Probability Theory
EUI September 2006

Tobias Broer

European University Institute

Examples

- Example 1: Betting on an African dice
- Example 2: Betting on your post-PhD salary
- Example 3: A simple Macroeconomic model of Optimal Growth

$$max_{c_t, k_{t+1}} E_0 \sum_{t=0}^{\infty} \beta^t U(c_t) \; s.t. \; k_{t+1} = z_t F(k_t) - c_t$$

Aim of the course

- Provide the necessary background for the compulsory courses.
- Enable students to master more advanced texts in probability and measure theory.
- More particulary, the course wants to
  - provide tools to formally describe and analyse random experiments using axiomatic probability theory.
  - introduce the concepts of discrete and continuous random variables and their distributions.
  - provide an overview of the main univariate and multivariate probability distribution functions.
  - show students how probability theory is a special case of the theory of measures.
  - provide enough possibilities for practice.

Probability Theory - Statistics - Econometrics

- **Probability Theory** analyses characteristics of probability mechanisms on the basis of a limited number of definitions and axioms.
- On the basis of data on trials and some maintained assumptions about a probability mechanism **Statistics** "estimates" its parameters, or assesses "hypotheses" about them.
- **Econometrics** applies statistics to assess the likelihood of economic models and theories.

Outline

1. Probability Spaces and axiomatic probability theory
2. Conditional probability and combinatorics
3. Univariate random Variables
4. Integral theory, mathematical expectations, and moments of RVs
5. Some common univariate distributions
6. Multivariate random variables

References

- *Goldberger (1991)*
- *Hogg and Craig (various editions)*
- Appendix to *Hansens's Lecture Notes*
- *Benoît Champagne's class notes* "Probability and random signals I"
- *Ivan Wilde's script* "Measure, Integration and Proability"
- *Spanos (1986)*

Section I Probability spaces and axiomatic probability theory

1.1 Random Experiment, Events, and Sigma-Algebras

Working definition of Probability

In a random situation, **probability** aims to attach to possible statements about the future a number that describes their likelihood to be true in a consistent manner.

Random Experiment Ξ

- A random experiment Ξ is a situation with different possible *outcomes* (follow-on situations), such that
  1. There is always exactly one outcome.
  2. All possible outcomes are known a priori.
  3. In a particular trial, the outcome is not known a priori.
  4. The situation is repeatable.
- A particular realisation of a random experiment, yielding a particular outcome, is called a **trial**.
  - **Example 1:** Repeated coin toss
  - **Example 2:** Rainfall in Florence in August

Sample Space S of a random experiment

The set of all possible outcomes of a random experiment Ξ is called the "Sample Space" S.

Elements of S are called outcomes or "elementary events".

- **Example 1:** Repeated coin toss:
  $S = \{\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}\}$.
- **Example 2:** Rainfall in Florence in August: S $= R^+$.

Event

An event is "any proposition associated with Ξ which may occur or not at each trial" (Spanos). Since every proposition describes a subset of S, we get more formally:

Definition: Any collection of outcomes, or subset of the sample space S, is an event, including sure (S) and impossible event ($\emptyset$). An event "occurs" if one of the outcomes it comprises occurs.

- **Example 1:** At least one head in 2 coin tosses.
- **Example 2:** Rainfall in Florence in August of more than 20 liters (per m^2).

Events and "derived" events

For two events A_1 and A_2, the following are also events:

- "not A_1", which is the complementary set of A_1 relative to S, or A_1^c.
- "A_1 and/or A_2", which is the set equal to the union $A_1 \cup A_2$.
- "A_1 and A_2", which is the set equal to the intersection $A_1 \cap A_2 = (A_1^c \cup A_2^c)^c$.
- "A_1 but not A_2", which is the set $A_1 \backslash A_2 = (A_1^c \cup A_2)^c$. (Note: We have reduced 4 set operations to 2.)

Consistent sets of events Ω

This implies that, for every event A_i in Ω

1. $A_i^c = S \setminus A_i$ must be in Ω.
2. "$A_1 \cup A_2 \cup ...$" must be in Ω for a sequence of events $\{A_i\} : A_i \in \Omega, \forall i$.
3. From 1 and 2, the event "$A_1 \cup A_1^c = S$" is the "sure event" equal to the set of all possible outcomes, so $S \in \Omega$.
4. From 1 and 3, the "impossible event" $S^c = \emptyset \in \Omega$.

Sigma-algebra $\Im$ of S

Definition: A family $\Im$ of subsets of any set S is called a "Sigma-algebra" of S, if

1. For every $A \in \Im$, $A^c = \{s \in S : s \notin A\} \in \Im$. ($\Im$ is closed under complementation.)
2. For every sequence of $A_i \in \Im$, i = 1,2,..., $\bigcup_{i=1}^{\infty} A_i \in \Im$. ($\Im$ is closed under countable union.)

   This implies:
3. $S \in \Im$ (since $A \cup A^c \cup \emptyset \cup \emptyset \cup ... = S \in \Im$)
4. $\emptyset \in \Im$ (since $S^c = \emptyset \in \Im$)
5. $(\bigcup_{i=1}^{\infty} A_i)^c = \bigcap_{i=1}^{\infty} A_i^c \in \Im$

Sigma-algebra $\Im$ of S

- The pair (S,$\Im$) is called a **"measurable space"**.
- A consistent set of events Ω is a "Sigma-algebra" of a sample space S.

Power set $\mathbb{P}(S)$

For a set S with a finite, or countably infinite, number of elements, we can use as Sigma-Algebra the "**power set**" of S, written $\mathbb{P}(S)$.

Definition: The set of all subsets of S is called the power set of S.

- **Example 1:** Toss coin once, i.e. $S = \{H, T\}$
  1. One Sigma-algebra is the power set $\mathbb{P} = \{\emptyset, S, \{H\}, \{T\}\}$
  2. But $\Omega_1 = \{\emptyset, S\}$ is also a Sigma-algebra for S, called the "trivial Sigma-algebra".
- **Example 2:** Toss coin twice, i.e.

  $S = \{\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}\}$

  $\mathbb{P} = \{\emptyset, S, \{(H, T)\}, \{(H, H)\}, \{(T, T)\}, \{(T, H)\},$

  $\{(H, T), (H, H)\}, \{(T, T), (H, H)\}, \{(T, H), (H, H)\}, \{(H, T), (T, H)\},$

  $\{(T, T), (T, H)\}, \{(H, T), (T, T)\},$

  $\{(H, T), (T, H), (T, T)\}, \{(H, T), (T, H), (H, H)\}, \{(H, H), (T, T), (T, H$

Sigma algebra generated by family of subsets $C \subseteq \mathbb{P}(S)$

For C a collection of subsets of S, the smallest Sigma-algebra of S containing C (more accurately the intersection of all Sigma algebras containing C, $\Im(C) \doteq \bigcap_{\Im \supseteq C} \Im$), is called "Sigma-algebra generated by C".

- **Example 1:** Consider $S = \{(H,H),(H,T),(T,H),(T,T)\}$, the repeated coin toss. The Sigma-Algebra generated by $C = \{\{(HH),(TT)\}\}$ is
  $\Im(C) = \{\ \emptyset, S, \{(H,T),(T,H)\}, \{(H,H),(T,T)\}\ \}$

Borel algebra

For $S = \mathbb{R}^n$ the n dimensional Euclidian Space, the Borel Algebra $\mathbb{B}^n$ is defined as the Sigma-algebra generated by the open sets in $\mathbb{R}^n$, or smallest Sigma-algebra containing all open balls in $\mathbb{R}^n$. Moreover, any $B \in \mathbb{B}^n$ is a "Borel set".

Borel algebra

Proposition 1.1: $\mathbb{B}$, the Borel Algebra for the one-dimensional Euclidian space contains

- all open intervals $(-\infty, b)$,(a, ∞),(a, b),$(-\infty, \infty)$ (by definition of $\mathbb{B}$)
- all closed and half-closed intervals $(-\infty, b]$,$[a, \infty)$,$[a, b]$, *etc.* (by complementation and intersection of open sets)
- $\mathbb{R}$ (by countably infinite union of open sets)
- $\emptyset$ (by complementation of $\mathbb{R}$)

Borel algebra

Proposition 1.2: $\mathbb{B}(\mathbb{R})$ can be generated by any of the families C_i of subsets of $\mathbb{R}$ defined by the following intervals, where $a, b \in \mathbb{R},\ a < b$:

1. $(a, b),\ a < b$
2. $(-\infty, a)$
3. (a, ∞)
4. $[a, b],\ a \leq b$
5. $(-\infty, a]$
6. $[a, \infty)$
7. $(a, b],\ a < b$
8. $[a, b),\ a < b$
9. *any closed subset of* $\mathbb{R}$

"It is a deep and difficult result of measure theory that the Borel field of the real line is in fact different from the power set of the real line." (Gray and Davisson 2004)

Roadmap

- So far,we
  - defined events as subsets of a sample space S
  - defined consistent sets of events as Sigma-algebras Ω of S (families of subsets that are closed under complementation and countable union)
  - showed that often we can use the power set of S as Sigma-algebra
  - but defined a smaller Sigma-algebra for the real line, the Borel-Algebra $\mathbb{B}(\mathbb{R})$
- Now:
  - need to find a definition for "Probabiliy of Event A"
  - show how this is a special case of a "measure" on the Sigma-algebra of a sample space

Section I Probability spaces and axiomatic probability theory

I.2 Probability functions as measures

Working definition of Probability

In a random situation, **probability** aims to attach to possible statements about the future a number that describes their likelihood to be true in a consistent manner.

Working definition of Probability

Moreover, probability should at least have the following features

1. It is defined for all elements in a consistent set of events/ in a Sigma-algebra of a sample space S.
2. The probability of every event is greater or equal to 0 and less or equal to 1.
3. For any two events A and B that do not share any outcomes (rainfall tomorrow vs. a dry day), we want the probability of either of the two occuring to be the sum of their individual probabilities.
4. The probability that anything occurs, or the probability of the sample space S is 1 (i.e. there is always some outcome).

Axiomatic Definition of Probability

"Probability of an event A" is a set function $P(\cdot) : \Omega \longrightarrow [0, 1]$ s.t.

- $P(A) \geq 0, \ \forall A \in \Omega$
- $P(S) = 1$
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for all sequences of disjoint events $\{A_i\}$, i.e. $A_i \cap A_j = \emptyset, \forall i \neq j$

  Note:

  By the definition of the probability set function we can write the probability of any event as the sum of the probabilities of its elementary events, as these are by definition disjoint.

Measure μ on $\Im$

Definition: For a given measurable space $(S, \Im)$, a measure $\mu(\cdot)$ is a function $\mu : \Im \longrightarrow \mathbb{R}^+ \cup \infty$, s.t. if $\{A_n\}_{n=1}^{\infty}$ is a countable, disjoint sequence of subsets in $\Im$, then $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ ("countable additivity" or "σ additivity").

- The Triple $(S, \Im, \mu)$ is called a "**measure space**".
- If $\mu(A)$ is finite for all $A \in \Im$, then μ is called a "**finite measure**".
- If $\mu(S) = 1$, then $\mu(\cdot)$ is called a "**probability measure**", and $(S, \Im, \mu)$ is called a "**probability space**".

Examples

- **Example 1:** "Number of students" on (EUI students, $\mathbb{P}()$)
- **Example 2:** "Weight" on (Stones on the beach, $\mathbb{P}()$).

Another example: "Length" L defined on $\mathbb{B}(\mathbb{R})$ as

- $L(A) = a - b$ for all open and closed intervals in $\mathbb{R}$, i.e. $A = (a, b), [a, b]$, etc., with $a \geq b$
- $\mathrm{L}(\mathrm{A}) = \infty$ for $A = (\infty, a), [a, \infty)$, etc.
- $L(\emptyset) = 0$
- $L(\bigcup_{i=1}^{N}(a_i, b_i)) = \sum_{i=1}^{N}(b_i - a_i)$, for all disjoint intervals (a_i, b_i)

  is a measure on $\mathbb{B}(\mathbb{R})$.

  Similarly, area, volume, etc. are measures on higher dimensional Borel sets.

Proposition: Properties of finite measures

1. $\mu(\emptyset) = 0$ (by noting that the union of empty sets is empty, i.e. $\emptyset \cup \emptyset \cup ... = \emptyset$, so $\mu(\emptyset \cup \emptyset \cup ...) = \sum_{i=1}^{\infty} \mu(\emptyset) = \mu(\emptyset)$, which only holds if $\mu(\emptyset) = 0$)
2. $\mu(A) \geq 0, \ \forall A \in \Im$ (by the definiton of the range of μ).
3. $\mu(A_1 \cup A_2 \cup .... \cup A_n) = \sum_{i=1}^{n} \mu(A_i)$ for any finite sequence of disjoint subsets $\{A_i\}$ of $\Im$ (by setting $A_{n+1}, A_{n+2}, ... = \emptyset$)
4. $\mu(A) \leq \mu(B)$ *if* $A \subseteq B$ *and* $A, B \in \Im$ (by noting that $\mu(B) = \mu(A \cup B \setminus A) = \mu(A) + \mu(B \setminus A)$ *and* $\mu(B \setminus A) \geq 0$)

Measures on countable measure spaces

For a countable set $S = \{s_1, s_2, ..., s_n\}$ (i.e. a set with a finite or countably infinite number of elements), we can define a finite measure on its power set $\mathbb{P}(S)$ using *any* sequence of non-negative numbers $\{p_i\}_{i=1}^{n}$ with $\sum_i p_i$ finite, as
$\mu(A) = \sum_{i \in I_A} p_i$, *for* $A \in \mathbb{P}(S)$ *and* $I_A = \{i : s_i \in A\}$.

Properties of probability functions

...follow immediately from those of finite measures:

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$
4. $P(B \cap A^c) = P(B) - P(B \cap A)$
5. For $A_1 \subseteq A_2$, $A_1, A_2 \in \Omega$, $P(A_1) \leq P(A_2)$
6. $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
7. If $\{A_i\}_{i=1}^{N}$ is a monotone sequence of events in Ω, then $P(lim(\{A_i\})) = lim(P(A_i))$.

Roadmap

So far,we looked at the axiomatic definitions of ...

- the **sample space** of a RE as the set of all outcomes S
- **events** A_i as subsets of S, and showed how we can derive events "not A", "$A_1 or A_2$", etc. by 2 set operations
- consistent sets of events as **Sigma-algebras** Ω of S
- the **Borel-algebra** as the Sigma-Algebra of $\mathbb{R}$ "generated by" the open sets, or equivalently the open intervals (a, b), half-closed intervals $(-\infty, a]$, etc.
- **probability** as a set function $P() : \Omega \longrightarrow [0, 1]$, with 3 properties: $P \geq 0, P(S) = 1, P(\bigcup_i A_i) = \sum_i A_i$ for disjoint $\{A_i\}$
- showed that P is just a special case of a σ-additive mapping from a Sigma-algebra to $R^+ \cup \infty$, called "measure".
- Defined $(S, \Omega, P(\cdot))$ as a "Probability Space"

Roadmap

Today we will look at:

- Conditional probability: What is the probability of "2 Heads", once one head has been observed?
- Combinatorics: if all outcomes are equally likely, $P(A)$=no. of outcomes in A/no. of all outcomes. How can we count outcomes in A, or S?
- Random Variables:
  - Can we define probabilities of numbers attached to outcomes by some function X?
  - What functions $X : S \longrightarrow \mathbb{R}$ preserve the probability and event structure?
  - Can we summarise the probabilities of X conveniently?

Section II Conditional Probability, Independence and Combinatorics

II.1 Conditional Probability

Example

You ring the bell of a house where a couple lives with their two children.

1. What is the probability that a boy opens the door?
2. The door opens, and a boy says hello to you. What is the probability that the other child is also a boy?

Definition of Conditional Probability

If for a probability space $(S, \Im, P())$ $A, B \in \Im$ and $P(B) \geq 0$, the "conditional probability of event A, given event B", is defined as $P(A|B) = P(A \cap B)/P(B)$.

This implies the **"Law of Multiplication"**
$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$.

Proposition 2.1

$P(\cdot|B)$ is a probability set function where we replace S with B, and $\Im$ with $\Im_B$, the Sigma-algebra generated by $\{A_i \cap B : \ A_i \in \Im\}$. That is:

1. $P(A|B) \geq 0$ for any $A \in \Im$
2. $P(B|B) = 1$
3. $P(\bigcup_i A_i|B) = \sum_i P(A_i|B)$ for any sequence of disjoint events $A_i \in \Im$

Proposition 2.1: Proof

For all $A \in \mathfrak{S}$

1. $P(A \cap B), P(B) \geq 0$, so $P(A \cap B)/P(B) \geq 0$.
2. $P(B \cap B) = P(B)$, so $P(B \cap B)/P(B) = 1$.
3. For any sequence of disjoint events $A_1, A_2, ...$, $(A_1 \cap B)$, $(A_2 \cap B)$,... are also disjoint events. So $P(A_1 \cap B \cup A_2 \cap B) = P(\ (A_1 \cap B)\ ) + P(\ (A_2 \cap B)\ )$ from the definition of probability. The rest follows immediately.

Law of total probability

For $\{C_i\}_{i=1}^N$ a partition of the sample space, i.e. $C_i \cap C_j = \emptyset, \forall\, j \neq i$ and $\bigcup_{i=1}^N C_i = S$, the probability of any event C is given by

$P(C) = \sum_{i=1}^N P(C|C_i) \cdot P(C_i)$.

Note: This implies

1. If C has a partition $\{C_i\},\ i = 1, ..., k$ (for example k elementary events) with probabilities $P(C_i)$, $P(C) = \sum_{i=1}^K 1 \cdot P(C_i)$.
2. If C_is are "equi-likely" elementary events with probability p=1/N, $P(C) = \sum_{i=1}^K 1 \cdot p = K/N$.

Bayes' Rule

For any set B and any partition $A_1, A_2, \ldots$ of the sample space S,
$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j) \cdot P(A_j)}$.

Example (Champagne 2003)

An urn contains 10 white balls and 5 black balls. We draw two balls from the urn at random, without replacement. Given the second ball is white, what is the probability that the first one was also white?

Solution

Define events

1. W_1="First ball is white"
2. B_1="First ball is black"
3. W_2="Second ball is white"

and use Bayes' law noting that W_1 and B_1 partition the sample space. Thus

$$P(W_1|W_2) = \frac{P(W_2|W_1)P(W_1)}{P(W_2|W_1)P(W_1) + P(W_2|B_1)P(B_1)} = \frac{9/14 * 2/3}{9/14 * 2/3 + 10/1}$$

Section II Conditional Probability, Independence and Combinatorics theory

II.2 Independence

Definition of independence

Two events A and B are "statistically independent", if $P(A \cap B) = P(A)P(B)$, implying $P(A|B) = P(A)$.
A collection of events $A_1, ..., A_n$ is "mutually independent" when for any of its subsets $P(\bigcap_{i \in I_A} A_i) = \prod_{i \in I_A} P(A_i)$ for $I_A \subseteq \{1, ..., n\}$.

- **Example:** The probability to get (H) in second toss is 0.5, no matter the outcome of the first toss.

Section II Conditional Probability, Independence and Combinatorics theory

II.2 Independence

Independence

If A and B are two independent events, then so are A and B^c, A^c and B, A^c and B^c.

Section II Conditional Probability, Independence and Combinatorics theory

II.3 Combinatorics

Probability of events consisting of K out of N equilikely elementary events

$$P(C) = \frac{\textit{Number of elementary outcomes where C occurs}}{\textit{Number of all elementary outcomes N}} = \frac{K}{N}.$$

So need rules to determine K and N.

Example

What is the probability of 2 aces when drawing twice from a deck of 52 cards?

1. $N = \frac{\textit{No. of possible first cards drawn} \times \textit{No. of possible second cards}}{\textit{No. of undistinguishable pairs}} = \frac{52 \times 51}{2} = 2652$
2. $K = \frac{\textit{No. of distinguishable pairs of aces}}{=} \frac{4 \times 3}{2} = 6$
3. $P = 6/1326 < 0.5\%$

Counting rules

More generally, when counting the number of **selections** from a set (draws from the balls in an urn), we need to make 2 distinctions:

1. ordered (i.e. it matters at which draw we get a particular ball) vs. unordered
2. with vs. without replacement

Counting rules

When choosing r objects from a set A of n objects, we get the following numbers of selections

- Ordered, with replacement: n^r
- Ordered, without replacement (also called "r-element permutation of A"): P(n,r)$=n!/(n-r)!$
- Unordered, without replacement (also called "r-element combination of A"): C(n,r)$=n!/(r! * (n-r)!) =$ "*noverr*"
- Unordered, with replacement: $(n+r-1)!/(r! * (n-1)!)$

Exercises

1. What is the probability of winning the first prize of a lottery of 6 from 49 numbers, i.e. getting all 6 numbers right, where the order of the draws does not matter?
2. What is the probability of winning the second prize of a lottery of 6 from 49 numbers, i.e. of getting all but 1 number correct?

Roadmap

- So far,we defined the ingredients of $(S, \Omega, P(\cdot)$, a "Probability Space"
- But: Often we are more interested in numbers attached to outcomes (e.g. salary). How do we define their probability?

- Can we summarise the probabilities of values of X by a simpler function than P_X on $\mathbb{B}$?

Roadmap

- So far,we defined the ingredients of $(S, \Omega, P(\cdot)$, a "Probability Space"
- But: Often we are more interested in numbers attached to outcomes (e.g. salary). How do we define their probability?

- Can we summarise the probabilities of values of X by a simpler function than P_X on $\mathbb{B}$?

Roadmap

- So far,we defined the ingredients of $(S, \Omega, P(\cdot)$, a "Probability Space"
- But: Often we are more interested in numbers attached to outcomes (e.g. salary). How do we define their probability?

- Can we summarise the probabilities of values of X by a simpler function than P_X on $\mathbb{B}$?

Roadmap

- So far,we defined the ingredients of $(S, \Omega, P(\cdot)$, a "Probability Space"
- But: Often we are more interested in numbers attached to outcomes (e.g. salary). How do we define their probability?
- Also, S is an arbitrary set. Often we need to tabulate outcomes, events, probabilities. Can we transform S into more standard $\mathbb{R}$ but keep the probability characteristics of $(S, \Omega, P(\cdot)$?
- Can we summarise the probabilities of values of X by a simpler function than P_X on $\mathbb{B}$?

Section III Univariate Random Variables

III.1 Definition

Random Variables: Intuition

- Often more interest in numbers attached to outcomes (e.g. salary)
- Suppose we can assign a real number to every elementary event in S, e.g. "the number of heads in n coin tosses". Call this mapping X. X transforms non-standard S into $\mathbb{R}$.
- But: To be useful, X has to preserve the event structure of the measurable space (S, Ω), i.e.
  - there must be an event E in Ω that corresponds to any subset M of the range of X
  - to their unions, intersections and complements must correspond the unions, etc. of the corresponding events in Ω
- Then the probability of any $M \subseteq \mathbb{R}$ is simply the probability of the corresponding event $E \in \Omega$.

Definition: Random Variable

Given $(S, \Omega, P(\cdot))$, a RV is a function $X : S \longrightarrow \mathbb{R}$, which satisfies the condition that for every half-closed interval
$I_x = (-\infty, x], x \in \mathbb{R}$, the inverse image
$X^{-1}(I_x) \doteq \{s \in S : X(s) \leq x\}$ is an event in Ω.

In (other) words, every half-closed interval $I_x = (-\infty, x], x \in \mathbb{R}$ has a corresponding subset of S in Ω, given by the set of elements in S that X maps into I_x.

- **Note**: To check that X is a valid random variable, we have to show that for every half-closed interval I in $\mathbb{R}$, there is an event in Ω the elements of which X maps into I.

Points to note:

- $X^{-1}((-\infty, x]) \in \Omega,\ \forall x \in \mathbb{R} \iff X^{-1}(B) \in \Omega,\ \forall B \in \mathbb{B}$.
- A RV $X : S \longrightarrow \mathbb{R}$ is always defined with respect to some Sigma-Algebra Ω of S, so "RV X on (S, Ω)".
- Distinguish the random variable X from x, the value it takes in a particular trial of a random experiment .
- To decide whether $X(\cdot) : S \longrightarrow \mathbb{R}$ is a random variable, one needs to proceed from half-closed intervals in $\mathbb{R}$ to the elements of Ω, the Sigma-Algebra of S, not the other way.
- A random variable is a real-valued function on a sample space with certain properties. It is neither "random", nor "variable".

Example:

Is $X \doteq$ *Number of Heads*, a RV for $(\{H, T\}, \mathbb{P}(S))$?

- X(H) = 1; X(T)=0
- So

  $X^{-1}(1) = H;\ X^{-1}(0) = T;\ X^{-1}(a) = \emptyset,$ *for all* $a \notin \{1, 0\}$.
- So for every $B = (-\infty, a]$ we have
  - If $a < 0,\ X^{-1}(B) = \emptyset \in \Omega$.
  - If $0 \leq a < 1,\ X^{-1}(B) = \{T\} \in \Omega$.
  - If $1 \leq a,\ X^{-1}(B) = \{T, H\} \in \Omega$.
- So X is a random variable.

Exercise:

Consider $S = \{(H,H),(H,T),(T,H),(T,T)\}$, the repeated coin toss, and the two Sigma-algebras

1. $\Omega = \{\ \emptyset, S, \{(H,H)\}, \{(T,T)\}, \{(H,H),(T,H),(H,T)\}, \{(T,H),(H,T),(T,T)\}, \{(H,H),(T,T)\}\{(H,T),(T,H)\}\ \}$
2. $\Omega = \{\ \emptyset, S, \{(H,T),(H,H)\}, \{(T,H),(T,T)\}\ \}$

- Take the Random Variable X = "number of heads". Write down the sets $\{s \in \Omega : X(s) = a\}$ where $a \in \{0,1,2\}$.
- Now consider the half-open intervals defined by $(-\infty, a]$, $a \in \mathbb{R}$. Write down the sets $\{s \in \Omega : X(s) \leq a\}$ where $a \in \{-1,0,1,2,3\}$.
- Using this, show that X is a random variable with respect to the first but not the second Sigma-Algebra.

Exercise continued:

Consider $S = \{(H,H),(H,T),(T,H),(T,T)\}$, the repeated coin toss, and the two Sigma-algebras

1. $\Omega = \{\ \emptyset, S, \{(H,H)\}, \{(T,T)\}, \{(H,H),(T,H),(H,T)\},$ $\{(T,H),(H,T),(T,T)\}, \{(H,H),(T,T)\}\{(H,T),(T,H)\}\ \}$
2. $\Omega = \{\ \emptyset, S, \{(H,T),(H,H)\}, \{(T,H),(T,T)\}\ \}$

1. Consider the random variable Y defined by $Y(\{H,H\}) = Y(\{H,T\}) = 1, Y(\{T,T\}) = Y(\{T,H\}) = 0.$ With respect to which of the two Sigma-Algebras is Y a random variable?

Measurable functions

Definition: Given a measurable space $(S, \Im)$ a real-valued function $g(\cdot) : S \longrightarrow \mathbb{R}$ is "Borel measurable with respect to $\Im$" if for all open sets $A \subseteq \mathbb{R}$ the sets $g^{-1}(A) = \{s \in S : g(s) \in A\}$ are in $\Im$.

- Let C_i be a collection of subsets of $\mathbb{R}$ s.t. $\Im(C_i) = \mathbb{B}$. Then $X^{-1}(C) \in \Im,\ \forall C \in C_i \Longleftrightarrow X^{-1}(B) \in \Omega,\ \forall B \in \mathbb{B}$.
- So for $(S, \Im)$ a measurable function $g(\cdot) : S \longrightarrow \mathbb{R}$ has an inverse image in $\Im$ for **all** Borel sets.
- Thus a random variable is simply a measurable function on a probability space.

Exercise:

For $S = \{0, 1\}$, consider $\Im_1 = \{\emptyset, \{1\}, \{0\}, S\}$ and $\Im_2 = \{\emptyset, S\}$. Show that all functions on S are $\Im_1$-measurable, but only constant functions are $\Im_2$-measurable.

Proposition: Properties of measurable functions

- All monotone functions $g(\cdot) : (a, b) \longrightarrow \mathbb{R}$ are measurable on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$.
- All continuous functions $g(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}$, are measurable on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$.
- For Borel-measurable functions $f : S \longrightarrow \mathbb{R}$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ the composite function $g \circ f : S \longrightarrow \mathbb{R}$ is Borel-measurable on $(S, \Im)$.

Proposition: Properties of measurable functions continued

- For $(S, \Im)$ a measurable space and $f(\cdot) : S \longrightarrow \mathbb{R}$,$g(\cdot) : S \longrightarrow \mathbb{R}$ two Borel functions, the following are Borel functions
  1. $af + b,\ a, b \in \mathbb{R}$
  2. $f + g$
  3. $|f|^{\alpha},\ \forall \alpha \geq 0$
  4. if f never vanishes, 1/f
  5. f*g
  6. max{f,g}, min{f,g}

Proof: see script and Wilde, p. 6-8

Properties of random variables

Again, properties of measurable functions translate of course to those of random variables, i.e.

- A measurable function $g : \mathbb{R} \longrightarrow \mathbb{R}$ of a random variable $X : S \longrightarrow \mathbb{R}$ is itself a random variable.
- The sum of n random variables is a random variable, so is their mean.
- etc.

Roadmap: So far ...

- Defined the elements of the probability space $(S, \Omega, P(\cdot))$.
- showed that P is just a special case of a Sigma-additive mapping $\mu : \Im \longrightarrow \mathbb{R}^+ \cup \infty$, called "measure".
- Defined a RV as a mapping $X : S \longrightarrow \mathbb{R}$, such that we can assign to all half-closed intervals I_x in $\mathbb{R}$ events in Ω.
- ... or formally: X is a RV if $X^{-1}(I_x) = \{s \in S : X(s) \in (-\infty, x]\} \in \Omega$, for all $I_x \subseteq \mathbb{R}$
- Stated without proof that this implies that the inverse image $X^{-1}(B)$ is an event in Ω for all Borel sets in $\mathbb{R}$.
- Showed that RVs are simply measurable functions on a probability space.

Roadmap: So far ...

- Showed that monotone and continuous functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ are measurable w.r.t. the Borel-sets, and that sums, linear transformations and certain composites of measurable functions are measurable.

Roadmap: Now ...

- Define a probability measure P_X on the Borel sets for a given random variable X on a given probability space .
- Summarise these probabilities conveniently using distribution functions.

Section III Univariate Random Variables

III.2 Distribution functions of random variables

Defining probability measures for random variables

- We know: Given a probability space $((S, \Omega, P_\Omega())$, a random variable $X : S \longrightarrow \mathbb{R}$ associates to every Borel set B, an Event in Ω.
- This establishes a new probability measure $P_X : \mathbb{B}(\mathbb{R}) \longrightarrow [0, 1]$ on the Borel sets, with $P_X(B) = P_\Omega(X^{-1}(B)) = P_\Omega(\{s \in S : X(s) \in B\})$ for all $B \in \mathbb{B}(\mathbb{R})$.
- So a RV transforms a probability space (S, Ω, P_Ω) into a new one $(\mathbb{R}, \mathbb{B}, P_X)$.

Summarising probability measures for random variables by their cumulative distribution

- We want: A way to summarise $P_X : \mathbb{B}(\mathbb{R}) \longrightarrow [0,1]$. If possible a point function $F : \mathbb{R} \longrightarrow [0,1]$.
- Try as summary the probabilities of X to fall in half-closed intervals $\mathbb{C} = \{(-\infty, x] : x \in \mathbb{R}\}$, i.e. $P_X(\ (-\infty, x]\ ), \forall x \in \mathbb{R}$, a **set function** on the half-cosed intervals.
- But: Every set $\{(-\infty, x] : x \in \mathbb{R}\}$ is completely characterised by its upper bound x. So we get a point function $F_X : \mathbb{R} \longrightarrow [0,1]$ with $F_X(x) = P_X(\ (-\infty, x]\ ) = Prob(X \leq x)$ the "Cumulative distribution function of X".

Cumulative Distribution function (CDF)

The function $F_X(x) \doteq Prob(X \leq x)$ is called the "Cumulative Distribution function of random variable X" or simply its "Distribution function".

Proposition 4.1: A function $F(\cdot)$ is a CDF **if and only if** it satisfies the following three properties:

1. $lim_{x \longrightarrow -\infty} = 0$, $lim_{x \longrightarrow \infty} = 1$
2. F(x) is nondecreasing in x
3. F(x) is right-continuous

Cumulative Distribution function and measures

- The measure $P_X(B) = P_\Omega(X^{-1}(B))$ on the Borel-sets assigns probabilities to all half-closed intervals, so we can define the CDF on the basis of P_X. But does the reverse hold?
- $F_X : \mathbb{R} \longrightarrow [0, 1]$, a measure on the half-closed intervals, uniquely characterises a measure $v(\cdot)$ on **ALL** Borel sets, called the "Lebesgue-measure on $(\mathbb{R}, \mathbb{B})$ given by $F(\cdot)$". This is a result from measure theory called "Carathéodory's extension theorem".
- **Corollary:** Whenever we know $F(\cdot)$, it gives us all the information we need about probabilities of values of X in $\mathbb{R}$.

Discrete random variables and their distribution

A random variable X is called "discrete" if its CDF is a step function.

- For X a discrete RV, the "**probability mass function (PMF)**" of X is $p_X(x) = P(X = x)$, with
  1. $p_X(x) > 0$ for all $\{x \in \mathbb{R} : F(x^-) - F(x) > 0\}$
  2. $p_X(x) = 0$ otherwise, and
  3. $\sum_{x \in \mathbb{R}} p_X(x) = 1$.

Discrete random variables and their distribution

Note:

- Any random variable on a discrete (i.e. finite or countably infinite) sample space is discrete.
- More generally, any random variable with discrete range is discrete.
- The "support of X" D is defined for a discrete random variable as the set of "jumps", i.e. the points with positive mass, i.e. $D = \{x \in \mathbb{R} : P_X(x) > 0\}$.

Example The random variable "number of heads in the repeated coin toss" with the sigma algebra the Power set.

Continuous random variables and their distribution

A random variable X is called "continuous" if its CDF is continuous for all $x \in \mathbb{R}$. In this case, P(X=x)=0, for all $x \in \mathbb{R}$.

- $F(\cdot)$ is "absolutely continuous" if there is a "**probability density function**" (pdf) $f_X(x)$ that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(x)dx$, such that
  - $f(x) = \frac{d}{dx} F_X(x)$, and
  - $Prob(a \leq x \leq b) = \int_a^b f_X(x)dx$

Properties of a pdf

- A function $f_X(x)$ is a pdf if and only if
  1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$
  2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- The "support of X" D is $D = \{x \in \mathbb{R} : f_X(x) > 0\}$.

Example: Uniform Distribution on [0,2]

- The uniform distribution with support $[a, b]$ is defined as one that has a constant pdf $f(x) = c$, for all $x \in [a, b]$, $f(x) = 0$ otherwise.
- Find the constant c for a=0,b=2.
- Find its CDF.
- What is $P(x \leq 1.5)$?

Mixed random variables

If a random variable is neither discrete, nor continuous, it is called a "mixed random variable".

Summarizing distribution functions: measures of location and spread

- Three numbers to summarise the location of $F(\cdot)$ on the real line are
  - Its expectation, or mean, E(X) (see next section)
  - The "middle observation" or "median" m_x defined by $\int_{-\infty}^{m_x} f_X(x)dx = 1/2$ for continuous variables
  - The "mode" defined as the most frequent observation, or value with highest density
- The spread around its location can be summarised e.g. by the variance, or standard deviation.

Distributions of functions of RVs and the change of variables formula

- Suppose we know the pdf of a continuous RV $X : S \longrightarrow \mathbb{R}$, but want the pdf of $Y = g(x)$, with $g(\cdot)$ measurable.
- A measurable function of a RV is a RV, so has a distribution.
- But $F_Y(y) = P(\{s \in S : g(X(s)) \leq y\})$ can be very complicated.
- So can we get $F_Y(\cdot) directly from F_X(\cdot)$?
- Yes, if g one-to-one, differentiable and invertible.
- Distinguish 2 cases: g increasing vs. decreasing.

Change of variables formula Case 1: g(X) increasing

1. Find F(y):
   - $g(x) \leq y \Leftrightarrow x \leq h(y)$, where $h(\cdot) = g^{-1}$.
   - So $F_Y(y) = Prob(Y \leq y) = Prob(g(X) \leq y) = Prob(X \leq h(y)) = F_X(h(y))$
2. Find f(y): Differentiating both sides w.r.t. y and applying Leibniz rule, we get
   $f(y) = \frac{d}{dy} F_X(h(y)) = \frac{d}{dy} \int_{-\infty}^{h(y)} f_X(t))dt = f_X(h(y)) \frac{d}{dy} h(y)$.

Change of variables formula Case 2: g(X) decreasing

1. $g(X) \leq y \Leftrightarrow X \geq h(y)$. So $F_Y(y) = Prob(Y \leq y) = Prob(g(X) \geq y) = 1 - Prob(X \leq h(y)) = 1 - F_X(h(y))$
2. So $f(y) = -\frac{d}{dy} F_X(h(y)) = -f_X(h(y)) \frac{d}{dy} h(y)$
3. Combine both cases:

   Given that in the first case $\frac{d}{dy} h(y)$ is positive, in the second negative, we can write $f(y) = f_X(h(y)) |\frac{d}{dy} h(y)|$.

Change of variables formula - step-by-step

- Step 1: Write y=g(x), and get $x = g^{-1}(y) = h(y)$.
- Step 2: Get $\frac{d}{dy}h(y)$.
- Step 3: Calculate $f_Y(y) = f_X((h(y))|\frac{d}{dy}h(y)|$.
- Step 4: Transform the domain of X $D_X = \{x : x \in A\}$ into that of y by writing $D_Y = \{y : h(y) \in A\}$.
- Step 5: Check that $\int_{D_Y} f_Y(t)dt = 1$.

Change of variables formula - Example 1

Suppose X is distributed uniformly on $[0, 1]$, and define $Y = 2 - 3x$. What is the pdf of Y?

- Step 1: Write y=g(x), and get $x = g^{-1}(y) = h(y)$.
- Step 2: Get $\frac{d}{dy}h(y)$.
- Step 3: Calculate $f_Y(y) = f_X((h(y))|\frac{d}{dy}h(y)|$.
- Step 4: Transform the domain of X $D_X = \{x : x \in A\}$ into that of y by writing $D_Y = \{y : h(y) \in A\}$.
- Step 5: Check that $\int_{D_Y} f_Y(t)dt = 1$.

Change of variables formula - Example 2

Suppose X is distributed uniformly on $[0, 1]$, and define $Y(x) = -ln(x)$. What is the pdf of Y?

- Then we have $h(y) = e^{-y}$, and $| \frac{\delta}{\delta y}(h(y)) | = e^{-y}$.
- $f_X(x) = 1$ for $x \in [0, 1]$, so $f_X(h(y)) = 1$ for $y \in [0, \infty]$
- Thus $f_Y(y) = e^{-y}$, $y \in [0, \infty]$, an exponential density.
- Y maps values in $D_X = [0, 1]$ into $D_Y = [0, \infty]$.

Roadmap

- So far,we
  - defined the probability set function on a probability space $(S, \Omega, P()$ as $P(\cdot) : \Omega \longrightarrow [0, 1]$
  - Defined a RV as a function $X : S \longrightarrow \mathbb{R}$ with inverse images in Ω for all half-closed intervals (and thus all Borel-sets)
  - Summarised the probabilities of the Borel sets by the Distribution function X
    $F_X(x) = Prb(X \leq x) = P(\{s \in S : X(s) \in (-\infty, x]\})$
  - Saw that for every random variable on $(S, \Omega, P())$ as $P(\cdot) : \Omega \longrightarrow [0, 1]$ there is a unique measure $v(\cdot)$ on $(\mathbb{R}, \mathbb{B},$ where v() is the Lebesgue-Stieltjes measure with respect to X's CDF $F(\cdot)$

Roadmap

Now we want to characterise RVs further:

- Calculate the "mean" of a random variable, that weighs values of X in $\mathbb{R}$ by their measure μ.
- Calculate other "moments", weighted averages of functions of RVs.
- So need "weighting scheme" to weigh (functions of) values of a RV $X : S \longrightarrow \mathbb{R}$ by the probability measure P.
- The integral does this.

Integration theory

- ...
- (See script for a very basic introduction.)

Section V Integration theory, mathematical expectation, and moments of random variables

V.2 Mathematical expectation

Mathematical expectation - General and simplified

The expectation of a random variable $X : S \longrightarrow \mathbb{R}$ *on a probability space* (S, Ω, P) *is defined as*
$E(X) = \int_S X(s)dP.$

- This is called the "(Lebesgue) integral of X over the sample space S with respect to the probability measure P".
- Note: Our high-school ("Riemann") integral $\int_{(a,b)} Xdx$ only works on $(a, b) \subseteq \mathbb{R}$, and weighs by "length" of (very small) subintervals, not on general measurable spaces S, Ω with any probability measure P.

Mathematical expectation - General and simplified

The expectation of a random variable $X : S \longrightarrow \mathbb{R}$ *on a probability space* (S, Ω, P) *is defined as*
$E(X) = \int_S X(s)dP.$

- But: P is uniquely characterised by the CDF of X. So we can use a simpler measure based on F (" Lebesgue-Stieltjes measure") and write
  $E(X) = \int_S XdP = \int_{\mathbb{R}} xdF_X$, where F_X denotes the CDF of X.
- And if X is absolutely continuous, i.e. $F_x(x) = \int_{\infty}^{x} f(x)dx$ for some PDF $f(\cdot)$, then
  $E(X) = \int_S XdP = \int_{\mathbb{R}} xdF_X = \int_{\mathbb{R}} xf(x)dx$
- So we recover the usual Riemann integral!!

Expectation of functions of a continuous random variable

Measurable functions $g(\cdot)$ of RV are a RV. So more generally:

Definition: For a probability space (S, Ω, P), the expectation of a measurable function $g(\cdot)$ of an absolutely continuous RV $X : S \longrightarrow \mathbb{R}$ is defined as

$E(X) = \int_{-\infty}^{\infty} xf(x)dx$

- X is said to have "finite expectation" if the integral exists.
- If the integral is not bounded, evaluate the integral for negative and positive values of g separately:
  - If $I^+ = \infty$ and $I^- < \infty$, $E(g(x)) = \infty$.
  - If $I^+ < \infty$ and $I^- = \infty$, $E(g(x)) = -\infty$.
  - If both integrals are ∞, the expectation is not defined.

Expectation of functions of a discrete random variable

- **Definition:** For a probability space (S, Ω, P) with finite S, the expectation of a Ω measurable function $g(\cdot)$ of a discrete random variable X with support $D = x_1, ..., x_n$ is defined as $E(g(x)) = \sum_{i=1}^{n} g(x_i) p_X(x_i)$, where p_X is the pmf of X.

Section V Integration theory, mathematical expectation, and moments of random variables

V.3 Moments of Random Variables

Mean and other raw moments

The "expectation" or "mean" of a random variable X, defined as

- $E(x) = \int_{-\infty}^{\infty} x f_X(x) dx$
- or $E(x) = \sum_i x_i p_X(x_i)$ for discrete random variables, is also called the "first moment about 0", "first raw moment", or simply "first moment" of X.
- The mean of X is often denoted μ_x.
- More generally, for any positive m the **mth (raw) moment** of X is defined as $E(x^m)$.

Properties of the mean

- E(a) = a
- By the linearity of the sum or integral operators E(a+bg(X)) = a+bE(g(X)).
- E(XY)=?

Variance and other central moments

- The mth "central" moment of X is defined as $E((x - \mu_x)^m)$.
- The "variance of X" is the second central moment, denoted as $Var(X)$ or σ_x^2.
- Variance in terms of central moments:
  $Var(X) = E((x - \mu_x)^2 = E(x^2 - 2x\mu_x + \mu_x^2) = E(x^2) - \mu_x^2$.
  "The variance is the expectation of the square minus the square of the expectation."
- The square root of the variance, σ, is called the standard deviation of x.
- The third central moment is called "skewness", the fourth "kurtosis".

Properties of the variance

- $Var(a) = 0$
- $Var(a + bX) = E([a + bX - (a + b\mu_x)]^2) = b^2 Var(X)$.

Moment generating function (MGF)

- The MGF of a random variable X is defined for both discrete and continuous distributions as $M(\lambda) = E(e^{\lambda X})$.
- It does not exist for all random variables.
- If $\exists h > 0$ *s.t. for* $-h < \lambda < h$ $E(e^{\lambda X})$ is defined and finite, then $\frac{d^m}{d^m \lambda} M(\lambda)|_{\lambda=0} = E(X^m)$.

Moment generating function (MGF)

- The MGF of a random variable X is defined for both discrete and continuous distributions as $M(\lambda) = E(e^{\lambda X})$.
- It does not exist for all random variables.
- If $\exists h > 0$ *s.t. for* $-h < \lambda < h$ $E(e^{\lambda X})$ is defined and finite, then $\frac{d^m}{d^m \lambda} M(\lambda)|_{\lambda=0} = E(X^m)$.

Moment generating function (MGF)

- The MGF of a random variable X is defined for both discrete and continuous distributions as $M(\lambda) = E(e^{\lambda X})$.
- It does not exist for all random variables.
- If $\exists h > 0$ *s.t. for* $-h < \lambda < h$ $E(e^{\lambda X})$ is defined and finite, then $\frac{d^m}{d^m \lambda} M(\lambda)|_{\lambda=0} = E(X^m)$.

Moment generating function (MGF)

- The MGF of a random variable X is defined for both discrete and continuous distributions as $M(\lambda) = E(e^{\lambda X})$.
- It does not exist for all random variables.
- If $\exists h > 0$ *s.t. for* $-h < \lambda < h$ $E(e^{\lambda X})$ is defined and finite, then $\frac{d^m}{d^m \lambda} M(\lambda)|_{\lambda=0} = E(X^m)$.
- Thus, the mth derivative of the moment generating function evaluated at 0 gives the mth raw moment of the random variable X.

Moment generating function - Example

Consider X uniform on [0,10], i.e. $f(x) = 1/10$, $x \in [0, 10]$, 0 otherwise.

- $E(e^{\lambda x}) = \int_{[0,10]} e^{\lambda x} 1/10 dx$.
- $E(X) = M_1(X) = \frac{d}{d\lambda} E(e^{\lambda x}|_{\lambda=0}) = \int_{[0,10]} x e^{0 \cdot x} 1/10 dx = 1/10 \int_{[0,10]} x dx = 1/10(1/2x^2)|_0^{10} = 5$
- $M_2(X) = \frac{d^2}{d\lambda^2} E(e^{\lambda x}|_{\lambda=0}) = \int_{[0,10]} x^2 e^{0 \cdot x} 1/10 dx = 1/10 \int_{[0,10]} x dx = 1/10(1/3x^3)|_0^{10} = 1000/30$
- $Var(X) = E(X^2) - E(x)^2 = 1000/30 - 750/30 = 25/3$

Exercises

- Using the MGF, show that the expectation and variance of the **Poisson distribution** $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ are equal to λ.
- Using the MGF, show that the expectation and variance of the **exponential distribution** $f(x) = \frac{1}{\theta}e^{-\theta x}$ are equal to θ and θ^2 respectively.

Section VI Common univariate distribution functions

Summary Criteria

When looking at distribution functions, we consider the following criteria:

- What is its support, i.e. what are the values of the underlying random variable X with positive probability density or mass? These may be the integers, the positive real numbers, etc.
- What parameters characterise the distribution, or its pdf / pmf?
- What are its moments?

Summary Criteria

When looking at distribution functions, we consider the following criteria:

- What is its support, i.e. what are the values of the underlying random variable X with positive probability density or mass? These may be the integers, the positive real numbers, etc.
- What parameters characterise the distribution, or its pdf / pmf?
- What are its moments?

Summary Criteria

When looking at distribution functions, we consider the following criteria:

- What is its support, i.e. what are the values of the underlying random variable X with positive probability density or mass? These may be the integers, the positive real numbers, etc.
- What parameters characterise the distribution, or its pdf / pmf?
- What are its moments?

Summary Criteria

When looking at distribution functions, we consider the following criteria:

- What is its support, i.e. what are the values of the underlying random variable X with positive probability density or mass? These may be the integers, the positive real numbers, etc.
- What parameters characterise the distribution, or its pdf / pmf?
- What are its moments?

Summary Criteria

When looking at distribution functions, we consider the following criteria:

- What is its support, i.e. what are the values of the underlying random variable X with positive probability density or mass? These may be the integers, the positive real numbers, etc.
- What parameters characterise the distribution, or its pdf / pmf?
- What are its moments?
- Not generally the CDF.

Section VI Common univariate distribution functions

VI.1 Discrete univariate distributions

Bernoulli Distribution

- Takes value 1 ("success") and 0 ("failure") only.
- Defined by one parameter $0 \leq p \leq 1$ such that $Prob(X = 1) = p$, and $Prob(X = 0) = 1 - p$
- Or more compactly $P(X = x) = p^x(1 - p)^{(1-x)},\ x \in \{0, 1\}$
- **Moments**
  - E(x)=p
  - Var(x)=p(1-p)

Binomial Distribution

- Sum of N independent Bernoulli distributed random variables with identical p.
- Defined by the two parameters N and p. The support of X are the integers from 0 to N.
- Its PMF:
  $Prob(X = x) = Prob(\text{"}x \text{ times } 1 \text{ and } N - x \text{ times } 0\text{"}) = (\frac{N!}{x! * (N-x)!}) * p^x * (1-p)^{(N-x)}$.
- **Exercise:** Show that the moments of the Binomial distribution are given as $E(X) = n \cdot p$, Var(X)=np(1-p).

Poisson Distribution

- Has PMF $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, *for* $x = 0, 1, 2, 3, ...$
- It is characterised entirely by the parameter lambda.
- **Exercise**: Using the definition of the exponential function, show that $E(X) = Var(X) = \lambda$.

Section VI Common univariate distribution functions

VI.1 Continuous univariate distributions

Uniform Distribution

- Has pdf $f(x) = \frac{1}{b-a}$, *for* $x \in [a, b]$. It is characterised by the two parameters a,b.
- **Exercise:** Show that the moments of the uniform distribution are $E(X) = \frac{b-a}{2}$, $Var(X) = \frac{(b-a)^2}{12}$.

Exponential Distribution

- Has pdf $f(x) = \frac{1}{\theta} e^{\frac{x}{\theta}}$, for $\infty > x \geq 0$ and $\theta > 0$ its only parameter.
- **Exercise** Using the moment generating function show that $E(x) = \theta$ and $Var(x) = \theta^2$.

Standard normal Distribution

- The standard normal distribution X has pdf $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \ x \in \mathbb{R}$.
- Its CDF has no closed form solution.
- As it is symmetric around 0, all odd central moments are 0.
- Also, $E(X^2) = Var(X) = 1$, so one writes often $X \sim N(0, 1)$

General univariate normal Distribution

Exercise: Show, using the change of variables technique, that $Y = \mu + \sigma X$ has

- $f(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \ x \in \mathbb{R}$
- using iterated integration by parts, that $E(X) = \mu$, and $Var(X) = \sigma^2$.
- Y has a general univariate normal distribution, also written as $Y \sim N(\mu, \sigma^2)$.
- It is entirely characterised by its mean and variance.

Roadmap: So far, we have looked at

- General probability spaces (S, Ω, P)
- RVs $X : S \longrightarrow \mathbb{R}$ that transform this into $(\mathbb{R}, \mathbb{B}, v(\cdot))$, where $v(\cdot)$ is a probability measure that can be summarised by X's CDF
- Properties of CDF and its link to PDF (for abs. continuous RVs) and PMF (for discrete RVs)
- Moving from the pdf of X to that of g(x) for invertible and differentiable g
- Mathematical Expectations and their properties
- (Raw and Central) "Moments" of RVs
- The moment generating function $M(\lambda) = E(e^{\lambda x})$, with $\frac{d^m}{d\lambda^m} M(\lambda) = E(x^m)$
- Some common univariate distributions and their characteristics

Roadmap

Now we would like to

- Look at the probability of "My salary in 5 years is lower than yours" - 2 random variables on the same probability space (S, Ω, P).
- More general: vectors of k RVs on the same (S, Ω, P).
- Look at some new objects:
  - Distribution of (X,Y).
  - Distribution of Y.
  - Distribution of X given Y=y.
  - "Covariance" of X and Y.
  - etc.

Section VII Multivariate Random Variables

VII.1 Bivariate Random Variables

Bivariate Random Variables - Definition

A pair of random variables (X,Y) on the same probability space $(S, \Omega, P(\cdot))$ make a measurable function from the sample space into $\mathbb{R}^2$, and are called a "bivariate random variable".

Joint distribution function

- Given a probability space $(S, \Omega, P())$, and two RVs $X : S \longrightarrow \mathbb{R}$, $Y : S \longrightarrow \mathbb{R}$, $\{s \in S : (X(s), Y(s)) \in B^2\} = \{s \in S : X(s) \in B_1\} \cap \{s \in S : Y(s) \in B_2\}$ is a well-defined event for all Borel-sets $B^2 \in \mathbb{B}^2$.
- Again, we can summarize the probability measure $P_{X,Y}(B^2) = \{s \in S : (X(s), Y(s)) \in B^2\}$ by the probabilities of the half-closed sets, or their **joint cumulative distribution function** $F(x, y) = P(X \leq x, Y \leq y)$.

Properties of the joint cumulative distribution function

1. $0 \leq F(x, y) \leq 1$ for all $(x, y) \in \mathbb{R}^2$
2. $F(\cdot)$ is a non-decreasing function in both of its arguments.
3. $F(\cdot)$ is right-continuous in both of its arguments.
4. $F(-\infty, y) = F(x, -\infty) = F(-\infty, -\infty) = 0$
5. $F(\infty, \infty) = 1$

Joint probability mass function

If X and Y are both discrete random variables on the same probability space, or "jointly discrete", with positive mass on the support $D_X = \{x_1, x_2, ...\}, D_Y = \{y_1, y_2, ...\}$, we can define the bivariate probability mass function as $p(x, y) = P(X = x, Y = y)$, with properties similar to those of univariate PMFs, i.e.

1. $1 \geq p(x, y) \geq 0$
2. $p(x, y) = 0 \ \forall \ x \notin D_X, y \notin D_Y$
3. $\sum_{x \in D_x} \sum_{y \in D_y} p(x, y) = 1.$

Joint probability density function

If X and Y are both absolutely continuous random variables on the same probability space, or "jointly continuous", there is a bivariate pdf $f(x, y)$, that satisfies $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y) dxdy$ with the properties

1. $f(x, y) \geq 0$
2. $\int \int_{\mathbb{R}^2} f(x, y) dxdy = 1$

Marginal distribution function

- For every Borel set A, $X \in A$ is a well-defined event with probability $P(X \in A) = P(s \in S : X(s) \in A) = P(s \in S : X(s) \in A \text{ and } Y \in \mathbb{R})$.
- The marginal CDF of X is obtained from the joint CDF by calculating the limit $Y \longrightarrow \infty$, for any given x: $F(x, \infty) = P(X \leq x, Y \leq \infty) = P(X \leq x) = F_X(x)$.

Marginal PMF of X

For jointly discrete X and Y this yields the **marginal PMF of X** $\sum_{y \in D_Y} p(x, y) = P(X = x) = p_X(x)$ and marginal CDF of X $F_X(x) = \sum_{s_x \in D_X : s_x \leq x} \sum_{y \in D_Y} p(s_x, y)$, where D_X, D_Y are the supports of X and Y.

Marginal PDF of X

- For jointly (absolutely) continuous X and Y, the **marginal probability density of X** is $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$.
- The marginal CDF of X then satisfies $F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y)dxdy = lim_{y \longrightarrow \infty} F(x, y)$.

Conditional distribution function

- For a bivariate random variable (X,Y) on S, Ω, P and given Borel sets A and C, the event $\{s \in S : X \in A, Y \in C\}$ is well-defined.
- So whenever $Y(s) \in C$ for some $s \in S$ we can calculate the conditional probability of $X \in A$ given $Y \in C$ as $P(X \in A \mid Y \in C) = \frac{P(X \in A, Y \in C)}{P(Y \in C)}$.

Conditional PMF of X, given Y=y

- For jointly discrete random variables X and Y, we can calculate the **conditional PMF of X, given Y=y**
  $p_{X|Y}(x \mid y) = P(X = x, \mid Y = y) = \frac{p(x,y)}{p_Y(y)}$.
- The **conditional CDF** is simply
  $F(x \mid y) = \sum_{s_x \in D_X : s_x \leq x} p_{X|Y}(s_x \mid Y = y)$.

Jointly continuous RVs: Conditional Distribution of X given Y=y

- Problem: $P_Y(Y = y) = 0$, so using simple definition of conditional probability impossible.
- Using limits, we can write
  $F_{X|Y}(x \mid Y = y) = lim_{\epsilon \longrightarrow 0^+} P(X \leq x \mid y - \epsilon < Y < y + \epsilon)$.
- The **conditional PDF** is then $f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$
- **Proof**
- $f_{X|Y}(x \mid y) = \frac{\delta}{\delta x} F_{X|Y}(x \mid y)$
  $= \frac{\delta}{\delta x} lim_{\epsilon \longrightarrow 0^+} \frac{P(X \leq x, y-\epsilon < Y < y+\epsilon)}{P(y-\epsilon < Y < y+\epsilon)}$
  $= \frac{\delta}{\delta x} lim_{\epsilon \longrightarrow 0^+} \frac{F(x,y+\epsilon) - F(x,y-\epsilon)}{F_Y(y+\epsilon) - F_Y(y-\epsilon)}$
  $= \frac{\delta}{\delta x} lim_{\epsilon \longrightarrow 0^+} \frac{\frac{F(x,y+\epsilon) - F(x,y-\epsilon)}{\epsilon}}{\frac{F_Y(y+\epsilon) - F_Y(y-\epsilon)}{\epsilon}} = \frac{\delta}{\delta x} \frac{\frac{\delta}{\delta y} F(x,y)}{\frac{\delta}{\delta y} F_Y(y)} = \frac{f(x,y)}{f_Y(y)}$

Example: Uniform distribution on the plane

- Two jointly distributed continuous RVs have the bivariate uniform distribution on $(0, a) \times (0, b)$ if their joint density is a constant, i.e. $f(x, y) = c, c \in \mathbb{R}^{++}$.
- Determining c: $\int_0^a \int_0^b c dy dx = 1$, so $c = \frac{1}{ab}$.
- Joint CDF of x and y, $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \frac{1}{ab} dx dy$; 5 cases:
  1. $F(x, y) = 0$ for $x, y : x < 0$ *or* $y < 0$
  2. $F(x, y) = \frac{1}{ab} xy$ for $\{x, y : 0 \leq x \leq a, 0 \leq y \leq b\}$
  3. $F(x, y) = \frac{1}{a} x$ for $\{x, y : 0 \leq x \leq a, b < y\}$
  4. $F(x, y) = \frac{1}{b} y$ for $\{x, y : a < x, o \leq y \leq b\}$
  5. $F(x, y) = 1$ for $\{x, y : a < x, b < y\}$

Example: Uniform distribution on the plane

- Marginal pdf of X is $f(x) = \frac{1}{a}$ on $\{x : 0 \leq x \leq a\}$ and 0 otherwise. So the marginal CDF of X is simply $F(x, y) = \frac{1}{a}x$ for $\{x : 0 \leq x \leq a\}$, 0 for $x < 0$ and 1 for $x > a$.
- Conditional pdf of X given Y=y is $f_{X|Y}(x \mid Y = y) = \frac{f(x,y)}{f(y)} = \frac{1}{a}$. The conditional CDF follows as above.

Exercise (Champagne 2003)

A rope of length L is cut into three pieces in the following way:

- The first piece of length X is obtained by cutting the rope at random (with uniform probability for all points $x \in [0, L]$).
- The second piece of length Y is obtained by cutting the remaining segment of length $L - X$ at random.
- The third piece is obtained as the remaining segment of length $L - X - Y$.

1. Find $f_{Y|X}(y \mid x)$, the conditional PDF of Y given $X = x,\ (0 < x < L)$.
2. Find f(x, y), the Joint PDF of X and Y , and illustrate the region of the plane where it takes on non-zero values.
3. What is the probability that both X and Y be less than $L = 2$?

Expectations and moments of bivariate random variables

- The expectation of a bivariate RV is simply the expectations of the individual random variables written as a vector.
- The expectation of a measurable function $g(\cdot) : \mathbb{R}^2 \longrightarrow \mathbb{R}$ of two jointly distributed random variables X and Y is:
  - For jointly discrete RVs with support $D_X = \{x_1, x_2, ...\}, D_Y = \{y_1, y_2, ...\}$ we have $E(g(x,y)) = \sum_{D_X} \sum_{D_Y} g(x,y) p_{XY}(x,y)$.
  - For jointly (absolutely) continuous random variables we have $E(g(x,y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dxdy$.

Covariance -Definition

For two jointly distributed RVs X and Y with unconditional means μ_X and μ_Y, the covariance of X and Y is defined as $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

Properties of the covariance

1. $Cov(X, X) = Var(X) = \sigma_X^2$

   (by the definition of variance and covariance)

2. $Cov(X, Y) = Cov(Y, X)$

   (by the commutativity of the product operator)

3. $Cov(aX + b, cY + d) = acCov(X, Y)$

   (by the linearity of the expectations operator)

4. $Cov(X, Y) = E(XY) - E(X)E(Y)$

   - (as $E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) = E(XY) - 2E(X)E(Y) + E(X)E(Y)$)

5. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

   - (as $E((X + Y - \mu_X - \mu_Y)^2) = E((X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)) = Var(X) + Var(Y) + 2Cov(X, Y)$)

Definition: Correlation coefficient ρ

The correlation coefficient ρ of two jointly distributed RVs is defined on the basis of their covariance as $\rho(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, where σ_X and σ_Y are the standard deviations of X and Y.

Properties of the correlation coefficient

1. ρ is dimensionless even for RVs that have units.
2. $-1 \leq \rho \leq 1$
   (as $\mid Cov(X,Y) \mid \leq \sqrt{Var(X)Var(Y)} = \sigma_X \sigma_Y$)
3. $\rho(X,X) = 1$
   (as $Cov(X,X) = Var(X) = \sigma_X \sigma_X$)
4. $\rho(X,Y) = \rho(Y,X)$
   (as $Cov(X,Y) = Cov(Y,X)$)
5. $\rho(aX+b, cY+d) = sign(ac)\rho(X,Y)$
   - (as $Cov(aX+b, cY+d) = acCov(X,Y)$, for $Z = aX+b$ $\sigma_Z = \mid a \mid \sigma_X$, etc.)

Properties of the correlation coefficient continued

1. $\rho(X, Y) = 1 \leftrightarrow Y = aX + b$ for any $a > 0$, *and any* $b \in \mathbb{R}$
   $\rho(X, Y) = -1 \leftrightarrow Y = aX + b$ for any $a < 0$, *and any* $b \in \mathbb{R}$
   (by a similar argument)
2. **Note**: The correlation coefficient is a measure of linear association between two RVs. From 1. it equals 1 or -1 whenever one random variable is a linear affine function of the other, e.g $Y = a + bX$.

Independence of 2 jointly distributed RVs

- The RVs X and Y are defined to be independent if the events $X \in A$ and $Y \in B$ are independent for any pair of Borel sets A and B, i.e.
  $P(X \in A, Y \in B) = P(X \in A)P(Y \in B),\ \forall A, B \in \mathbb{B}.$
- Applied to half-open intervals:
  $F(x, y) = P(X \leq x, Y \leq y) = F_X(x)F_Y(y)$. This implies:
  1. PMF of independent jointly discrete RVs
     $p(x, y) = p_X(x)p_Y(y)$.
  2. PDF of independent jointly continuous RVs
     $f(x, y) = f_X(x)f_Y(y)$.

Properties of independent RVs

For any 2 independent jointly distributed RVs X,Y

1. The random variables defined by measurable function g(x) and h(y) are also independent.
2. Conditional distributions are equal to marginal distributions, e.g. $F_{X|Y}(x \mid y) = F_X(x)$, and equivalently for PMF, PDF.
3. $E(g(x)h(y)) = E(g(x))E(h(y))$ for all measurable functions $g,\ h$.
4. Particularly, $E(xy) = E(x)E(y)$.
5. So the moment generating function for the sum of two random variables is the product of the individual moment generating functions.
6. X,Y indep. $\rightarrow Cov(X, Y) = corr(X, Y) = 0$. But not vice versa.

Conditional expectation

Unconditional vs. conditional expectation for bivariate RVs:

1. The **unconditional expectation of X** is a real number defined by
   $E(X) = \int_X \int_Y x f(x,y) dx dy = \int_X x f_X(x) dx = \mu_X$, or
   $E(X) = \sum_i \sum_j x_i p(x_i, y_j) = \sum_i x_i p_X(x_i) = \mu_X$
2. The **conditional expectation of X given** $Y = y$ is a function $\mu_{X|Y}(\cdot) : D_Y \longrightarrow D_X$ that maps every value y into the conditional expectation of X given Y=y.
   $E_{X|Y}(X \mid y) = \int_X x f_{X|Y}(x \mid y) dx = \mu_{X|Y}$
   $E_{X|Y}(X \mid y) = \sum_i x_i p_{X|Y}(x_i \mid y) dx = \mu_{X|Y}$

Conditional variance

Unconditional vs. conditional expectation for bivariate RVs:

1. Equally, the **conditional Variance of X given Y=y** is defined as
   $\sigma^2_{X|Y} = E_{X|Y}((X - E_{X|Y}(X \mid y))^2 \mid y) = E_{X|Y}(X^2 \mid y) - (E_{X|Y}(X \mid y))^2$.
2. Replacing the particular value y with the RV Y, $E_{X|Y}(X \mid \underline{\mathbf{Y}})$ and $\sigma^2_{X|\underline{\mathbf{Y}}}$ are measurable functions of the RV Y, and thus RVs themselves.

Law of iterated expectations and Decomposition of Variance

- The simple law of iterated expectations:
  $E_Y(E_{X|Y}(X \mid Y)) = E(X)$.
- In words: the unconditional expectation of X equals the "expectation over Y of the conditional expectation function of X given Y=y".
- Thus:
  1. $E(X) = \int_Y E_{X|Y}(X|Y)dy$
     Proof $E(X) = \int_X \int_Y xf(x,y)dxdy =$
     $\int_X \int_Y xf_{X|Y}(x|y)f_Y(y)dxdy = \int_Y \int_X xf_{X|Y}(x|y)dxf_Y(y)dy =$
     $\int_Y E_{X|Y} f_Y(y)dy = E_Y(E_{X|Y}(X|Y))$
  2. $E(X) = \sum_j E_{X|Y}(X|Y = y_j)$

Decomposition of Variance

The Variance decomposes into

$$Var(X) = Var_Y(E_{X|Y}[X \mid Y = y]) + E_Y(Var_{X|Y}(X \mid Y = y)).$$

Conditioning Theorem

$$E_{X|Y}(g(y)X \mid y) = g(y)E_{X|Y}(X \mid y).$$

Example (Champagne 2003)

Let RVs X and Y be jointly uniform over the region
$D = \{(x, y)\} : 0 < x < y < 1$.

1. Draw the region D.
2. Find the real number u such that x,y are jointly uniform on D with pdf $f(x, y) = u,\ \forall x, y \in D$, and $f(x, y) = 0$ otherwise.
3. Find $E(X),\ E(Y),\ E(X^2),\ E(XY)$

Section VII Multivariate Random Variables

VII.1 Multivariate Random Variables

Definition: Random Vector

A list of n random variables (X,Y) on the same probability space $(S, \Omega, P(\cdot))$ make a measurable function from the sample space into $\mathbb{R}^n$, and is called an "n-dimensional random vector".

Joint distribution function

- The event $\{s \in S : (x_1, x_2, ...x_n) \in B\}$ is a well-defined event for all n-dimensional Borel-sets $B = (B_1, B_2, ..., B_n) \in \mathbb{B}^n$
- So the joint probability $P(x_1 \in B_1, x_2 \in B_2, ..., x_n \in B_n)$ is well defined.
- As in the bivariate case, it can be summarized by the **joint CDF** of $X1, ..., Xn$, defined as
  $F(x_1, x_2, ...x_n) = P(X1 \leq x1, X2 \leq x2, ..., Xn \leq xn)$.

Properties of the Joint CDF

These are equivalent to those of the bivariate case, i.e.

1. $0 \leq F(\cdot) \leq 1$
2. $F(\cdot)$ is non-decreasing in all of its arguments.
3. $F(\cdot)$ is right-continuous in all of its arguments.
4. $lim_{x_i \longrightarrow -\infty} F(x_1, x_2, ...x_n) = 0, \ \ i \in \{1, ..., n\}$
5. $lim_{x_1 \longrightarrow \infty, x_2 \longrightarrow \infty, \ ...} F(x_1, x_2, ...x_n) = 1$

Joint PDF and PMF

- For jointly (absolutely) continuous $X_1, X_2, ...$ there exists a multivariate pdf $f(x_1, x_2, ...x_n)$ satisfying $f(x_1, x_2, ...x_n) = \frac{\delta^n}{\delta x} F(x_1, x_2, ...x_n)$ with $\delta x = (\delta x_1, \delta x_2, ..., \delta x_n)$.
- The joint PMF for jointly discrete random variables is equivalent to the bivariate case.

Marginal distribution function

1. The **marginal CDF** of a subvector X_I, $I \subseteq \{1, 2, ..., n\}$ is obtained by letting all $x_j \longrightarrow \infty$, for $X_j \in X_J, J = \{1, 2, ..., n\} \backslash I$.
2. For n jointly continuous RVs, the **marginal PDF** of a subvector X_I is obtained from the joint PDF by integrating over all $X_j \in X_J, J = \{1, 2, ..., n\} \backslash I$, i.e. $f_{X_I} = \int_J f(x_1, x_2, ..., x_n) dx_J$.
3. For n jointly discrete RVs, the marginal PMF of a subvector X_I is obtained by summing over all possible values of $X_j \in X_J, J = \{1, 2, ..., n\} \backslash I$, i.e. $p_{X_I} = \sum_J p(x_1 1, x_2, ..., x_n)$.

Conditional distribution function

1. The conditional CDF of subvectors X_I given values x_j of subvector X_J is $F_{X_I|X_J}(x_I \mid x_J) = P(X_I \leq x_I \mid X_J = x_J) = lim_{\epsilon \longrightarrow 0^+} \frac{P(X_I \leq x_I, x_J - \epsilon \leq X_J \leq x_J + \epsilon)}{P(x_J - \epsilon \leq X_J \leq x_J + \epsilon)}$.
2. The conditional PMF of n jointly discrete random variables is $p_{X_I|X_J}(x_J \mid x_J) = \frac{p(x_I, x_J)}{p_{X_J}(x_J)}$, wherever $p_{X_J}(x_J) > 0$.
3. The conditional PDF of n jointly continuous random variables is $f_{X_I|X_J}(x_I \mid x_J) = \frac{f(x_I, x_J)}{f_{X_J}(x_J)}$.

Expectation and covariance of n random variables

- The expectation of a measurable function $g(\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}$ of an n-dimensional random vector is $E(g(X_1, X_2, ..., X_n)) = \int_{X_1} \int_{X_2} \cdots \int_{X_n} g(x_1, x_2, ..., x_n) f((x_1, x_2, ..., x_n)) dx_1 dx_2 ... dx_n$.
- Or for the special case of jointly discrete random variables $E(g(X_1, X_2, ..., X_n)) = \sum_{X_1} \sum_{X_2} \cdots \sum_{X_n} g(x_1, x_2, ..., x_n) p((x_1, x_2, ..., x_n))$.

Expectation and covariance of n random variables

- The covariance being a bivariate concept we can extend it to the n-variable case by defining the **Variance-Covariance Matrix of the multivariate random variable** $X = (X_1, X_2, ..., X_n)$ as

$$\Sigma = \begin{bmatrix} Var_{X_1} & Cov_{21} & ... & Cov_{n1}\, Cov_{21} & Var_{22} & ... & Cov_{1n} \\ ... & & ... & & & & \\ Cov_{1n} & Cov_{12} & ... & Var_{nn} & & & \end{bmatrix}$$

Independence of n random variables

The random variables $X_1, X_2, ..., X_n$ are called independent if for all Borel sets $B_i \in \mathbb{B}, i = 1, ..., n$, the events $(X_1 \in B_1), (X_2 \in B_2), ..., (X_n \in B_n)$ are mutually independent.

- This implies $P((X_1 \in B_1), (X_2 \in B_2), ..., (X_n \in B_n)) = P(X_1 \in B_1)P(X_2 \in B_2)...P(X_n \in B_n)$.
- For jointly discrete RVs the joint PMF is $p(x_1, x_2, ..., x_n) = p_{X_1}(x_1)p_{X_2}(x_2)...p_{X_n}(x_n)$
- For jointly continuous random variables, the joint PDF is $f(x_1, x_2, ..., x_n) = f_{X_1}(x_1)f_{X_2}(x_2)...f_{X_n}(x_n)$.
- The VCM of n independent random variables is diagonal.

General law of iterated expectations

A more general version of the law of iterated expectations is

$$E_{Y|Z}(E_{X|Y,Z}(X \mid y, z)) = E_{X|Z}(X \mid z).$$

Multivariate Transformations

- The transformation theorem, or "change-of-variables formula", extends to functions of n-dimensional random vectors, $g(\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ that are one-to-one, differentiable and invertible.
- Defining a new random vector Y as $Y = g(x)$, and the inverse function $h(y) = g(x)^{-1}$, and using the definition of the Jacobian matrix as the determinant of the matrix of first partial derivatives of a vector valued funtion, i.e. $J(y) = det(\frac{\delta}{\delta y'} h(y))$, the probability distribution function of Y is then $f_Y(y) = f_X(h(y)) \mid J \mid$.

Multivariate Transformations: Example

Let random variables X and Y be jointly uniform over the region $A = \{(x, y)\} : 0 < x, y < 1$. Let R and Q be defined by $R = x + y$, and $Q = x - y$. What is the joint pdf of R and Q?

1. The constant pdf of (x,y) equals 1 for all points in the domain.
2. Note that $h(r, q) = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} r \\ q \end{bmatrix}$ So the absolute value of the Jacobian is $|-1/2| = 1/2$.
3. Thus we get $f(r, q) = f(h(r, q)1/2$, so r and q are uniformly distributed with parameter $1/2$.

Multivariate Transformations: Example

Let random variables X and Y be jointly uniform over the region $A = \{(x, y)\} : 0 < x, y < 1$. Let R and Q be defined by $R = x + y$, and $Q = x - y$. What is the joint pdf of R and Q?

1. The Domain of (R,Q) is the set

$$\text{B}=\{\text{r,q}: \begin{bmatrix} -1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} r \\ q \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} r \\ q \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \}.$$

   - Draw this set in r,q space.
   - Check that the integral of c=1/2 over this set is 1.

Multivariate Normal Distribution

For non-singular Σ_X, the PDF of an n-dimensional normal random variable X is

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} det(\Sigma_X)^{1/2}} e^{-\frac{(x-\mu_X)'\Sigma_X^{-1}(x-\mu_X)}{2}}.$$

- Its domain is $x \in \mathbb{R}^n$.
- Its expectation μ_X and the $(n+1)n/2$ distinct elements of the variance-covariance matrix Σ_X are the parameters that completely characterise its distribution, so we write $X \sim N(\mu_X, \Sigma_X)$.

Multivariate Normal Distribution - equivalent Definitions

The following are equivalent definitions of an n-dimensional normal random vector X, allowing for singular Σ_X

- There is an m-dimensional random vector Z whose elements are independent normal random variables, a vector of real numbers $\mu = (\mu_1, ..., \mu_n)$, and an nxm matrix A such that $X = \mu + Az$.
- There is a vector of real numbers $\mu = (\mu_1, ..., \mu_n)$, and a positive semi-definite matrix Σ_X, such that the Moment generating function of X is $e^{\lambda'\mu + \lambda'\Sigma_X\lambda/2}$.

Multivariate Normal Distribution - equivalent Definitions

The following are equivalent definitions of an n-dimensional normal random vector X, allowing for singular Σ_X

- Every linear combination $Y = a_1X_1 + ... + a_nX_n$ is normally distributed.
- There is an m-dimensional random vector Z whose elements are independent normal random variables, a vector of real numbers $\mu = (\mu_1, ..., \mu_n)$, and an nxm matrix A such that $X = \mu + Az$.
- There is a vector of real numbers $\mu = (\mu_1, ..., \mu_n)$, and a positive semi-definite matrix Σ_X, such that the Moment generating function of X is $e^{\lambda'\mu+\lambda'\Sigma_X\lambda/2}$.

Conditional and marginal distributions of subvector

$$x_I = (x_1, ..., x_k), \ k \leq n$$

- The conditional distribution of X_I given $X_J = x_J$ is multivariate normal $X_I \mid X_J = x_J \sim N(\mu_{I|J}, \Sigma_{I|J})$, where $\mu_{I|J} = \mu_I + \Sigma_{12}\Sigma_{22}^{-1}(x_J - \mu_J)$ and $\Sigma_{I|J} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.
- So knowing the value of x_J alters mean and variance of X_I.
- The marginal distribution of X_I is simply $N(\mu_I, \Sigma_{11})$.

Conditional and marginal distributions of subvector $x_I = (x_1, ..., x_k),\ k \leq n$

- Partition X into a kx1 and an (n-k)x1 vector as $X = (X_I, X_J)$, and accordingly $\mu_X = (\mu_I, \mu_J)$ and $\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ with Σ_{11} kxk, Σ_{21} kx(n-k), etc.
- The conditional distribution of X_I given $X_J = x_J$ is multivariate normal $X_I \mid X_J = x_J \sim N(\mu_{I|J}, \Sigma_{I|J})$, where $\mu_{I|J} = \mu_I + \Sigma_{12}\Sigma_{22}^{-1}(x_J - \mu_J)$ and $\Sigma_{I|J} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.
- So knowing the value of x_J alters mean and variance of X_I.
- The marginal distribution of X_I is simply $N(\mu_I, \Sigma_{11})$.

Example (Banerjee Sample questions)

Suppose that y, x1 and x2 have a joint normal distribution with parameters $\mu' = [1, 2, 4]$ and covariance matrix $\Sigma = \begin{bmatrix} 2 & 3 & 1 \\ 3 & 5 & 2 \\ 1 & 2 & 6 \end{bmatrix}$

Compute $E(y|x1)$; $var(y|x1)$; $E(y|x1 = 2,5; x2 = 3,3)$; $var(y = x1 = 2,5; x2 = 3,3)$

Exercises

- Using the fact that the covariances of independent RVs are zero, show that the joint pdf of n independent jointly normal RVs is the product of the univariate normal pdfs. This also shows, that for normal RVs, independence implies 0 covariance AND vice versa.
- Using the multivariate change-of-variables formula, show for an n-dimensional normal vector X, and $Y = A + Bx$, that linear affine transformations of multivariate normal RVs are also normally distributed, with

$$f_Y(y) = \frac{1}{2\pi^{\frac{n}{2}} det(B\Sigma_X B')^{1/2}} e^{-\frac{(y-B\mu_X)'(B\Sigma_X B')^{-1}(x-B\mu_X)}{2}} =$$

$$\frac{1}{2\pi^{\frac{n}{2}} det(\Sigma_Y)^{1/2}} e^{-\frac{(x-\mu_Y)'\Sigma_Y^{-1}(x-\mu_Y)}{2}}.$$