x^2$.

So $\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\lambda^4}{\lambda^2}\right)$ and

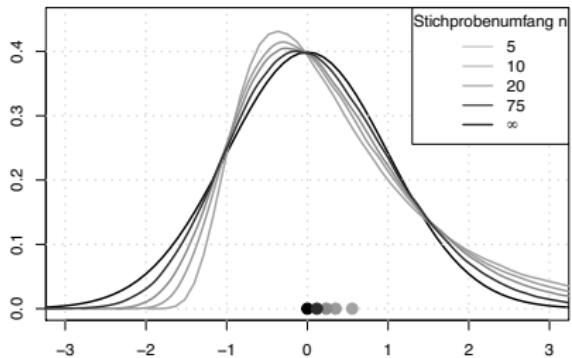
$$\frac{1}{\bar{X}_n} \stackrel{\text{approx}}{\sim} \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right).$$



Right: exact distributions, $\lambda = 1$.

Top: $\sqrt{n\lambda^2}(\bar{X}_n - 1/\lambda)$

Bottom: $\sqrt{n/\lambda^2}\left(\frac{1}{\bar{X}_n} - \lambda\right)$



The normal only approximates!

Some final remarks

The delta method is based on a first-order Taylor approximation.

- We note that we may work analogously with a vector function $\mathbf{g}(\mathbf{x})$ (the limit then involves the Jacobian of \mathbf{g}).
- If \mathbf{g} is smooth enough, a higher-order Taylor approximation can tell you what the approximation error looks like.
- A higher-order approximation also helps when $\partial\mathbf{g}/\partial\mathbf{x} = \mathbf{0}$ at $\mathbf{x} = \boldsymbol{\mu}$.
- (By the way, such a stationary point of \mathbf{g} simply means that $\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu}) = o_p(n^{-1/2})$.)
- And we can even work out approximations for moments of $\mathbf{g}(\mathbf{X}_n)$.

If \mathbf{g} is not smooth (or not continuous) at $\boldsymbol{\mu}$, then we need a case-by-case discussion, but these are fortunately rarely encountered.

Outline

1 Weak Laws of Large Numbers

2 Central Limit Theorems

3 The delta method

4 Up next

Coming up

Inferential Statistics (*fka* Adv Stat II)