BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 1

Sets

- Sets:
  - Sets are basic concepts in mathematics and probability.
  - Crudely defined as: “a collection of some elements”.
  - Usually denoted with a capital letter (e.g. A, B, S)
- Special sets:
  - $\mathbb{N}$ = Natural numbers (1, 2, 3, 4, ...)
  - $\mathbb{Z}$ = Integers (0, +1, -1, +2, -2, ...)
  - $\mathbb{Q}$ = Rational numbers
  - $\mathbb{R}$ = Real numbers

Common Notation

- $\emptyset$: empty set
- $w \in A$: w is an element of the set A
- $\{w\} \subset A$ or $\{w\} \subseteq A$: the set consisting of the singleton $w \in A$ is a subset of A
- (a, b), where $a, b \in \mathbb{R}$, is the set of real numbers between (but not including) a and b
- $[a, b]$, where $a, b \in \mathbb{R}$, is the set of real numbers between and including a and b
- $[a, b)$, where $a, b \in \mathbb{R}$, is the set of real numbers between and including a but not b
- $(a, b]$, where $a, b \in \mathbb{R}$, is the set of real numbers between and including b but not a
- $\{w$: a statement$\}$: the set of elements w for which the statement holds. Example: the open interval (a, b) can be defined as $\{w : a < w < b\}$.
- $A = B$ if A and B contain exactly the same elements (this can be shown by showing (1) $A \subset B$ and (2) $B \subset A$)

The Sample Space:

- Need to define an "abstract space", often denoted by Ω, as a non-empty set of all the elements concerned. These elements are called "points" and often denoted with lower case letter
- In probability we often use Ω to denote the **sample space**, which is the collection of all possible distinct realizations of a non-deterministic experiment. (Usually idealized)
- Choice of sample space is the first stem in formulating a probabilistic model for an experiment. Examples of sample spaces:
  1. Draw a card from a deck of 52 cards: $\Omega = \{1, 2, \ldots, 52\}$, which is a **finite** sample space
  2. Toss a coin until one gets two successive heads and record the number of tosses performed: $\Omega = \{2, 3, 4, \ldots, \infty\}$, which is a **countably infinite** sample space.
  3. Two components in an electrical system record their failure times: $\Omega = \{(x, y) : x \geq 0, y \geq 0\}$, which is an **uncountably infinite** sample space.

Events:

- An event is always a subset of Ω, "events" .....
- ... but not all subsets are necessarily events!
- If Ω is countable (i.e. finite or countably infinite), then any subset of Ω is an event
- If Ω is uncountable, we cannot handle all possible subsets (not all are events). Instead, we restrict events to be a "well-behaved" class of subsets. More on this later.
- Individual points in Ω are called "simple events"

Union

Definition: For two sets A and B, their **union** is denoted as $A \cup B = \{w : w \in A \text{ or } w \in B\}$

Intersection

Definition: For two sets A and B, their **intersection** is denoted as $A \cap B = \{w : w \in A \text{ and } w \in B\}$

Complementation

Definition: For a set A in Ω, the complement of A (w.r.t. Ω) is denoted by $A^c = \{w \in \Omega : w \notin A\}$

Note: One can show that under complementation, $\subset$ and $\supset$ are swapped.

Difference

Definition: For two sets A and B, their **difference** is denoted as
$A - B = \{w : w \in A, w \notin B\} = A \cap B^c$

Symmetric Difference

Definition: For two sets A and B, their **symmetric difference** is denoted as
$A \Delta B = (A - B) \cup (B - A) = \{w : w \in$ exactly one of A and $B\}$

Disjoint Union

Definition: Two sets A and B are called **disjoint** if $A \cap B = \phi$.

Definition: For two disjoint sets A and B, their **disjoint union** is denoted as $A \cup B = A + B$

Properties of the Union/Intersection

For sets A_1, A_2, A_3:

- Properties of Union:
  1. Associative: $(A_1 \cup A_2) \cup A_3 = A_1 \cup (A_2 \cup A_3)$
  2. Distributive: $(A_1 \cup A_2) \cap A_3 = (A_1 \cap A_3) \cup (A_1 \cap A_3)$
  3. Commutative: $A_1 \cup A_2 = A_2 \cup A_1$
- Properties of Intersection:
  1. Associative: $(A_1 \cap A_2) \cap A_3 = A_1 \cap (A_2 \cap A_3)$
  2. Distributive: $(A_1 \cap A_2) \cup A_3 = (A_1 \cup A_3) \cap (A_1 \cup A_3)$
  3. Commutative: $A_1 \cap A_2 = A_2 \cap A_1$

Countable Unions/Intersections

- Finite sequence of sets: Let $\{A_1, \ldots, A_k\}$ be a finite sequence of k sets in Ω. We define

$$\sup_{1 \le n \le k} A_n = \bigcup_{n=1}^{k} A_n = \{w : w \in A_n \text{ for some } 1 \le n \le k\}$$

$$\inf_{1 \le n \le k} A_n = \bigcap_{n=1}^{k} A_n = \{w : w \in A_n \text{ for any } 1 \le n \le k\}$$

- Countable sequence of sets: Let $\{A_n\}$ be an infinite sequence of sets in Ω. We define

$$\sup_{n \ge 1} A_n = \bigcup_{n=1}^{\infty} A_n = \{w : w \in A_n \text{ for some } n\}$$

$$\inf_{n \ge 1} A_n = \bigcap_{n=1}^{\infty} A_n = \{w : w \in A_n \text{ for any } n \ge 1\}$$

Uncountable Unions/Intersections

We extend the intersection over a set of integers to any arbitrary set:

Definition: For $\{A_t, t \in T\}$, where T is a (possibly uncountable) index set, $\bigcup_{t\in T} A_t = \{w : w \in A_t \text{ for some } t \in T\}$. The definition for intersection is similar.

DeMorgan's Rule:

DeMorgan's Rule (for 2 sets):

$$(A \cup B)^c = A^c \cap B^c$$

and

$$(A \cap B)^c = A^c \cup B^c$$

DeMorgan's Rule (general):

$$\left(\bigcup_{t\in T} A_t\right)^c = \bigcap_{t\in T} A_t^c$$

and

$$\left(\bigcap_{t\in T} A_t\right)^c = \bigcup_{t\in T} A_t^c$$

where T is any index set (finite, countably infinite, uncountably infinite).

Proof of DeMorgan's Rule (2 sets):

$$(A \cap B)^c = A^c \cup B^c$$

Proof:

Proof of DeMorgan's Rule (continued):

The proof that $(A \cup B)^c = A^c \cap B^c$ is similar.

Limsup and Liminf of Sets

- **Definition:** Let $\{A_n\}$ be a sequence of sets in Ω. Define

$$A^* = \limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

and

$$A_* = \liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

these are the upper and lower limits of the sequence.
- **Theorem:** (a) $w \in \limsup_n A_n$ if and only if w is in infinitely many of the A_n.
(b) $w \in \liminf_n A_n$ if and only if there is an m such that $w \in A_n$ for all $n \geq m$ (i.e. w is in all but the first m) [Proof omitted]
- **Theorem:** $\liminf_n A_n \subset \limsup_n A_n$ (proof for homework)

Limit of Sets

If $\limsup A_n = \liminf A_n$ then we say that $\{A_n\}$ is convergent and write
$\lim_{n\to\infty} A_n = \limsup A_n = \liminf A_n$

Example: let $\Omega = [0, 1], A_n = [0, 1/n]$ if n is even and $A_n = [1 - 1/n, 1)$ if n is odd.

- Then by definition $A_* = \emptyset$ and $A^* = \{0\}$.
- Since $A_* \neq A^*$, then $\lim A_n$ does not exist.
- On the other hand, if we let $A_n = [0, 1/n]$ for all n, then $A_* = A^* = \{0\}$ and the limit exists.

More Examples

What is A_* and A^* for:

- Example: let $\Omega = \{0, 1\}$, then consider the sequence $\{0\}$, $\{1\}$, $\{0\}$, $\{1\}$, $\{0\}$, $\ldots$
- Example: let $\Omega = \mathbb{Z}$, Consider the sequence of sets: $\{0\}$, $\{0, 1\}$,$\{0\}$,$\{0, 1\}$,$\{0\}$,$\{0, 1\}$,$\{0\}$, $\ldots$
- Example: let $\Omega = \mathbb{Z}$, Consider the sequence of sets: $\{0\}$, $\{0, 1\}$, $\{0, 1, 2\}$, $\{0, 1, 2, 3\}$, $\{0, 1, 2, 3, 4, \}$, $\ldots$
- Example: let $\Omega = [0, 1], A_n = [0, 1/n]$ if n is even and $A_n = [1 - 1/n, 1)$ if n is odd.

Monotonicity

- **Monotonicity:** A monotine sequence of sets is defined as:
  - $\{A_n\}$ is called monotone increasing iff $A_n \subset A_{n+1}$ for any n
  - $\{A_n\}$ is called monotone decreasing iff $A_n \supset A_{n+1}$ for any n
- **Theorem:** A monotone sequence of sets is convergent
  Proof for increasing sequence:

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 2

Intro: More on Events

- **Casella & Berger Definition**: An event is any collection of possible outcomes of an experiment, that is, any subset of Ω (including Ω itself). $\leftarrow$ not strictly true!
  Previously:
  - An event is necessarily a collection of possible outcomes of a random experiment (a set of elements in Ω, i.e. a subset of Ω)
  - For discrete (finite and countable) sample space, the set of possible events is the power set (2^{Ω}), i.e. any subset of Ω.
  - For continuous sample spaces (uncountably infinite spaces), the set of possible events is NOT the power set, there are subsets that are not considered events.
- Strictly speaking, **events** are the subsets of the sample space for which a probability is defined.

Intro: On Measures and Probability

A **measure** is a set function that assigns a number $\mu(A)$ to each set A in a certain class of sets.

Examples:

- Length
- Area
- Volume
- Probability

Some structure must be imposed on the class of sets in which the set function μ is defined (i.e. cannot take probability of any set in 2^{Ω}).

Classes of Sets

- **Definition:** A **class** is a collection of sets (set of sets) that satisfy some conditions. Usually denoted with script characters ($\mathcal{S}$, $\mathcal{X}$, $\mathcal{A}$, etc.)
- **Definition:** A class of sets $\mathcal{X}$ is closed under an operation (e.g. union, intersection, etc.) if when performed on any members of $\mathcal{X}$ yields a set which also belongs to the class.
- Example: $\mathcal{X} = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, A \cup B \cup C\}$ is closed under Union operation.

Fields

Definition: A class $\mathcal{X}$ of sets in Ω is called a **field** if

1. $\mathcal{X}$ is non-empty
2. $\mathcal{X}$ is closed under finite union
3. $\mathcal{X}$ is closed under complementation.

Examples:

- $\mathcal{X} = \{\emptyset, \Omega\}$ (trivial class)
- $\mathcal{X}$ = all subsets of Ω
- Let $\Omega = (-\infty, \infty)$, $\mathcal{X}_3$ = class of all finite intervals (a, b) where $a, b \in \mathbb{R}$.

Some Properties

Properties of a field $\mathcal{X}$:

1. $\mathcal{X}$ is also closed under finite intersection
2. $\emptyset \in \mathcal{X}$ and $\Omega \in \mathcal{X}$

σ-Fields

- **Definition:** A class $\mathcal{X}$ of sets is a σ**-field** if
  1. $\mathcal{X}$ is non-empty
  2. $\mathcal{X}$ is closed under countable unions
  3. $\mathcal{X}$ is closed under complementation
- Examples: (1) $\mathcal{X} = \{\emptyset, \Omega\}$ (2) $\mathcal{X}$ = all subsets of Ω are σ-fields
- Obviously, a σ-field is necessarily a field, but the converse does not hold: Consider class of all finite sets and sets whose complement is finite.
- **Theorem:** $\mathcal{X}_1$ and $\mathcal{X}_2$ are σ-fields, then $\mathcal{X}_1 \cap \mathcal{X}_2 = \{A : A \in \mathcal{X}_1 \text{ and } \mathcal{X}_2\}$ is also a σ- field.

Example

Let $\Omega = \mathbb{R}$ and $\mathcal{A} = \{A : \ A \text{ or } A^c \text{ is finite}\}$. Show that $\mathcal{A}$ is a field, but that $\mathcal{A}$ is not a σ-field. Note that $\mathcal{A}$ is the finite-cofinite field.

- Let $A_1 = \{1\}, A_2 = \{2\}, A_3 = \{3\}, \cdots,$
- Then $\bigcup_{n=1}^{\infty} A_n = \mathbb{N} \notin \Omega$.

σ-field Generated by a Class of Sets

- **Theorem:** Given a class of sets, $\mathcal{S}$, not necessarily a σ-field, there is a minimum σ-field (denoted $\sigma(\mathcal{S})$ containing) it, i.e. $\sigma(\mathcal{S})$ is a class of sets that:
  - is a σ-field
  - it contains $\mathcal{S}$: if $A \in \mathcal{S}$, then it is also in $\sigma(\mathcal{S})$.
  - if $\mathcal{X}$ is another σ-field that contains $\mathcal{S}$, then $\mathcal{X}$ contains $\sigma(\mathcal{S})$.

  $\sigma(\mathcal{S})$ is also called the **σ-field generated by** $\mathcal{S}$.
- $\sigma(\mathcal{S})$ also defined as the intersection of all σ-fields that contain $\mathcal{S}$
- Properties:
  - $\sigma(\mathcal{S})$ is itself a σ-field
  - $\mathcal{S} \subset \sigma(S)$
  - $\mathcal{S}_1 \subset \mathcal{S}_2$ implies $\sigma(\mathcal{S}_1) \subset \sigma(\mathcal{S}_2)$
  - if $\mathcal{S}$ is itself a σ-field, then $\sigma(\mathcal{S}) = \mathcal{S}$

Borel σ-fields ($\mathcal{B}$)

- Let $\Omega = \mathbb{R} = (-\infty, \infty)$. We have 4 types of finite intervals:
  1. $\mathcal{S}_1 = \{[a, b) : a < b, a, b \in \mathbb{R}\}$,
  2. $\mathcal{S}_2 = \{(a, b) : a < b, a, b \in \mathbb{R}\}$,
  3. $\mathcal{S}_3 = \{(a, b] : a < b, a, b \in \mathbb{R}\}$,
  4. $\mathcal{S}_4 = \{[a, b] : a < b, a, b \in \mathbb{R}\}$
- Let $\mathcal{S} = \bigcup_{i=1}^{4} \mathcal{S}_i$ = a class of all finite intervals. Note that $\mathcal{S}$ is neither a field nor a σ-field.
- We extend $\mathcal{S}$ to a σ-field through the following definition:

  Definition: The σ-field generated by $\mathcal{S}$ is called the **Borel σ-field on** $\mathbb{R}$ and is denoted by $\mathcal{B} = \sigma(\mathcal{S})$. Any set in $\mathcal{B}$ is called a **Borel Set**.
- Important: Do not try to characterize $\mathcal{B}$.
- We will return to this σ-field when we start talking about Random Variables.

Events... Again!

- Why do we care about σ-fields?
  They are key to understanding the definition of an event.
- **Definition:** A **measurable space** is a set Ω endowed with a σ-field $\mathcal{F}$ of subsets of Ω denoted by the pair $(\Omega, \mathcal{F})$.
- Sets in $\mathcal{F}$ are defined as **events**!
- In this class, we will often deal with the triplet $(\Omega, \mathcal{F}, P)$ where P is a probability measure on the space $(\Omega, \mathcal{F})$.
- For discrete (finite and countably infinite) Ω, $\mathcal{F} = 2^{\Omega}$, and for uncountable Ω, $\mathcal{F} = \mathcal{B}$.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 2.1

Examples for Sample Space and σ-fields

Definition: A σ-field on a set Ω is a collection of subsets of Ω that includes the empty subset, is closed under complement, and is closed under countable unions and countable intersections.

Examples for Sample Space and σ-fields:

1. **Experiment**: Toss a coin with a "Head" and a "Tail".
   Sample space Ω: $\{$Head, Tail$\}$
   σ**-field on** Ω: $\{\emptyset, \{\text{Head}\}, \{\text{Tail}\}, \{\text{Head, Tail}\}\}$.

2. **Experiment**: In a health survey among cancer patients, we ask each patient to report their perceived quality of life, categorized as $\{$Very Good, Good, Poor$\}$.
   Sample space Ω: $\{$Very Good, Good, Poor$\}$
   σ**-field on** Ω: $\{\emptyset, \{\text{Very Good}\}, \{\text{Good}\}, \{\text{Poor}\}, \{\text{Very Good, Good}\},$
   $\{\text{Very Good, Poor}\}, \{\text{Good, Poor}\}, \{\text{Very Good, Good, Poor}\}\}$.

Examples (cont.)

3. **Experiment**: Toss 2 coins and record the results from both tosses.
   Sample space Ω: $\{$HH, HT, TH, TT$\}$
   σ**-field on** Ω: $\{$ $\emptyset$, $\{$HH$\}$, $\{$HT$\}$, $\{$TH$\}$,$\{$HT$\}$, $\{$HH, HT$\}$, $\{$HH, TH$\}$,
   $\{$HH, TT$\}$, $\{$HT, TH$\}$, $\{$HT, TT$\}$,$\{$TH, TT$\}$,$\{$HH, HT, TH$\}$,
   $\{$HH, HT, TT$\}$, $\{$HH, TH, TT$\}$, $\{$HT, TH, TT$\}$, $\{$HH, HT, TH, TT$\}$ $\}$.

4. **Experiment**: Toss 2 coins and record the number of heads.
   Sample Space Ω:$\{$0, 1, 2$\}$
   σ**-field on** Ω: $\{$ $\emptyset$, $\{$0$\}$,$\{$1$\}$,$\{$2$\}$,$\{$0,1$\}$,$\{$0,2$\}$,$\{$1,2$\}$,$\{$0,1,2$\}$ $\}$.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 3

Intuitive probability

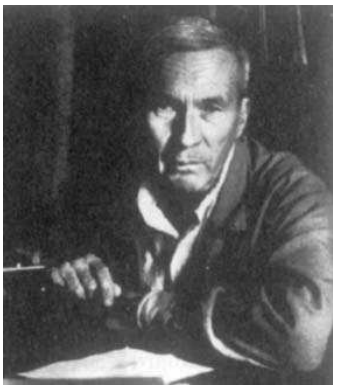
- **Random experiment**: Can be repeated indefinitely but future outcomes cannot be exactly predicted.
- **Stabilization of relative frequency**: Suppose we perform "independent" repetitions of a random experiment and count how many times an "event" E occurs. Let $f_n(E)$ be the number of occurrences in n repetitions. Then the *relative frequency* of E

$$r_n(E) = \frac{f_n(E)}{n}$$

  "converges" to some number as $n \to \infty$. We call this number $P(E)$, the probability of E.
- Examples: rolling of a die, spin the wheel.
- But what is "event", "independent", "convergence"?

Axiomatic probability

Andrey Kolmogorov (1903–1987)



Grundbegriffe der Wahrscheinlichkeitsrechnung (1933)
Basic Concepts of Probability Calculation
Foundations of the Theory of Probability

Kolmogorov's Axioms of Probability

Suppose $E \subset \Omega$ and we want to assign a probability to E.
$P(E)$ is a real-valued function of the event E, called the probability of E, that satisfies the following:

Axioms

i. Regularity: $P(\Omega) = 1$

ii. Non-negativity: $P(E) \geq 0$

iii. Countable Additivity: If $E_1, E_2, \ldots$ are mutually exclusive $(E_i E_j = \emptyset, \ \ i \neq j)$, then $P\big(\bigcup_{i=1}^{\infty} E_i\big) = \sum_{i=1}^{\infty} P(E_i)$ (countable additivity)

The last axiom contains a hidden assumption, namely that if $E_1, E_2, \ldots$ is an infinite sequence of events, then $\bigcup_{i=1}^{\infty} E_i$ is also an event. This is guaranteed if the events $E_1, E_2, \ldots$ belong to a σ-field.

Axioms continued

Definition: A probability measure P defined on a $\sigma-$field of subsets of Ω is a real-valued set function satisfying Kolmogorov's Axioms.

Definition - formal: A probability space is denoted by $(\Omega, \mathcal{A}, P)$ where Ω is the sample space, $\mathcal{A}$ refers to a $\sigma-$field of subsets of Ω, and P is a probability measure.

Probability calculus

If P is a probability function and A is any set in $\mathcal{A}$:

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$

If P is a probability function and A, B are sets in $\mathcal{A}$:

1. $P(A - B) = P(A) - P(A \cap B)$ (Note: C&B uses $\setminus$ instead of '-')
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. $A \subset B \Rightarrow P(A) \leq P(B)$

Law of Total Probability: If P is a probability function and $\{C_1, C_2, \dots\}$ is a partition of Ω:

$$P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$$

Monotone Sequence of Events

(i) If $\{E_n\}$ is an increasing sequence of sets and $E_n \in \mathcal{A}$, then $P(\lim_n E_n) = \lim_n P(E_n)$;
(ii) If $\{E_n\}$ is a decreasing sequence of sets, then $P(\lim_n E_n) = \lim_n P(E_n)$.

Proof of (i):

Boole's inequality (Union bound)

Suppose $E_1, E_2, \ldots, E_n$ are events.

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i)$$

Counting: Preliminaries 10 / 25

Discrete Probability

- **Discrete Sample Space:** A sample space is called **discrete** if it contains only finitely many points or infinitely many points which can be arranged into a simple sequence e_1, e_2, ..., i.e. a countable number of elements.
- $\mathcal{F} = 2^{\Omega}$.
- $P(\{e_1\}) + P(\{e_2\}) + \ldots = P(\Omega) = 1$, and more generally, for a set A that contains the points $a_1, a_2, \ldots, a_k$, then $P(A) = P(\{a_1\}) + P(\{a_2\}) + \ldots + P(\{a_k\})$.

Simplest Setting

Simplest setting:

1. Finite sample space with n elements, $\Omega = \{e_1, e_2, \ldots, e_n\}$
2. All sample points are equally probable.

In this case, $P(\{e_j\}) = \frac{1}{n}$, and for an event $\{a_1, \ldots, a_k\} = A \subset \Omega$,
$P(A) = kP(\{a_j\}) = \mu(A)/\mu(\Omega) = k/n$ where μ is the counting measure (counts the number of elements in a set).
$P(A)$ is simply the sum of the probabilities of each of the elements in A.

Counting and combinatorics will be concerned with identifying the number of individual elements within A.

Simple Examples

Examples: Suppose we have a standard deck of cards and we draw a single card. (probability of drawing any card is the same so they are equiprobable)

- Probability of drawing Jack of Hearts is 1/52
- Probability of drawing a Club is 13/52 = 1/4
- Probability of drawing an Ace is 4/52 = 1/13
- Probability of drawing an Ace or Heart is 16/52 = 4/13

More Preliminaries

- **Pairs:** With m elements $a_1, \ldots, a_m$, and n elements $b_1, \ldots, b_n$, it is possible to form nm pairs (a_j, b_k) containing one element from each group.
- **Multiplets:** Given r experiments that are to be performed such that the r^{th} experiment may result in any of n_r possible outcomes, then there is a total of $n_1 n_2 \ldots n_r$ possible outcomes of the r experiments.
- Examples:
  1. College committee consists of 3 freshman, 4 sophomores, 5 juniors, and 2 seniors. A subcommittee consisting of 1 person from each class is to be chosen. How many subcommittees are possible?
  2. How many 7-place license plates are possible if the first 3 places are to be occupied by letters and the final 4 by numbers?

Counting: Ordered Samples 15 / 25

Ordered Samples

- Many probability problems use these same ideas of counting in settings where repeated selections are taken from the same set of objects. The number of possible selection patterns (and hence the probability of a particular pattern) depends on
  1. whether or not items are replaced after selection
  2. whether or not the order of selection matters
- Suppose we have a "population" of n elements $a_1, \ldots, a_n$. Any ordered arrangement $a_{j_1}, \ldots, a_{j_r}$ of r symbols is called an *ordered sample of size* r.
  1. Sampling with replacement: each selection is made from the entire population (each of the r elements can be chosen in n ways. there are n^r possible samples)
  2. Sampling without replacement: an element, once chosen, is removed from the population. (sample cannot exceed size n)

Sampling with Replacement

Draw	Set
1	$A_1, A_2, \ldots, A_n$
2	$A_1, A_2, \ldots, A_n$
$\vdots$	
r	$A_1, A_2, \ldots, A_n$

Theorem: For a population of n elements and a prescribed sample size r, there exist n^r different samples with replacement

Example: The Braille alphabet has 6 locations that can be raised or not raised. How many combinations of raised/not raised locations are there?

Why does this differ from the actual number of Braille letters?

Sampling without Replacement

Draw	Set	
1	$A_1, A_2, \ldots, A_n$	
2	$B_1, B_2, \ldots, B_{n-1}$	$B_i's$ are $A_i's$ not drawn
3	$C_1, C_2, \ldots, C_{n-2}$	$C_i's$ are $B_i's$ not drawn
$\vdots$		
r	$X_1, X_2, \ldots, X_{n-(r-1)}$	$X_i's$ are remaining items after first $r-1$ draws

- **Theorem:** For a population of n elements and a prescribed sample size r, there exist $n(n-1)(n-2)...(n-r+1) = \frac{n!}{(n-r)!}$ different samples without replacement.
- An important special case occurs when $r = n$ (define 0!=1):

$$\frac{n!}{(n-r)!} = \frac{n!}{(n-n)!} = n(n-1)(n-2)\cdots 2 \cdot 1 = n!$$

This is the number of **permutations** of a set of size n

Examples

1. You want to take a photo of six of your friends. How many ways can you arrange them in a row?
2. A competition has 10 competitors of which 4 are Russian, 3 are from the USA, 1 is from Brazil, and 2 are from China. If the tournament result lists just the nationalities of the players in which they placed, how many outcomes are possible?
3. You want to invite those same 6 friends to dinner. How many ways can you arrange them at your round dining table?
4. 4 Martians, 3 Plutonians and 5 Jupiterians are to be seated in a row at an interplanetary conference. How many seating arrangements are possible?
5. With indistinguishable elements: How many different letter arrangements can be formed using the letters B-A-N-A-N-A?

Examples (1)

Suppose one has a population of size n. What is probability that a particular element will not be in sample of size r?

Solution *Without* Replacement:

$$\frac{(n-1)!}{(n-1-r))!} = \text{No. of samples without particular element.}$$

$$\frac{n!}{(n-r)!} = \text{No. of Samples}$$

$$\begin{aligned} q &= \text{Probability element is not in sample} \\ &= \frac{(n-1)!/(n-(r+1))!}{n!/(n-r)!} \\ &= \frac{(n-1)(n-2)\dots(n-r)}{n(n-1)(n-2)\dots(n-r+1)} \\ &= \frac{n-r}{n} = 1 - \frac{r}{n} \\ p &= \text{Probability element in the sample} = 1 - q = r/n \end{aligned}$$

Examples (2)

Suppose one has a population of size n. What is probability that a particular element will not be in sample of size r?

Solution *With* Replacement:

$$\begin{aligned} q &= \frac{(n-1)^r}{n^r} = \left(1 - \frac{1}{n}\right)^r \\ p &= 1 - q = 1 - \left(1 - \frac{1}{n}\right)^r \end{aligned}$$

Examples (3)

Suppose one samples r items with replacement from a total of n. What is probability of no repetitions in the sample?

$$\begin{aligned} q &= \frac{\text{No. of Samples with No Repetition}}{\text{No. of Samples}} = \frac{n!/(n-r)!}{n^r} \\ &= \frac{n \cdot (n-1) \cdot (n-2) \ldots (n-r+1)}{n \;\; n \;\; n \ldots n} \\ &= 1(1 - \frac{1}{n})(1 - \frac{2}{n}) \cdot \cdot (1 - \frac{(r-1)}{n}) \\ &\sim (1 - \frac{1+2+\ldots+r-1}{n}) \\ &= 1 - \frac{r(r-1)}{2n} \end{aligned}$$

because $1 + 2 + \ldots + r - 1 = \frac{r(r-1)}{2}$

Note: $p =$ Probability of repetitions $= 1 - q \sim \frac{(r-1)r}{2n}$

Examples (4): Birthday Problem

Application - the Birthday Problem. If there is a class of r people, what is the probability that some have birthdays on the same day? $(n = 365)$

$$p \sim \frac{r(r-1)}{2 \cdot 365}$$

Suppose $r = 25$

$$p \sim \frac{25 \cdot 24}{2 \cdot 365} = 0.82$$

Suppose $p = 1/2$. What is value of r?

$$r(r-1)/730 = 1/2 \qquad \text{Solution is } \; r \approx \frac{39}{2} \sim 20$$

Additional Reading 24 / 25

Additional Reading

See Chapter 1.2 in Casella and Berger.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 4

Combinations (Subpopulations)

How many ways can we select r objects from n paying no attention to order? Continue to suppose that the n items are distinct.

$n = 4$	A, B, C, D
$r = 2$	Possible ordered samples: $(A,B),(A,C),(A,D),(B,C),(B,D),(C,D)$ $(B,A),(C,A),(D,A),(C,B),(D,B),(D,C)$
$n!/(n-r)!$	= $4!/2! = 4 \cdot 3 = 12$

6 = No. of ways of choosing 2 objects from among 4 objects with no ordering... Why?

There are 4*3 = 12 ways of choosing ordered pairs. Then there are 2! = 2*1 orderings that we don't really care about, so 12/2 = 6.

Combinations (2)

No. of ordered samples = No. of Combinations $\times$ No. of ways of permuting each ordered sample.

$$\begin{aligned}
\frac{n!}{(n-r)!} &= c \cdot r! \\
c &= \frac{n!}{(n-r)!r!} = \binom{n}{r} = \binom{n}{n-r} \\
\binom{n}{r} &= \text{No. of ways of choosing } r \text{ objects from among } n \\
&= \text{we say "}n\text{ choose }r\text{"}
\end{aligned}$$

Theorem: A population of n elements possesses $\binom{n}{r}$ different combinations (subpopulations) of size $r \leq n$, i.e. $\binom{n}{r}$ represents the number of possible combinations of n objects taken r at a time.

Conventions

Conventions:

$$0! = 1; \binom{n}{0} = 1\ ; \quad \binom{n}{r} = 0 \qquad \text{if } r > n,\ r < 0$$

Examples

Applications of $\binom{n}{r} = \frac{n!}{r!(n-r)!}$

- From a group of 5 women and 7 men, how many different committees consisting of 2 women and 3 men can be formed?
- How many different poker hands with 5 cards are possible?
- What is the *probability* of 4 aces in a poker hand with 5 cards?

Examples(2)

- What is the *probability* of 4 of a kind (e.x. 4 Kings, 4 Queens, etc)?

- What is the *probability* of (exactly) 3 of a kind with no other pairs?
  Size of Event:

Binomial Theorem

Suppose n is a positive integer. The Binomial Theorem says

$$(a+b)^n = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) a^r b^{n-r}$$

Proof: Homework, by induction.

Note that if $a+b=1$ one has

$$\sum_{r=0}^{n} \binom{n}{r} a^r (1-a)^{n-r} = 1,$$

and that if $a=b=1$,

$$\sum_{r=0}^{n} \binom{n}{r} = 2^n.$$

Note: the left hand side is the number of all possible subsets from $\{X_1, \ldots, X_n\}$. The above equality tells us that the number of all possible subsets from $\{X_1, \ldots, X_n\}$ is 2^n, which motivates the notation 2^{Ω} for indicating power set .

Pascal's Triangle (1)

Pascal's triangle is an illustration of the following result:

$$\begin{pmatrix} n+1 \\ r \end{pmatrix} = \begin{pmatrix} n \\ r-1 \end{pmatrix} + \begin{pmatrix} n \\ r \end{pmatrix}$$

Proof: Homework.

Pascal's Triangle (2)

Pascal's triangle

$$\begin{array}{ccccccccccc}
 & & & & \binom{0}{0}=1 & & & & & & \\
 & & & \binom{1}{0}=1 & & & \binom{1}{1}=1 & & & & \\
 & & \binom{2}{0}=1 & & & \binom{2}{1}=2 & & & \binom{2}{2}=1 & & \\
 & \binom{3}{0}=1 & & & \binom{3}{1}=3 & & & \binom{3}{2}=3 & & & \binom{3}{3}=1 \\
 & & & & 1 & & & & & & \\
 & & & 1 & & 1 & & & & & \\
 & & 1 & & 2 & & 1 & & & & \\
 & 1 & & 3 & & 3 & & 1 & & & \\
1 & & 4 & & 6 & & 4 & & 1 &
\end{array}$$

Partitions

- Beyond a single subpopulation...
- **Theorem:** Let $r_1, r_2, \cdots, r_k$ be nonnegative integers such that

$$r_1 + r_2 + \cdots + r_k = n$$

The number of ways in which a population of n elements can be divided into k ordered groups (partitioned into k-subpopulations) of which the first contains r_1 elements, the second r_2 elements, etc., is

$$\frac{n!}{r_1! r_2! \cdots r_k!}$$

- Example: Ten children are being divided into an A team and a B team of 5 each. How many different divisions are possible?

Occupancy Problem

Suppose we have n indistinguishable balls which we wish to place in r distinguishable urns. How many different outcomes are possible?

- We can describe the outcome by a vector $(x_1, \ldots, x_r)$ where $x_j \geq 0$ is the number of balls in the jth urn and $\sum_{j=1}^{r} x_j = n$. We are therefore looking for the number of possible vectors.
- We can further think of this as taking n indistinguishable items and dividing them into r groups.
- First consider the situation that each urn should have at least 1 ball. In this situation, we are dealing with $x_j > 0$ ($j = 1, \cdots, r$). There are $n - 1$ spaces between the objects, and we select $r - 1$ of them to choose our dividing points. Therefore, there are $\binom{n-1}{r-1}$ different partitions.
- Now suppose that some of the urns are allowed to be empty. In this case we can imagine that we have n objects plus $r - 1$ dividers. The number of possible partitions is therefore the number of (non-redundant) orderings of these $n + r - 1$ items which is equal to $(n + r - 1)!/[n!(r - 1)!] = \binom{n+r-1}{r-1}$

Occupancy: Results

- **Theorem:** there are $\binom{n-1}{r-1}$ distinct positive integer valued vectors $(x_1, \ldots, x_r)$ satisfying

$$x_1 + \cdots + x_r = n, \quad x_j > 0, j = 1, \ldots, r$$

- **Theorem:** there are $\binom{n+r-1}{r-1}$ distinct nonnegative integer-valued vectors $(x_1, \ldots, x_r)$ satisfying

$$x_1 + \cdots + x_r = n, \quad x_j \geq 0, j = 1, \ldots, r$$

Examples

1. How many distinct nonnegative integer-valued solutions of $x_1 + x_2 = 3$ are possible?

2. The state health department has 20 cases of flu vaccine to distribute among 4 possible hospitals. If all of the vaccine must distributed, how many different ways can the 20 cases be distributed if not every hospital has to receive some? What if the department wants to have the flexibility to hold some in reserve and not every hospital has to receive some?

Stirling's Approximation

- **Stirling's Formula:**

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$$

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 5

Conditional Probability

If $P(B) > 0$ we define the *conditional probability* of the event A given B as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Intuitively, conditioning on B means reducing the original sample space S to B, which becomes the new, reduced sample space. All probabilities are computed with respect to B. Notice:

$$P(A|\Omega) = \frac{P(A \cap \Omega)}{P(\Omega)} = P(A)$$

Disjoint events: If $A \cap B = \emptyset$, then $P(A|B) = 0$ and $P(B|A) = 0$.

cont.

Conditional probability satisfies the axioms of probability:

1. $P(\Omega|B) = 1$
2. $P(A|B) \geq 0$
3. If $A_1, A_2, \ldots$ are mutually exclusive events, then $P\big(\bigcup_{i=1}^{\infty} A_i \big| B\big) = \sum_{i=1}^{\infty} P(A_i|B)$

and all the other properties:

1. $P(\emptyset|B) = 0$
2. $P(A|B) \leq 1$
3. $P(A^c|B) = 1 - P(A|B)$

etc.

Independence

Two events A and B are said to be *independent* if

$$P(A \cap B) = P(A)P(B). \tag{1}$$

Why? Because then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$
$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$$

This could have been taken as the definition, but (1) is easier to generalize.
If A and B are independent, then so are

- A^c and B
- A and B^c
- A^c and B^c

* Can two disjoint events be independent and vice versa? (HW)

Independence of many events

The events $A_1, A_2, \ldots, A_n$ are mutually independent if for *every* subcollection $A_{i_1}, \ldots, A_{i_k}$ of size $k = 2, \ldots, n$

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Notice that is a very strong condition. But it is necessary to ensure that

$$P(A_j|A_{i_1}, \ldots, A_{i_k}) = P(A_j)$$

for every j and every subcollection $A_{i_1}, \ldots, A_{i_k}$ that does not include A_j. Pairwise independence is not enough.

Independence of many events (cont.)

Example: Toss a coin three times. Sample space Ω={HHH, HHT, HTH HTT, THH, THT, TTH, TTT}. Define the events:

- H_1 = { The outcome of the first toss is heads }
  ={HHH, HHT, HTH, HTT}
- H_2 = { The outcome of the second toss is heads }
  ={HHH, HHT, THH, THT}
- H_3 = { The outcome of the third toss is heads }
  ={HHH, HTH, THH, TTH}

Suppose every outcome is equally likely. Then these events are independent.
$H_1 \cap H_2 = \{HHH, HHT\}$; $H_1 \cap H_3 = \{HHH, HTH\}$; $H_2 \cap H_3 = \{HHH, THH\}$;
$P(H_1) = P(H_2) = P(H_3) = 4/8 = 1/2$; $P(H_1 \cap H_2) = 2/8 = 1/4 = P(H_1)P(H_2)$;
$H_1 \cap H_2 \cap H_3 = \{HHH\}$; $P(H_1 \cap H_2 \cap H_3) = 1/8 = P(H_1)P(H_2)P(H_3)$)

Independence of many events (cont.)

Now define the events:

- A_{12} = { The outcome of the first toss equals the second }
- A_{13} = { The outcome of the first toss equals the third }
- A_{23} = { The outcome of the second toss equals the third }

These events are pairwise independent but not mutually independent.

(A_{12}= { HHH, HHT, TTH, TTT }; A_{13}= {HHH, HTH, THT, TTT }; A_{23}={HHH, HTT, THH, TTT};
$P(A_{12}) = P(A_{13}) = P(A_{23}) = 4/8 = 1/2$;
$A_{12} \cap A_{13}$= {HHH, TTT };
$P(A_{12} \cap A_{13}) = 2/8 = 1/4 = P(A_{12})P(A_{13})$.

On the other hand,
$A_{12} \cap A_{13} \cap A_{23}$={HHH, TTT};
$P(A_{12} \cap A_{13} \cap A_{23}) = 2/8 = 1/4 \neq P(A_{12})P(A_{13})P(A_{23})$.)

Sequential conditioning

By the definition of conditional probability:

$$P(A \cap B) = P(A)P(B|A)$$
$$P(A \cap B) = P(B)P(A|B)$$

This is useful for computing probabilities of sequential events.
E.g: What is the probability of dealing two aces in a row?

More generally:

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\ldots P(A_n|A_1\ldots A_{n-1})$$

The Birthday Problem

In a group of n students in a class, what is the probability that at least two have the same birthday?

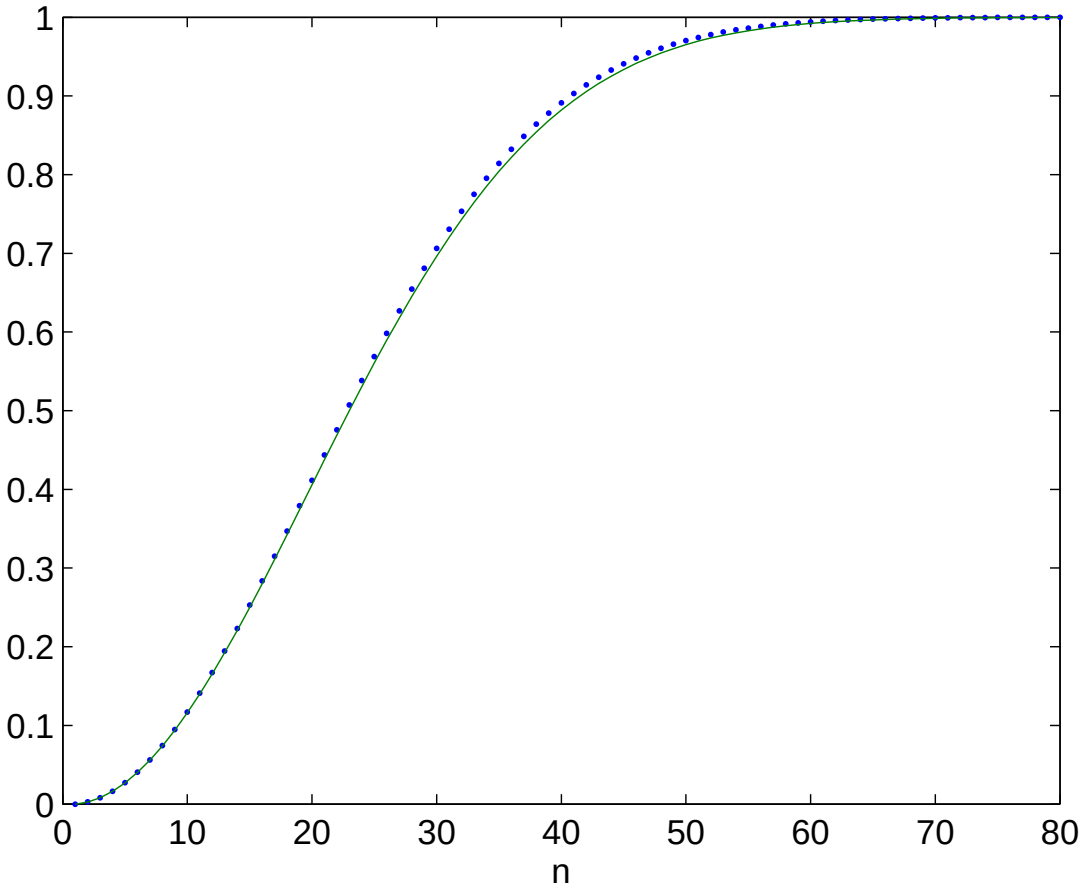
Solution:
Suppose we order the n students in an arbitrary order. Let D_j be the event that the first j have different birthdays. Based on page 22 in Notes 3, we have

$$P(D_j) = \frac{\text{No. of Samples with No Repetition}}{\text{No. of Samples}} = \frac{365!/(365-j)!}{365^j}$$

Let $D_j = \{A_1, A_2, A_3, \ldots, A_j\}$, where A_1 is the birth day of the first person, A_2 is the birth day of the second person, etc., and all the A_i ($i = 1, 2, \ldots, j$) are different. Based on sequential conditioning,

$$\begin{aligned} & P(D_j) = P(\{A_1, A_2, A_3, \ldots, A_j\}) \\ = \; & P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\ldots P(A_j|A_1\ldots A_{j-1}) \\ = \; & \frac{365}{365}\frac{365-1}{365}\cdots\frac{365-j+1}{365} = \frac{365!/(365-j)!}{365^j} \end{aligned}$$

The Birthday Problem



Turning around probabilities

Also by the definition of conditional probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

This is useful for computing conditional probabilities when the reverse conditioning is easier to compute.
E.g.: Prob. that it will rain given that it is thundering vs. prob. that it thundered given that it is raining.

(2) is called sometimes **Bayes' rule**. It is often used in a context where we want to know the probability that a particular hypothesis is true. We have an *a priori* belief in whether or not the hypothesis is true, then update that probability by collecting data.
E.g.: Suppose a priori boys are equally likely to be born as girls. Say 90% of boys play with trucks. Baby X plays with trucks. What is the probability that Baby X is a boy?

Decomposition Formula (Total Probability)

Let $\{A_1, A_2, \ldots\}$ be a partition of Ω. Let B be any subset in Ω. Then

$$P(B) = \sum_{i=1}^{\infty} P(A_i)P(B|A_i)$$

Proof:

(So called) Bayes' Theorem

Let $\{A_1, A_2, \ldots\}$ be a partition of Ω. Let B be any subset in Ω. Then

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}$$

Proof:

Bayes and Screening

An important application of Bayes' theorem is screening.
Notation:

Let D be the disease:

D means diseased
$\overline{D}$ means "no disease"

and T be the diagnostic test:

T^+ means a positive test
T^- means a negative test.

Then we have that the **positive predictive value**

$$\begin{aligned} P(D|T^+) &= \frac{P(D)\,P(T^+|D)}{P(D)\,P(T^+|D) + P(\overline{D})\,P(T^+|\overline{D})} \\ &\equiv \frac{\text{prevalence} \times \text{sensitivity}}{\text{prev.} \times \text{sens.} + (1 - \text{prev.}) \times (1 - \text{specificity})} \end{aligned}$$

Papanicolaou Example

Let D be cervical cancer.
$P(D)$ we'll take to be 1 in 21,000, which is the approximate annual incidence rate in the US (SEER 2002 estimate).

$$P(D) = .00004762$$

Let us take the sensitivity ($P(T^+|D)$) to be 0.71 and the specificity ($1 - P(T^+|\overline{D})$) to be 0.75.
Thence the positive predictive value is:

$$\begin{aligned} P(D|T^+) &= \frac{0.00004762 \times 0.71}{0.00004762 \times 0.71 + (1 - 0.00004762) \times (1 - 0.75)} \\ &= 0.000135 \end{aligned}$$

That means that for every 1,000,000 positive results, about 135 truly have cervical cancer. But for any particular patient, testing positive increases the probability of having the disease by a factor of 2.8!

Relative risks and relative odds

Given two conditions–smokers (S) and non-smokers ($\overline{S}$), say–then we say the *relative risk* of a disease (D)–lung cancer, say–due to smoking is:

$$RR = \frac{P(D|S)}{P(D|\overline{S})}$$

The *relative odds* of the disease (D) due to smoking is:

$$OR = \frac{\frac{P(D|S)}{1-P(D|S)}}{\frac{P(D|\overline{S})}{1-P(D|\overline{S})}}$$

Of course, if $P(D|S) \approx 0$ and $P(D|\overline{S}) \approx 0$, (*rare disease*) then

$$OR \approx \frac{\frac{P(D|S)}{1}}{\frac{P(D|\overline{S})}{1}} = RR$$

Bayes and Case Control Studies

Consider

$$OR(D|S) \equiv \frac{P(D|S)}{1-P(D|S)} / \frac{P(D|\overline{S})}{1-P(D|\overline{S})}$$

Now consider the numerator, which from Bayes theorem,

$$\frac{\frac{P(D)P(S|D)}{P(S)}}{\frac{P(\overline{D})P(S|\overline{D})}{P(S)}} = \frac{P(D)P(S|D)}{P(\overline{D})P(S|\overline{D})}$$

Do the same for the denominator, and get,

$$OR(D|S) = \frac{\frac{P(D)P(S|D)}{P(\overline{D})P(S|\overline{D})}}{\frac{P(D)P(\overline{S}|D)}{P(\overline{D})P(\overline{S}|\overline{D})}} = \frac{\frac{P(S|D)}{P(\overline{S}|D)}}{\frac{P(S|\overline{D})}{P(\overline{S}|\overline{D})}} = OR(S|D)$$

Additional Reading

See Chapter 1.2-1.3 in Casella and Berger.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 6

Random Variables

Suppose we start with a probability space $(\Omega, \mathcal{A}, P)$. Instead of referring to outcomes and events observed from the sample space Ω, it is often convenient to assign a number to each possible outcome and record that instead.
Example: Flip a coin three times.

- $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- $\mathcal{A} = 2^{\Omega}$
- Define a random variable Y to be the number of heads.
- $Y(HHH) = 3, Y(HHT) = 2, Y(HTT) = 1$, etc.

Random Variables (Formal Definition)

Definition: A random variable Y is a real-valued and measurable function defined on a probability space. That is, $Y : \Omega \to \mathbb{R}$.
Every point ω in Ω maps to a point in $\mathbb{R}$, namely $Y(\omega)$.
Conversely, we define the *inverse image* under Y of a subset B of $\mathbb{R}$ as

$$Y^{-1}(B) = \{\omega : Y(\omega) \in B\}$$

The definition of a random variable requires that the inverse image of every Borel set $B \subset \mathbb{R}$ is an element of $\mathcal{A}$. This property allows us to assign probabilities to random variables. More precisely,

$$P\{Y \in B\} = P\{Y^{-1}(B)\}$$

Conventions

A Random Variable is a set function which takes values on the real line (for now). Often the argument is omitted and one writes Y instead of $Y(\omega)$.

Random variables are usually denoted by capital letters (e.g. Y).
Values which random variable can take on are denoted by lower case letters (e.g. y).

Example: Coin Tosses

$$\Omega : \{H, T\}$$
$$Y(H) = 1, \quad Y(T) = 0$$

If $P(\text{head}) = .5$, then $P(Y = 1) = .5$

Examples

Roll of a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
Define $Y(\omega) = \omega$
$P\{Y(\omega) = \omega\} = P(\omega) = 1/6$

An artificial example:
Suppose $\Omega = \{\omega : 0, \ \pm 1, \ \pm 2, \ \ldots\}$
Define

$$\begin{aligned} Y(\omega) = a \quad &\text{if} \quad \omega \leq 0 \\ Y(\omega) = b \quad &\text{if} \quad \omega > 0 \\ P\{Y = a\} \quad &= \quad \sum_{i=-\infty}^{0} P(i) \quad , \\ P\{Y = b\} \quad &= \quad \sum_{i=1}^{\infty} P(i) \end{aligned}$$

Cumulative Distribution Functions

Distribution Functions are used to describe the behavior of a rv.

Definition The *cumulative distribution function* (*cdf*) of a random variable Y is a real valued function $F(y)$ defined by

$$F(y) = P\{Y \leq y\} = P\{\omega : Y(\omega) \leq y\}$$

Example: cdf of a die: $F(y) = y/6$

Definition The *survival function* of Y is defined by

$$S(y) = 1 - F(y) = P(Y > y)$$

Some Properties of $F(y)$

1. $0 \leq F(y) \leq 1$
2. $\lim_{y \to -\infty} F(y) = 0$
3. $\lim_{y \to \infty} F(y) = 1$
4. F is nondecreasing: i.e. if $a < b$, then $F(a) \leq F(b)$
5. F is right continuous: that is, for any b and any decreasing sequence $b_n, n \geq 1$ that converges to $b, \lim_{n \to \infty} F(b_n) = F(b)$
6. $P\{a < Y \leq b\} = F(b) - F(a)$

These properties can all be proved using the properties of probability measures.

Induced Probability Space

All probability questions about a random variable can be answered via its *cdf*.
Every random variable defined on a probability space induces a probability space on $\mathbb{R}$:

$$(\Omega, \mathcal{A}, P) \longrightarrow \quad Y(\omega) \longrightarrow \quad (\mathbb{R}, \mathcal{B}, F(\cdot))$$

- Points in Ω are transformed to points on $\mathbb{R}$ (Real line)
- Sets (events in $\mathcal{A}$) are mapped into intervals on real line, i.e., into members of the Borel sets, $\mathcal{B}$.
- P is replaced by $F(\cdot)$.

Because of this, the abstract notion of a sample space recedes, and attention is usually given primarily to random variables and their distributions.
We will sometimes refer to the 'sample space' of a random variable, which will be taken to be the values in $\mathbb{R}$ that a random variable takes on.

Identically distributed rvs

The cdf does not contain information about the original sample space.

Example: Toss a coin n times. The number of heads and number of tails have the same distribution.

Definition: Two rvs X and Y are identically distributed if for every Borel set $A \subset \mathbb{R}$, $P(X \in A) = P(Y \in A)$.

Theorem C&B 1.5.10 The following two statements are equivalent:

a. The rvs X and Y are identically distributed
b. $F_X(x) = F_Y(x)$ for every x.

Note that two rvs can have the same distribution even if they are not equal to one another.

The distinction between two rvs being equal and having the same distribution will become important later in questions of convergence.

Types of Random Variables

A random variable Y can be

- *discrete* - Y takes on a finite or countably infinite number of values
- *continuous* - the range of Y consists of subsets of the real line.
- *mixed* - best to see this with an example

Example of a mixed random variable:
Consider a random variable with *cdf* given by:

$$F(x) = \begin{cases} 0 & x < 0 \\ x/2 & 0 \leq x < 1 \\ 2/3 & 1 \leq x < 2 \\ 11/12 & 2 \leq x < 3 \\ 1 & 3 \leq x \end{cases}$$

Discrete random variables 13 / 26

Discrete random variables

If a random variable Y takes only a finite or countable number of values, then the *cdf* can be expressed as:

$$F(y) = P\{Y \leq y\} = \sum_{z \leq y} P\{Y = z\}$$

If the sample space of Y is $\Omega = \{y_1, y_2, ...\}$, then

$$F(y) = \sum_{y_i \leq y} P\{Y = y_i\}$$

Probability mass function

Definition The *prob. mass function* (*pmf*) or *frequency function* is a function $f(y)$ defined by

$$f(y) = P\{Y(\omega) = y\}$$

Thus, we can write $F(y) = P\{Y \leq y\} = \sum_{z \leq y} f(z)$.

If the sample space of Y is $\Omega = \{y_1, y_2, ...\}$, then

$$f(y_i) = P(Y = y_i) = P(y_{i-1} < Y \leq y_i) = F(y_i) - F(y_{i-1})$$

Example: Suppose Y is a random variable that takes the values 0, 1 or 2 with probability .5, .3, and .2, respectively. Then

$f(0) = 0.5, f(1) = .3$, and $f(2) = .2$.

Properties of the pmf

Definition: The *domain* of a random variable Y is the set of all values of y for which $f(y) > 0$.

Properties of the pmf:

1. $f(y) > 0$ for at most a countable number of values y. For all other values y, $f(y) = 0$.
2. Let $\{y_1, y_2, \ldots\}$ denote the domain of Y. Then

$$\sum_{i=1}^{\infty} f(y_i) = 1$$

An obvious consequence is that $f(y) \leq 1$ over the domain.

Example: What is the pmf of a deterministic rv (a constant)?

$f(x) = 1$ for $x = k$ and $f(x) = 0$ for $x \neq k$.

Example

In many applications, a formula can be used to represent the *pmf* of a random variable. Suppose Y can take values 1,2 ... with *pmf*

$$f(y) = \begin{cases} \frac{1}{y(y+1)} & y = 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

How would we determine if this is an allowable *pmf*?

$$\begin{aligned} \sum_{y=1}^{\infty} \frac{1}{y(y+1)} &= \sum_{y=1}^{\infty} \left(\frac{1}{y} - \frac{1}{y+1} \right) \\ &= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots \\ &= 1 - \lim_{y \to \infty} \frac{1}{y} \\ &= 1 \end{aligned}$$

Continuous Random Variables 18 / 26

Continuous Random Variabes

Recall that a random variable is a function that maps a probability space $\{\Omega, \mathcal{A}, P\}$ to the real line, thereby inducing a new probability space:

$$\{\Omega, \mathcal{A}, P\} \longrightarrow \quad Y \longrightarrow \quad (\mathbb{R}, \mathcal{F}, F(\cdot))$$

We have discussed the setting where Y is discrete (i.e. Y can take on a finite or countably infinite number of values).
A random variable Y is called *continuous* if its distribution function $F(y) = P(Y \leq y)$ is a continuous function.

Absolute Continuity

The distribution of a continuous random variable is characterized by the probability of falling in intervals, e.g. $P(Y \in (a, b])$.
We will focus on *absolutely continuous* random variables.

Definition: A function $F(y)$ is *absolutely continuous* if it can be written

$$F(y) = \int_{-\infty}^{y} f(x)dx,$$

where for now, you may interpret $\int$ as the usual Riemann integral.

A random variable is said to be absolutely continuous if its distribution function is absolutely continuous.

Note: Absolute continuity is stronger than continuity but weaker than differentiability. An example of an absolutely continuous function is one that is differentiable everywhere except for a countable number of points.

cdf — density relation

If $F(y)$ is absolutely continuous, $f(y)$ is called the *probability density function (pdf)* of Y and

$$F'(y) = \frac{dF(y)}{dy} = f(y).$$

Building on this idea,

$$P(a < Y \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

More generally, for a set B,

$$P(Y \in B) = \int_{B} f(x)dx$$

Note that of course B has to be an "allowable" subset of the real line $\mathbb{R}$, that is, a Borel set.

Properties

In general, a function $f(x)$ is a *pdf* iff

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Examples:

- Suppose $F(x) = 1 - e^{-\lambda x}$ for $x > 0$ and $F(x) = 0$ otherwise. Is $F(x)$ a cdf? What is the associated pdf?
  - $F(x)$ is continuous and nondecreasing, and $lim_{x\to-\infty}F(x) = 0,\ \ lim_{x\to+\infty}F(x) = 1$, so $F(x)$ is a cdf.
  - $f(x) = \frac{d}{dx}F(x) = \lambda e^{-\lambda x}$
- What about $f(x) = 1/x^r$ for $x > 1$ and $f(x) = 0$ otherwise?

$$\int_1^{\infty} \frac{dx}{x^r} = \left[\frac{x^{1-r}}{(1-r)}\right] = \frac{1}{r-1}$$

  Then $f(x)$ is not a pdf but $(r-1)f(x)$ is a pdf.

Notes

- $f(x)$ is not the probability that $Y = x$. In fact, if Y is an absolutely continuous random variable with density function $f(x)$, then $P(Y = x) = 0$. Why?

$$\begin{aligned} P(Y = x) &= \lim_{h\to 0}\int_{x-h}^{x+h} f(u)du \\ &= \lim_{h\to 0}[F(x+h) - F(x-h)] \\ &= F(x+) - F(x-) \\ &= 0 \end{aligned}$$

Notes (cont.)

- More generally, if B is a subset of $\mathbb{R}$ with

$$\int_B dx = 0,$$

  then if Y is an absolutely continuous random variable defined on $\mathbb{R}$, then $P(Y \in B) = 0$ also.
- Because $P(Y = a) = 0$, all the following are equivalent:

$$P(a \leq Y \leq b), \quad P(a \leq Y < b) \quad \text{and} \quad P(a < Y < b)$$

- Also, note that $f(x)$ can exceed one!

Notes (cont.)

- $f(x)$ can be interpreted as the *relative* probability that Y takes the value x. Why? By the mean value theorem, we can say

$$P(x < Y \leq x + \Delta) \approx f(x)\Delta$$

  Thus

$$P(Y \in \text{interval of width } \Delta \text{ centered at } a) \approx f(a)\Delta$$

  and

$$P(Y \in \text{interval of width } \Delta \text{ centered at } b) \approx f(b)\Delta$$

  Hence, if $f(b) > f(a)$, we can say that it is more likely for Y to take the values near b rather than near a.

Stochastic ordering

Suppose Y is a rv and define $X = Y + 2$. Then $X > Y$ always.

Now suppose

$$X \sim F_X(t) = (1 - e^{-t})\ 1(t > 0)$$
$$Y \sim F_Y(t) = (1 - e^{-2t})\ 1(t > 0)$$

Then X is not always greater than Y, but it is likely to be.

Definition: X is *stocastically greater* than Y if

$$F_X(t) \leq F_Y(t) \text{ for all } t$$
$$F_X(t) < F_Y(t) \text{ for some } t$$

or equivalently

$$P(X > t) \geq P(Y > t) \text{ for all } t$$
$$P(X > t) > P(Y > t) \text{ for some } t$$

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 7

Functions of Random Variables

(C-B Chap 2.1 & Gut I.2)

If X is a rv with sample space $\mathcal{X} \subset \mathbb{R}$ and cdf $F_X(x)$ then any function of X, say $Y = g(X)$ is also a random variable. The new random variable Y has a new sample space $\mathcal{Y} = g(\mathcal{X}) \subset \mathbb{R}$. The objective is to find the cdf $F_Y(y)$ of Y.

Example: Suppose X is an exponential random variable with parameter 1, i.e. $F_X(x) = 1 - e^{-x}$, $f_X(x) = e^{-x}$. What is the distribution of $Y = X/\lambda$?

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X/\lambda \leq y) = P(X \leq \lambda y) \\ &= F_X(\lambda y) = 1 - e^{-\lambda y} \end{aligned}$$

Y is distributed $\exp(\lambda)$ with density $f_Y(y) = \lambda e^{-\lambda y}$.

Simple example

Example: Change units – Fahrenheit to Celsius

$$C = \frac{5}{9}(F - 32)$$

Ranges:

$$20^\circ C < C < 30^\circ C \;\Leftrightarrow\; 68^\circ F < F < 86^\circ F$$

$$(\mathcal{Y}, \mathcal{B}, F_Y) \leftarrow g(\cdot) \leftarrow (\mathcal{X}, \mathcal{B}, F_X)$$

Probability mapping

For any Borel set A:

$$\begin{aligned} P(Y \in A) &= P(g(X) \in A) \\ &= P(\{x \in \mathcal{X} : g(x) \in A\}) \\ &= P(X \in g^{-1}(A)). \end{aligned}$$

where we have defined

$$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}$$

Notice that $g^{-1}(A)$ is well defined even if $g(\cdot)$ is not bijective (one-to-one).

Example: Let $g(x) = x^2$.
Then

$$g^{-1}([-1, 1]) = [-1, 1]$$

But

$$g(g^{-1}([-1, 1])) = [0, 1]$$

Discrete RVs

Suppose that X is a discrete random variable with probability mass function $p(x) = P(X = x)$. Then, the *pmf* of a 1-1 transformation $Y = g(X)$ is given by

$$P(Y = y) = P(g(X) = y) = P(\{x : g(x) = y\}) = \sum_{x:g(x)=y} p(x)$$

In practice, one never sees many general results about transformations of discrete random variables because the results are so simple!

Continuous RVs

Consider the transformation $Y = g(X)$ where $g(x)$ is strictly increasing (consequently a one-to-one transformation), and suppose g is differentiable. This means that we can also define the *inverse function*, $g^{-1}(y)$.

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\{g(X) \leq y\} \\ &= P\{X \leq g^{-1}(y)\} = F_X(g^{-1}(y)). \end{aligned}$$

The *pdf* of Y is thus,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = F'_X[g^{-1}(y)] \frac{dg^{-1}(y)}{dy} = f_X(g^{-1}(y)) \frac{dx}{dy}$$

since

$$x = g^{-1}(y), \text{ so that } \frac{dx}{dy} = \frac{dg^{-1}(y)}{dy}$$

Continuous RVs

Suppose $Y = g(x)$ is still one-to-one, but decreasing instead of increasing.

$$F_Y(y) = P\{g(X) \leq y\} = P\{X > g^{-1}(y)\} = 1 - F_X(g^{-1}(y))$$

and

$$\begin{aligned} f_Y(y) &= -F'_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} = -f_X(g^{-1}(y)) \frac{dx}{dy} \\ &= f_X(g^{-1}(y)) \left| \frac{dx}{dy} \right| \end{aligned}$$

The last step follows because $\frac{dx}{dy}$ is negative.
Therefore, regardless of whether $Y = g(x)$ is increasing or decreasing, so long as it is monotonic, we have

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dx}{dy} \right|$$

Linear Transformation

Given X with pdf $f_X(x)$, let

$$Y = a + bX, \qquad \frac{dy}{dx} = b$$

Then

$$f_Y(y) = f_X[g^{-1}(y)] \left|\frac{dx}{dy}\right| = f_X\left(\frac{y-a}{b}\right) \frac{1}{|b|}$$

This transformation is often used when X has mean 0 and standard deviation 1. The linear transformation above creates a rv Y with a distribution that has the same shape as that of X but has mean a and standard deviation b.
Conversely, if Y has mean a and standard deviation b, then $X = (Y - a)/b$ has mean 0 and standard deviation 1. This is called sometimes the *"Studentized"* transform.

Normal Distribution

Let $X \sim N(0, 1)$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty$$

The transformation

$$Y = \mu + \sigma X, \qquad X = \frac{Y - \mu}{\sigma}$$

yields

$$f_Y(y) = f_X\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

More generally, a distribution is a member of the class of *location-scale distributions* if the distribution of a linear transformation of a random variable with that distribution has the same distribution, but with different parameters.

Square root of an exponential RV

We have already seen that a constant times an exponential random variable leads to another exponential random variable. Suppose $X \sim \exp(\lambda)$, so that

$$f_X(x) = \lambda e^{-\lambda x}\, 1(x \geq 0)$$

and consider the distribution of $Y = \sqrt{X}$.

The transformation

$$y = g(x) = \sqrt{x}, \qquad x \geq 0$$

is one-to-one and has an inverse $x = y^2$ with $dx/dy = 2y$. Thus

$$f_Y(y) = f_X(y^2)2y = 2\lambda y e^{-\lambda y^2}, \qquad y \geq 0$$

This distribution is a particular form of the Rayleigh distribution and is a special case of the χ, Rice and Weibull distributions.

Probability Integral Transform

Let $X \sim F_X(x)$. Define the transformation

$$Y = F_X(X) \in [0, 1], \qquad X = F_X^{-1}(Y)$$

Here

$$\frac{dy}{dx} = F_X'(x) = f_X(x)$$
$$f_Y(y) = f_X[F_X^{-1}(y)]\, \frac{1}{f_X[F_X^{-1}(y)]} = 1$$

i.e. Y is uniform over $[0, 1]$.

Another way to see it is through the distribution function:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(F_X(X) \leq y) \\ &= P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y \end{aligned}$$

Probability Integral Transform (cont.)

The probability integral transform is useful in statistics for checking goodness of fit of a distribution to a set of data.

Example:

$$X \sim \exp(\lambda) \quad \Rightarrow \quad Y = 1 - \exp(-\lambda X) \sim U[0,1]$$

If one has data $X_1, \ldots, X_n$, one could compute the transformed data $Y_i = 1 - \exp(-\lambda X_i)$, $i = 1, \ldots, n$, and check whether the Y_i's appear uniformly distributed over the interval $[0,1]$.

Inverse Probability Integral Transform

We can also start from the uniform distribution and do the inverse procedure.
Suppose $X \sim U[0,1]$, so that $f_X(x) = 1$ and $F_X(x) = x$ for $x \in [0,1]$. Let

$$Y = F^{-1}(X), \qquad X = F(Y)$$

where $F(\cdot)$ is a non-decreasing absolutely continuous function $F : \mathbb{R} \to [0,1]$, $F(y) = \int_{-\infty}^{y} f(x)\,dx$. Then

$$\frac{dx}{dy} = F'(y) = f(y) \quad \Rightarrow \quad f_Y(y) = f(y)$$

i.e. Y has the *pdf* corresponding to F.
Another way to see it is through the distribution function:

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(F^{-1}(X) \le y) \\ &= P(X \le F(y)) = F(y) \end{aligned}$$

Inverse Probability Integral Transform (cont.)

The Inverse Probability Integral Transform is used extensively in simulation of random variables.

Example: Most random number generators generate random numbers uniformly in the interval $[0, 1]$. Suppose we want to generate random numbers from an exponential distribution with parameter λ.

Let $X \sim U[0, 1]$. We want Y with distribution function $F(y) = 1 - \exp(-\lambda y)$. Then we need the transformation

$$Y = F^{-1}(X) = \frac{-1}{\lambda} \log(1 - X)$$

The recipe is

1. Generate random numbers X_i uniformly over $[0, 1]$.
2. Compute $Y_i = -\log(1 - X_i)/\lambda$.

Example: Cauchy Distribution

Let θ be distributed uniformly between $(-\pi/2, \pi/2)$:

$$f(\theta) = \frac{1}{\pi}, \qquad -\pi/2 < \theta < \pi/2$$

Consider $Y = \tan\theta$.

$$\begin{aligned} \frac{dy}{d\theta} = \sec^2\theta &= 1 + \tan^2\theta = 1 + y^2 \\ f_Y(y) = \frac{1}{\pi}\left|\frac{d\theta}{dy}\right| &= \frac{1}{\pi}\frac{1}{(1+y^2)} \qquad -\infty < y < \infty \end{aligned}$$

The distribution with this density is known as the Cauchy distribution.

One-to-many

What if the transformation is not 1-1? The trick is to start with the cdf of the transformed random variable.

Example: Let $Y = |X|$, and assume X is continuous.

$$\begin{aligned} F_Y(y) &= P\{Y \le y\} = P\{-y \le X \le y\} = F_X(y) - F_X(-y) \\ f_Y(y) &= F'_X(y) - F'_X(-y)(-1) = f_X(y) + f_X(-y) \end{aligned}$$

Suppose

$$X \sim N(0,1), \qquad f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Then

$$f_Y(y) = \frac{2}{\sqrt{2\pi}} e^{-y^2/2}, \qquad 0 < y < \infty$$

Quadratic transformation

Let

$$Y = X^2, \qquad \frac{dy}{dx} = 2x, \qquad \left|\frac{dy}{dx}\right| = 2\sqrt{y}$$

Then

$$\begin{aligned} F_Y(y) &= P\{Y \le y\} = P\{X^2 \le y\} = P\{-\sqrt{y} < X \le \sqrt{y}\} \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$$

$$\begin{aligned} f_Y(y) &= F'_X(\sqrt{y}) \left(\frac{1}{2} y^{-\frac{1}{2}}\right) - F'_X(-\sqrt{y}) \left(-\frac{1}{2} y^{-\frac{1}{2}}\right) \\ &= \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})], \qquad y > 0 \end{aligned}$$

Example

Suppose $X \sim N(0,1)$, $Y = X^2$:

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad -\infty < x < \infty \\ f_Y(y) &= \frac{1}{2\sqrt{y}} \left[\frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}y} \right] \\ &= \frac{(y/2)^{-\frac{1}{2}} e^{-\frac{y}{2}}}{2\sqrt{\pi}}, \quad y > 0 \end{aligned}$$

This is the density of a χ^2 distribution with 1 degree of freedom.

Result: If $X \sim N(0,1)$, then $X^2 \sim \chi^2(1)$.

General Result–Theorem C-B 2.1.8

Suppose $Y = g(X)$ is not 1-1, but there are disjoint sets $A_1, ...A_k$ that span the domain (sample space) of X such that $g(.) = g_j(.)$ is continuous and 1-1 on each A_j. This means that the inverse, $x = g_j^{-1}(y)$ exists on each A_j. Then

$$f_Y(y) = \sum_{j=1}^{k} f(g_j^{-1}(y)) \left| \frac{dg_j^{-1}(y)}{dy} \right|$$

Example: A wrapped distribution

Suppose $X \in \mathbb{R}$ with density $f_X(x)$ represents a random angle of rotation (in radians) from the x-axis on the unit circumference. The observed angle is

$$\Theta = X \mod 2\pi, \qquad \Theta \in [0, 2\pi)$$

because it is impossible to tell if the rotation involved more than one full turn. In this case

$$f_\Theta(\theta) = \sum_{j=-\infty}^{\infty} f(\theta + 2\pi j), \qquad 0 \leq \theta < 2\pi$$

If $X \sim N(0, \sigma^2)$, then

$$f_\Theta(\theta) = \sum_{j=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\theta + 2\pi j)^2}{2\sigma^2}\right)$$

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 8

Expectation

- The *expected value* or *mean* of a rv X, denoted $\mathsf{E}(X)$ or $\mathsf{E}X$ is

$$\mathsf{E}X = \begin{cases} \int_{-\infty}^{\infty} x\, f_X(x)\, dx, & X \text{ continuous} \\ \sum_{x \in \mathcal{X}} x\, f_X(x), & X \text{ discrete} \end{cases}$$

  Provided the integral or summation exists.

- This is generalized for a function of a random variable $g(X)$ as

$$\mathsf{E}g(X) = \begin{cases} \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx, & X \text{ continuous} \\ \sum_{x \in \mathcal{X}} g(x)\, f_X(x), & X \text{ discrete} \end{cases}$$

  Notice we could also find the pdf or pmf of Y and use the first definition. Both give the same answer (HW).

Expectation of some continuous variables

- Let $X \sim \mathsf{U}[a, b]$. Then

$$\mathsf{E}X = \int_a^b x \frac{1}{b-a}\, dx = \frac{a+b}{2}$$

- Let $X \sim \mathsf{Exp}(1)$, $f_X(x) = e^{-x}\, 1(x > 0)$. Then

$$\mathsf{E}X = \int_0^{\infty} x e^{-x}\, dx = -x e^{-x} \Big|_0^{\infty} + \int_0^{\infty} e^{-x}\, dx = 1$$

- Let $X \sim \mathsf{N}(0, 1)$. Then

$$\mathsf{E}X = \int_{-\infty}^{\infty} x\, \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 0$$

  (since the above integral is an odd function)

Expectation of some discrete variables

- Let $X = 1(A)$, where A is a Borel set. Then

$$\mathsf{E}X = 0 \cdot P(A^c) + 1 \cdot P(A) = P(A)$$

- Let $X \sim$ Binomial(n, p) for n positive integer and $0 < p < 1$ (n is the number of independent identical binary trials and p is the probability of success). Then

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, \ldots, n$$

and

$$\begin{aligned}
\mathsf{E}X &= \sum_{x=0}^{n} x \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!(x-1)!} p^{x-1}(1-p)^{n-x} \\
(\text{let } y = x-1) &= np \sum_{y=0}^{n-1} \frac{(n-1)!}{(n-1-y)!y!} p^y (1-p)^{n-1-y} \\
&= np
\end{aligned}$$

An easier way later.

cont.

- Let $X \sim$ Geom(p) (e.x. toss till first success), $f_X(x) = pq^{x-1}$, $x = 1, 2, \ldots$, where $q = 1 - p$. Then

$$\begin{aligned}
\mathsf{E}X &= \sum_{x=1}^{\infty} x \cdot pq^{x-1} = \sum_{x=1}^{\infty} p \frac{d}{dq}(q^x) \\
&= p \frac{d}{dq} \sum_{x=1}^{\infty} q^x = p \frac{d}{dq} \left(\frac{1}{1-q} - 1 \right) \\
&= p \frac{1}{(1-q)^2} = \frac{1}{p}
\end{aligned}$$

Note: The differentiation operator can be moved outside the summation sign because the geometric series converges uniformly.

Mean = area under survival curve

For a non-negative random variable X (i.e. $f(x) = 0$ for $x < 0$),

$$\mathsf{E}(X) = \begin{cases} \int_0^\infty (1 - F(x))dx, & X \text{ continuous} \\ \sum_{x=0}^\infty (1 - F(x)), & X \text{ discrete} \end{cases}$$

Proof: Homework! (Exercise 2.14)

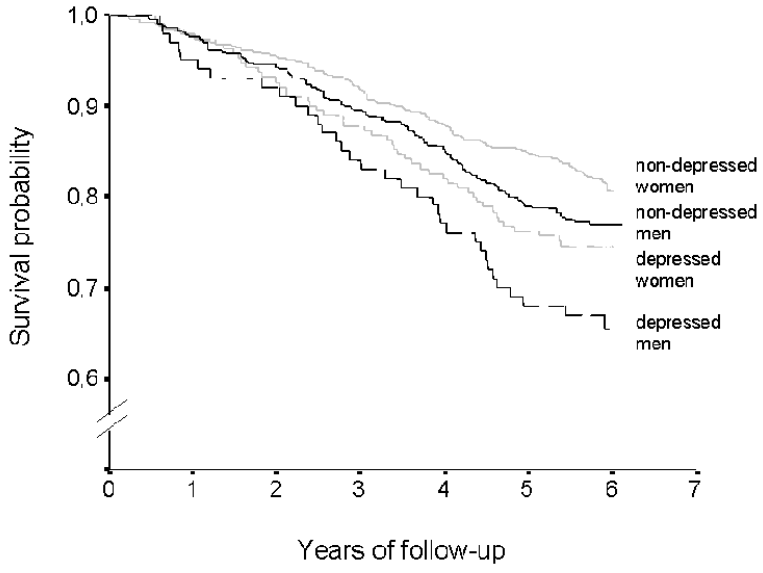
Example: Let $X \sim \text{Exp}(\lambda)$ with

$$F_X(x) = (1 - e^{-\lambda x})\, 1(x > 0)$$

Then

$$\mathsf{E}X = \int_0^\infty e^{-\lambda x}\, dx = \lambda$$

Survival curves



Properties of expectation

$$\begin{aligned} \mathsf{E}[ag(X)+c] &= a\mathsf{E}[g(X)]+c, \qquad a, c \text{ constants} \\ \mathsf{E}[g_1(X)+g_2(X)] &= \mathsf{E}[g_1(X)]+\mathsf{E}[g_2(X)] \\ \text{If } g_1(x) \geq 0 \; \forall x, &\quad \text{then} \quad \mathsf{E}g_1(X) \geq 0 \\ \text{If } a \leq g(X) \leq b \; \forall x, &\quad \text{then} \quad a \leq \mathsf{E}g(X) \leq b \end{aligned}$$

(Proofs are immediate page 57)

Note: The linearity property above applies to two random variables X and Y that have the same distribution:

$$\mathsf{E}[g_1(X)+g_2(Y)] = \mathsf{E}[g_1(X)]+\mathsf{E}[g_2(Y)]$$

We will see later that this is true even if X and Y do not have the same distribution.

Method of indicators

An example of how the above properties are useful.

Let $X \sim$ Binomial(n,p) for n positive integer and $0<p<1$ (n is the number of independent identical binary trials and p is the probability of success). We can write

$$X = \sum_{i=1}^{n} I_i$$

where I_i is the indicator that the i^{th} trial is a success. We have

$$\mathsf{E}I_i = p$$

Therefore

$$\mathsf{E}X = \sum_{i=1}^{n} \mathsf{E}I_i = \sum_{i=1}^{n} p = np$$

Minimal property of the mean

Assuming all integrals exist,

$$\min_b \mathsf{E}(X-b)^2 = \mathsf{E}(X-\mathsf{E}X)^2 = \mathsf{Var}X$$

Two proofs:

- By differentiation with respect to b (homework). Requires more assumptions.
- By sum-of-squares decomposition.

Moments

For a random variable X, the expectation of the polynomials $g(X) = X^r$, $r = 0, 1, 2, \ldots$ are called the *moments* of X:

$$m_r = \mathsf{E}(X^r), \qquad r = 0, 1, 2, \ldots$$

These are sometimes called *non-central* moments or *moments about the origin*.

Notes:

- $m_0 = 1$.
- m_1 is the *mean*, usually denoted by $m_1 = \mu$.

Central moments

The r^{th} *central moment* of X is

$$\mu_r = \mathsf{E}[X - \mathsf{E}X]^r = \mathsf{E}[X - \mu]^r, \qquad r = 0, 1, 2, \ldots$$

These are sometimes called *moments about the mean*.

Notes:

- $\mu_0 = 1$.
- $\mu_1 = 0$.
- μ_2 is the *variance*.
- μ_3 is related to the *skewness*.
- μ_4 is related to the *kurtosis*.

Variance

The second central moment μ_2 is called the *variance of* X and is usually denoted by σ^2:

$$\sigma^2 = \mu_2 = \mathsf{Var}(X) = \mathsf{E}[X - \mu]^2 = \mathsf{E}(X^2) - \mu^2$$

Notes:

- Useful property: For $a, b \in \mathbb{R}$,
$$\mathsf{Var}(aX + b) = a^2\mathsf{Var}(X)$$
- Notice that $\mathsf{E}(X^2) \neq [\mathsf{E}(X)]^2$. In fact, $\mathsf{E}(X^2) \geq [\mathsf{E}(X)]^2$ because
$$\mathsf{Var}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 \geq 0$$

Skewness

The *skewness* of a rv X is defined as

$$\alpha_3 = \frac{\mu_3}{(\mu_2)^{3/2}}, \qquad \mu_3 = \mathsf{E}[(X-\mu)^3]$$

Skewness measures the symmetry of a distribution.

$$\begin{aligned} \mu_3 = 0 \quad &\Rightarrow \quad \text{symmetric} \\ \mu_3 \geq 0 \quad &\Rightarrow \quad \text{right skew} \\ \mu_3 \leq 0 \quad &\Rightarrow \quad \text{left skew} \end{aligned}$$

Central and non-central moments

Central moments can be written as a function of non-central moments and vice-versa:

$$\begin{aligned} \mu_2 &= m_2 - m_1^2 \\ \mu_3 &= m_3 - 3m_1m_2 + 2m_1^3 \end{aligned}$$

etc.

Method of indicators (again)

Recall that $X \sim$ Binomial(n, p) can be written as $X = \sum_{i=1}^{n} I_i$ where I_i is the indicator that the i^{th} trial out of n is a success. We have

$$\mathsf{E}I_i = p, \qquad \mathsf{E}I_i^2 = p, \qquad \mathsf{Var}I_i = p(1-p)$$

Therefore

$$\begin{aligned}\mathsf{E}X^2 &= \mathsf{E}\left(\sum_{i=1}^{n} I_i\right)^2 = \mathsf{E}\sum_{i=1}^{n}\sum_{j=1}^{n} I_i I_j \\ &= \mathsf{E}\left(\sum_{i=1}^{n} I_i + \sum_{i\neq j} I_i I_j\right) = \sum_{i=1}^{n} p + \sum_{i\neq j} p^2 = np + (n^2-n)p^2\end{aligned}$$

and

$$\mathsf{Var}X = \mathsf{E}X^2 - \mathsf{E}^2 X = np + (n^2-n)p^2 - (np)^2 = np(1-p)$$

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 9

Moment Generating Function

(C-B 2.3, Gut III.3)

The *moment generating function (mgf)* of the rv X is defined to be

$$M_X(t) = \mathsf{E}(e^{tX})$$

provided that the expectation exists in a neighbourhood $(-h, h)$ of $t = 0$.

Theorem: Suppose the mgf $M_X(t)$ of X exists for $t \in (-h, h)$ for some $h > 0$. Then for any positive integer n,

$$\mathsf{E}(X^n) = M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}$$

Notice that $M_X(0) = 1$ always.

cont.

Proof: Assuming that we can interchange expectation and differentiation,

$$\begin{aligned} & \frac{d}{dt} M_X(t) = \frac{d}{dt}\mathsf{E}(e^{tX}) = \mathsf{E}\left(\frac{d}{dt} e^{tX}\right) = \mathsf{E}(Xe^{tX}) \\ \Rightarrow \quad & \frac{d}{dt} M_X(t)\bigg|_{t=0} = \mathsf{E}(X) \\ & \frac{d^2}{dt^2} M_X(t)\bigg|_{t=0} = \mathsf{E}(X^2 e^{tX})\bigg|_{t=0} = \mathsf{E}(X^2) \end{aligned}$$

Another way to see this is

$$M_X(t) = \mathsf{E}(e^{tX}) = \mathsf{E}\left(\sum_{n=0}^{\infty} \frac{t^n}{n!} X^n\right) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathsf{E}X^n$$

so the moments can be obtained from a Taylor expansion of $M_X(t)$ around $t = 0$.

Example: Continuous

Mgf of an exponential rv: Let $f_X(x) = \lambda e^{-\lambda x}\ 1(x > 0)$. Then

$$M_X(t) = \mathsf{E}(e^{tX}) = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{-(\lambda - t)x}\, dx$$
$$= \lambda \frac{-1}{\lambda - t} e^{-(\lambda - t)x} \Big|_0^\infty = \begin{cases} \frac{\lambda}{\lambda - t} & \text{if } t < \lambda \\ \infty & \text{otherwise} \end{cases}$$

This is fine as we only need the mgf to be defined near zero. To obtain the moments, assume $|t| < \lambda$:

$$M_X(t) = \frac{1}{1 - t/\lambda} = \sum_{n=0}^\infty \frac{t^n}{\lambda^n} = \sum_{n=0}^\infty \frac{t^n}{n!} \mathsf{E}(X^n) \quad \Rightarrow \quad \mathsf{E}(X^n) = \frac{n!}{\lambda^n}$$

In particular,

$$\mathsf{E}X = 1/\lambda$$
$$\mathsf{Var}X = \mathsf{E}X^2 - \mathsf{E}^2 X = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$$

Example: Discrete

Mgf of a geometric rv: Let $f_X(x) = pq^{x-1}$, $x = 1, 2, \ldots$ ($q = 1 - p$).

$$M_X(t) = \mathsf{E}(e^{tX}) = \sum_{x=1}^\infty e^{tx} \cdot pq^{x-1} = \frac{p}{q} \sum_{x=1}^\infty (qe^t)^x$$
$$= \frac{p}{q} \left(\frac{1}{1 - qe^t} - 1 \right) = \frac{pe^t}{1 - qe^t}$$

The sum converges if $e^t q < 1$, that is, $t < \log(1/q)$.

The moments can be obtained by differentiation:

$$\mathsf{E}X = \frac{d}{dt} \left[\frac{p}{e^{-t} - q} \right]_{t=0} = \frac{pe^{-t}}{(e^{-t} - q)^2} \Big|_{t=0} = \frac{1}{p}$$

Linear transformations

For any constants a, b, the mgf of the rv $g(X) = aX + b$ is

$$M_{aX+b}(t) = e^{bt} M_X(at)$$

Proof:

$$\begin{aligned} M_{aX+b}(t) &= \mathsf{E}(e^{t(aX+b)}) = \mathsf{E}(e^{taX} e^{bt}) \\ &= e^{bt} \mathsf{E}(e^{(at)X}) = e^{bt} M_X(at) \end{aligned}$$

Assign as HW

Example: Continuous

Mgf of a Gaussian: Let $X \sim N(0, 1)$. Then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = e^{t^2/2} \int_{-\infty}^{\infty} \frac{e^{-(x-t)^2/2}}{\sqrt{2\pi}}\, dx = e^{t^2/2}$$

Also,

$$e^{t^2/2} = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{t^2}{2}\right)^n = \sum_{n=0}^{\infty} \frac{t^{2n}}{2^n n!} = \sum_{m:\ \text{even}} \frac{t^m}{m!} \frac{m!}{2^{m/2}(m/2)!}$$

Matching coefficients of $t^m/m!$ we get that all the odd moments are zero and that the even moments are $E(X^m) = m!/(2^{m/2}(m/2)!)$.

Now let $Y = \mu + \sigma X$ so that $Y \sim N(\mu, \sigma^2)$,

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

Existence of moments

Too hard. May eliminate. **Theorem**: Suppose the mgf $M_X(t)$ of X exists for $t \in (-h, h)$ for some $h > 0$. Then all moments exist: $|\mathsf{E}X^r| < \infty$ for all $r > 0$.

Proof: Fix $t \in (-h, h)$. There exists $C > 0$ such that

$$|x^r| \leq Ce^{|tx|}, \qquad \forall x \in \mathbb{R}$$

(What is C?)
so

$$\begin{aligned} |\mathsf{E}X^r| &\leq \mathsf{E}|X^r| \\ &\leq C\mathsf{E}e^{|tX|} \\ &\leq C\mathsf{E}(e^{tX} + e^{-tX}) \\ &\leq C\left[M_X(t) + M_X(-t)\right] < \infty \end{aligned}$$

Can moments not exist?

Example: Cauchy distribution

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \qquad x \in \mathbb{R}$$

The mean is

$$\begin{aligned} \mathsf{E}X &= \int_{-\infty}^{\infty} x\,\frac{1}{\pi}\frac{1}{1+x^2}\,dx \\ &= \int_{-\infty}^{0} \frac{1}{\pi}\frac{x}{1+x^2}\,dx + \int_{0}^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\,dx \\ &= \int_{0}^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\,dx - \int_{0}^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\,dx = \ ?? \end{aligned}$$

because

$$\int_0^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\,dx = \int_0^{\infty} \frac{1}{2\pi}\frac{dy}{1+y} = \frac{1}{2\pi}\log(1+y)\Big|_0^{\infty} = \infty$$

Cont.

One intuitive explanation for this is that the Cauchy distribution has infinite variance:

$$\begin{aligned}\mathsf{E}X^2 &= \int_{-\infty}^{\infty} \frac{1}{\pi}\frac{x^2}{1+x^2}\,dx = \int_{-\infty}^{\infty} \frac{1}{\pi}\left(1 - \frac{1}{1+x^2}\right)dx \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty} dx - 1 = \infty\end{aligned}$$

Can mgf's not exist?

Example: Cauchy distribution
If the 1st and 2nd moments do not exist, certainly the mgf does not either!
(How would you prove this statement is true?)

Example: Log-normal distribution
If $X \sim N(0,1)$ then $Y = e^X$ is called log-normal.

$$f_Y(y) = \frac{1}{y\sqrt{2\pi}} e^{-(\log y)^2/2}, \qquad y > 0$$

For $n = 0, 1, 2, \ldots$ the moments exist, but the mgf does not, i.e. the integral $\mathsf{E}(e^{tY})$ does not converge. (Homework).

Characterizing distributions

For rvs with unbounded support, the moments do not specify the distribution: there exist distributions with different pdfs and yet have all the same moments (see Example C-B 2.3.10).

However, moments uniquely identify distributions when the rvs have bounded support.

Also, mgfs uniquely identify distributions when the mgfs exist.

Theorem: Let $F_X(x)$ and $F_Y(y)$ be cdfs all of whose moments exist.

1. If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u iff $\mathrm{E}X^n = \mathrm{E}Y^n$ for all $n = 0, 1, 2, \ldots$.
2. If the mgfs exist and $M_X(t) = M_Y(t)$ for all t in a neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.

Convergence of mgfs

Convergence of mgfs implies convergence of cdfs.

Theorem 2.3.12 C-B: Let $X_1, X_2, \ldots$ be a sequence of rvs with corresponding mgfs $M_{X_1}(t), M_{X_2}(t), \ldots$ such that

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t), \qquad \forall t \in (-h, h), \;\; h > 0.$$

Then $\exists!$ a unique cdf $F_X(t)$ whose moments are given by $M_X(t)$ and for all x where $F_X(x)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Application

Example: Normal approximation to Poisson.
Let $X \sim Poisson(\lambda)$, then $M_X(t) = \exp[\lambda(e^t - 1)]$
(Exercise C-B 2.33), with E$X = \lambda$, Var$X = \lambda$.

Let $Y = (X - \lambda)/\sqrt{\lambda}$. Then

$$M_Y(t) = \mathsf{E}(e^{tY}) = \mathsf{E}(t(X - \lambda)/\sqrt{\lambda}) = e^{-\sqrt{\lambda}t} M_X(t/\sqrt{\lambda}).$$

Hence,

$$\begin{aligned} log(M_Y(t)) &= -t\sqrt{\lambda} + \lambda(e^{t/\sqrt{\lambda}} - 1) \\ \text{(when } \lambda \text{ is large)} &= -t\sqrt{\lambda} + \lambda(\frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{3!\lambda^{3/2}} + \cdots) \\ &= \frac{t^2}{2} + \frac{t^3}{3!\lambda^{1/2}} + \cdots \end{aligned}$$

Therefore,

$$\lim_{\lambda \to \infty} M_Y(t) = e^{t^2/2},$$

which is the mgf of a $N(0, 1)$ variable.

Relationship to other transforms

For continuous rvs:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$$

is similar to the two-sided *Laplace transform* of the function $f_X(x)$.

The transform

$$S_X(t) = \log M_X(t) = \log \mathsf{E} e^{tX}$$

is called *cumulant generating function*. The derivatives at $t = 0$ are called *cumulants*. In particular, (Homework)

$$S_X(0) = 0, \qquad S_X^{(1)}(0) = \mathsf{E}X, \qquad S_X^{(2)}(0) = \mathsf{Var}X$$

E.g. $X \sim N(\mu, \sigma^2)$,

$$S_X(t) = \log(e^{\mu t + \sigma^2 t^2/2}) = \mu t + \sigma^2 t^2/2$$

Characteristic Function

(Gut III.4)

The *characteristic function (cf)* of the rv X is defined as

$$\phi_X(t) = \mathsf{E}e^{itX} = \mathsf{E}[\cos(tX) + i\sin(tX)]$$

where $i^2 = -1$.

- The cf is complex-valued, $\phi_X(t) \in \mathbb{C}$.
- The cf always exists because
$$|\mathsf{E}e^{itX}| \leq \mathsf{E}|e^{itX}| = \mathsf{E}1 = 1$$
- For calculations, the cf can often be obtained from the mgf replacing the argument t by it (as long as the mgf exists!).

Examples

$X \sim Exp(\lambda)$:

$$\phi_X(t) = \frac{\lambda}{\lambda - it}$$

$X \sim N(\mu, \sigma^2)$:

$$\phi_X(t) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right)$$

$X \sim Geom(p)$:

$$\phi_X(t) = \frac{pe^{it}}{1 - qe^{it}}$$

And the range is $t \in \mathbb{R}$ in all cases.

Also, for X with the Cauchy distribution

$$f_X(x) = \frac{1}{\pi(1 + x^2)} \qquad \Rightarrow \qquad \phi_X(t) = e^{-|t|}$$

Exists!

Properties

1. $|\phi_X(t)| \leq \phi_X(0) = 1$
2. Complex conjugate: $\overline{\phi_X(t)} = \phi_X(-t)$ (Homework)
3. Linear transformations:
$$\phi_{aX+b}(t) = e^{ibt}\phi_X(at)$$
4. The distribution is symmetric about 0, $f_X(x) = f_X(-x)$, iff $\phi_X(t)$ is real.
5. Moment generation:
$$\phi_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0} = i^n \mathsf{E}(X^n)$$

Characterizing distributions

Characteristic functions uniquely identify distributions, always.

Theorem: If X and Y are rvs with cfs $\phi_X(t) = \phi_Y(t)$, then $F_X(u) = F_Y(u)$ for all u.

Relationship to other transforms

For continuous rvs:

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx$$

is similar to the *Fourier transform* of the function $f_X(x)$.

For discrete rvs:

$$\phi_X(t) = \sum_{x=-\infty}^{\infty} e^{itx} f_X(x)$$

is similar to the *discrete Fourier transform* of the sequence $f_X(x)$ (when x is sampled at equal intervals).

Thus Fourier transform tables can be helpful for finding cfs.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 10

Why parametric models?

- *Parametric models* or *distribution families* have a specific form but can change according to a fixed number of parameters.
- The objective is to model a population. Parametric models are often appropriate in common situations with similar mechanisms.
- Parametric models have many known and useful properties and are easy to work with. When fitting a population, only a few parameters need to be estimated: *parametric inference*.
- Sometimes one does not want to make parametric assumptions and would rather work with non-parametric models. But non-parametric models can be infinite dimensional. E.g. $f_X(x),\ x = 0, 1, 2, \ldots$ or $F_X(x),\ x \in \mathbb{R}$.
- In this course we emphasize parametric models.

Discrete uniform

X has the *discrete uniform*$(1, N)$ distribution if X is equally likely to be one of $\{1, 2, \ldots, N\}$.

sample space: $\{1, 2, \ldots, N\}$

pmf:

$$f_X(x) = \frac{1}{N}, \qquad x = 1, 2, \ldots, N$$

cdf:

$$F_X(x) = P(X \leq x) = \frac{x}{N}, \qquad x = 1, 2, \ldots, N$$

moments:

$$\mathsf{E}X = \frac{N+1}{2}$$

This definition can be extended to the range $N_0, \ldots, N_1$ (consecutive intergers starting at any integer N_0 and ending with N_1) with $f_X(x) = 1/(N_1 - N_0 + 1)$.

Bernoulli distribution

Consider an experiment where outcomes are binary (say, Success or Failure) and the probability of success is p.
Define the following random variable

$$Y = \begin{cases} 1 & \text{outcome is success} \\ 0 & \text{outcome is failure} \end{cases}$$

Then, Y has a Bernoulli Distribution.

sample space: $\{0, 1\}$

pmf: $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$. We can write this as:

$$f(y) = P(Y = y) = \begin{cases} p^y(1-p)^{(1-y)} & y = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

What are the cdf, mean and variance?

Binomial Distribution

Now consider a *series of* n *Bernoulli trials* where

1. trials are independent
2. Prob of success or failure is the same for each trial, i.e.

$$P(S_i) = p \text{ and } P(F_i) = q = 1 - p \text{ for the } i^{th} \text{ trial.}$$

More concisely, consider n *iid* (independent, identically distributed) Bernoulli rvs Y_i.

A *binomial(n,p)* random variable X is defined as the number of successes in n iid Bernoulli trials, each with probability p of success:

$$X = \sum_{i=1}^{n} Y_i$$

Example: Coin Tossing

Toss 3 coins ($n = 3$) $P(H) = p, \quad P(T) = q$

$$\begin{array}{lcll} P(HHH) & = & p^3 & P(TTT) = q^3 \\ P(THH) & = & p^2q & P(HTT) = pq^2 \\ P(HTH) & = & p^2q & P(THT) = pq^2 \\ P(HHT) & = & p^2q & P(TTH) = pq^2 \end{array}$$

The binomial distribution is concerned with the distribution of the **number** of successes (heads):

$$\begin{array}{lcll} P(0H) & = & P(3T) & = q^3 \\ P(1H) & = & P(2T) & = 3pq^2 \\ P(2H) & = & P(1T) & = 3p^2q \\ P(3H) & = & P(0T) & = p^3 \end{array}$$

Binomial distribution

For any particular sequence of s successes and $n - s$ failures

$$P(SFS \ldots F) = p^s q^{n-s}$$

However there are $\binom{n}{s}$ ways to get s successes from n trials.
Formally, the *binomial distribution* has:

sample space: $\{0, 1, ...n\}$

pmf:

$$f_Y(s) = \begin{cases} \binom{n}{s} p^s q^{n-s} & s = 0, 1, ...n \\ 0 & \text{otherwise} \end{cases}$$

Is it easy to check that the pmf sums to 1?

Binomial cont.

cdf:

$$F_Y(y) = \sum_{s=0}^{y} \binom{n}{s} p^s q^{n-s}$$

i.e.

$$\begin{aligned} F(y) &= 0 \quad \text{for} \quad y < 0 \\ F(0) &= q^n \\ F(1) &= q^n + npq^{n-1} \\ F(2) &= q^n + npq^{n-1} + \frac{n(n-1)}{2} p^2 q^{n-2} \\ &\vdots \\ F(y) &= 1 \quad \forall y \geq n \end{aligned}$$

Binomial example

A physician treats $n = 10$ people having a particular disease where P(success)$= 1/4$. What are probabilities of outcomes?

$$f(s) = \binom{10}{s} p^s q^{10-s}, \quad p = 1/4$$

s	$f(s)$	$F(s)$	s	$f(s)$	$F(s)$
0	.056		6	.016	
1	.188		7	.003	
2	.282		8	.0004	
3	.250		9	.00003	
4	.146		10	.00000009	
5	.059				

Poisson Distribution

The Poisson distribution was derived by the French mathematician Poisson in 1837 as a limiting version of the binomial distribution.

Suppose Y has a binomial(n, p) distribution, but consider what happens when n becomes large, but p is small enough so that np stays constant, and equal to a fixed value λ.

$$\lim_{n\to\infty} \binom{n}{y} p^y q^{n-y} = \frac{e^{-\lambda}\lambda^y}{y!}$$

Proof:

$$\begin{aligned} f_Y(y) &= \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \\ &= \frac{n!}{y!(n-y)!} \left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y} \\ &= \frac{\lambda^y}{y!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-y} \frac{(n-y+1)\ldots n}{n^y} \end{aligned}$$

Binomial vs Poisson

Comparison of binomial and Poisson *pmf's*

$n = 5,\ p = 1/5\ \ (\lambda = 1)$

y	binomial	poisson
0	.328	.368
1	.410	.368
2	.205	.184
3	.051	.061
4	.006	.015
5	.000	.003
6+	0	.001

Binomial to Poisson via cfs

Another way to see the convergence of a binomial to a Poisson is via convergence of the cf.

The cf of $X \sim Binomial(n, p)$ is

$$\phi_X(t) = (pe^{it} + 1 - p)^n$$

Let $p = \lambda/n$:

$$\phi_X(t) = \left(\frac{\lambda}{n}e^{it} + 1 - \frac{\lambda}{n}\right)^n = \left(1 + \frac{\lambda(e^{it} - 1)}{n}\right)^n$$

As $n \to \infty$,

$$\phi_X(t) \to \exp[\lambda(e^{it} - 1)]$$

which is the cf of a Poisson. Since cfs characterize distributions, the distribution fo the rv X converges to a Poisson.

Poisson example

The death rate among females (Age 15-44) from pulmonary embolism is 4 per million. In a city of one million females (Age 15-44), what is the probability distribution of number of cases; i.e. $\lambda = 4$.

Note that $\lambda = np$ when $p = \frac{4}{1{,}000{,}000}$ is probability for a woman $\lambda = np = 10^6 \cdot 4/10^6 = 4$.

s	$e^{-4}4^s/s!$	s	$e^{-4}4^s/s!$	s	$e^{-4}4^s/s!$
0	.018	4	.20	8	.03
1	.07	5	.16	9	.01
2	.15	6	.10		
3	.20	7	.06		

Uses of the Poisson

The Poisson distribution is often used to describe:

- incidence of rare events in time or space
- failure of equipment
- incidence of rare diseases
- mortality
- pixel intensity in CT and PET imaging
- readings of molecular binding experiments (e.g. gene transcription)

Also arises in queueing theory (cashiers and internet), survival analysis, etc.

Hypergeometric Distribution 18 / 27

Example: Capture-Recapture Method

Suppose we wish to estimate how many fish there are in a lake.
Consider the following technique:
Capture a certain number of fish, 100 say, mark them, and return them to the lake.
Wait a sufficient amount of time for them to intermix again with the other fish.
Capture a new batch, 120 say, and see how many of these were previously marked.
Suppose there are 10 marked. Then the argument goes: the proportion marked that were recaptured should be in the same proportion as those amongst the non-captured. And so we can estimate the total number of fish in the lake to be the solution N to:

$$\frac{100-10}{N-120} = \frac{10}{120} \qquad \Rightarrow \qquad N =$$

This is a situation where we might apply a hypergeometric distribution.

Hypergeometric Distribution

Suppose a population of N entities is made up of two types, and there are M of the first type; and so $N - M$ of the second type.
Suppose we take a sample of size K, and we wish to know X, the number in the sample of the first type.
The probability mass function of X is given by:

$$f_X(x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$

for $x = \max(0, M - N + K), \ldots, \min(M, K)$.
The sample space is defined so that all binomial coefficients are valid.
We must have:

$$0 \le x \le K, \qquad 0 \le x \le M, \qquad , 0 \le K - x \le N - M$$

Often $K < M$ and $K < N - M$ so the range becomes $0 \le x \le K$.

Hypergeometric Moments

Mean:

$$EX = \sum_{x=0}^{K} x\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} = \sum_{x=1}^{K} \text{ditto}$$

We need the identity

$$x\binom{M}{x} = x\frac{M!}{x!(M-x)!} = M\frac{(M-1)!}{(x-1)!(M-x)!} = M\binom{M-1}{x-1}$$

or

$$\binom{M}{x} = \frac{M}{x}\binom{M-1}{x-1}$$

assuming everything is legit, i.e. all the numbers are positive integers, etc.. So

$$EX = \sum_{x=1}^{K} \frac{M}{\frac{N}{K}} \frac{\binom{M-1}{x-1}\binom{N-M}{K-1-(x-1)}}{\binom{N-1}{K-1}} = \frac{MK}{N}$$

Random sampling

Example: Vaccines are manufactured in batches of size N. Suppose n vials are sampled.
Decision Rule: If no vials are defective, batch is accepted.
What is the probability that batch with M defective vials is accepted?

Example: Experience suggests that a treatment for liver cancer should be considered effective if 20% of treated patients respond. A hospital plans to run a trial of a new treatment in 12 patients, and will consider the drug ineffective (no better than standard) if less than 2 patients respond. What is the probability that a drug with a true efficacy rate of 30% is classified as ineffective?

cont.

If we think of N being very large compared with n, then it makes sense to approximate this probability by thinking of sampling **with** replacement, in which case, we have

$$P(\text{Accept batch}) = (1 - M/N)^n$$

Table of $P(A)$ (approximate)

		M/N		
		.001	**.01**	**.1**
	5	.995	.951	.59
n	10	.990	.904	.349
	50	.95	.605	.005

Hypergeometric vs. Binomial

We can show that the limiting form of the hypergeometric pmf is the binomial pmf

$$\begin{aligned} P(s) &= \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{n}} \\ &= \frac{\frac{M!}{s!(M-s)!}\frac{(N-M)!}{(n-s)!(N-M-n+s)!}}{\frac{N!}{n!(N-n)!}} \\ &= \frac{\frac{n!}{s!(n-s)!}\frac{M!}{(M-s)!}\frac{(N-M)!}{(N-M-n+s)!}}{\frac{N!}{(N-n)!}} \end{aligned}$$

cont.

Note

$$\begin{aligned} \frac{M!}{(M-s)!} &= \frac{M(M-1)(M-2)\ldots(M-s)!}{(M-s)!} \\ &= M^s\left[1(1-\frac{1}{M})\ldots(1-\frac{s-1}{M})\right] \\ \frac{N!}{(N-n)!} &= N^n\left[1(1-\frac{1}{N})\ldots(1-\frac{n-1}{N})\right] \\ \frac{(N-M)!}{[(N-M)-(n-s)]!} &= (N-M)^{n-s} \\ &\quad \left[1\cdot(1-\frac{1}{N-M})\ldots(1-\frac{n-s-1}{N-M})\right] \end{aligned}$$

cont.

Letting $N \to \infty$, $M \to \infty$, $\frac{M}{N} \to p$,

$$\begin{aligned} P(s) &= \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{n}} \\ &\sim \binom{n}{s}\frac{M^s(N-M)^{n-s}}{N^n} \\ &= \binom{n}{s}\left(\frac{M}{N}\right)^s\left(1-\frac{M}{N}\right)^{n-s} \\ &\to \binom{n}{s}p^s(1-p)^{n-s} \qquad \text{Binomial Distribution} \end{aligned}$$

Summary

Hypergeometric $\to$	Binomial $\to$	Poisson
$N \to \infty$,	$n \to \infty$	$\lambda = np$
$M \to \infty$	$p \to 0$	
$\frac{M}{N} \to p$	$np \to \lambda$	

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 11

Geometric Distribution

Consider a series of iid Bernoulli Trials with p=probability of success in each trial. Define a random variable X representing the number of trials until first success. *Note*: X includes the trial at which the success occurs. Then, X has a geometric distribution.

sample space: $\{1, 2, ...\}$

pmf:

$$f(X) = P\{X = x\} = \begin{cases} p(1-p)^{x-1} & x = 1, 2... \\ 0 & \text{otherwise} \end{cases}$$

cdf:

$$F(x) = P\{(X \leq x)\} = 1 - (1-p)^x$$

Moments:

$$\mathsf{E}(X) = 1/p$$
$$\mathsf{Var}(X) = (1-p)/p^2$$

Memoryless property

Suppose $k > i$, then

$$P(X > k | X > i) = P(X > k - i)$$

Proof:

$$\begin{aligned} P(X > k | X > i) &= \frac{P(X > k)}{P(X > i)} = \frac{(1-p)^k}{(1-p)^i} \\ &= (1-p)^{k-i} = P(X > k - i) \end{aligned}$$

Example: Suppose X is # years you live

$$\begin{aligned} P(\text{survive two more years}) &= P(X > \text{current age } + 2 | X > \text{current age}) \\ &= P(X > 2) \end{aligned}$$

This model is clearly too simple for human populations (since we do age).

Note

Some texts use the geometric distribution to describe the distribution of the number of trials **before** the first success. In this case, the pdf changes to:

$$f(x) = P(X = x) = \begin{cases} p(1-p)^x & x = 0, 1, 2... \\ 0 & \text{otherwise} \end{cases}$$

The moments also change. Be careful!

Example

In studies of human fertility, researchers are often interested in exposures that increase the time it takes for a couple to successfully conceive. If a woman has a $p = 25\%$ chance of conceiving in a particular menstrual cycle, what is the probability that she becomes pregnant on the 3rd cycle?

What is the expected number of cycles to pregnancy?

What is her probability of not becoming pregnant within 12 cycles? (this is the definition of clinical infertility)

Negative Binomial Distribution

Still in the context of iid Bernoulli trials, define a random variable corresponding to the number of trials required to have s successes. We say $X \sim$ *Negbin*(s, p).

sample space: $\{s, (s+1), ...\}$
pmf: for $x = s, s+1, s+2, \ldots,$

$$f(x) = \left(\begin{array}{c} x-1 \\ s-1 \end{array} \right) p^{s-1} q^{x-s} \cdot p$$

$$= \left(\begin{array}{c} x-1 \\ s-1 \end{array} \right) p^{s} q^{x-s}$$

cdf: no closed form.
expectation: $\mathsf{E}(X) = s/p$
Variance: $\mathsf{Var}(X) = s(1-p)/p^2$

Notes

- Why the name? See C-B p.95.
- $X \sim$ *Negbin*$(1, p)$ is the same as $X \sim$ *geometric*(p)
- *Negbin*(n, p) is the same as the sum of n *geometric*(p) random variables
- Note that the outcome $SFFSFSSFSS$ may have been generated from a $Binomial(10, p)$ or from a $NegBinom(6, p)$. Need to know the experimental design to compute probabilities.

Negative binomial sampling

Example: Suppose a proportion p of the population possess a certain characteristic (e.g., have a certain disease). How many people should we expect to sample in order to collect r people with that characteristic? (Suppose the population is big enough so that you can assume sampling with replacement)

How many should we sample in order to be 95% sure of getting r people with the characteristic?

Other parametrizations

As with the geometric, some books define a negative binomial random variable as the number of **failures before the s^{th} success**. This is equal to the previous definition minus s.

sample space: $\{0, 1, 2, \dots\}$

pmf:

$$f(x) = \binom{s+x-1}{x} p^s q^x, \qquad x = 0, 1, 2, \dots$$

cdf: no closed form.

expectation: $\mathsf{E}(X) = s(1-p)/p$

Variance: $\mathsf{Var}(X) = s(1-p)/p^2$

Negative binomial vs. Poisson

The negative binomial distribution is often good for modeling count data as an alternative to the Poisson.

In the previous parametrization, define

$$\lambda = \frac{s(1-p)}{p} \quad \Leftrightarrow \quad p = \frac{s}{s+\lambda}$$

Then we have

$$\mathsf{E}X = \lambda$$
$$\mathsf{Var}X = \frac{\lambda}{p} = \lambda\left(1+\frac{\lambda}{s}\right) = \lambda + \frac{\lambda^2}{s}$$

For the Poisson we had that the variance equals the mean.
For the negative binomial, the variance is equal to the mean plus a quadratic term. Thus the negative binomial can capture overdispersion in count data.

cont.

In the previous parametrization, the pmf becomes

$$f(x) = \binom{s+x-1}{x} p^s q^x = \frac{(s+x-1)!}{x!(s-1)!}\left(\frac{s}{s+\lambda}\right)^s\left(\frac{\lambda}{s+\lambda}\right)^x$$
$$= \frac{\lambda^x}{x!}\frac{s(s+1)\ldots(s+x-1)}{(s+\lambda)^x}\left(1+\frac{\lambda}{s}\right)^{-s}$$

Letting $s \to \infty$ we get that

$$f(x) \to \frac{\lambda^x}{x!}e^{-\lambda}$$

So for large s, the negative binomial can be approximated by a Poisson with parameter $\lambda = s(1-p)/p$.
We can see that also from convergence of the moments (Homework).

Uniform Distribution

A random variable X having a *pdf*

$$f(x) = \begin{cases} 1 & \text{for } 0 < x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

is said to have a *uniform distribution* over the interval $(0, 1)$.
The *cdf* is:

$$F(y) = \int_{-\infty}^{y} f(x)dx = \int_{0}^{y} dx = \begin{cases} 0 & \text{for } y \leq 0 \\ y & \text{for } 0 \leq y \leq 1 \\ 1 & \text{for } y > 1 \end{cases}$$

Uniform Distribution (cont.)

Uniform: $Y \sim U[a, b]$:

sample space $[a, b]$
pdf:

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{for } a < y \leq b \\ 0 & \text{elsewhere} \end{cases}$$

cdf:

$$F(y) = \int_{a}^{y} \frac{1}{b-a} dx = \begin{cases} 0 & \text{for } y < a \\ \frac{y-a}{b-a} & \text{for } a < y \leq b \\ 1 & \text{for } y > b \end{cases}$$

moments

$$\begin{aligned} \mathsf{E}(Y) &= (a+b)/2 \\ \mathsf{Var}(Y) &= \frac{(b-a)^2}{12} \end{aligned}$$

Notes

- The uniform extends to the continuous case the idea of equally likely outcomes.
- If $Y \sim U[0,1]$, then $a + (b-a)Y \sim U[a,b]$.
- Useful for settings where individuals arrive at a destination, enter a study, etc. randomly over time.

Example: Suppose that during rush hour buses leave every 10 minutes, starting at 7:00am. Suppose you arrive at the bus stop according to a uniform distribution between 7am and 7:30am. What is the probability that you have to wait more than 5 minutes? What about 2 minutes?

Exponential Distribution 17 / 22

Exponential Distribution

Denoted $X \sim Exp(\lambda)$:

sample space: $y \geq 0$
pdf:

$$f(y) = \begin{cases} \lambda e^{-\lambda y} & \text{for } y \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

cdf:

$$F(y) = \int_0^y \lambda e^{-\lambda y} dx = \begin{cases} 1 - e^{-\lambda y} & \text{for } y \geq 0 \\ 0 & \text{for } y < 0 \end{cases}$$

moments:

$$\begin{aligned} \mathsf{E}(Y) &= 1/\lambda \\ \mathsf{Var}(Y) &= 1/\lambda^2 \\ M_X(t) &= \lambda/(\lambda - t) \end{aligned}$$

Interpretation

The exponential can be derived as the waiting time between Poisson events. Suppose that the number of events in a unit interval of time follows a Poisson(λ) distribution. Then, let Y be the time until the first event.

$$P(Y > t) = P(\text{0 events in } [0, t])$$

But, the number of events in $[0, t]$ follows a Poisson distribution with parameter λt. Hence,

$$P(Y > t) = e^{-\lambda t}$$

The *cdf* corresponding to Y is

$$F(t) = 1 - P(Y > t) = 1 - e^{-\lambda t}$$

and hence the density is

$$f(t) = \lambda e^{-\lambda t}$$

Notes

- Many books write the density as

$$f(y) = \begin{cases} \frac{1}{\theta} e^{-y/\theta} & \text{for } y \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

  so that E$(Y) = \theta$ and Var$(Y) = \theta^2$

- The exponential has a *memorylessness property*, just like the geometric.

$$P(Y > s + t | Y > t) = P(Y > s)$$

  Same interpretation as the geometric for continuous time: the probability of an event in a time interval depends only on the length of the interval, not the absolute time of the interval.
  The underlying Poisson process is stationary: the rate λ is constant. (In the geometric, the prob. p of getting an event in every discrete time unit is constant.)

Shifted exponential

Let $X \sim Exp(\lambda)$ and $Y = X + v$, $v \in \mathbb{R}$.
Y has the *shifted exponential* distribution with pdf:

$$f(y) = \begin{cases} \lambda e^{-(y-v)\lambda} & \text{for } y \geq v \\ 0 & \text{elsewhere} \end{cases}$$

Double Exponential

The *double exponential* distribution is formed by reflecting the exponential distribution around zero. It has pdf:

$$f(x) = 0.5\lambda e^{-\lambda|x|}, \qquad x \in \mathbb{R}$$

Suppose X has the above distribution with $\lambda = 1$.
Now let $Y = \sigma X + \mu$, $\mu \in \mathbb{R}$ (shifting) and $\sigma > 0$ (scaling).
Then Y has the *double exponential distribution* with pdf:

$$f_Y(y) = \frac{1}{2\sigma} \exp\left(- \frac{|y - \mu|}{\sigma} \right)$$

and moments

$$\mathrm{E}Y = \mu, \qquad \mathrm{Var}Y = 2\sigma^2$$

The double exponential distribution provides an alternative to the normal for centered data with fatter tails but all finite moments.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 12

Normal Distribution

Introduced by De Moivre (1667 - 1754) in 1733 as an approximation to the binomial. Later studied by Laplace and others as part of the Central Limit Theorem. Gauss derived the normal as a suitable distribution for outcomes that could be thought of as sums of many small deviations.

sample space: $R = (-\infty, \infty)$

pdf: For $Y \sim N(\mu, \sigma^2)$,

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2} \qquad -\infty < y < \infty$$

If $\mu = 0$ and $\sigma^2 = 1$ is referred to as the *standard normal*.

cdf: There is no closed form. The notation $\Phi(x)$ is often used for $F(x) = P(Y \leq x)$ for the standard normal case. Many books have tables of its values for $x > 0$. Values for $x < 0$ can be obtained by the formula $\Phi(-x) = 1 - \Phi(x)$.

Normal Moments

Mean:

$$\mathsf{E}Y = \mu$$

Variance:

$$\mathsf{Var}(Y) = \mathsf{E}(Y-\mu)^2 = \sigma^2$$

Higher central moments:

$$\mathsf{E}(Y-\mu)^m = \begin{cases} \frac{m!}{2^{m/2}(m/2)!}\sigma^m & m \text{ even} \\ 0 & m \text{ odd} \end{cases}$$

In particular:

$$\mu_3 = \mathsf{E}(Y-\mu)^3 = 0 \quad \text{(Skewness)}$$
$$\mu_4 = \mathsf{E}(Y-\mu)^4 = 3\sigma^4$$

Standartization

Standartization:

$$Y \sim N(\mu, \sigma^2) \quad \Leftrightarrow \quad Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

Shifting and scaling:

$$Z \sim N(0, 1) \quad \Leftrightarrow \quad Y = \sigma Z + \mu \sim N(\mu, \sigma^2)$$

Easy to prove using the mgf.

Density integrates to 1

Theorem:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = 1$$

Proof:

This is not as easy as one might think. Call the integral I. Then,

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{(x^2+y^2)}{2}} dx dy \end{aligned}$$

Now make a change of variables to polar coordinates, i.e. put

$$y = r \sin\theta \quad x = r\cos\theta, \quad 0 < \theta \leq 2\pi, \quad 0 < r < \infty$$

cont.

Now, $dxdy \rightarrow rdrd\theta$, because

$$\text{Jacobian} = J = \left| \begin{array}{cc} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \theta} & \frac{\partial y}{\partial \theta} \end{array} \right| = \left| \begin{array}{cc} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{array} \right| = r$$

$$\begin{aligned} I^2 &= \int_{r=0}^{\infty} \int_{\theta=0}^{2\pi} \frac{e^{-r^2/2}}{2\pi} r dr d\theta \\ &= \int_0^{\infty} e^{-r^2/2} r dr = -e^{-r^2/2}\Big|_0^{\infty} = 0 - (-1) = 1 \end{aligned}$$

Notes

- Normal distribution useful in many practical settings
- Plays an important role in *sampling distributions in large samples*, since the Central Limit theorem says that sums of independent identically distributed random variables are approximately normal
- There are many important distributions that can be derived from functions of normal random variables (e.g. χ^2, t, F). We will see much more on this later, for now, we will briefly present the *pdf's* and sample spaces of these distributions.

χ^2 distribution

If $Z \sim N(0,1)$, then $X = Z^2$ has the χ^2 *distribution* with 1 degree of freedom.

More generally, we have the χ^2 *distribution* with ν degrees of freedom with pdf:

$$f(x) = \frac{(x/2)^{\frac{\nu}{2}-1}e^{-x/2}}{2\Gamma(\nu/2)}, \qquad x > 0$$

where $\Gamma(a)$ is the complete gamma function,

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

Note that if a is an integer, $\Gamma(a) = (a-1)!$.

The $\chi^2(\nu)$ distribution is a special case of the gamma distribution, so it is easier to derive its properties from the gamma.

Student's t and F distributions

Y has a t_k distribution (t with ν degrees of freedom) if its *pdf* can be written as:

$$f(y) = \frac{\Gamma[(\nu+1)/2)]}{\sqrt{\nu\pi}\Gamma(\nu/2)} \frac{1}{(1+y^2/\nu)^{(\nu+1)/2}}, \qquad -\infty < y < \infty$$

Y has an $F(\nu_1, \nu_2)$ distribution if its *pdf* can be written as:

$$f(y) = \frac{(\nu_1/\nu_2)\Gamma[(\nu_1+\nu_2)/2)](\nu_1 y/\nu_2)^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(1+\nu_1 y/\nu_2)^{(\nu_1+\nu_2)/2}}, \qquad 0 \le y < \infty$$

There are many important properties and relationships between these three distributions (e.g. χ^2_k is the distribution of the sum of the squares of k independent standard normals). We'll come back to these in a few weeks when we do *sampling distributions and transformations of the normal distribution*.

Gamma distribution

Notation: $Y \sim gamma(a, \lambda)$.
pdf:

$$f(y) = \frac{\lambda e^{-\lambda y}(\lambda y)^{a-1}}{\Gamma(a)}, \qquad y \geq 0$$

where $\Gamma(a)$ is the *complete gamma function*,

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

Note that if a is an integer, $\Gamma(a) = (a-1)!$.

cdf: In general, there is no closed form, unless a is an integer.

moments

$$\begin{aligned} \mathsf{E}(Y) &= a/\lambda \\ \mathsf{Var}(Y) &= a/\lambda^2 \end{aligned}$$

Notes

- The special case $a = 1$ corresponds to an exponential(λ)
- Can be thought of as a flexible generalization of the exponential (a can be interpreted as a *shape parameter*)
- The special case $gamma(n/2, 1/2)$, for integer n, corresponds to the χ^2 distribution with n degrees of freedom.
- We will see later in the class that the gamma distribution can be derived as the sum of a independent exponential(λ) distributions
- When a is an integer, the *gamma*(a, λ) distribution can be derived as the distribution of time until the occurence of the a^{th} event in a Poisson process.

Weibull distribution

This is another useful generalization of the exponential. It is useful to begin with the *cdf* instead of the *pdf*:
sample space: $[v, \infty]$
cdf:

$$F(y) = 1 - \exp\left[-\left(\frac{y-v}{\alpha}\right)^{\beta}\right], \qquad y \geq v$$

It follows that the *pdf* is:

$$f(y) = \frac{\beta}{\alpha}\left(\frac{y-v}{\alpha}\right)^{(\beta-1)} \exp\left[-\left(\frac{y-v}{\alpha}\right)^{\beta}\right], \qquad y \geq v$$

The usual case is $v = 0$.
If $\beta = 1$ we get an exponential with parameter $\lambda = 1/\alpha$.

Cauchy distribution

This is a famous distribution to mathematical statisticians, since it often serves as a useful counterexample.
pdf

$$f(y) = \frac{1}{\pi}\frac{1}{[1 + (y-\mu)^2/\sigma^2]} \quad \text{for } -\infty < y < \infty$$

The Cauchy with $\mu = 0$, $\sigma = 1$, corresponds to the t-distribution with 1 degree of freedom.

While the moments of the Cauchy are not defined, its quantiles are (HW).

The Cauchy is not just a pathological case. We'll see later that the ratio of two standard normals is Cauchy. So ratios of observations can be problematic (e.g. BMI).

Beta distribution

Notation: $Y \sim beta(a, b)$.
sample space: [0,1]
pdf:

$$f(y) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}, \qquad 0 \leq y \leq 1$$

where $B(a, b)$ is the (complete) Beta function,

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where $\Gamma(a)$ is the complete gamma function. The normalizing constant is required so that $\int_0^1 f(x)dx = 1$.
Note that if a and b are integers, then $B(a, b)$ can be calculated in closed form.

cont.

cdf: In general, there is no closed form, except if a and b are integers.

moments

$$\begin{aligned} \mathsf{E}(Y) &= \frac{a}{a+b} \\ \mathsf{Var}(Y) &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

The beta distribution is very flexible, and can take a wide variety of shapes by varying its parameters.

Special case: $beta(1, 1) = U(0, 1)$.

* Read C-B Section 3.3 (normal, beta, Cauchy, lognormal and double exponential)

Location and Scale families

Let $f(x)$ be any pdf. Then the family of pdfs

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

for $\mu \in \mathbb{R}$, $\sigma > 0$, is called a location-scale family.
If $\mu = 0$ we get a scale family; if $\sigma = 1$ we get a location family.

Examples: Normal, Laplace, Cauchy, exponential.

Properties: Let $Z \sim f(z)$ and $X = \sigma Z + \mu$. Then

1. X has pdf $f_{\mu,\sigma}(x)$.
2. $$\mathsf{E}(X) = \sigma \mathsf{E}(Z) + \mu, \qquad \mathsf{Var}(X) = \sigma^2 \mathsf{Var}(Z)$$

Group families

Let $\mathcal{G}$ be a class of 1-to-1 functions $g : \mathbb{R} \to \mathbb{R}$. The class of transformations $\mathcal{G}$ is called a *transformation group* if

1. $\mathcal{G}$ is closed under composition: $g_1, g_2 \in \mathcal{G}$ implies $g_2 \cdot g_1 \in \mathcal{G}$.
2. $\mathcal{G}$ is closed under inversion: $g \in \mathcal{G}$ implies $g^{-1} \in \mathcal{G}$.

Given a rv Z with cdf $F(z)$, the class

$$\{X = g(Z), g \in \mathcal{G}\}$$

is a group family.

Group families: Examples

- *Parametric*: Location-scale families, $g(z) = \sigma z + \mu$, $\sigma > 0$, $\mu \in \mathbb{R}$.
- *Non-parametric*: Let $\mathcal{G}$ is the class of all continuous strictly increasing functions $g(z)$ such that

$$\lim_{z \to -\infty} g(z) = -\infty, \qquad \lim_{z \to \infty} g(z) = \infty$$

  Let Z be a rv supported on $(-\infty, \infty)$. Then the class $\{X = g(Z), g \in \mathcal{G}\}$ is the class of all rvs suported on $(-\infty, \infty)$ whose cdfs are continuous and strictly increasing.

- *Non-parametric*: Same as before with the additional restriction that Z has a symmetric distribution about 0 and g is odd: $g(-z) = -g(z)$. The generated rvs are now the class of all rvs with symmetric distributions about 0.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 13

Multiple Possible Outcomes of a Trial

Suppose: Result of drug trial is Failure, Partial Success and Success. Let

$$P(F) = p_1, \quad P(PS) = p_2, \quad P(S) = p_3, \quad (p_1 + p_2 + p_3) = 1$$

Suppose in a sample of size n

$$\begin{aligned} s_1 &= \text{Number of Failures} \\ s_2 &= \text{Number of Partial Successes} \\ s_3 &= \text{Number of Successes} \end{aligned}$$

$$where \quad s_1 + s_2 + s_3 = n,$$

$$P(s_1, s_2, s_3) = \frac{n!}{s_1! s_2! s_3!} p_1^{s_1} p_2^{s_2} p_3^{s_3}$$

These are *multinomial probabilities*.

continued

The probability of any ordered arrangement resulting in s_1"F", s_2 "PS" and s_3 "S" is

$$p_1^{s_1} p_2^{s_2} p_3^{s_3}$$

However there are $\frac{n!}{s_1! s_2! s_3!}$ such ordered arrangements. Therefore

$$P(s_1, s_2, s_3) = \frac{n!}{s_1! s_2! s_3!} p_1^{s_1} p_2^{s_2} p_3^{s_3}$$

Multinomial distribution

The generalization to k classes gives us the *Multinomial Distribution*

$$p(s_1, s_2, \ldots, s_k) = \frac{n!}{s_1! s_2! \ldots, s_k!} \; p_1^{s_1} p_2^{s_2} \ldots p_k^{s_k}$$

where $\sum_{i=1}^{k} s_i = n$ and $\sum_{i=1}^{k} p_i = 1$.

More on the Poisson 6 / 56

The Poisson distribution again

sample space: $\{0, 1, 2...\}$
pmf:

$$P(s) = \begin{cases} e^{-\lambda}\lambda^s/s! & s = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

cdf:

$$F(y) = \sum_{s=0}^{y} e^{-\lambda}\lambda^s/s!$$

expectation:

$$\mathsf{E}(Y) = \lambda$$

Variance:

$$\mathsf{Var}(Y) = \lambda$$

continued

Note: we can write

$$\frac{P(X=i+1)}{P(X=i)} = \frac{\lambda}{i+1}$$

Poisson: Examples and generalizations

Last time, we had an example concerning pulmonary embolism among young women. The rate is 4 per million, and we had looked at the pdf for the number of cases in a city with 1,000,000 women. But, suppose we are interested in a city that only has 100,000 women. How does the probability distribution change?

$$p = \frac{4}{1,000,000}, \quad n = 100,000, \quad np = \frac{4}{10} = .4 = \lambda$$

$P(s) = e^{-.4}(.4)^s/s!$

$\underline{s}$	$\underline{p(s)}$
0	.67
1	.27
2	.05
3	.007
4	.0007

continued

It is often useful to write

$P(s) = e^{-\lambda n}(\lambda n)^s/s!$ λ = rate/unit population
n = size of population

where λn is the expected number of events.
Note: "population" can be in units of 10, 100, 1000, etc. λ is the mean number of events per unit population.

Example: Failure of Equipment

A computer has a failure rate of 1 failure per 1,000 hours of use. How many failures would be expected in 500 hours of use?

What is the probability distribution of the number of failures?

$\underline{s}$	$\underline{P(s) = e^{-.5}(.5)^s/s!}$
0	.61
1	.30
2	.08
3	.01

More generally, the distribution may be written

$$P(s) = e^{-\lambda t}(\lambda t)^s/s!$$

where λt = number of failures (events) in t units of time.

Now suppose there are n computers, each being observed for t time units and each having a failure rate of λ, what is $P(s)$?

Continued

$P(s) = e^{-\lambda nt}(\lambda nt)^s/s!$

λ	$=$	rate per unit time per unit individual
n	$=$	number of units (individuals)
t	$=$	time frame
λnt	$=$	Expected number of events for n units and time t

Example: Leukemia

Suppose the incidence rate for childhood leukemia is 3.9 per 100,000 per year for children less than 4 years old. In a population of 5,000 children observed for 10 years, what would the probability distribution be for number of cases?

$$\lambda nt = \frac{3.9}{100,000} \times 5,000 \times 10 = 1.95$$

$$P(s) = e^{-1.95}(1.95)^s/s!$$

$\underline{s}$	$\underline{P(s)}$
0	.14
1	.28
2	.27
3	.18
4	.08
5	.03
6	.01
7	.003

Example: Safety Testing of Vaccine

Suppose a vaccine contains m live virus per cm^3. Suppose a sample of v cm^3 of vaccine is tested. The expected number of virus in v cm^3 is thus equal to mv. What is probability that vaccine tested will be free of a virus?

$$P(s) = e^{-mv}(mv)^s/s!$$

$$P(0) = e^{-mv}$$

e.g. Suppose $m = .005$ and $v = 600cc$. Then, $mv = 3$ and $P(0) = e^{-3} = .05$

Example: Leukemia in Woburn, MA

During a 19 year period 15 leukemias were observed. Is this an unusual event?

Age	Population (n)	Rate per $10^5/yr(\lambda)$	Expected Number
$0-4$	2120	6.27	.133
$5-9$	2191	3.09	.068
$10-14$	2969	2.04	.061
$15-19$	3592	2.19	.079
	$10,872$		.341

During a 19 year period, we expect (19)(.341) = 6.5

Therefore, $P(s) = e^{-6.5}(6.5)^s/s!$ and $\sum_{s=15}^{\infty} e^{-6.5}(6.5)^s/s! = .007$

Probability of observing 15 or more leukemias = .007.

continued

Note: For this example, we are glossing over the fact here that we are really combining several different populations that have different Poisson rates. We will see that if X_1 and X_2 are two independent Poisson random variables with parameters λ_1 and λ_2, then their sum is Poisson with parameter $\lambda_1 + \lambda_2$.

More Negative Binomial 17 / 56

Negative Binomial

In the context of iid Bernoulli trials, define a random variable corresponding to the number of trials required to have s successes. We say $Y \sim$ *Negbin*(s, p):
sample space: $\{s, (s+1), ...\}$
pmf: for $y = s, s+1, s+2, \ldots,$

$$f(y) = \left(\begin{array}{c} y-1 \\ s-1 \end{array} \right) p^s q^{y-s}$$

Why? If y is number of trials, then first $(y-1)$ trials resulted in $(s-1)$ successes and the last trial is success.
cdf: no closed form.
expectation: E$(Y) = s/p$
Variance: Var$(Y) = s(1-p)/p^2$
Recall that the negative binomial is the sum of s independent geometrics with parameter p such that these follow immediately.

Example

The Red Sox and the Atlanta Braves are playing in the world series. The winning team is the first one to win 4 games. Suppose each game is independent of the others, and that the Red Sox win a game with probability p. What is the probability that the Red Sox win?

$$\begin{aligned} P(\text{Red Sox wins}) &= \binom{3}{3} p^4 + \\ &\quad \binom{4}{3} p^4(1-p) + \\ &\quad \binom{5}{3} p^4(1-p)^2 + \\ &\quad \binom{6}{3} p^4(1-p)^3 \end{aligned}$$

Example (continued)

What is the probability that the series goes to 7 games?
If the Red Sox wins, then we need to have three Braves wins and three Red Sox wins (in any order) followed by a Red Sox win:

$$\binom{6}{3} p^4(1-p)^3.$$

Similarly, the probability that the Braves wins in 7 games is

$$\binom{6}{3} p^3(1-p)^4.$$

Hence the total probability is

$$\binom{6}{3} [p^4(1-p)^3 + p^3(1-p)^4].$$

Comments

- One of the most important things to get out of this class is to understand the different distributions and when/where you would choose to use them.
- You need to be very familiar with all of these distributions as well as simple transformations of these distributions e.g. what happens if you scale an exponential? what about a gamma?
- You will also need to be very familiar with what happens if there are multiple random variables (next class) and need to understand what happens if they are combined (transformed), e.g. what happens if you add Poissons, normals, exponentials? what happens if you take the ratio of a binomial to a Poisson?

5

Exponential Families

Exponential Families

A family of pdfs or pmfs with vector parameter $\boldsymbol{\theta}$ is called an *exponential family* if it can be expressed as

$$f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta})\, \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta})\, t_j(x)\right), \qquad x \in S \subset \mathbb{R}$$

where

- S is not defined in terms of $\boldsymbol{\theta}$ (i.e., **the support of the distribution does not depend on the unknown parameter**)
- $h(x), c(\boldsymbol{\theta}) \geq 0$ and the functions are just functions of the parameters specified; i.e. h is free of $\boldsymbol{\theta}$, $c(\boldsymbol{\theta})$ is free of x, etc...

Example: Binomial

Let $X \sim Binom(n, p)$, $0 < p < 1$.

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n \left[\frac{p}{1-p}\right]^x \\ &= \binom{n}{x} (1-p)^n \exp\left[\log\left(\frac{p}{1-p}\right) x\right] \end{aligned}$$

Thus,

$$\begin{aligned} h(x) &= \binom{n}{x}, \;\; x = 0, \ldots, n \qquad & w_1(p) &= \log\left(\frac{p}{1-p}\right) \\ c(p) &= (1-p)^n, \;\; 0 < p < 1 \qquad & t_1(x) &= x \end{aligned}$$

Note that this works when p is considered the parameter, while n is fixed. If n is not fixed, then the support depends on the unknown parameter. Also, p cannot be 0 or 1.

Example: Gaussian

Let $X \sim N(\mu, \sigma^2)$.

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right) \end{aligned}$$

Thus

$$\begin{aligned} h(x) &= \frac{1}{\sqrt{2\pi}} \qquad & c(\mu, \sigma) &= \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \\ w_1(\mu, \sigma) &= -\frac{1}{2\sigma^2} \qquad & w_2(\mu, \sigma) &= \frac{\mu}{\sigma^2} \\ t_1(x) &= x^2 \qquad & t_2(x) &= x \end{aligned}$$

The parameter space is $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$.

Theorem C-B 3.4.2

If X is a rv from the exponential family, then

$$\begin{aligned} \mathsf{E}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) &= -\frac{\partial}{\partial \theta_j} \log c(\boldsymbol{\theta}) \\ \mathsf{Var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) &= -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) \\ &\quad - \mathsf{E}\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(X)\right) \end{aligned}$$

Proof: Homework

Example: Binomial

Recall:

$$w_1(p) = \log \frac{p}{1-p}, \quad c(p) = (1-p)^n$$

Relevant derivatives:

$$\frac{\partial}{\partial p} w_1(p) = \frac{\partial}{\partial p} \log \frac{p}{1-p} = \frac{1}{p(1-p)}$$

$$\frac{\partial}{\partial p} \log c(p) = \frac{\partial}{\partial p} n \log(1-p) = \frac{-n}{1-p}$$

So

$$\mathsf{E}\left[\frac{1}{p(1-p)} X\right] = \frac{n}{1-p} \quad \Rightarrow \quad \mathsf{E}(X) = np$$

Indicator function

Definition 3.4.5. The *indicator function* of a set A, most often denoted by $I_A(x)$, is the function

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0, & \text{if } x \notin A. \end{cases}$$

Also denoted as $I(x \in A)$, $1_A(x)$, or $1(x \in A)$.

Note on exponential family:

- The set of x values for which $f(x|\boldsymbol{\theta}) > 0$ cannot depend on $\boldsymbol{\theta}$ in an exponential family.
- The entire definition of the pdf or pmf must be incorporated into the form for the exponential family.
- Incorporate the range of x into the expression for $f(x|\boldsymbol{\theta})$ through the use of an indicator function.

Example. Normal pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) I_{(-\infty,\infty)}(x).$$

More examples

Other exponential families are: Poisson, negative binomial, gamma, beta.

Some densities that are not exponential families: t, F.

Uniform: $X \sim U(0, \theta)$

$$f_X(x) = \theta^{-1} I(0 < x < \theta)$$

Truncated exponential:

$$f_X(x) = \theta^{-1} \exp(1 - x/\theta)\, I(\theta, \infty)$$

What about $X \sim U(0, 1)$?

Natural parameters

An exponential family can be reparametrized as

$$f(x|\boldsymbol{\eta}) = h(x)\, c^*(\boldsymbol{\eta})\, \exp\left(\sum_{j=1}^{k} \eta_j\, t_j(x)\right), \qquad x \in S \subset \mathbb{R}$$

where $\boldsymbol{\eta}$ is called the natural parameter vector.

This parametrization is often more useful. We have the following property:

$$\begin{aligned} \mathsf{E}\left[t_j(X)\right] &= -\frac{\partial}{\partial \eta_j} \log c^*(\boldsymbol{\eta}) \\ \mathsf{Var}\left[t_j(X)\right] &= -\frac{\partial^2}{\partial \eta_j^2} \log c^*(\boldsymbol{\eta}) \end{aligned}$$

Proof: Homework

Example: Gaussian

Let $X \sim N(\mu, \sigma^2)$. Recall:

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

and

$$w_1(\mu, \sigma) = -\frac{1}{2\sigma^2}, \quad w_2(\mu, \sigma) = \frac{\mu}{\sigma^2}$$

Define

$$\eta_1 = -\frac{1}{2\sigma^2} < 0, \qquad \eta_2 = \frac{\mu}{\sigma^2} \in \mathbb{R}$$

then

$$\sigma^2 = -\frac{1}{2\eta_1} > 0, \qquad \mu = -\frac{\eta_2}{2\eta_1} \in \mathbb{R}$$

The parameter space is now $(\eta_1, \eta_2) \in (-\infty, 0) \times \mathbb{R}$.

Curved exponential families

Let d be the dimension of the parameter space of the exponential family with k terms. The exp. family is called

$$\begin{array}{ll} \text{full} & \text{if} \quad d = k \\ \text{curved} & \text{if} \quad d < k \end{array}$$

Example: $X \sim N(\mu, \sigma^2)$. Suppose $\sigma^2 = \mu^2$, i.e. the coefficient of variation is constant equal to 1. The parameter space $(\mu, \sigma^2) = (\mu, \mu^2)$ is now a parabola . For the natural parameters:

$$\eta_1 = -\frac{1}{2\mu^2}, \quad \eta_2 = \frac{1}{\mu} \qquad \Rightarrow \qquad \eta_1 = -\frac{\eta_2^2}{2}$$

Example: Let $X_1, \ldots, X_n$ be an iid sample from $Po(\lambda)$. Let $\bar{X} = \sum_{i=1}^n X_i/n$. Then for large n (by the Central Limit Theorem (you will learn this in bios 661/673)),

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda}} \Rightarrow N(0,1) \qquad \text{so} \qquad X \dot{\sim} N(\lambda, \lambda)$$

This is a curved exp. family with parameter space $(\mu, \sigma^2) = (\mu, \mu)$.

Probability Inequalities 33 / 56

Chebychev Inequality

Let X be a random variable and let $g(x)$ be a non-negative function. Then for any $r > 0$,

$$P[g(X) \geq r] \leq \frac{\mathsf{E}g(X)}{r}$$

Proof:

$$\begin{aligned} \mathsf{E}g(X) &= \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx \\ &\geq \int_{\{x: g(x) \geq r\}} g(x)\, f_X(x)\, dx \\ &\geq r \int_{\{x: g(x) \geq r\}} f_X(x)\, dx \\ &= r\, P\{g(X) \geq r\} \end{aligned}$$

Application

Let

$$g(x) = \frac{(x-\mu)^2}{\sigma^2}$$

where $\mu = \mathsf{E}(X)$, $\sigma^2 = \mathsf{Var}(X)$.

Let $r = t^2$, then

$$P\left[\frac{(x-\mu)^2}{\sigma^2} \geq t^2\right] \leq \frac{1}{t^2}\mathsf{E}\left[\frac{(x-\mu)^2}{\sigma^2}\right]$$

$$P[|X-\mu| \geq t\sigma] \leq \frac{1}{t^2}$$

The probability that X is more than $t\sigma$ away from μ cannot be more than $1/t^2$, no matter what the distribution of X. E.g. $t = 2$.

Normal tail bound

Let $Z \sim N(0,1)$:

$$\begin{aligned} P[Z \geq t] &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \\ &\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t} \end{aligned}$$

and so

$$P[|Z| \geq t] = 2P[Z \geq t] \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}$$

For $t = 2$, the bound is 0.054.

Multiple random measurements

Multiple endpoints in health studies:

- Cancer: survival, quality of life, toxicity
- Reproductive health: time to pregnancy, birth defects
- AIDS: time from infection to AIDS, time from AIDS to death
- Carcinogenicity studies: time to tumor, time to death
- Health care: cost, hospital duration of stay

Multiple time points/spatial locations:

- Environmental monitoring: daily temperature, humidity, CO_2/ozone concentration, geographically located monitoring stations.
- Finance: daily stock prices, portfolios.

Massively multivariate:

- Genomics and other omics: thousands of gene expression values, SNPs, protein concentrations, metabolite concentrations.
- Imaging: thousands of voxels (volume pixels) measuring brain activity (fMRI), tumor metabolism (PET); satellite remote sensing.

Random Vectors

Suppose we start with a probability space $(\Omega, \mathcal{A}, P)$.

Defintion: An n-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a function from a sample space Ω into $\mathbb{R}^n$.

- Each coordinate X_i is a random variable.
- The random vector is associated with a probability space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), F)$.
- For every Borel set B,

$$P\{\boldsymbol{X} \in B\} = P\{\boldsymbol{X}^{-1}(B)\}$$

where

$$\boldsymbol{X}^{-1}(B) = \{\omega : \boldsymbol{X}(\omega) \in B\}$$

Example: Bivariate

A fair coin is flipped 3 times. Define the random vector (X, Y) where X represents the number of heads on the last toss and Y the total number of heads. Then, the probabilities of various outcomes are given in the following table:

Outcome	(x, y)	P(outcome)
(H,H,H)	(1,3)	1/8
(H,H,T)	(0,2)	1/8
(H,T,H)	(1,2)	1/8
(H,T,T)	(0,1)	1/8
(T,H,H)	(1,2)	1/8
(T,H,T)	(0,1)	1/8
(T,T,H)	(1,1)	1/8
(T,T,T)	(0,0)	1/8

Discrete Bivariate RVs

Two random variables X and Y are said to be jointly *discrete* if there is an associated *joint probability mass function*,

$$f_{X,Y}(x, y) = P\{X = x,\ Y = y\}$$

which sums to 1 over a finite or possibly countable combinations of x and y for which $f_{X,Y}(x, y) > 0$, i.e.,

$$\sum_{x,y} f_{X,Y}(x, y) = 1$$

From this, one can also obtain the marginal pmfs of X and Y as follows:

$$f_X(x) = P(X = x) = \sum_y f_{X,Y}(x, y)$$

$$f_Y(y) = P(Y = y) = \sum_x f_{X,Y}(x, y)$$

Example

Back to the fair coin example again. From the definition we can construct the joint pdf of X and Y:

		Y			
		0	1	2	3
X	0	1/8	1/4	1/8	0
	1	0	1/8	1/4	1/8

The marginal distributions of X and Y are also easy to find.

Note: Marginals do not determine joint pmf.

Bivariate cdfs

Regardless of whether they are discrete or continuous or some combination of the two, we can always define the *joint cumulative distribution function*.

For $n = 2$, the *bivariate cumulative distribution function* is

$$F_{X,Y}(x, y) = P\{X \leq x, Y \leq y\}$$

Properties:

- $F_{X,Y}(x, y) \geq 0$
- $F_{X,Y}(\infty, \infty) = 1$
- $F_{X,Y}(-\infty, y) = F(x, -\infty) = 0$
- $F_{X,Y}(-\infty, -\infty) = 0$
- F is non-decreasing and right-continuous in each variable separately

Marginal distributions

From $F_{X,Y}$, we can derive the univariate distribution functions for X and Y. These are generally called *marginal distributions*.

$$F_X(x) = P\{X \leq x\} = P\{X \leq x, Y < \infty\} = F_{X,Y}(x, \infty)$$
$$F_Y(y) = P\{Y \leq y\} = P\{X < \infty, Y \leq y\} = F_{X,Y}(\infty, y)$$

Note: Although we can obtain $F_X(x)$ and $F_Y(y)$ from the joint *cdf*, we cannot do the reverse.

Joint probabilities

All joint probability statements about X and Y can be answered in terms of their joint *cdf*. For example,

$$P(X > x, Y > y) = 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y)$$

More generally,

$$P(x_1 < X \leq x_2,\ y_1 < Y \leq y_2) =$$
$$F_{X,Y}(x_2, y_2) + F_{X,Y}(x_1, y_1) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1)$$

Comments:

Continuous Bivariate Random Variables

Continuous Bivariate RVs

The random variables X and Y are said to be *jointly (absolutely) continuous* if there exists a function $f_{X,Y}(x,y)$, such that for any Borel set B of 2-tuples in $\mathbb{R}^2$,

$$P\{(X,Y) \in B\} = \int\int_{(x,y)\in B} f_{X,Y}(x,y)dxdy$$

The function $f_{X,Y}(x,y)$ is called the *joint probability density function* for X and Y.

It follows in this case that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)dsdt,$$

$$f_{X,Y}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$$

Properties of the bivariate pdf

- $f_{X,Y}(x,y) \geq 0$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$
- $f_{X,Y}(x,y)$ is **not a probability**, but can be thought of as a relative probability of (X,Y) falling into a small rectangle located at (x,y):

$$P\{x < X \leq x+dx,\ \ y < Y \leq y+dy\} \approx f(x,y)dxdy$$

- The *marginal probability density functions* for X and Y can be obtained as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Example 1

$$F_{XY}(x,y) = xy \qquad 0 < x \leq 1, \quad 0 < y \leq 1$$

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x\, \partial y} = 1$$

$$f_X(x) = \int_0^1 dy = 1$$

$$f_Y(y) = \int_0^1 dx = 1$$

Example 2

$$F_{XY}(x,y) = x - x\ \log(\tfrac{x}{y}) \qquad 0 < x \leq y \leq 1$$

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}}{\partial x \partial y} = \frac{\partial}{\partial x}[-x(\tfrac{y}{x})(-\tfrac{x}{y^2})] = \frac{\partial}{\partial x}\frac{x}{y} = \frac{1}{y}$$

$$f_X(x) = \int_x^1 \frac{dy}{y} = -\log(x)$$
$$f_Y(y) = \int_0^y \frac{dx}{y} = 1$$

Note: Once we have $f_X(y)$ and $f_Y(y)$, we can obtain $F_X(x)$ and $F_Y(y)$ directly.

Double check:

$$\begin{aligned} F_X(x) &= F_{X,Y}(x,1) = x - x\log(x); \\ &\quad \frac{d}{dx}[x - x\log(x)] = -\log(x). \\ F_Y(y) &= F_{X,Y}(y,y) = y; \\ &\quad \frac{d}{dy}y = 1. \end{aligned}$$

Conditional Distributions - Discrete

Recall if A and B are two events, the probability of A conditional on B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{or} \quad \frac{P(AB)}{P(B)}$$

Defining the events $A = \{Y = y\}$ and $B = \{X = x\}$, it follows that

$$\begin{aligned} P\{Y = y|X = x\} &= \frac{P(X = x,\ Y = y)}{P(X = x)} \\ &= \frac{f_{X,Y}(x, y)}{f_X(x)} \\ &= f_{Y|X}(y|x) \end{aligned}$$

This is called the **conditional probability mass function** of Y given X.

Example: Discrete

A fair coin is flipped 3 times. Define the random vector (X, Y) where X represents the number of heads on the last toss and Y the total number of heads. From the joint pmf of X and Y we can derive all the conditional pmfs: $f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x)$ and $f_{X|Y}(x|y) = f_{X,Y}(x, y)/f_Y(y)$.

Examples: $f_{Y|X}(0|0) = f_{X,Y}(0, 0)/f_X(0) = \frac{1/8}{1/2} = 1/4$;
$f_{X|Y}(1|2) = f_{X,Y}(1, 2)/f_Y(2) = \frac{1/4}{3/8} = 2/3$.

		Y				
		0	1	2	3	Sum
X	0	1/8	1/4	1/8	0	1/2
	1	0	1/8	1/4	1/8	1/2
	Sum	1/8	3/8	3/8	1/8	1

Conditional Distributions - Continuous

If $F(x, y)$ is absolutely continuous, we define the conditional density of Y given X as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \text{ if } f_Y(y) > 0$$

Because Y is continuous, we cannot directly evaluate this probability, since the denominator will be zero. Instead, think of small dx, dy:

$$\begin{aligned} & Pr(x \leq X < x + dx \,|\, y \leq Y < y + dy) \\ = & \frac{Pr(x \leq X < x + dx, y \leq Y < y + dy)}{Pr(y \leq Y < y + dy)} \\ \approx & \frac{f(x, y)dxdy}{f_Y(y)dy} \\ = & f_{X|Y}(x|y)dx \end{aligned}$$

Show that it satisfies the conditions for a density.

Example 1

$$\begin{array}{lll} F_{XY}(x, y) = xy & 0 < x \leq 1, \quad 0 < y \leq 1 & \\ f_{XY}(x, y) = 1 & 0 < x < 1, \quad 0 < y < 1 & \\ f_X(x) = 1 & 0 < x < 1 & \\ f_Y(y) = 1 & 0 < y < 1 & \\ f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = 1 & 0 < x < 1 & (0 < y < 1) \\ f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = 1 & 0 < y < 1 & (0 < x < 1) \end{array}$$

In this particular case, we get that the conditional densities are the same as the marginals. This means X and Y are independent.

Example 2

$$F_{XY}(x,y) = x - x\ \log\tfrac{x}{y} \qquad 0 < x \le y \le 1$$

$$f_{XY}(x,y) = 1/y \qquad 0 < x \le y \le 1$$

$$f_X(x) = -\log x \qquad 0 < x \le 1$$

$$f_Y(y) = 1 \qquad 0 < y \le 1$$

$$f_{X|Y}(x|y) = \tfrac{f_{XY}(x,y)}{f_Y(y)} = 1/y \qquad 0 < x \le y \qquad (0 < y \le 1)$$

$$f_{Y|X}(y|x) = \tfrac{f_{XY}(x,y)}{f_X(x)} = -\tfrac{1}{y \log x} \qquad x \le y \le 1 \qquad (0 < x \le 1)$$

- Y is marginally uniform, but not conditionally
- X is conditionally uniform, but not marginally

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 14

Independence

The random variables X and Y are said to be *independent* if for any two Borel sets A and B,

$$P(X \in A,\ Y \in B) = P(X \in A)\, P(Y \in B)$$

All events defined in terms of X are independent of all events defined in terms of Y.

Using the Kolmogorov axioms of probability, it can be shown that X and Y are independent if and only if $\forall(x, y)$ (*except possibly for sets of prob. 0*)

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

or in terms of *pmfs* (discrete) and *pdf*'s (continuous)

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

* Check previous examples.

Checking independence

1. A necessary condition for independence of X and Y is that their joint pdf/pmf has positive probability on a rectangular domain.
2. If the domain is rectangular, one can try to write the joint pdf/pmf as a product of functions of x and y only.

Lemma C-B 4.2.7: Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ such that for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f(x, y) = g(x)h(y)$$

Proof:

Example

Two points are selected randomly on a line of length a so as to be on opposite sides of the mid-point of the line. Find the probability that the distance between them is less than $a/3$.

Solution:
Let X be the coordinate of a point selected randomly in $[0, a/2]$ and Y the coordinate of a point selected randomly in $[a/2, a]$. Assume X and Y are independent and uniform over its interval. The joint density is:

$$f_{X,Y}(x, y) = 4/a^2, \qquad 0 \leq x \leq a/2, \;\; a/2 \leq y \leq a$$

Hence, the solution is

$$P(Y - X < a/3) = \int_{a/6}^{a/2} \int_{a/2}^{a/3+x} \frac{4}{a^2} dy dx = 2/9$$

Example: Buffon's Needle

A table is ruled with lines distance 1 unit apart. A needle of length $L \leq 1$ is thrown randomly on the table. What is the probability that the needle intersects a line?

Solution: Define two random variables:

- X: distance from low end of the needle to the nearest line above
- θ: angle from the vertical to the needle.

Example (cont.)

By "random", we assume X and θ are independent, and

$$X \sim U(0,1) \quad \text{and} \quad \theta \sim U[-\pi/2, \pi/2].$$

This means that

$$f_{X,\Theta}(x,\theta) = 1/\pi, \qquad 0 \le x \le 1, \quad -\pi/2 \le \theta \le \pi/2$$

For the needle to intersect a line, we need $X < L\ \cos(\theta)$. So,

$$\begin{aligned} P(\text{needle intersects a line}) &= P(X < L\cos(\theta)) \\ &= \int_{-\pi/2}^{\pi/2} \int_0^{L\cos\theta} \frac{1}{\pi} dx d\theta \\ &= \frac{2L}{\pi} \end{aligned}$$

Expectations of Independent RVs

Theorem C-B 4.2.10 Let X and Y be independent rvs.

- For any $A \subset \mathcal{R}$ and $B \subset \mathcal{R}$,

  $$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

  i.e. the events $\{X \in A\}$ and $\{Y \in B\}$ are independent.
  (C-B write it as a theorem, we took it as definition.)
- Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then

  $$\mathsf{E}(g(X)h(Y)) = (\mathsf{E}g(X))(\mathsf{E}h(Y))$$

Example: X, Y indep.

$$\mathsf{E}(X^2Y^3) = (\mathsf{E}X^2)(\mathsf{E}Y^3)$$
$$\mathsf{E}(Y^2Y^3) \neq (\mathsf{E}Y^2)(\mathsf{E}Y^3)$$

Proof

$$\begin{aligned}
\mathsf{E}(g(X)h(Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_{X,Y}(x,y)dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dx\,dy \\
&= \left(\int_{-\infty}^{\infty} g(x)f_X(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)dy\right) \\
&= (\mathsf{E}\,g(X))(\mathsf{E}\,h(Y))
\end{aligned}$$

Bivariate Transformations 10 / 29

Functions of random vectors

Let (X, Y) be a bivariate rv with known distribution. Define (U, V) by

$$U = g_1(X, Y), \qquad V = g_2(X, Y)$$

Probability mapping: For any Borel set $B \subset \mathbb{R}^2$,

$$P[(U, V) \in B] = P[(X, Y) \in A]$$

where A is the inverse mapping of B, i.e.

$$A = \left\{(x, y) \in \mathbb{R}^2 : \ \ (g_1(x, y), g_2(x, y)) \in B\right\}$$

The inverse is well defined even if the mapping is not bijective.

Example: Let $g_1(x, y) = x$, $g_2(x, y) = x^2 + y^2$.

Discrete RVs

Suppose that (X, Y) is a discrete rv, i.e. the pmf is positive on a countable set $\mathcal{A}$. Then (U, V) is also discrete and takes values on a countable set $\mathcal{B}$. Define

$$A_{uv} = \big\{(x, y) \in \mathcal{A} : \;\; g_1(x, y) = u, \;\; g_2(x, y) = v\big\}$$

Then

$$f_{U,V}(u, v) = P(U = u, V = v) = \sum_{(x,y)\in A_{uv}} f_{X,Y}(x, y)$$

Sum of two independent Poissons

Let $X \sim Po(\lambda_1)$, $Y \sim Po(\lambda_2)$, independent, and define

$$U = X + Y, \qquad V = Y$$

- (X, Y) takes values in $\mathcal{A} = \{0, 1, 2, \dots\}^2$.
- (U, V) takes values on $\mathcal{B} = \{(u, v) : \;\; v = 0, 1, 2, \dots, \;\; u = v, v + 1, v + 2, \dots\}$.
- For a particular (u, v), $A_{uv} = \{(x, y) \in \mathcal{A} : \;\; x + y = u, \;\; y = v\} = (u - v, u)$.

The joint pmf of U and V is

$$f_{U,V}(u, v) = f_{X,Y}(u - v, v) = \frac{e^{-\lambda_1}\, \lambda_1^{u-v}}{(u - v)!} \; \frac{e^{-\lambda_2}\, \lambda_2^{v}}{v!}$$

cont.

The distribution of $U = X + Y$ is the marginal

$$\begin{aligned} f_U(u) &= \sum_{v=0}^{u} \frac{e^{-\lambda_1}\ \lambda_1^{u-v}}{(u-v)!}\ \frac{e^{-\lambda_2}\ \lambda_2^{v}}{v!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{u!} \sum_{v=0}^{u} \left(\begin{array}{c} u \\ v \end{array} \right) \lambda_1^{u-v}\ \lambda_2^{v} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{u!} (\lambda_1 + \lambda_2)^u \end{aligned}$$

We obtain that U is Poisson with parameter $\lambda = \lambda_1 + \lambda_2$.

Bivariate Transformations of Continuous RVs 15 / 29

Continuous RVs

Suppose (X, Y) is continuous and the joint transformation

$$u = g_1(x, y), \qquad v = g_2(x, y)$$

is one-to-one and differentiable. Define the inverse mapping

$$x = h_1(u, v), \qquad y = h_2(u, v)$$

Then

$$f_{UV}(u, v) = f_{XY}\big(h_1(u, v), h_2(u, v)\big)\ |J(u, v)|$$

where $J(u, v)$ is the Jacobian of the transformation $(x, y) \to (u, v)$ given by

$$J(u, v) = \frac{\partial(x, y)}{\partial(u, v)} = \left| \begin{array}{cc} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{array} \right|$$

Rotation of a bivariate normal vector

Let $X \sim N(0,1)$, $Y \sim N(0,1)$, independent. Define the rotation

$$U = X\cos\theta - Y\sin\theta$$
$$V = X\sin\theta + Y\cos\theta$$

for fixed θ. Then $U \sim N(0,1)$, $V \sim N(0,1)$, independent.

Proof: The range of (X,Y) is $\mathbb{R}^2$. The range of (U,V) is $\mathbb{R}^2$. Need the inverse transformation

$$X = U\cos\theta + V\sin\theta$$
$$Y = -U\sin\theta + V\cos\theta$$

with Jacobian

$$J(u,v) = \left| \begin{array}{cc} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{array} \right| = \left| \begin{array}{cc} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{array} \right| = 1$$

cont.

The joint pdf of (X,Y) is

$$f_{XY}(x,y) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}}e^{-y^2/2} = \frac{1}{2\pi}e^{-(x^2+y^2)/2}$$

The joint pdf of (U,V) is

$$\begin{aligned} f_{UV}(u,v) &= \frac{1}{2\pi}e^{-[(u\cos\theta+v\sin\theta)^2+(-u\sin\theta+v\cos\theta)^2]/2} \cdot |1| \\ &= \frac{1}{2\pi}e^{-(u^2+v^2)/2} = \frac{1}{\sqrt{2\pi}}e^{-u^2/2} \cdot \frac{1}{\sqrt{2\pi}}e^{-v^2/2} \end{aligned}$$

so $U \sim N(0,1)$, $V \sim N(0,1)$, independent.

Functions of independent random variables

Theorem C-B 4.3.5: Let X and Y be independent rvs. Let $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ be functions. Then the random variables $U = g(X)$ and $V = h(Y)$ are independent.

Proof:

Extensions of previous example

- Suppose $X \sim N(0, \sigma^2)$, $Y \sim N(0, \sigma^2)$, independent

$$U = a(X \cos\theta - Y \sin\theta)$$
$$V = a(X \sin\theta + Y \cos\theta)$$

  Then $U \sim N(0, a^2\sigma^2)$, $V \sim N(0, a^2\sigma^2)$, independent.

- Above, take $\theta = \pi/4$, $a = \sqrt{2}$:

$$U = \sqrt{2}(X/\sqrt{2} - Y/\sqrt{2}) = X - Y$$
$$V = \sqrt{2}(X/\sqrt{2} + Y/\sqrt{2}) = X + Y$$

  We get $U \sim N(0, 2\sigma^2)$, $V \sim N(0, 2\sigma^2)$, independent.

Ratio of two independent normals

Let $X \sim N(0,1)$, $Y \sim N(0,1)$, independent. The ratio X/Y has the Cauchy distribution.

Proof: Define the variables

$$U = X/Y, \qquad V = Y$$

with inverse

$$X = UV, \qquad Y = V$$

The Jacobian is

$$J(u,v) = \left| \begin{array}{cc} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{array} \right| = \left| \begin{array}{cc} v & u \\ 0 & 1 \end{array} \right| = v$$

The range of (X,Y) is $\mathbb{R}^2$. The range of (U,V) is $\mathbb{R}^2$.

cont.

The joint pdf of (X,Y) is

$$f_{XY}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$$

The joint pdf of (U,V) is

$$f_{UV}(u,v) = \frac{1}{2\pi} e^{-[(uv)^2+v^2]/2} \cdot |v| = \frac{|v|}{2\pi} e^{-(u^2+1)v^2/2}$$

The marginal of U is

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_{UV}(u,v)\, dv = 2 \int_0^{\infty} \frac{v}{2\pi} e^{-(u^2+1)v^2/2}\, dv \\ &= \frac{1}{\pi} \int_0^{\infty} e^{-(u^2+1)z}\, dz = \frac{1}{\pi(u^2+1)} \end{aligned}$$

Sum of Two Independent RVs

Suppose X and Y are independent. What is the distribution of $Z = X + Y$?

In general:

$$F_Z(z) = P(X + Y \leq z) = P(\{(x, y) \text{ such that } x + y \leq z\})$$

Various approaches:

- Bivariate transformation method (continuous and discrete)
- Discrete convolution

$$f_Z(z) = \sum_{x+y=z} f_X(x) f_Y(y) = \sum_x f_X(x) f_Y(z - x)$$

- Continuous convolution (C-B Section 5.2)
- Mgf/cf method (continuous and discrete)

Sum of two independent Poissons

Define X, Y to be two independent random variables having Poisson distributions with parameters $\lambda_i, \ \ i = 1, 2$. Then:

$$f_{X,Y}(x, y) = \frac{e^{-\lambda_1} \lambda_1^x}{x!} \frac{e^{-\lambda_2} \lambda_2^y}{y!} \qquad x, y = 0, 1, 2, ...$$

The distribution of $S = X + Y$ is

$$\begin{aligned} f_S(s) &= \sum_{x=0}^{s} \frac{e^{-\lambda_1} \lambda_1^x}{x!} \frac{e^{-\lambda_2} \lambda_2^{s-x}}{(s-x)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{s!} \sum_{x=0}^{s} \binom{s}{x} \lambda_1^x \lambda_2^{s-x} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{s!} (\lambda_1 + \lambda_2)^s \end{aligned}$$

Thus S is Poisson with parameter $\lambda = \lambda_1 + \lambda_2$.

Characteristic Function

Theorem 4.2.12 Let X and Y be independent rvs with characteristic functions $\phi_X(\cdot)$ and $\phi_Y(\cdot)$, respectively. Then the characteristic function of $Z = X + Y$ is

$$\phi_Z(\theta) = \phi_X(\theta)\,\phi_Y(\theta)$$

Proof:

$$\begin{aligned}
\phi_Z(\theta) &= \mathsf{E}\exp(iZ\theta) &&= \mathsf{E}\exp\left[i(X+Y)\theta\right] \\
&= \mathsf{E}\exp(iX\theta)\exp(iY\theta) &&= \mathsf{E}\exp(iX\theta)\mathsf{E}\exp(iY\theta) \\
&= \phi_X(\theta)\,\phi_Y(\theta)
\end{aligned}$$

Corollary If X and Y independent and $Z = X - Y$,

$$\phi_Z(\theta) = \phi_X(\theta)\,\phi_Y(-\theta)$$

Sum of two independent Poissons

Suppose $X \sim \text{Poisson}(\lambda_X)$ and $Y \sim \text{Poisson}(\lambda_Y)$ and put $Z = X + Y$. Then, $Z \sim \text{Poisson}(\lambda_X + \lambda_Y)$.

Proof:

$$\begin{aligned}
\phi_Z(\theta) &= \exp\left[\lambda_X\left(e^{\theta}-1\right)\right]\exp\left[\lambda_Y\left(e^{\theta}-1\right)\right] \\
&= \exp\left[\left(\lambda_X+\lambda_Y\right)\left(e^{\theta}-1\right)\right]
\end{aligned}$$

Sum of Two Independent Normals

Suppose $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ and X and Y are independent and $Z = X + Y$ then

$$Z \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

Proof

$$\begin{aligned}
\phi_Z(\theta) &= \exp\left(i\,\mu_x\theta - \frac{1}{2}\sigma_x^2\theta^2\right)\,\exp\left(i\,\mu_y\theta - \frac{1}{2}\sigma_y^2\theta^2\right) \\
&= \exp\left[i\,(\mu_x + \mu_y)\theta - \frac{1}{2}(\sigma_x^2 + \sigma_y^2)\theta^2\right]
\end{aligned}$$

Sum of two independent gammas

Suppose $X \sim \Gamma(\alpha_y, \beta)$ and independently $Y \sim \Gamma(\alpha_y, \beta)$ and put $Z = X + Y$. Then, $Z \sim \Gamma((\alpha_x + \alpha_y), \beta)$

Proof:

$$\begin{aligned}
\phi_Z(\theta) &= \left(\frac{1}{1-\beta\theta}\right)^{\alpha_x}\left(\frac{1}{1-\beta\theta}\right)^{\alpha_y} \\
&= \left(\frac{1}{1-\beta\theta}\right)^{\alpha_x+\alpha_y}
\end{aligned}$$

Remember that

- If $\alpha = 1$ we have an exponential with parameter β.
- If $\alpha = n/2$ and $\beta = 2$, we have a $\chi^2(n)$ (with n d.f.). The above result states that $\chi^2(n_1) + \chi^2(n_2) = \chi^2(n_1 + n_2)$.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 15

Conditional Expectation

Suppose we have discrete rvs X and Y with conditional pmf $f_{Y|X}(y|x)$. The *conditional expectation* of $g(Y)$ given $X = x$ is

$$\mathsf{E}[g(Y)|X = x] = \sum_y g(y)\ f_{Y|X}(y|x)$$

Notice that this is a function $h(x)$ of x. We may define the random variable

$$h(X) = \mathsf{E}[g(Y)|X]$$

In particular, if $X = x$ then $\mathsf{E}[g(Y)|X] = \mathsf{E}[g(Y)|x]$.

For continuous rvs:

$$h(x) = \mathsf{E}[g(Y)|x] = \int_{-\infty}^{\infty} g(y)\ f_{Y|X}(y|x)\, dy$$

$$h(X) = \mathsf{E}[g(Y)|X]$$

Example

Let $X \sim Exp(\lambda)$, $Z \sim Exp(\lambda)$, independent, be the times to mitosis of a cell and its daughter. Define the total time $Y = X + Z$.

Given $X = x > 0$, $Y = x + Z$ is a shifted exponential rv

$$f_{Y|X}(y|x) = \lambda e^{-\lambda(y-x)}, \qquad y > x,$$

The conditional expectation of Y given $X = x$ is

$$\mathsf{E}[Y|X = x] = \int_x^{\infty} y\ \lambda e^{-\lambda(y-x)}\, dy = x + 1/\lambda$$

yielding the rv

$$\mathsf{E}[Y|X] = X + 1/\lambda$$

More directly

$$\mathsf{E}(Y|X) = \mathsf{E}(X + Z|X) = \mathsf{E}(X|X) + \mathsf{E}(Z|X) = X + 1/\lambda$$

Iterative expectation formula

If X and Y are any two random variables then

$$\mathsf{E}X = \mathsf{E}(\mathsf{E}(X|Y))$$

Proof:

$$\begin{aligned}
\mathsf{E}(\mathsf{E}(X|Y)) &= \int \mathsf{E}(X|y) f(y) dy \\
&= \int \left[\int x f(x|y) dx \right] f(y) dy \\
&= \int \int x f(x,y) dx dy \\
&= \mathsf{E}X
\end{aligned}$$

Notice that the expectations are with respect to different variables and densities. Careful!

Example

Back to the cells example:

$$\mathsf{E}(Y|X) = X + 1/\lambda$$

so

$$\mathsf{E}[\mathsf{E}(Y|X)] = \mathsf{E}[X + 1/\lambda] = \mathsf{E}X + 1/\lambda = 1/\lambda + 1/\lambda = 2/\lambda$$

This should be equal to $\mathsf{E}Y$, which is

$$\mathsf{E}Y = \mathsf{E}(X + Z) = \mathsf{E}X + \mathsf{E}Z = 1/\lambda + 1/\lambda = 2/\lambda$$

Hierarchical Model

Example: Suppose an insect lays eggs according to a Poisson(λ) and each egg survives with probability p. Assume that the survival of eggs is independent of each other, then what is the average number of eggs surviving?

Let's say $Y \sim$ Poisson(λ) and $X|Y \sim$ Binomial(Y, p) where X is the total number of eggs surviving.

$$\begin{aligned} P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) \\ &= \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \\ &= \sum_{y=x}^{\infty} \left[\binom{y}{x} p^x (1-p)^{y-x}\right] \left[\frac{e^{-\lambda}\lambda^y}{y!}\right] \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \end{aligned}$$

cont

$$\begin{aligned} &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \exp((1-p)\lambda) \\ &= \frac{(\lambda p)^x}{x!} \exp(-\lambda p) \end{aligned}$$

So $X \sim$ Poisson (λp) and E$X = \lambda p$.

Using the iterative expectation formula

$$\mathsf{E}X = \mathsf{E}(\mathsf{E}(X|Y)) = \mathsf{E}(Yp) = \lambda p$$

Three-layer herarchical model

Example: In the previous example, suppose that there are several insect mothers, each with different average number of eggs. Model λ of the Poisson as being an exponential rv Λ with mean parameter β. Now what is the expected number of eggs surviving?

$$\begin{aligned} \mathsf{E}X &= \mathsf{E}(\mathsf{E}(X|Y)) \\ &= \mathsf{E}(pY) \\ &= \mathsf{E}(\mathsf{E}(pY|\Lambda)) \\ &= \mathsf{E}(p\Lambda) \\ &= p\beta \end{aligned}$$

Conditional Variance

Suppose we have rvs X and Y. Recall that the marginal variance of $g(Y)$ is

$$\mathsf{Var}[g(Y)] = \mathsf{E}[g(Y) - \mathsf{E}(g(Y))]^2$$

The *conditional variance* of $g(Y)$ given X is

$$\mathsf{Var}[g(Y)|X] = \mathsf{E}\big\{[g(Y) - \mathsf{E}(g(Y)|X)]^2|X\big\}$$

where both expectations are taken with respect to the conditional pmf or pdf $f_{Y|X}(y)$.
Like the conditional expectation, the conditional variance of Y given X is a random variable whose value depends on the rv X.

Conditional variance hierarchical formula

For any two random variables X and Y,

$$\mathrm{Var}X = \mathrm{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(\mathrm{E}(X|Y))$$

Proof

$$\begin{aligned}
\mathrm{Var}X &= \mathrm{E}\{[X - \mathrm{E}X]^2\} \\
&= \mathrm{E}\{[X - \mathrm{E}(X|Y) + \mathrm{E}(X|Y) - \mathrm{E}X]^2\} \\
&= \mathrm{E}\{[X - \mathrm{E}(X|Y)]^2\} + \mathrm{E}\{[\mathrm{E}(X|Y) - \mathrm{E}X]^2\} \\
&\quad +2\mathrm{E}\{[X - \mathrm{E}(X|Y)][\mathrm{E}(X|Y) - \mathrm{E}X]\}
\end{aligned}$$

Study the three terms separately. 1st term:

$$\begin{aligned}
\mathrm{E}\{[X - \mathrm{E}(X|Y)]^2\} &= \mathrm{E}\Big(\mathrm{E}\{[X - \mathrm{E}(X|Y)]^2|Y\}\Big) \\
&= \mathrm{E}(\mathrm{Var}(X|Y))
\end{aligned}$$

cont

2nd term:

$$\begin{aligned}
\mathrm{E}\{[\mathrm{E}(X|Y) - \mathrm{E}X]^2\} &= \mathrm{E}\{[\mathrm{E}(X|Y) - \mathrm{E}(\mathrm{E}(X|Y))]^2\} \\
&= \mathrm{Var}(\mathrm{E}(X|Y))
\end{aligned}$$

3rd term:

$$\begin{aligned}
&\mathrm{E}\{[X - \mathrm{E}(X|Y)][\mathrm{E}(X|Y) - \mathrm{E}X]\} \\
&= \mathrm{E}\Big(\mathrm{E}\{[X - \mathrm{E}(X|Y)][\mathrm{E}(X|Y) - \mathrm{E}X]|Y\}\Big) \\
&= \mathrm{E}\Big([\mathrm{E}(X|Y) - \mathrm{E}X]\ \mathrm{E}\{[X - \mathrm{E}(X|Y)]|Y\}\Big) \\
&= \mathrm{E}\Big([\mathrm{E}(X|Y) - \mathrm{E}X]\ \{\mathrm{E}[X|Y] - \mathrm{E}(X|Y)\}\Big) \\
&= 0
\end{aligned}$$

Example

In the the Poisson-Binomial hierarchical model, we had

$$Y \sim \text{Poisson}(\lambda), \qquad X|Y \sim \text{Binomial}(Y, p)$$

and showed that $\text{E}X = \lambda p$. Using the conditional variance formula,

$$\begin{aligned} \text{Var}X &= \text{E}(\text{Var}(X|Y)) + \text{Var}(\text{E}(X|Y)) \\ &= \text{E}(Yp(1-p)) + \text{Var}(Yp) \\ &= \lambda p(1-p) + \lambda p^2 \\ &= \lambda p \end{aligned}$$

This is consistent with the result that $X \sim$Poisson(λp).

Covariance and Correlation 14 / 44

Definitions

Let X and Y be two random variables with respective means μ_X, μ_Y and variances $\sigma_X^2 > 0$ and $\sigma_Y^2 > 0$, all assumed to exist.

- The *covariance* of X and Y is

$$\text{Cov}(X, Y) = \text{E}[(X - \mu_X)(Y - \mu_Y)] = \sigma_{XY}$$

- The *correlation* between X and Y is

$$\begin{aligned} \rho_{XY} &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}X \ \text{Var}Y}} = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y} \\ &= \text{E}\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] \end{aligned}$$

- X and Y are called *uncorrelated* iff

$$\text{Cov}(X, Y) = 0 \qquad \text{or equivalently} \qquad \rho_{XY} = 0$$

Properties of Covariance

1. $\text{Cov}(X, X) = \text{Var}(X)$
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
3. For any constant c, $\text{Cov}(X, c) = 0$
4. $\text{Cov}(X, Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y)$
5. If X and Y are independent and $\text{Cov}(X, Y)$ exists, then $\text{Cov}(X, Y) = 0$.

Note: If X and Y are uncorrelated, this does not imply that they are independent. Example:

$$X \sim U[-1, 1], \qquad Y = \begin{cases} X, & \text{prob. } 1/2 \\ -X, & \text{prob. } 1/2 \end{cases}$$

Linear Combinations

If X, Y, and Z are rvs each with a variance, and a and b are constants, then

$$\begin{aligned} \text{Cov}(aX + bY, Z) &= a\text{Cov}(X, Z) + b\text{Cov}(Y, Z) \\ \text{Var}(aX + bY) &= a^2\text{Var}X + b^2\text{Var}Y + 2ab\text{Cov}(X, Y) \\ \text{Corr}(aX + b, cY + d) &= sign(ac)\text{Corr}(X, Y) \end{aligned}$$

Proof:

Correlation Coefficient

Theorem C-B 4.5.7 For any rvs X and Y,

1. $-1 \leq \rho_{XY} \leq 1$
2. $|\rho_{XY}| = 1$ if and only if $\exists\ a \neq 0$ and b such that

$$P(Y = aX + b) = 1.$$

If $\rho_{XY} = 1 \Rightarrow a > 0$, and if $\rho_{XY} = -1 \Rightarrow a < 0$.

Proof: Let $\tilde{X} = (X - \mu_X)/\sigma_X$, $\tilde{Y} = (Y - \mu_Y)/\sigma_Y$, so that $\rho_{XY} = \mathsf{E}(\tilde{X}\tilde{Y})$.

1.
$$0 \leq \mathsf{E}(\tilde{X} - \tilde{Y})^2 = 1 + 1 - 2\mathsf{E}(\tilde{X}\tilde{Y}) \quad \Rightarrow \quad \mathsf{E}(\tilde{X}\tilde{Y}) \leq 1$$
$$0 \leq \mathsf{E}(\tilde{X} + \tilde{Y})^2 = 1 + 1 + 2\mathsf{E}(\tilde{X}\tilde{Y}) \quad \Rightarrow \quad -1 \leq \mathsf{E}(\tilde{X}\tilde{Y})$$

2.
$$\rho_{XY} = 1 \text{ iff } P(\tilde{Y} = \tilde{X}) = 1 \quad \Rightarrow \quad a > 0$$
$$\rho_{XY} = -1 \text{ iff } P(\tilde{Y} = -\tilde{X}) = 1 \quad \Rightarrow \quad a < 0$$

Bivariate Normal 19 / 44

Standard Bivariate Normal

Given a number $-1 \leq \rho \leq 1$, define the *standard bivariate normal density* of $(X, Y) \in \mathbb{R}^2$ by

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1 - \rho^2)}\right]$$

Properties:

1. The marginal distribution of X is $N(0, 1)$.
2. The marginal distribution of Y is $N(0, 1)$.
3. The correlation of X and Y is ρ.
4. The conditional distributions are normal:

$$Y|X \sim N(\rho X, 1 - \rho^2), \qquad X|Y \sim N(\rho Y, 1 - \rho^2)$$

The means are the *regression lines* of Y on X and X on Y respectively.

Bivariate Normal

Let $\tilde{X}$ and $\tilde{Y}$ have a standard bivariate normal distribution with correlation ρ. Let

$$\begin{aligned} X &= \mu_X + \sigma_X \tilde{X}, \qquad \mu_X \in \mathbb{R}, \sigma_X > 0 \\ Y &= \mu_Y + \sigma_Y \tilde{Y}, \qquad \mu_Y \in \mathbb{R}, \sigma_Y > 0 \end{aligned}$$

Then (X, Y) has the *bivariate normal density*

$$\begin{aligned} f_{XY}(x,y) &= \left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \\ &\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right.\right. \\ &- \left.\left. 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} \end{aligned}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

Properties

1. The marginal distribution of X is $N(\mu_X, \sigma_X^2)$.
2. The marginal distribution of Y is $N(\mu_Y, \sigma_Y^2)$.
3. The correlation between X and Y is ρ.
4. The conditional distributions are normal:

$$Y|X \sim N\left[\mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X),\ \sigma_Y^2(1-\rho^2)\right]$$

   The mean is the *regression line* of Y on X.
5. For any constants a and b, the distribution of $aX + bY$ is

$$N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$$

Proof: HW
(Suggestion: standardize the variables first)

Multivariate Distributions

The n-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ defined on the triplet $(S, \mathcal{B}, P)$ takes values on the hyperspace $\mathcal{R}^n$.

If the variables $X_1, \ldots, X_n$ are *discrete* then we have a discrete random vector, if the X's are *continuous*, then we have a continuous random vector.

If $\mathbf{X}$ is discrete then

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x})$$

where $f(\mathbf{x})$ denotes the *joint pmf*.

If $\mathbf{X}$ is continuous then

$$P(\mathbf{X} \in A) = \int \ldots \int_A f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

where $f(\mathbf{x})$ denotes the *joint pdf*.

Marginals and Conditionals

Definition The *marginal pdf* or *pmf* of any subset of the coordinates of $(X_1, \ldots, X_n)$ can be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates.

Definition The *conditional pdf* or *pmf* of a subset of the coordinates of $(X_1, \ldots, X_n)$ given the values of the remaining coordinates is obtained by dividing the full joint pdf or pmf by the joint pdf or pmf of the conditioning variates:

$$f(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_k)}$$

Multinomial Distribution

Let n and m be positive integers and let $p_1, \ldots, p_n$ be probabilities summing to one. Then the random vector $(X_1, \ldots, X_n)$ has a *multinomial distribution with m trials and cell probabilities* $p_1, \ldots, p_n$ if its joint pmf is

$$\begin{aligned} f(x_1, \ldots, x_n) &= \binom{m}{x_1 \cdots x_n} p_1^{x_1} \ldots p_n^{x_n} \\ &= \frac{m!}{x_1! \ldots x_n!} p_1^{x_1} \ldots p_n^{x_n} = m! \prod_{j=1}^{n} \frac{p_j^{x_j}}{x_j!} \end{aligned}$$

for $x_i = 0, \ldots, m$, $i = 1, \ldots, n$, and $x_1 + \cdots + x_n = m$.

The proof that this is a pmf is called the Multinomial Theorem.

Multinomial Theorem 4.6.4

Let m and n be positive integers. Let $\mathcal{A}$ be the set of all vectors $\mathbf{x} = (x_1, \ldots, x_n)$ which are such that the sum of their nonnegative integer components is m, i.e. $\sum_{j=1}^{n} x_j = m$ and $x_j \geq 0$.Then, for any real numbers $p_1, \ldots, p_n$

$$(p_1 + \ldots + p_n)^m = \sum_{\mathbf{x} \in \mathcal{A}} \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}$$

E.g. for $n = 2$ we have the binomial theorem.

E.g. for $n = 3$:

$$(p_1 + p_2 + p_3)^m = \sum_{x_1=0}^{m} \sum_{x_2=0}^{m-x_1} \frac{m!}{x_1!\, x_2!\, (m - x_1 - x_2)!} p_1^{x_1}\, p_2^{x_2}\, p_3^{m-x_1-x_2}$$

Multinomial Distribution: Properties

- Marginals are multinomials. E.g.

$$\begin{aligned} X_1 &\sim \text{Binomial}(m, p_1) \\ (X_1, X_2, m - X_1 - X_2) &\sim \text{Multinomial}(m, p_1, p_2, 1 - p_1 - p_2) \end{aligned}$$

- Conditionals are multinomials. E.g.

$$(X_1, \ldots, X_{n-1}) | [X_n = x_n] \sim \text{Multinomial}\left(m - x_n, \frac{p_1}{1 - p_n}, \ldots, \frac{p_{n-1}}{1 - p_n}\right)$$

- Variance and covariance:

$$\text{Cov}(X_j, X_k) = \text{E}[(X_j - mp_j)(X_k - mp_k)] = \begin{cases} mp_j(1 - p_j), & j = k \\ -mp_j p_k, & j \neq k \end{cases}$$

Multivariate Independence 29 / 44

Independent Random Vectors

Let $\mathbf{X_1}, \ldots, \mathbf{X_n}$ be random vectors with joint pdf or pmf $f(\mathbf{x_1}, \ldots, \mathbf{x_n})$. Let $f_{\mathbf{X_j}}(\mathbf{x_j})$ denote the marginal pdf or pmf of $\mathbf{X_j}$. Then $\mathbf{X_1}, \ldots, \mathbf{X_n}$ are called *mutually independent random vectors* if, for every $(\mathbf{x_1} \ldots, \mathbf{x_n})$,

$$f(\mathbf{x_1}, \ldots, \mathbf{x_n}) = \prod_{j=1}^{n} f_{\mathbf{X_j}}(\mathbf{x_j})$$

cont.

Theorem C-B 4.6.6 (Generalization of 4.2.10) Let $X_1, \ldots, X_n$ be independent rvs. Let $g_1, \ldots, g_n$ be real-valued functions such that $g(x_j)$ is only a function of x_j. Then

$$\mathsf{E}\left[g_1(X_1)\ldots g_n(X_n)\right] = \mathsf{E}g_1(X_1)\ldots \mathsf{E}g_n(X_n)$$

Theorem C-B 4.6.7 (Generalization of 4.2.12) Let $X_1, \ldots, X_n$ be mutually independent rvs with characteristic functions $\phi_{X_1}(\theta), \ldots, \phi_{X_2}(\theta)$. Let $Z = X_1 + \ldots + X_n$. Then the characteristic function of Z is

$$\phi_Z(\theta) = \prod_{j=1}^{n} \phi_{X_j}(\theta)$$

Note simplification if $\phi_{X_j}(\theta) = \phi_X(\theta)$.

Specific Characteristic Functions

	mgf	cf
Bernoulli(p)	$pe^t + q$	$pe^{it} + q$
Binomial(n,p)	$(pe^t + q)^n$	$(pe^{it} + q)^n$
Poisson(λ)	$e^{\lambda(e^t-1)}$	$e^{\lambda(e^{it}-1)}$
Geometric(p)	$pe^t/(1-qe^t)$	$pe^{it}/(1-qe^{it})$
Negbin(n,p)	$\left[\frac{pe^t}{1-qe^t}\right]^n$	$\left[\frac{pe^{it}}{1-qe^{it}}\right]^n$
Uniform(a,b)	$\frac{e^{tb}-e^{ta}}{t(b-a)}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
Normal(μ,σ^2)	$e^{t\mu+\frac{1}{2}\sigma^2t^2}$	$e^{it\mu-\frac{1}{2}\sigma^2t^2}$
Exponential(λ)	$\frac{\lambda}{\lambda-t} = (1-\frac{t}{\lambda})^{-1}$	$(1-\frac{it}{\lambda})^{-1}$
Gamma(a,λ)	$(1-\frac{t}{\lambda})^{-a}$	$(1-\frac{it}{\lambda})^{-a}$

Multivariate moments

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ be a random vector.
The *mean vector* $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{X}) = \begin{pmatrix} \mathsf{E}X_1 \\ \vdots \\ \mathsf{E}X_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

We can also define the second moment matrix

$$\mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top) = \begin{pmatrix} \mathsf{E}X_1^2 & \mathsf{E}X_1X_2 & \cdots & \mathsf{E}X_1X_n \\ \mathsf{E}X_2X_1 & \mathsf{E}X_2^2 & \cdots & \mathsf{E}X_2X_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{E}X_nX_1 & \mathsf{E}X_nX_2 & \cdots & \mathsf{E}X_n^2 \end{pmatrix}$$

Multivariate covariance

The *variance-covariance matrix* $\boldsymbol{\Sigma}$ is defined as

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathsf{Cov}\left[\boldsymbol{X} - \boldsymbol{\mu}, (\boldsymbol{X} - \boldsymbol{\mu})^\top\right] \\ &= \mathsf{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right] \\ &= \begin{pmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_2, X_1) & \mathsf{Var}(X_2) & \cdots & \mathsf{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \mathsf{Cov}(X_n, X_2) & \cdots & \mathsf{Var}(X_n) \end{pmatrix} \end{aligned}$$

Notice that $\boldsymbol{\Sigma}$ is a symmetric matrix.

Bivariate normal

Let X and Y are bivariate normal, then

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

The joint density of the vector $\boldsymbol{X} = (X, Y)^\top$ can be written as

$$f_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right]$$

Proof:

Linear functions

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ be a random vector with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. For a vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^\top \in \mathbb{R}^n$, define

$$Y = \boldsymbol{c}^\top \boldsymbol{X} = \sum_{i=1}^{n} c_i X_i$$

Then

$$\begin{aligned} \mathsf{E}(Y) &= \boldsymbol{c}^\top \boldsymbol{\mu} \\ \mathsf{Var}(Y) &= \boldsymbol{c}^\top \Sigma \boldsymbol{c} \end{aligned}$$

Proof:

Positive definiteness

Definition:

- An $n \times n$ symmetric matrix $\boldsymbol{\Lambda}$ is called *positive semi-definite* or *nonnegative definite* iff for every vector $\boldsymbol{c} \in \mathbb{R}^n$, $\boldsymbol{c}^\top \boldsymbol{\Lambda} \boldsymbol{c} \geq 0$.
- An $n \times n$ symmetric matrix $\boldsymbol{\Lambda}$ is called *positive definite* iff for every vector $\boldsymbol{c} \in \mathbb{R}^n$, $\boldsymbol{c}^\top \boldsymbol{\Lambda} \boldsymbol{c} > 0$.

Properties:

- A positive definite matrix:
  - is invertible
  - its determinant is positive
  - all its eigenvalues are positive
- The variance-covariance matrix is positive semi-definite.

Multivariate linear transformations

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\top$ have joint pdf $f_{\boldsymbol{X}}(\boldsymbol{X})$ and let

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ...a_{1n}X_n \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ...a_{2n}X_n \\ &\vdots \\ Y_n &= a_{n1}X_1 + a_{n2}X_2 + ...a_{nn}X_n \end{aligned}$$

Using vector and matrix notation, we can write this transformation as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$$

where

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

Multivariate linear transformations

The Jacobian of the transformation

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$$

is

$$\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}^\top} = \begin{pmatrix} \partial y_1/\partial x_1 & \dots & \partial y_1/\partial x_n \\ \vdots & \ddots & \vdots \\ \partial y_n/\partial x_1 & \dots & \partial y_n/\partial x_n \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} = \boldsymbol{A}$$

with determinant $\det(\boldsymbol{A})$.

The inverse transformation is

$$\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{Y}$$

with Jacobian determinant $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$.

Multivariate normal

Let $\boldsymbol{\mu}$ be a vector in $\mathbb{R}^n$ and $\boldsymbol{\Sigma}$ be an $n \times n$ symmetric positive definite matrix. The vector $\boldsymbol{X} = (X_1, \dots, X_n)$ has a *multivariate normal distribution* with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ if it has joint density

$$f_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right]$$

In particular:

$$\begin{aligned} \boldsymbol{\mu} &= \mathsf{E}\boldsymbol{X} \\ \boldsymbol{\Sigma} &= \mathsf{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top] \end{aligned}$$

Construction

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ be i.i.d. $N(0,1)$ variables. Then

$$f_{\boldsymbol{Z}}(z_1, \ldots, z_n) = \prod_{i=1}^{n} \frac{e^{-z_i^2/2}}{\sqrt{2\pi}} = \frac{e^{-\sum_{i=1}^{n} z_i^2/2}}{(2\pi)^{n/2}} = \frac{e^{-\boldsymbol{Z}^\top \boldsymbol{Z}/2}}{(2\pi)^{n/2}}$$

Now define

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is a vector and $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an invertible matrix.
The inverse transformation is

$$\boldsymbol{Z} = \boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

with Jacobian

$$J = \frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}^\top} = \boldsymbol{A}^{-1}$$

and determinant $|J| = 1/|\boldsymbol{A}|$.

Construction (cont.)

The joint pdf of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{e^{-(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})/2}}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$. We can confirm that

$$\mathsf{E}(\boldsymbol{X}) = \mathsf{E}(\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{\mu}$$

and

$$\begin{aligned}\mathsf{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top] &= \mathsf{E}[(\boldsymbol{A}\boldsymbol{Z})(\boldsymbol{A}\boldsymbol{Z})^\top] = \mathsf{E}(\boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A}^\top) \\ &= \boldsymbol{A}\mathsf{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{I}\boldsymbol{A}^\top = \Sigma\end{aligned}$$

Properties

Theorem:

1. Let X and Y be jointly normal random variables. Then X and Y are independent if and only if they are uncorrelated.
2. Let $X_1, \ldots, X_n$ be jointly normal random variables. Then they are mutually independent if and only if they are pairwise uncorrelated.

Theorem: If $\boldsymbol{X}$ has a multivariate normal distribution, then

1. All marginal distributions are normal.
2. All conditional distributions are normal.
3. For any constants $\boldsymbol{A} = (a_1, \ldots, a_n)^\top$ the random variable $Y = \boldsymbol{A}^\top \boldsymbol{X}$ has a normal distribution.

BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 16

Recall: Chebychev Inequality

Let X be a random variable and let $g(x)$ be a non-negative function. Then for any $r > 0$,

$$P[g(X) \geq r] \leq \frac{\mathsf{E}g(X)}{r}$$

Proof:

$$\begin{aligned} \mathsf{E}g(X) &= \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx \\ &\geq \int_{\{x:g(x)\geq r\}} g(x)\, f_X(x)\, dx \\ &\geq r \int_{\{x:g(x)\geq r\}} f_X(x)\, dx \\ &= r\, P\{g(X) \geq r\} \end{aligned}$$

Corollaries

1. Suppose X is a non-negative and g is a positive, non-decreasing function, with $\mathsf{E}[g(X)] < \infty$. Then

$$P\{X \geq a\} \leq \frac{\mathsf{E}(g(X))}{g(a)}$$

2. Suppose g is a non-negative symmetric function, increasing on $\mathbb{R}^+$, with $\mathsf{E}[g(X)] < \infty$. Then

$$P\{|X| \geq a\} \leq \frac{\mathsf{E}[g(X)]}{g(a)}$$

Proof: $P\{X \geq a\} = P\{g(X) \geq g(a)\}$, so the results follow from the inequality on the previous slides.

Special cases

Provided that the appropriate expectations exist, for $a > 0$:

$$X \geq 0: \qquad P\{X \geq a\} \leq \frac{\mathsf{E}(e^{tX})}{e^{ta}} \tag{1}$$

$$X \in \mathbb{R}: \qquad P\{|X| \geq a\} \leq \frac{\mathsf{E}(|X|)}{a} \tag{2}$$

$$X \in \mathbb{R},\ p > 0: \qquad P\{|X| \geq a\} \leq \frac{\mathsf{E}(|X|^p)}{a^p} \tag{3}$$

$$\sigma^2 = \mathsf{Var}(X): \qquad P\{|X - \mathsf{E}X| \geq a\sigma\} \leq \frac{1}{a^2} \tag{4}$$

Note:

- (1) is called Chernoff bound, useful when the mgf is easier to compute than the cdf.
- (3) is sometimes called Markov's inequality.
- (4) is sometimes called Chebychev's inequality.

Functional inequalities

First a couple of useful items:

- L^p *spaces*:
  The space called L^p consists of all random variables whose p^{th} absolute power is integrable, i.e., $E(|X|^p) < \infty$.
- *Triangle inequality*:
  For two real or complex numbers a and b,

  $$|a + b| \leq |a| + |b|$$

  Proof: HW

Convex functions

Definition: Let I be an interval on $\mathbb{R}$. A function $g : I \to R$ is *convex* on I if for any $\lambda \in [0, 1]$, and any points x and y in I

$$g\left[\lambda x + (1-\lambda)y\right] \leq \lambda g(x) + (1-\lambda)g(y)$$

Properties:

- A differentiable function g is convex iff it lays above all its tangents.
- A twice differentiable function g is convex iff its second derivative is non-negative.

Definition: Let I be an interval on $\mathbb{R}$. A function $g : I \to R$ is *concave* on I if $-g$ is convex on I.

Examples:

- $g(x) = x^2$ is a convex function for all x.
- $g(x) = \log(x)$ is concave for $x > 0$.

Jensen's Inequality

Let $X \in L^1$ and $g(x)$ be a convex function where $\mathsf{E}[g(X)]$ exists. Then,

$$\mathsf{E}[g(X)] \geq g[\mathsf{E}X]$$

with equality if and only if for every line $a + bx$ tangent to $g(x)$ at $x = \mathsf{E}X$, $P[g(X) = a + bX] = 1$.

Examples:

$$\begin{aligned} g(x) = x^2 &\rightarrow \mathsf{E}(X^2) \geq \mathsf{E}^2(X) \\ g(x) = 1/x,\ x > 0 &\rightarrow \mathsf{E}(1/X) \geq 1/\mathsf{E}(X),\ X > 0 \end{aligned}$$

Note: The direction of the inequality is reversed if g is concave.

Jensen's Inequality (proof)

Let $l(x) = a + bx$ be the tangent line to $g(x)$ at $g(\mathsf{E}X)$. Then

$$\begin{aligned} \mathsf{E}g(X) &\geq \mathsf{E}(a + bX) \\ &= a + b\mathsf{E}X \\ &= l(\mathsf{E}X) \\ &= g(\mathsf{E}X) \end{aligned}$$

Example

Let $a_1, \ldots, a_n > 0$. Then

$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}\right)^{-1} \leq \left(\prod_{i=1}^{n} a_i\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} a_i$$

Proof: Let X be a rv such that $P(X = a_i) = 1/n$. Since $\log(x)$ is concave,

$$\begin{aligned} \log\left(\prod_{i=1}^{n} a_i\right)^{1/n} &= \frac{1}{n}\sum_{i=1}^{n}\log(a_i) \\ &= \mathsf{E}(\log(X)) \\ &\leq \log(\mathsf{E}(X)) \\ &= \log\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) \end{aligned}$$

The proof of the second inequality is similar.

Young's Inequality

Let $a, b > 0$ and $p, q > 1$ with $1/p + 1/q = 1$. Then

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab$$

With equality only if $a^p = b^q$.

Proof: Consider

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

To minimize $g(a)$, differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \rightarrow a^{p-1} - b = 0 \rightarrow a = b^{1/(p-1)}.$$

Since $g(b^{1/(p-1)}) = 0$, the result follows.

Hölder's inequality

Suppose $X \in L^p$, $Y \in L^q$ where $p, q > 1$ and $1/p + 1/q = 1$. Then

$$\mathsf{E}|XY| \leq [\mathsf{E}|X|^p]^{1/p} \, [\mathsf{E}|Y|^q]^{1/q}$$

with equality if $X^p = cY^q$ for some $c \in \mathbb{R}$.

Proof: Let

$$a = \frac{|X|}{(\mathsf{E}|X|^p)^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{(\mathsf{E}|Y|^q)^{1/q}}$$

By Young's Inequality,

$$\frac{|X|^p}{p\mathsf{E}|X|^p} + \frac{|Y|^q}{q\mathsf{E}|Y|^q} \geq \frac{|XY|}{(\mathsf{E}|X|^p)^{1/p}(\mathsf{E}|Y|^q)^{1/q}}$$

The result follows by taking the expected value of both sides and noting that the expected value of the left-hand side is 1.

Corollaries

- **Cauchy-Schwartz inequality:** Special case where $p = q = 2$.

$$\mathsf{E}|XY| \leq [\mathsf{E}|X|^2]^{1/2}[\mathsf{E}|Y|^2]^{1/2} = \sqrt{\mathsf{E}[X^2]\mathsf{E}[Y^2]}$$

  with equality if $X = cY$ for some $c \in \mathbb{R}$.

- **Lyapunov's inequality:** For $1 \leq r \leq s$ and $X \in L^s$,

$$[\mathsf{E}|X|^r]^{1/r} \leq [\mathsf{E}|X|^s]^{1/s}$$

  Proof:
  Apply Hölder's inequality to $|X|^r$ with $Y = 1$ and $p = s/r$.

Application of Cauchy-Schwartz:

Let ρ represent the correlation between two rvs X and Y, ie,

$$\rho = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\ \mathsf{Var}(Y)}}.$$

Then, $|\rho| \leq 1$, with equality iff $Y - \mu_Y = c(X - \mu_X)$ for some $c \in \mathbb{R}$.

Proof: By the Cauchy-Schwartz Inequality,

$$\mathsf{E}|(X - \mu_x)(Y - \mu_Y)| \leq \{\mathsf{E}(X - \mu_X)^2\}^{\frac{1}{2}}\{\mathsf{E}(Y - \mu_Y)^2\}^{\frac{1}{2}}.$$

Squaring both sides, we obtain

$$(\mathsf{Cov}(X, Y))^2 \leq \sigma_X^2\sigma_Y^2.$$

Thus, $|\rho| \leq 1$.

Minkowski's inequality

Suppose $X, Y \in L^p$, $p \geq 1$. Then $(X+Y) \in L^p$ and

$$[\mathsf{E}|X+Y|^p]^{1/p} \leq [\mathsf{E}|X|^p]^{1/p} + [\mathsf{E}|Y|^p]^{1/p}$$

Proof:
For $p = 1$, the proof follows almost immediately from the triangle inequality (HW).
The case $p > 1$ is more complicated. Consider

$$\begin{aligned}
\mathsf{E}|X+Y|^p &= \mathsf{E}\left(|X+Y|\,|X+Y|^{p-1}\right) \\
&\leq \mathsf{E}\left(|X|\,|X+Y|^{p-1}\right) + \mathsf{E}\left(|Y|\,|X+Y|^{p-1}\right) \\
&\leq [\mathsf{E}|X|^p]^{1/p}\left[\mathsf{E}|X+Y|^{(p-1)q}\right]^{1/q} \\
&+ [\mathsf{E}|Y|^p]^{1/p}\left[\mathsf{E}|X+Y|^{(p-1)q}\right]^{1/q}
\end{aligned}$$

where the last row follows by Hölder's inequality with $1/p + 1/q = 1$.

Order Statistics (*C-B*, Section 5.4; *Gut*, Chapter IV)

Distribution of the Maximum

The *cdf* of $Z = \max(Y_1, \ldots, Y_n)$ is

$$\begin{aligned}
F_Z(z) &= Pr\{Z \leq z\} \\
&= Pr\{Y_1 \leq z, Y_2 \leq z, \ldots, Y_n \leq z\} \\
&= \prod_{j=1}^{n} Pr\{Y_j \leq z\} \quad \text{(indep)} \\
&= F_Y(z)^n \quad \text{(ident. distrib.)}
\end{aligned}$$

and thus the density (or pmf) is:

$$f_Z(z) = nF_Y(z)^{n-1} f_Y(z)$$

Distribution of the Minimum

Similarly, consider $W = \min(Y_1, Y_2, \ldots, Y_n)$.

$$\begin{aligned} 1 - F_W(w) &= Pr\{W > w\} \\ &= Pr\{Y_1 > w, Y_2 > w, \ldots, Y_n > w\} \\ &= \prod_{j=1}^{n} Pr\{Y_j > w\} \qquad \text{(indep)} \\ &= (1 - F_Y(w))^n \qquad \text{(ident. distrib)} \end{aligned}$$

Thus

$$F_W(w) = 1 - (1 - F_Y(w))^n$$

and the corresponding density (or pmf) is:

$$f_W(w) = n(1 - F_Y(w))^{n-1} f_Y(w)$$

Example

Suppose $Y_i \sim$ *exp*(λ):

$$f_Y(y) = \lambda e^{-\lambda y} \qquad \text{for } y > 0, \quad 1 - F(y) = e^{-\lambda y}$$

Maximum:

$$f_Z(z) = n(1 - e^{-\lambda z})^{n-1} \lambda e^{-\lambda z} = n\lambda e^{-\lambda z}(1 - e^{-\lambda z})^{n-1}$$

Minimum:

$$f_W(w) = n(e^{-\lambda w})^{n-1} \lambda e^{-\lambda w} = \quad (n\lambda) e^{-n\lambda w}$$

$\Rightarrow$ exponential with parameter $n\lambda$

The next obvious statistic is the range defined as the difference of the maximum and the minimum, but to get its distribution we need the joint distribution of the maximum and the minimum.

Order Statistics

Let $Y_1, Y_2, \ldots, Y_n$ be *iid* with *pdf* $f_Y(x)$.
Order the observations; i.e.

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

The $Y_{(i)}$ are called *order statistics*. For example, the *minimum* is $Y_{(1)}$ and the *maximum* is $Y_{(n)}$.

We are interested in finding the distribution of an arbitrary $Y_{(i)}$, as well as the joint distributions of sets of $Y_{(i)}$s and $Y_{(j)}$s.

e.g. Range $= Y_{(n)} - Y_{(1)}$
or interquartile range, or joint of median and interquartile range, etc....

rth order statistic

We need to find the density of $Y_{(r)}$ at a value y:

$$\overbrace{\qquad}^{r-1}\ \overbrace{\qquad}^{1}\ \overbrace{\qquad}^{n-r}$$
$$y \qquad\qquad y+dy$$

Consider 3 intervals $(-\infty, y), [y, y+dy), [y+dy, \infty)$. The number of observations in each of the intervals follows the tri-nomial distribution

$$f(s_1, s_2, s_3) = \frac{n!}{s_1! s_2! s_3!}\ p_1^{s_1}\ p_2^{s_2}\ p_3^{s_3}$$

The event that $y \leq Y_{(r)} < y + dy$ is the event that we have
$(r-1)$ observations are less than y
$(n-r)$ observations are greater than y
1 observation is in interval; $y, y+dy$
In the trinomial distribution, this corresponds to

$$s_1 = r-1, \quad s_2 = 1, \quad s_3 = n-r$$
$$p_1 = F_Y(y), \quad p_2 = f_Y(y)dy, \quad p_3 = 1 - F_Y(y+dy)$$

cont.

Taking the limit as $dy \to 0$, we get:

$$f_{Y_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!} F_Y(y)^{r-1}[1 - F_Y(y)]^{n-r} f_Y(y)$$
$$= \frac{F_Y(y)^{r-1}[1 - F_Y(y)]^{n-r} f_Y(y)}{B(r, n-r+1)}$$

Gut has a more formal derivation based on deriving the joint density of the order statistics, then integrating out all but the r^{th} order statistic. (see also Casella and Berger, p.228).

Example

$F_Y(y) = y$, that is, $Y \sim$ *Uniform* $(0, 1)$

$$f_{Y_{(r)}}(y) = \frac{y^{r-1}(1-y)^{n-r}}{B(r, n-r+1)}$$

hence, $Y_{(r)}$ follows a *Beta Distribution* with parameters r and $n - r + 1$.

Note: $E[Y_{(r)}] = \frac{r}{n+1}$

Distribution of the median

To simplify, suppose the sample size is odd, $n = 2m + 1$, so that the median corresponds to the $(m + 1)^{\text{th}}$ order statistic.
Setting $r = m + 1$ and $n = 2m + 1$ into the formula derived earlier

$$f_{\text{med}}(y) = f_{Y_{(m+1)}}(y) = \frac{F_Y(y)^m (1 - F_Y(y))^m f_Y(y)}{B(m + 1, m + 1)}$$

If the density $f_Y(y)$ is symmetric around zero, so that $\text{E}Y = 0$, then

$$F_Y(-y) = 1 - F_Y(y)$$

and so the density of the median is also symmetric around zero, so that

$$\text{E}[\text{med}(Y_1, \ldots, Y_n)] = 0$$

Joint distribution of $Y_{(r)}, Y_{(s)}$, $r < s$

	Interval	Prob.	# obs $= s_i$
1.	$(-\infty, u]$	$p_1 = F_Y(u)$	$r - 1$
2.	$(u, u + du]$	$p_2 = f_Y(u)du$	1
3.	$(u + du, v]$	$p_3 = F_Y(v) - F_Y(u + du)$	$s - r - 1$
4.	$(v, v + dv]$	$p_4 = f_Y(v)dv$	1
5.	$(v + dv, \infty)$	$p_5 = 1 - F_Y(v + dv)$	$n - s$

This is a multinomial with 5 cells:

$$f(s_1, \ldots, s_5) = \frac{n!}{\prod s_i!} \prod_{i=1}^{5} p_i^{s_i}$$

cont.

Taking limits as du and dv approach 0, we get

$$f_{Y_{(r)},Y_{(s)}}(u,v) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F_Y(u)^{r-1}$$
$$\times [F_Y(v) - F_Y(u)]^{s-r-1} (1 - F_Y(v))^{n-s} f_Y(u) f_Y(v)$$

Example: Suppose $F_Y(x) = x$ (Uniform)

$$f_{Y_{(r)},Y_{(s)}}(u,v) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$
$$\times u^{r-1}(v-u)^{s-r-1}(1-v)^{n-s}$$

for $u < v$.

Joint distribution of all order statistics

Multinomial with $2n+1$ cells, where we have one observation in each interval $[u_i, u_i + du_i)$, $i = 1, \ldots, n$, and zero on the others.

$$f_{Y_{(1)},Y_{(2)},\ldots,Y_{(n)}}(u_1, \ldots, u_n) = n! \prod_{i=1}^{n} f_Y(u_i)$$

for $u_1 < \cdots < u_n$.

Example: Suppose $F_Y(x) = x$ (Uniform)

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(u_1, \ldots, u_n) = n! \qquad u_1 < \cdots < u_n$$

Distribution of the range

Setting $r = 1$ and $s = n$ in the joint dist. of the rth and sth order statistics gives the joint dist. of the *min* and *max*:

$$f_{Y_{(1)},Y_{(n)}}(u, v) = \frac{n!}{(n-2)!}[F_Y(v) - F_Y(u)]^{n-2} f_Y(u) f_Y(v)$$

Now, do a transformation to $R = Y_{(n)} - Y_{(1)}$ and $W = Y_{(1)}$. Note that the Jacobian is 1. What is the range?

Hence,

$$f_{W,R}(w, r) = n(n-1)[F_Y(w + r) - F_Y(w)]^{n-2} f_Y(w) f_Y(w + r)$$

The density of R can be obtained by integrating out W:

$$f_R(r) = \int_{-\infty}^{\infty} f_{W,r}(w, r)\, dw$$

Example:

Suppose $Y \sim U[0, 1]$, i.e. $F_Y(x) = x$

$$\begin{aligned} f_R(r) &= \int_0^{1-r} n(n-1) r^{n-2}\, dw \\ &= n(n-1) r^{n-2}(1-r) \end{aligned}$$

Note that R has a Beta distribution.

$$\begin{aligned} \mathsf{E}(R) &= n(n-1) \int_0^1 r \cdot r^{n-2}(1-r) dr \\ &= n(n-1) \left[\frac{1}{n} - \frac{1}{n+1}\right] \\ &= \frac{n-1}{n+1} \end{aligned}$$

What happens when $n = 2$ and $n \to \infty$?

Agenda

In the last part of the course, we discuss

- Convergence of random variables. Several different kinds
  - Convergence in probability
  - Almost sure convergence
  - Convergence in distribution
  - Convergence in L^p
  - Complete convergence
- Weak law of large numbers
- Strong law of large numbers
- Central limit theorems

The moment inequalities will be useful in proving these results.

Material is in *C-B*, Section 5.5, and *Gut*, Chapter VI.

Modes of Convergence

There are five modes of convergence. If $X_n \to X$ by any of these modes, then X is unique (see Section 2, Chapter VI of *Gut*).

1. *Convergence in Probability* $\quad X_n \xrightarrow{P} X$

   For any $\epsilon > 0, \quad \lim_{n\to\infty} P\{|X_n - X| < \epsilon\} = 1$

   Or equivalently,
   for any $\epsilon > 0, \quad \lim_{n\to\infty} P\{|X_n - X| > \epsilon\} = 0$

2. *Convergence "almost surely" (a.s.)*, denoted $X_n \xrightarrow{a.s.} X$.
   Also called *Convergence with Prob. 1*

   For any $\epsilon > 0, \quad P\{\lim_{n\to\infty} |X_n - X| < \epsilon\} = 1$
   Or

   $$P\Big\{\lim_{n\to\infty} X_n = X\Big\} = 1$$

   Or

   $$P\Big\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\} = 1$$

Notes

The distinction between *Convergence almost surely* and *Convergence in probability* is subtle.

We'll see how the Markov inequality and Chebychev's inequality can often be used to establish convergence in probability. Establishing convergence a.s. is often more difficult.

Almost sure convergence is stronger than convergence in probability. Or equivalently,

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X.$$

Convergence in Distribution

3. *Convergence in Distribution* $X_n \xrightarrow{d} X$

   Also called *convergence in law* or *weak convergence*.

   $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all continuity points of $F_X(x)$

Notes:

- Recall that cdfs can have at most a countable number of discontinuities.
- Theorem (no proof):

$$X_n \xrightarrow{d} X \quad \Leftrightarrow \quad \forall \text{ bounded continuous functions } g, \\ \mathrm{E}g(X_n) \to \mathrm{E}g(X) \text{ as } n \to \infty.$$

Convergence in Distrib. cont.

- Convergence in distribution is the weakest convergence and does not imply the other modes. E.g. $X_n \sim N(0,1)$ and $Y = -X$.
- An exception is the following important special case:

Suppose $X_n \xrightarrow{d} X$ where X has the degenerate distribution (i.e. $P\{X = a\} = 1$). Then, $X_n \xrightarrow{d} X \;\Rightarrow\; X_n \xrightarrow{P} X$.

Proof:

$$\begin{aligned} P\{|X_n - a| < \epsilon\} &= F_{X_n}(a+\epsilon) - F_{X_n}(a-\epsilon) \\ \lim_{n\to\infty} P\{|X_n - a| < \epsilon\} &= F(a+\epsilon) - F(a-\epsilon) \\ &= 1 - 0 = 1 \end{aligned}$$

Other modes of convergence

4. *Convergence in r^{th} mean* $(r \geq 1)$ $X_n \xrightarrow{r} X$
   If $E|X_n^r| < \infty$ for all n and
   $$\lim_{n\to\infty} (E|X_n - X|^r) = 0.$$
   Also called *convergence in* L^r and sometimes referred to as *convergence in the* L^r norm. (Some books use L^p)

5. *Complete convergence* (see Chapter VI, section 4 in *Gut*), defined as,
   $$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty \qquad \forall\, \epsilon > 0,$$
   is slightly stronger than a.s. convergence, but much easier to verify.

   Hence, it sometimes provides a relatively easy way to establish a.s. convergence. Some books use this as the definition of a.s. convergence.

Example 1

1. Let X_n be *gamma*(n, n) Then $X_n \xrightarrow{p} 1$.
   Proof:
   Since $E(X_n) = 1$ and Var$(X_n) = 1/n$, we can apply Chebychev's inequality to obtain

$$P(|X_n - 1| > \epsilon) \leq \frac{1}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty.$$

   Therefore $X_n \xrightarrow{p} 1$.

Example 2

2. Suppose $X_n \sim$ *binom*$(n, \lambda/n)$. Then $X_n \xrightarrow{d} X$, where X has a *Poisson*(λ) distribution.
   Proof:

$$F_{X_n}(x) = \sum_{y=0}^{x} \left(\begin{array}{c} n \\ y \end{array} \right) \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \to \sum_{y=0}^{x} e^{-\lambda} \frac{\lambda^y}{y!},$$

   as $n \to \infty$. (We saw this before). The RHS is the distribution function of the Poisson with parameter λ.

Example 3

3. Let $X_2, X_3...$ be a sequence of binary random variables defined by

$$P(X_n = 1) = 1 - \frac{1}{n} \quad \text{and} \quad P(X_n = n) = \frac{1}{n}$$

If we choose an ϵ smaller than 1, then

$$P(|X_n - 1| > \epsilon) = P(X_n = n) = 1/n \to 0,$$

hence $X_n \xrightarrow{P} 1$.
It turns out that X_n does not converge to 1 almost surely (see page 156 in *Gut*).

Example 4

4. Let $X_2, X_3...$ be a binary random variables defined by

$$P(X_n = 1) = 1 - \frac{1}{n^2} \quad \text{and} \quad P(X_n = n) = \frac{1}{n^2}$$

Now, for ϵ small enough,

$$\sum_{n=1}^{\infty} P(|X_n - 1| > \epsilon) = \sum_{n=1}^{\infty} \frac{1}{n^2}$$

which converges (the series $\sum_{n=1}^{\infty} \frac{1}{n^k}$ converges for $k > 1$). I.e., $X_n \to X$ in complete convergence, with $X = 1$. Hence $X_n \xrightarrow{a.s.} 1$.

Relationships among convergence modes

$$\begin{array}{lll} X_n \xrightarrow{Compl} X & \Rightarrow & X_n \xrightarrow{a.s.} X \quad \searrow \\ & & \qquad\qquad\qquad\quad X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{d} X \\ & & X_n \xrightarrow{r} X \quad \nearrow \end{array}$$

A silly mneumonic is *All Probabilists Drink*.

Also: If $r \geq s \geq 1$

$$X_n \xrightarrow{r} X \Longrightarrow X_n \xrightarrow{s} X.$$

(Try to prove this one).

In probability and distribution

Theorem: If $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$

Proof:
For any $\epsilon > 0$:

$$\begin{aligned} F_{X_n}(x) = P\{X_n \leq x\} &= P\{X_n \leq x \cap |X - X_n| \leq \epsilon\} \\ &\quad + P\{X_n \leq x \cap |X_n - X| > \epsilon\} \\ &\leq P\{X \leq x + \epsilon\} + P\{|X_n - X| > \epsilon\} \end{aligned}$$

because

$$\begin{aligned} \{|X - X_n| \leq \epsilon\} &= \{-\epsilon \leq X - X_n \leq \epsilon\} \\ &\subset \{X - X_n \leq \epsilon\} = \{X \leq X_n + \epsilon\} \end{aligned}$$

and $\{X_n \leq x\} \cap \{X \leq X_n + \epsilon\} \subset \{X \leq x + \epsilon\}$.

cont.

Hence,

$$F_{X_n}(x) \leq F_X(x+\epsilon) + P\{|X - X_n| > \epsilon\} \tag{5}$$

and as $n \to \infty$, this implies

$$\limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\epsilon)$$

Now, interchange the roles of X_n and X in (5) and repeat to get

$$F_X(x) \leq F_{X_n}(x+\epsilon) + P\{|X_n - X| > \epsilon\}$$

but apply inequality to $x = x - \epsilon$ instead of x, yielding

$$F_X(x-\epsilon) \leq F_{X_n}(x) + P\{|X_n - X| > \epsilon\}$$

or

$$F_X(x-\epsilon) - P\{|X_n - X| > \epsilon\} \leq F_{X_n}(x)$$

cont.

As $n \to \infty$, this implies

$$F_X(x-\epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x)$$

Putting these together, we have

$$F_X(x-\epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\epsilon)$$

for all $\epsilon > 0$. Therefore, for all x where $F_X(x)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Note: If $F_X(x)$ is not continuous at x, then all we can claim is

$$F_X(x-) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x)$$

rth moment and in Probability

Theorem: $X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{P} X$

Proof: By Markov's inequality,

$$\begin{aligned} P(|X_n - X| > \epsilon) &= P(|X_n - X|^r > \epsilon^r) \\ &\leq \frac{E(|X_n - X|^r)}{\epsilon^r} \to 0 \end{aligned}$$

Example: Let $Y_1, \ldots, Y_n$ be iid with common mean μ and variance σ^2. Let $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. Then

$$E(\bar{Y}_n - \mu)^2 = \text{var}(\bar{Y}) = \frac{\sigma^2}{n} \to 0$$

Therefore $\bar{Y}_n \xrightarrow{r=2} \mu$ and so $\bar{Y}_n \xrightarrow{P} \mu$.

This result is the *weak law of large numbers*.

Convergence properties 47 / 59

Convergence in probability

Theorem: If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then

1. $X_n + Y_n \xrightarrow{P} X + Y$
2. $X_n Y_n \xrightarrow{P} XY$
3. If $g(x)$ is a continuous function, then $g(X_n) \xrightarrow{P} g(X)$

Proof:

Slutsky's Theorem

Also known as *Cramer's Theorem* - **VERY VERY USEFUL**

Theorem: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a$, where a is a constant, then

1. $X_n + Y_n \xrightarrow{d} X + a$
2. $Y_n X_n \xrightarrow{d} aX$

Proof: Homework

Example

Let $X_1, \ldots, X_n$ be iid with mean μ, variance σ^2, and finite moments up to fourth order. The Central Limit Theorem (CLT) says that

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

But the empirical variance is a consistent estimator of σ^2, i.e. $S_n^2 \xrightarrow{P} \sigma^2$. Therefore

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$$

by Slutsky's Thereom.

This is useful for constructing confidence intervals for μ.

Example

Show that a t-distribution with n degrees of freedom converges in distribution to the standard normal as $n \to \infty$.

Proof: Let $Y_n \sim$ ChiSquare(n). Then $Y_n/n \to 1$ by the Weak Law of Large Number (WLLN).

By Slutsky's Theorem, if $X \sim$ Normal$(0, 1)$, then

$$\frac{X}{\sqrt{\frac{Y_n}{n}}} \xrightarrow{d} \text{Normal}(0, 1)$$

Convergence in distribution

Theorem: Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. Suppose further that X_n and Y_n are independent for all n, and that X and Y are independent. Then

$$X_n + Y_n \xrightarrow{d} X + Y$$

Proof: Omitted (use characteristic functions)

Example

Let $X_n \sim Bin(n_x, p_x)$ with $n_x p_x \to \lambda_x$ as $n_x \to \infty$.
Let $Y_n \sim Bin(n_y, p_y)$ with $n_y p_y \to \lambda_y$ as $n_y \to \infty$, indep. of X_n.
Then

$$X_n \xrightarrow{d} Po(\lambda_x), \qquad Y_n \xrightarrow{d} Po(\lambda_y)$$

and

$$X_n + Y_n \xrightarrow{d} Po(\lambda_x + \lambda_y)$$

The Delta Method 54 / 59

Approximate mean and variance

Suppose we know the distribution of X and want to get the distribution of $Y = g(X)$. The general method is to use the Jacobian transformation. But if the distribution of X is "well concentrated" around its mean $\mu = \mathsf{E}X$, we can approximate the mean and variance of Y as follows.

The Taylor expansion of $g(X)$ around μ is

$$g(X) = g(\mu) + g'(\mu)(X - \mu) + g''(\mu)(X - \mu)^2 + \ldots$$

Therefore

$$\begin{aligned} \mathsf{E}[g(X)] &= g(\mu) + \mathsf{E}g''(\mu)(X - \mu)^2 + \ldots \\ &\approx g(\mu) \end{aligned}$$

Similarly

$$\begin{aligned} \mathsf{Var}[g(X)] \approx \mathsf{E}[g(X) - g(\mu)]^2 &= \mathsf{E}[g'(\mu)(X - \mu)]^2 \\ &= \mathsf{E}[g'(\mu)]^2 \mathsf{Var}X \end{aligned}$$

Example

Let $X \sim N(\mu, \sigma^2)$ and $Y = \exp(X)$. The exact mean and variance of Y are

$$\mathsf{E}Y = e^{\mu+\sigma^2/2}, \qquad \mathsf{Var}Y = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

The first order Taylor expansion gives

$$\mathsf{E}(Y) = e^{\mu}, \qquad \mathsf{Var}(Y) = e^{2\mu}\sigma^2$$

The Delta method

Let Y_n be a sequence of rvs such that $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g'(\theta)$ exists and is nonzero. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d} N\big(0, \sigma^2[g'(\theta)]^2\big)$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + R_2$$

where $R_2 \to 0$ as $Y_n \to \theta$. Apply Slutsky's Theorem to

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta)$$

Example

Let X_i iid with mean μ and variance σ^2. The CLT gives that

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Now let $g(x) = e^x$, where $g'(x) = e^x > 0$ for all x. The Delta method gives that

$$\sqrt{n}\big(e^{\bar{X}_n} - e^{\mu}\big) \xrightarrow{d} N(0, \sigma^2 e^{2\mu})$$

Let $Y_i = \exp(X_i)$, then $e^{\bar{X}_n}$ is the geometric average of the Y_i. So we have an approximation for the distribution of the geometric average.

Second-order Delta method

Let Y_n be a sequence of rvs such that $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g'(\theta) = 0$ and $g''(\theta)$ exists and is nonzero. Then

$$n[g(Y_n) - g(\theta)] \xrightarrow{d} \frac{\sigma^2 g''(\theta)}{2} \chi^2_1$$

Proof: See C&B.