Statistical Inference (2nd Edition)

Chapter 1: Probability Theory

- Statistics builds upon the foundation of prob theory
- * outline some of the basic ideas of prob. theory fundamental to statistics
- Probability theory builds upon set theory:

1.1: Set Theory

Def (Sample space): the set, S, of all possible outcomes of a particular experiment
- can be (un)countable; dictates the way in which prob's can be assigned

countable: elmts of S can be put into a 1-1 correspondence w/ subset of integers

Recall: a set can be finite or infinite
→ an infinite set is either countable or uncountable
→ a countable set is either finite or countably infinite

Def (event): any collection of possible outcomes of an experiment, i.e. any subset of S (including S itself)

containment: $A \subset B \iff x \in A \Rightarrow x \in B$

equality: $A = B \iff A \subset B$ and $B \subset A$

- Given any two events (or sets) A and B, we have the following

elementary set operations (def):
- **union**: $A \cup B = \{x : x \in A \text{ or } x \in B\}$
- **intersection**: $A \cap B = \{x : x \in A \text{ and } x \in B\}$
- **complementation**: $A^c = \{x : x \notin A\}$

Remark: $\emptyset$ denotes the empty set

- (1) can be contained; being careful, we can treat sets as numbers

Theorem (Properties of set operations).
For any three events, A, B, and C, defined on a sample space S,

a. **commutativity**: $A \cup B = B \cup A$, $A \cap B = B \cap A$

b. **Associativity**: $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$

c. **Distributive laws**: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

d. **Demorgan's Laws**: $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$

Recall:

Def (Real number): any number that can be placed on the number line

Numbers: Real = Rational ∪ Irrational

Def (Rational Number): Any number that can be written as a fraction

$\frac{p}{q}$ where $\begin{cases} \cdot p \text{ and } q \text{ are integers (1)} \\ \cdot q \neq 0 \qquad (2) \end{cases}$,

e.g. integers, fractions, terminating & repeating decimals

Def (Irrational Number): real number that violates (1) and/or (2) s.t. its decimal expansion goes on forever,
e.g. π, e, square roots of non-perfect squares (√6)

Real
- Irrational ($\mathbb{Q}^c$)
- Rational ($\mathbb{Q}$)
  - Integers ($\mathbb{Z}$) $-2, -1, 0, 1, 2$
    - Natural ($\mathbb{N}$) $0, 1, 2, \ldots$

- $\mathbb{R} = \bigcup_{n=1}^{\infty} (-n, n)$ for $\mathbb{R}$, $n \in \mathbb{N}$
  $\mathbb{R} = \bigcup_{z \in \mathbb{Z}} (-z, z)$

the operations of union and intersection can be extended to infinite collections of sets as well:

Thm: If $A_1, A_2, A_3, \ldots$ is a collection of sets, all defined on a sample space S, then

countable unions & intersections

$$\bigcup_{i=1}^{\infty} A_i = \{x : x \in A_i \text{ for some } i \in \infty\} = \{x \in S : x \in A_i \text{ for some } i\} = \lim_{n \to \infty} \bigcup_{i=1}^{n} A_i$$

$$\bigcap_{i=1}^{\infty} A_i = \{x : x \in A_i \text{ for all } i = 1, \ldots, \infty\} = \{x \in S : x \in A_i \text{ for all } i\} = \lim_{n \to \infty} \bigcap_{i=1}^{n} A_i$$

Remark: It is also possible to define unions and intersections over uncountable collections of sets: (not being covered here)

corollary: If Γ is an index set (set of elements to be used as indices), then

uncountable unions & intersections:

$$\bigcup_{\alpha \in \Gamma} A_\alpha = \{x \in S : x \in A_\alpha \text{ for some } \alpha\}$$

$$\bigcap_{\alpha \in \Gamma} A_\alpha = \{x \in S : x \in A_\alpha \text{ for all } \alpha\}$$

note: don't play major role in statistics; sometimes provide a useful mechanism as seen later (Ch. 8).

- It can be verified that de Morgan's laws holds for infinite collections of subsets too, i.e.

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c$$

Proof. $w \notin \bigcup_i A_i \iff \forall i, w \notin A \iff \forall i, w \in A_i^c \iff w \in \bigcap_i A_i^c$. ∎

Def: two events A and B are disjoint (mutually exclusive) if $A \cap B = \emptyset$

the events $A_1, A_2, \ldots$ are pairwise disjoint (mutually exclusive) if $A_i \cap A_j = \emptyset \ \forall i \neq j$

- Disjoint sets are sets w/ no points in common

Def: If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, \ldots$ forms a partition of S
- Generally, partitions are useful, enabling us to divide the sample space into small, nonoverlapping pieces

Notes (countable unions):

Conceptual Layer	What it means	How it's practically used
Set Theory	$x \in \bigcup A_i \iff \exists i, x \in A_i$	logical membership
Prob. Theory	union: = "at least one event occurs"	aggregating likelihood

Notes (countable intersections):

	Set Theory	Prob. Theory
Conceptual	common elements across sets	joint occurrence of events
Abstract	$x \in \bigcap A_i \iff \forall i, x \in A_i$	$P(\bigcap A_i)$: prob. all occur

1.2: Basics of Probability Theory

- the 'frequency of occurrence' of an outcome can be thought of as a probability
- we describe outcomes of an experiment probabilistically to analyze the experiment statistically
- * Describe basics of prob. here
  → Define w.r.t. to a mathematically simpler axiomatic approach (instead of in terms of frequency)

+ concerned not w/ prob. interpretations but instead define prob's by a function satisfying the axioms

- Discuss subjective prob. interpretation as a belief in the chance of an event occurring

1.2.1: Axiomatic Foundations

- For each event A in S, we want to associate w/ A a number btwn 0 and 1 that we call the probability of A, i.e. $P(A)$

  + not as simple to define as follows:

    for each $A \subseteq S$, we define $P(A) :=$ prob. that A occurs

    → technical difficulties arise often more relevant to probabilists than statisticians

    → however, a firm understanding of statistics requires us at least a passing familiarity w/ the following:

Def: A collection of subsets of S is called a sigma algebra (or Borel field), denoted by $\mathcal{B}$, if it satisfies the following properties:

a. $\emptyset \in \mathcal{B}$

b. if $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ (closed under complementation)

c. if $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$ (closed under countable unions)

Remarks: · $S \in \mathcal{B}$

· $\mathcal{B}$ is also closed under countable intersections (by deMorgan's)

i.e. if $A_1, A_2, \ldots \in \mathcal{B}$ then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}$

· the collection of two sets $\{\emptyset, S\}$ is a sigma algebra, usually called the trivial sigma algebra

Ex(1): if S is a finite or countable, then these technicalities usually do not arise, for we define for a given sample space S,

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

Remarks: · If S has n elements, there are 2^n sets in $\mathcal{B}$, e.g.

if $S = \{1,2,3\}$, then $|\mathcal{B}| = 2^3 = 8$: $\{\{1\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\},\{2\},\{3\},\emptyset\}$.

Ex(2): Let $S = (-\infty, \infty)$ on the real line. Then choose $\mathcal{B}$ to contain all sets of the form

$[a,b]$, $(a,b]$, (a,b) and $[a,b)$ $\forall\, a \in \mathbb{R}$ and $b \in \mathbb{R}$.

Remark: From the properties of $\mathcal{B}$, it follows that $\mathcal{B}$ contains all sets that can be formed by taking (possibly countably infinite) unions and intersections of sets of the above varieties.

· We are now in a position to define a probability function.

Def: Given a sample space S and an associated sigma algebra $\mathcal{B}$, a probability function is a function P w/ domain $\mathcal{B}$ that satisfies:

1) $P(A) \geq 0$ for all $A \in \mathcal{B}$

2) $P(S) = 1$

3) if $A_1, A_2, \ldots$ are pairwise disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Remarks: · these three properties are called (Kolmogorov's) Axioms of Probability (W)

· Any function P that satisfies (W) is called a prob. function

· we assume finite additivity holds, i.e. **if $A \in \mathcal{B}$ and $B \in \mathcal{B}$ are disjoint, then $P(A \cup B) = P(A) + P(B)$**

* For any sample space, many different prob. functions can be defined

→ which one(s) reflect what is likely to be observed in a particular experiment is still to be discussed

Ex(Defining Probabilities-I): Tossing a fair coin: $S = \{H, T\}$.

$P(\{H\}) := P(\{T\}) := \frac{1}{2}$

Tossing an unfair coin: $\begin{cases} P(\{H\}) = q \\ P(\{T\}) = 1-q \end{cases}$ for some $q \in [0,1]$

Tossing two coins in a row:

$S = \{HH, HT, TH, TT\}$

Take $P(\{HH\}) = q_1, \ldots, P(\{TT\}) = q_4$

where $\begin{cases} q_1 + \cdots + q_4 = 1 \\ q_1, \ldots, q_4 \geq 0 \end{cases}$

* we need general methods of defining prob. functions that we know will always satisfy Kolmogorov's Axioms

→ the following gives a common method of defining a legitimate prob function

Thm. Let $S = \{s_1, \ldots, s_n\}$ be a finite set. Let $\mathcal{B}$ be any sigma algebra of subsets of S. Let $p_1, \ldots, p_n$ be nonnegative numbers that sum to 1.

for any $A \in \mathcal{B}$, define $P(A) := \sum_{i: s_i \in A} p_i$

Corollary: this remains true if $S = \{s_1, s_2, \ldots\}$ is a countable set.

· the physical reality of the experiment might dictate the prob. assignment, e.g.

Ex: Defining Probabilities-II: In a dart throwing example,

$P(\text{scoring } i \text{ points}) = \frac{\text{area on ring } i}{\text{area of dart board}} = \ldots = \frac{(6-i)^2 - (5-i)^2}{5^2}$, $i \in \{1, \ldots, 5\}$

→ the sum of the disjoint regions equals the area of the dart board

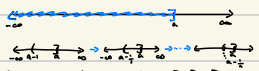
Remarks:

Q: How to show that sets are elements of $\mathcal{B}$?

A: Verify that the set in question can be obtained from known elements of $\mathcal{B}$ via countably many set operations

→ In particular, sometimes useful to use fact that any real number can be constructed as the limit of a sequence of rational numbers

Let $S = (-\infty, a)$ where $a \in \mathbb{Q} \Leftrightarrow \{x \mid x \leq a\}$, $a \in \mathbb{Q}^c$.



i) Let $a \in \mathbb{Q}$. I will show that $\{a\} \in \mathcal{B}$

For $n \in \mathbb{N}$, $a - \frac{1}{n} \in \mathbb{Q}$.

Then, $(-\infty, a) = \lim_{n\to\infty} \bigcup_{i=1}^{n} (-\infty, a - \frac{1}{n}) \in \mathcal{B}$.

Since $(-\infty, a] \in \mathcal{B}$,

$\{a\} = (-\infty, a] \setminus (-\infty, a) \in \mathcal{B}$. ■

ii) Let $a \in \mathbb{R}$, $n \in \mathbb{N}$

Pick rationals $q_n \downarrow a$ (decreasing sequence of rationals converging to a from above, i.e. $q_n > a \;\forall\, n$.)

Then, $(-\infty, a] = \lim_{n\to\infty} \bigcap_{i=1}^{n} (-\infty, q_n] \in \mathcal{B}$.

Pick rationals $r_n \uparrow a$ (increasing sequence of rationals converging to a from below, i.e. $r_n < a \;\forall\, n$)

Then, $(-\infty, a) = \lim_{n\to\infty} \bigcup_{i=1}^{n} (-\infty, r_n] \in \mathcal{B}$.

Thus, $\{a\} = (-\infty, a] \setminus (-\infty, a) \in \mathcal{B}$. ■

1.2.2: The calculus of Probabilities

* From the Axioms of Probability (AOP), we can build up many properties of the probability function, some of which are quite helpful in the calculation of more complicated probabilities

Thm. If P is a prob function and A is any set in $\mathcal{B}$, then
a. $P(\emptyset) = 0$
b. $P(A) \leq 1$
c. $P(A^c) = 1 - P(A)$

Remark: Fairly self-evident; applied to a single event

Thm. If P is a prob. function and A and B are any sets in $\mathcal{B}$, then
a. $P(B \cap A^c) = P(B) - P(A \cap B)$
b. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
c. (If $A \subset B$, then $P(A) \leq P(B)$)

Corollary: $P(A \cap B) \geq P(A) + P(B) - 1$ (Bonferoni's Inequality) (B)

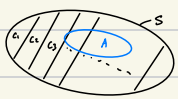
Proof (c): $P(B) \overset{\text{AOP 3}}{=} P(A) + P(B \cap A^c)$
$\Rightarrow P(B) \geq P(A)$, since $P(B \cap A^c) \geq 0$ by AOP 1.

Proof (a):

Proof (b): By (c), $P(A \cap B^c) = P(A) + P(B) - P(A \cup B)$
$\Rightarrow P(A \cap B) \geq P(A) + P(B) - 1$

Remarks:
- Allows us to bound the prob. of a simultaneous event (intersection) in terms of the prob.'s of the individual events
- Unless the probs of the individual events are sufficiently large, the (Bonferoni) bound is a useless (but correct) negative number

Thm. If P is a probability function, then
a. $P(A) = \sum_{c \in C} P(A \cap c_i)$ for any partition $C_1, C_2, \ldots$
b. $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any sets $A_1, A_2, \ldots$ (Boole's Inequality)



Proof (a): $P(\bigcup_{i=1}^{n} A_i^c) \leq \sum_{i=1}^{n} P(A_i^c)$, applying Boole's Inequality
$1 - P(\bigcap_{i=1}^{n} A_i) \leq n - \sum_{i=1}^{n} P(A_i)$, Using $P(A_i^c) = P((A_i)^c)$ and $P(A_i^c) = 1 - P(A_i)$
$P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$.

Remarks:
- A partition ($C_1, C_2, \ldots$) of S means $C_1, C_2, \ldots$ are pairwise disjoint, and $\bigcup_{i=1}^{\infty} C_i = S$
- Boole's inequality is a more general version of the Bonferoni inequality, i.e.

$$P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1) \quad (**)$$

1.2.3: Counting

- elementary task of counting can become sophisticated for a statistician
- most often, methods of counting are used to construct prob. assignments on finite sample spaces

* Break down counting problems into a series of simple tasks that are easy to count, and employ known rules of combining tasks

(Fundamental) Thm. of counting (FTOC): If a job consists of k separate tasks, the i^{th} of which can be done in n_i ways, $i = 1, \ldots, k$, then the entire job can be done in $n_1 \times n_2 \times \cdots \times n_k = n^k$ ways

Proof: Let $k = 2$. The first task can be done in n_1 ways, and for each of these ways, we have n_2 choices for the second task. Thus, we can do the job in $\underbrace{(1 \times n_2) + (1 \times n_2) + \cdots + (1 \times n_2)}_{n_1 \text{ terms}} = n_1 \times n_2$ ways, establishing the theorem for $k = 2$.

Remark: Although the FTOC is a reasonable place to start, in applications there are usually more aspects of a problem to consider, i.e.

counting
- with / without replacement
- whether / not order matters

- As an aside, we first discuss helpful notation below:

Def: For a positive integer n, $n!$ (n factorial) is the product of all of the positive integers less than/equal to 1, i.e.
$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$ where $0! = 1$.

Remark: # of distinct ways the original n objects can be shuffled

Def: For nonnegative integers n and r, where $n \geq r$, we define n permute r: P_r^n
$P_r^n = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$

Remark: an ordered arrangement of r distinct objects out of a total n

Def: For nonnegative integers n and r, where $n \geq r$, we define n choose r $\binom{n}{r}$ as
$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

Remark: the # of (unordered) combinations of n objects chosen r at a time

Ex (Lottery ticket): From the numbers $1, 2, \ldots, 44$ a person may pick any six for their ticket

a) ordered w/o replacement: $44 \times 43 \times 42 \times 41 \times 40 \times 39 = \frac{44!}{38!} = 5,082,517,440$

b) ordered w/ replacement: $\underbrace{44 \times 44 \times \cdots \times 44}_{6 \text{ times}} = 44^6 = 7,256,313,856$

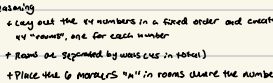
c) unordered w/o replacement
- must divide out the redundant orderings, e.g. here six numbers arranged in $6!$ ways

$$\left.\begin{matrix}\\ \\ \\ \end{matrix}\right\} \frac{44 \times 43 \times 42 \times 41 \times 40 \times 39}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{44!}{6!38!} = \frac{44!}{6!(44-6)!} = \binom{44}{6} = 7,059,052$$

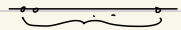
d) Unordered w/ replacement
- common mistake (but too small) is $\frac{44^6}{6!}$
- consider placing 6 markers on the 44 numbers, i.e. count all arrangements of 43 walls (44 bins -> 45 walls, disregard two ends) -> thus have 43+6 objects that can be arranged 49! ways, eliminating the redundant orderings, i.e.

$$\frac{49!}{6!43!} = \frac{(44+6-1)!}{6!(44-6)!} = 13,983,816$$

Reasoning
+ Lay out the 44 numbers in a fixed order and create 44 "rooms", one for each number
+ Rooms are separated by walls (43 in total)
+ Place the 6 markers "M" in rooms where the numbers have been chosen



$\Leftrightarrow$ Given $\left\{\begin{matrix}6 \text{ rooms} \\ 43 \text{ inner walls}\end{matrix}\right.$



Above: $\{1, 10, 14, 24, 41, 42\}$ were chosen

* Key obs: # tickets $\equiv$
# unique ways to place 6 markers to 44 bins
= # ways to arrange 6 markers w/ 43 inner walls
$= \binom{6+43}{6}$.

Possible Methods of counting (Summary)

Number of possible arrangements of size r from n objects

	w/o replacement	w/ replacement
ordered	$P_r^n = \frac{n!}{(n-r)!}$	n^r
unordered	$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!}$	$\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$

Remark: • $\binom{n}{r}$ are referred to as binomial coefficients

• $C_r^n = \frac{P_r^n}{r!} = \frac{n!}{r!(n-r)!}$; dividing by redundant orderings

Prop. Since $r! \geq 1$, it follows that $C_r^n \leq P_r^n$

1.2.4: Enumerating options

* the counting techniques of the previous section are useful when the sample space S is a finite set and all of the outcomes are equally likely

→ then prob.'s of events can be calculated by simply counting the number of outcomes in the event

Prop. Suppose $S = \{s_1, \ldots, s_N\}$ is a finite sample space.
Assume that all outcomes are equally likely s.t.

$$P(\{s_i\}) = \frac{1}{N} \quad \forall\, s_i \in S$$

Then, using Ax P(3),

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{N} = \frac{\#\text{ of elements of } A}{\#\text{ of elements in } S}$$

Remark: For large sample spaces, the counting techniques might be used to calculate both the numerator and denominator of this expression

Ex (Poker): suppose the events do not depend on order, so we use the unordered outcomes

Q: How many 5-card hands can be chosen from a 52-card deck

A: $\binom{52}{5} \Rightarrow P(\{\text{random hand}\}) = 1/\binom{52}{5} = 2{,}598{,}960$ ($\#$)

Q: $P(\{4 \text{ aces}\}) = \frac{48}{\#}$ where 48 different ways of specifying fifth card

A A A A *

Q: $P(\{4 \text{ of a kind}\}) = \frac{13 \cdot 48}{\#}$ where 48 as above

13 - which denomination there will be four of

2D 2H 2S 2C 2*

Q: $P(\{\text{exactly one pair}\}) = \frac{1}{\#} \cdot 13 \cdot \binom{4}{2} \cdot \binom{12}{3} \cdot 4^3$

e.g. {3D, 3S, 2S, 2C, 10C}

where 13 as above; $13 - 1 = 12$

$\binom{4}{2}$ - # ways to specify the two cards from that denomination

$\binom{12}{3}$ - # of ways of specifying the other three denominations

4^3 - # of ways of specifying the other three cards from those denominations

- When sampling w/o replacement, if we want to calculate the prob. of an event that does not depend on the order, we can either use the ordered or unordered sample space
  → Each outcome in the unordered sample space corresponds to $r!$ outcomes in the ordered sample space

- When sampling w/ replacement, each outcome in the unordered sample space corresponds to some outcomes in the ordered sample space, but the number of outcomes differs
  → the formula for the number of outcomes in the unordered sample space is useful for enumerating the outcomes, but ordered outcomes must be counted to correctly calculate prob.'s, e.g.

Ex (Sampling w/ replacement)
"Uniform Sampling" $r = 2$ items from $n = 3$ items w/ replacement

ordered	(1,1)	(2,2)	(3,3)	(1,2)/(2,1)	(1,3)/(3,1)	(2,3)/(3,2)
unordered	{1,1}	{2,2}	{3,3}	{1,2}	{1,3}	{2,3}
Prob.	1/9	1/9	1/9	2/9	2/9	2/9

1.3: Conditional Probability and Independence

- In many instances, we are in a position to update the sample space based on new info
  → In such cases, we want to be able to update prob. calculations, i.e. to calculate conditional prob.'s

Def: If A and B are events in S, and $P(B) > 0$, then the conditional probability (of A given B) is:

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

Remarks: • Equivalently, $P(A \text{ true} | B \text{ true}) = \frac{P(\text{both } A \text{ and } B \text{ true})}{P(B \text{ true})}$; relative prob. of overlap inside B

where (*) is 'renormalizing' under (new) info B

• Equivalently $P(A \cap B) = P(B) \cdot P(A|B)$
(joint prob. = marginal × conditional)

Prop. Generally, $P(A \cap B \cap C) = P(A) \times P(B \cap C | A)$
$= P(A) \times P(B|A) \times P(C|B,A)$

→ Hence, **Chain Formula:** $= P(A_1 \cap \cdots \cap A_n) = \prod_{i \geq 1} P(A_i | A_1 \cap \cdots \cap A_{i-1})$

Ex (Four Aces): $P(\{4 \text{ aces}\}) = \frac{1}{\binom{52}{4}} = \left(\frac{52 \times 51 \times 50 \times 49}{4 \times 3 \times 2 \times 1}\right)^{-1} = \overbrace{P(1^{st} \text{ ace}) \cdot P(2^{nd} \text{ ace} | 1^{st} \text{ ace})}^{(*)} = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}$
$\cdot P(2^{nd} \text{ ace} | \text{first two aces}) \cdot P(4^{th} \text{ ace} | \text{first three aces})$

s.t. (*) updates the sample space after each draw of a card

Prop. If $P(A)>0$, then $P(B\cap A) = P(A)P(B|A)$; joint. prob. of both happening
$= P(B)P(A|B)$

corollary: Hence, $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$ for $P(B)>0$, $A,B\in\mathcal{F}$.

Remarks:

- conditional probabilities provide the means to quantify how we may have different assessments of the same event
  - due to different info we use or have access to
  - what event/info we choose to condition on can affect the answer
- conditional prob.'s provide the mathematical machinery of Bayesian statistics which bases the inference on the posterior distribution, i.e. cond. prob. of quantities of interest conditionally given observed data
- cond. Prob.'s are obtained usually via the Bayes' formula

Ex (Three Prisoners): three prisoners A, B, and C
one of them chosen (randomly) to be pardoned.

A asks the warden, who is supposed to keep secret:
- which among B and C will be executed?

to which the warden responds
- B is to be executed.

warden's thinking: the prob. that either A or B or C gets pardoned is 1/3.
Between B and C, at least one is executed.
so, "I" gave A no new information on his life.

A's thinking: since either C or I gets the pardon, my chance of being alive has gone up to 1/2

Let A, B, C be pardon events, i.e.
$P(A)=P(B)=P(C)=1/3$.

IF Prisoner is Pardoned	Then, warden tells A	Prob
A	$\to$ B dies	1/2
	$\to$ C dies	1/2
B	$\to$ C dies	1
C	$\to$ B dies	1

Let W: event that warden tells A that B dies
- the warden's reasoning is as follows:

$$P(A|W) = \frac{P(A\cap W)}{P(W)} = \frac{P(A)P(W|A)}{P(W)} = \frac{(1/3)(1/2)}{(\frac{1}{3})(\frac{1}{2})+(\frac{1}{3})(1)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

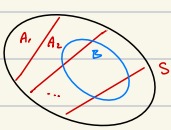
- A's reasoning comes from

$$P(A|B^c) = \frac{P(A\cap B^c)}{P(B^c)} \overset{A\subset B^c}{=} \frac{P(A)}{1-P(B)} = \frac{1/3}{2/3} = \frac{1}{2}.$$

Prop. Let $A_1, A_2, \ldots$ be a partition of S, i.e. $S=\bigcup_{i=1}^{\infty} A_i$

Then $P(B) = \sum_i P(B\cap A_i) = \sum_i P(B|A_i)P(A_i)$

$$\Rightarrow P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}$$



Remarks: (*) is the law of total probability
- 'turning around' conditional prob's

corollary: $\forall i$ $P(A_i|B) \propto P(B|A_i)P(A_i)$

Remarks:
- $\propto$: "is proportional to" (up to multiplying constant)
- In Bayesian statistics:

$P(A_i|B)$ or $P(B|A_i)$ $P(A_i)$
|| ||
posterior dist. likelihood prior dist.

Def: two events are statistically independent if $P(A\cap B)=P(A)P(B)$

Remarks:
- does not require $P(A)$ or $P(B)>0$
- A & B independent does not mean $A\cap B=\emptyset$ (disjoint def.)

Ex (Dice Rolling): $S=\{1,\ldots,6\}^4$. Here, $\tilde{S}=\{(s_1,\ldots,s_4)\}\in\{1,\ldots,6\}^4$

Q: $P($of throwing at least one 6 in 4 rolls$)$
Assuming outcome of rolls are independent

$= 1-P($no 6 in 4 rolls$)$
$= (1-\prod_{i=1}^{4} P($no 6 in roll $i$$)$
$= 1-(1-\frac{1}{6})^4 = 1-(5/6)^4 = 0.518$

Proof. We want to be able to convey that an event B has no effect on A by insisting that

$$P(A|B)=P(A) \quad (\text{provided } P(B)>0)$$

$$\Rightarrow P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B), \text{ provided } P(A)>0.$$

Moreover,
$P(A\cap B) = P(A|B)P(B)$
$= P(A)P(B)$. ∎

- Independence of A and B implies independence of complements, i.e.

Thm. (ind.) If $A \perp\!\!\!\perp B$, then
i) $A \perp\!\!\!\perp B^c$
ii) $A^c \perp\!\!\!\perp B$
iii) $A^c \perp\!\!\!\perp B^c$

Proof (i): $P(A\cap B^c) = P(A) - P(A\cap B)$, as ii)
$= P(A) - P(A)P(B)$, ind.
$= P(A)(1-P(B))$
$= P(A)P(B^c)$. ∎

Def: A collection of events $A_1, \dots, A_n$ are <u>mutually independent</u>, if for any subcollection $A_{i_1}, \dots, A_{i_k}$, we have

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j})$$

Remark: mutual independence much stronger than pairwise ind.

corollary: Pairwise ind. does <u>not</u> imply mutual ind., i.e.
given 3 events A, B, C s.t.

$$\left.\begin{aligned} P(A)P(B) &= P(A \cap B) \\ P(A)P(C) &= P(A \cap C) \\ P(B)P(C) &= P(B \cap C) \end{aligned}\right\} \Rightarrow P(A \cap B \cap C) = P(A)P(B)P(C)$$

conversely, in general,

$$P(A \cap B \cap C) = P(A)P(B)P(C) \not\Rightarrow P(A \cap B) = P(A)P(B)$$

Ex (Letters): Let $S = \{aaa, abb, ccc, abc, bca, cba, acb, bac, cab\}$ s.t. each element in S is equally (1/9) probable.
Let $A_i = \{i^{th}$ place in the triple is occupied by a $a\}$

Then, $P(A_1) = P(A_2) = P(A_3) = 1/3$,

$$P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = P(A_1 \cap A_2 \cap A_3) = P(\{aaa\}) = 1/9.$$

Thus, A_1, A_2, A_3 are pairwise independent, but

$$P(A_1 \cap A_2 \cap A_3) \neq P(A_1)P(A_2)P(A_3)$$ and, therefore, not mutually independent.

Ex (Three coin tosses): Assume three coin tosses, each of which produces H or T.

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$ Suppose equal prob. (1/8) on each element of S.
Let $A_i :=$ i^{th} toss is H, $i = 1, 2, 3$.
Then, $P(A_1) = P(\{HHH, HHT, HTH, HTT\}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$.

$$P(A_1 \cap A_2) = (\{HHH, HHT\}) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

$$P(A_1 \cap A_2 \cap A_3) = P(\{HHH\}) = \frac{1}{8}.$$

Likewise $P(A_i) = 1/2 \quad \forall\, i = 1, 2, 3$

$$P(A_i \cap A_j) = 1/4 \quad \forall\, i, j \in \{1, 2, 3\}$$

So $$\left.\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_i \cap A_j) &= P(A_i)P(A_j) \ \forall\, i \neq j \end{aligned}\right.$$

Hence, A_1, A_2, A_3 are pairwise <u>and</u> mutually independent

1.4: Random Variables

- In many experiments, it is easier to deal w/ a summary variable than w/ the original prob. structure
  - ★ In defining the quantity X, we have a mapping (a function) from the original sample space to a new sample space, usually a set of real numbers.

Def: A <u>random variable</u> is a function from a sample space S into the real numbers

Ex (RV's):

Experiments	Random Variables
1) Toss 2 dice $S = \{1, \dots, 6\}^2 = \{1, \dots, 6\}$	$X =$ sum of 2 numbers

- Suppose $s = (s_1, s_2)$. Then $X = X(s) = s_1 + s_2$

2) Toss a coin 25 times $S = \{H, T\}^{25}$	$X =$ # heads

- Suppose $s = \{(c_1, \dots, c_{25}) : c_i \in \{H, T\}\}$. Then, $X = X(s) = \sum_{i=1}^{25} \mathbb{1}(c_i = H)$

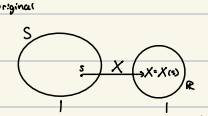
<u>note</u>: indicator function $\mathbb{1}(A) = \begin{cases} 1, & A \text{ true} \\ 0, & \text{otherwise} \end{cases}$

3) Apply different amt of fertilizer to flower plants	$X =$ yield/acre

- In defining a RV, we have also defined a new sample space (the range of the RV)
  - → we now must formally check that our prob. function, which is defined on the original sample space, can be used for the RV:

Prop. (★) Suppose we have a sample space $S = \{s_1, \dots, s_n\}$ w/ prob. function P and we define a RV X w/ range $\mathcal{X} = \{x_1, \dots, x_m\}$. We can define a prob function P_X on $\mathcal{X}$ in the following way:

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$$



Remarks:
- we observe $X = x_i$ i.f.f. the outcome of the random experiment is an $s_j \in S$ s.t. $X(s_j) = x_i$ (equivalence in words)
- RV's will always be denoted w/ uppercase letters and the realized values of the variable (its range) will be denoted by corresponding lowercase letters

Ex (Three coin tosses II): Let X: number of heads out of tossing coin 3 times

$$s \in S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$
$$X(s) = \quad 3 \quad 2 \quad 2 \quad 1 \quad 2 \quad 1 \quad 1 \quad 0$$

So, the range of (RV) X is:

$\mathcal{X}$	0	1	2	3
P_X	1/8	3/8	3/8	1/8

Ex (D'sty. of a RV): Let X = # heads after tossing coin 50 times

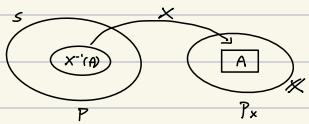
Then, $\mathcal{X} = \{0,\dots,50\}$. So $P_X(X=i) = \dfrac{\binom{50}{i}}{2^{50}} \quad \forall\, i \in \mathcal{X}$.

- Previous definition for $P_X(\cdot)$ applies to settings w/ a finite S and $\mathcal{X}$
  → more straightforward; also the case when $\mathcal{X}$ is countable
- If $\mathcal{X}$ is uncountable, we define the induced prob. function P_X as follows:

Def: Suppose the range of random variable X is $\mathcal{X}$. For any subset $A \subset \mathcal{X}$, the probability distribution of X is a probability function, denoted P_X, on (a sigma algebra of) subsets of $\mathcal{X}$ s.t.

$\forall$ such subset $A \subset \mathcal{X}$

$$P_X(X \in A) = P(X^{-1}(A)) = P(\{s \in S : X(s) \in A\})$$



Remark: For complicated space (e.g. uncountable), we need to specify prob.'s on "simple" sets which generate the associated sigma algebra

1.5: Distribution Functions

- With every RV X, we associate a function called the cumulative distribution function of X:

Def: The cumulative distribution function (cdf) of a RV X is defined by $F_X(x) = P(X \le x)$ for all $x \in \mathcal{X}$

- Every cdf satisfies certain properties, some of which are obvious when we think of the definition of $F_X(x)$ in terms of prob.'s:

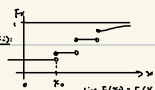
Thm. The function $F(x)$ is a cdf i.f.f. the following three conditions hold:

a) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

b) $F(x)$ is a nondecreasing function of x

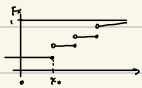
c) $F(x)$ is right-continuous, i.e. $\lim_{x \downarrow x_0} F(x) = F(x_0)$

e.g. (satisfying i-iii):



and thus right-continuous

whereas



$\lim_{x \uparrow x_0} F(x) = F(x_0)$ and thus left-continuous, but

$\lim_{x \downarrow x_0} F(x) \neq F(x_0) \Rightarrow$ not right-continuous

Ex (tossing for a head): Let p := prob. that a coin turns head

X := # independent coin tosses needed to get a head

(≡ # (ind.) trials until first head

≡ # (ind.) trials on which first head occurs

note: $X \sim Geo(p)$. ; Aside: some authors define $X \sim Geo(p)$ as { X^* := # failures before first success; X := # trials on which first success occurs

$\Rightarrow X^* = X - 1 \Rightarrow P(X = x_0) = P(X^* = x_0 - 1)$. ■

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

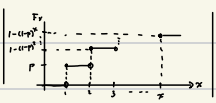
$\Rightarrow F_X(x) = P(X \le x), \ \forall\, x \ge 0$

$= \sum_{i=1}^{x} (1-p)^{i-1} p$

$= p \sum_{k=0}^{x-1} (1-p)^k$, letting $k = i-1$

$= p \dfrac{1-(1-p)^x}{1-(1-p)}$

$= 1-(1-p)^x, \ x \ge 1$



recall: (Geometric series)

Prop: $\forall\, r \neq 1 \in \mathbb{R}$, $\sum_{k=0}^{n} r^k = \dfrac{1-r^{n+1}}{1-r}$

Prop: $\forall\, |r| < 1$, i.e. $r \in (-1,1)$, $\sum_{k=0}^{\infty} r^k = \dfrac{1}{1-r}$; otherwise, series diverges.

Proof (cdf). i) $\lim_{x \to -\infty} F_X(x) = 0$ since $F_X(x) = 0 \ \forall\, x < 0$

$\lim_{x \to \infty} F_X(x) = \lim_{x \to \infty} 1-(1-p)^x = 1$ for $x \in \mathbb{Z}$.

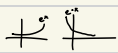
ii) $\sum_{i=1}^{x} (1-p)^{i-1} p$ contains more positive terms as x increases

iii) $\Rightarrow \lim_{\varepsilon \to 0} F_X(x+\varepsilon) = F_X(x)$. ■

Ex (continuous cdf): Let $F_X(x) = \dfrac{1}{1+e^{-x}}$ which satisfies

i) $\lim_{x \to -\infty} F_X(x) = 0$ since $\lim_{x \to -\infty} e^{-x} = \infty$

$\lim_{x \to \infty} F_X(x) = 1$ since $\lim_{x \to \infty} e^{-x} = 0$.



ii) $\frac{d}{dx} F_X(x) = \frac{d}{dx}(1+e^{-x})^{-1} = \dfrac{e^{-x}}{(1+e^{-x})^2} > 0$

$\Rightarrow$ increasing $\Rightarrow$ nondecreasing.

iii) $\Rightarrow$ (logistic function is right/left) continuous

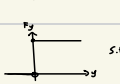
Def: A (RV X is) continuous if $F_X(x)$ is a continuous function of x

A (RV X is) discrete if $F_X(x)$ is a step function of x.

Remark: cdf can also be a "mixture" of continuous segments and jumps

e.g.

where $X \sim$ Uniform s.t. $F_X(x) = \begin{cases} x, & x \in [0,1] \\ 0, & x < 0 \\ 1, & x > 1 \end{cases}$



s.t. $F_Y(y) = \begin{cases} 0, & y < 0 \\ 1, & y \ge 0 \end{cases}$

Notice: $F_Z(z) = \frac{1}{2}F_X(z) + \frac{1}{2}F_Y(z)$

still satisfies (1)-(iii) ⇒ valid cdf.

and in context,
flip a fair coin $C = \{H, T\}$ s.t. $\begin{cases} \text{if } C=H, & Z=X \\ \text{if } C=T, & Z=Y \end{cases}$

Remark: motivates RV model construction from constituent RV mixtures

- We close w/ a thm. formally stating that F_X completely determines the prob. dist. of a RV X
  ⇒ this is true if $P(X \in A)$ is defined only for events $A \in \mathcal{B}^1$, the smallest sigma algebra containing all of the intervals of real numbers of the form $(a,b), [a,b), (a,b]$, and $[a,b]$
  ⇒ we don't consider a larger class of events than above; avoids such pathological cases

Def: the RVs X and Y are *identically distributed* if,
$$P(X \in A) = P(Y \in A) \quad \forall \text{ set } A \in \mathcal{B}^1$$

Remark: $X \overset{d}{=} Y$ does not necessarily imply $X = Y$

Ex (Identically Distributed RV's): toss a fair coin 3 times.
then X = # heads and Y = # tails
$\Rightarrow X \overset{d}{=} Y$ since $P(X=i) = P(Y=i) \ \forall i \in S$.

Note that, $X \neq Y$ since for no sample points do we have $X(\omega) = Y(\omega)$.

Thm: The following two statements are equivalent:
a) the RVs X and Y are identically distributed
b) $F_X(x) = F_Y(x) \ \forall x \in \mathbb{R}$

1.6: Density and Mass Functions

- Associated w/ a RV X and its cdf F_X is another function, called either the probability density function (pdf) or probability mass function (pmf).
  ⇒ respectively refer to the continuous & discrete case
  ⇒ concerned w/ "point probabilities" of RV's

Def: the probability mass function (pmf) of a discrete RV X is given by
$$f_X(x) = P(X = x) \text{ for all } x$$

Ex (Geometric prob's): Let $X \sim Geo(p)$, $0 < p < 1$
has $f_X(x) := P(X=x) := \begin{cases} (1-p)^{x-1}p, & x = 1,2,\dots \\ 0, & \text{otherwise} \end{cases}$

s.t. $P(a \le X \le b) = \sum_{k=a}^{b} f_X(k) = \sum_{k=a}^{b} (1-p)^{k-1}p$ (**)

$$P(X \le b) = \sum_{k=1}^{b} f_X(k) = F_X(b)$$

Prop: the probability density function (pdf), $f_X(x)$, of a continuous RV X is the function that satisfies
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \text{ for all } x$$

Remarks: • if F_X is differentiable, then the pdf always exists and it follows: $f_X(x) = \frac{d}{dx}F_X(x)$

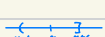
• F_X, or f_X contain all info there is abt the dist. of the RV X, we denote
$$X \sim F_X \text{ or } X \sim f_X \quad \text{(equivalently)}$$
"X is distributed as"
* can use either F_X or f_X to solve problems; try to choose simpler one

• we can instead understand the purpose of the pdf, i.e. instead of summing (as seen in the discrete case - **), substitute integrals for sums:
$$P(X \le x) = F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

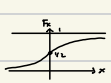
Prop: $P(X = x) = 0$ whenever X - continuous RV

Corollary: $P(a < X \le b) = P(a \le X \le b) = P(a \le X < b) = P(a < X < b)$

Proof (sketch): We must be careful in our definition of a pdf in the continuous case
If we naively try to calculate $P(X=x)$ for a continuous RV:
since $\{X=x\} \subset \{x-\varepsilon < X \le x\}$ for any $\varepsilon > 0$,
$$P(X=x) \le P(x-\varepsilon < X \le x) = F_X(x) - F_X(x-\varepsilon) \text{ for any } \varepsilon > 0.$$
Therefore,
$$0 \le P(X=x) \le \lim_{\varepsilon \to 0}[F_X(x) - F_X(x-\varepsilon)] = 0 \text{ by continuity of } F_X. \ \square$$



Ex (Logistic probabilities): Recall the logistic cdf:
$$F_X(x) = \frac{1}{1+e^{-x}}$$



which gives
$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

symmetry since $f_X(x) = f_X(-x)$, i.e.

$$f(-x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}e^{-2x}}{(1+e^{-x})^2 e^{-2x}} = \frac{e^{-3x}}{(1+e^{-x})^2} = f(x). \blacksquare$$

Then, $P(X \in (a,b)) = P(a < X < b) = F_X(b) - F_X(a)$

$$= \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)\,dx$$

$$= \int_{a}^{b} f_X(x)\,dx = \int_{a}^{b} \frac{e^{-x}}{(1+e^{-x})^2}\,dx$$

- There are really only two requirements for a pdf (pmf), both of which are immediate consequences of the definition:

Thm. A function $f_X(x)$ is a pdf (pmf) of a RV X i.f.f.

a) $f_X(x) \geq 0$ for all x

b) $\sum_x f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ (pdf)

Chapter 2: Transformations and Expectations

- often, if we are able to model a phenomenon in terms of a RV X w/ cdf $F_X(x)$, we also want to study the behavior of functions of X

* Ch. 2 covers techniques that allow us to gain info abt functions of X
  → can range from very complete (dist. of these functions) to more vague (avg. behavior)

2.1: Distributions of Functions of RV's

Def. Let X - RV with cdf $F_X(x)$. Then, any function of X, denoted $g(X)$ is also a RV

Remark: often write $Y = g(X)$ to denote new RV of interest

Def. Since Y is a function of X, we can describe the probabilistic behavior, i.e.

$$P(Y \in A) = P(g(X) \in A)$$

- Depending on the choice of g, it is sometimes possible to obtain a tractable expression for this prob.

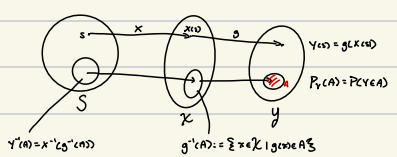
Prop. Formally, let $g: \mathcal{X} \to \mathcal{Y}$ for S: sample space. Then,

$$Y(s) = g(X(s)) = g \circ X(s)$$

and the following properties hold for the prob. dist. of $Y(s)$:

By def., $\forall A \subset \mathcal{Y}$ (prob dist of Y: =>)

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



Remarks:
- notation change from $\mathcal{X}$ to $\mathcal{X}$.
- If there is only one x for which $g(x) = y$, then $g^{-1}(y)$ is the point set $\{x\}$ and we write $g^{-1}(y) = x$, i.e. $g^{-1}(\{y\}) = \{x \in \mathcal{X} : g(x) = y\}$

Proof. Show that prob. dist. for Y satisfies Kolmogorov's Axioms.

Let X - discrete RV $\Rightarrow \mathcal{X}$: countable.

Then the sample space for $Y = g(X)$ is

$$\mathcal{Y} = \{y : y = g(x),\ x \in \mathcal{X}\},$$ also a countable set.

Thus, Y is also a discrete RV s.t.

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} f_X(x) \quad \text{for } y \in \mathcal{Y},$$

and $f_Y(y) = 0$ for $y \notin \mathcal{Y}$.

$\Rightarrow$ Finding $f_Y(y)$ involves simply identifying $g^{-1}(y)$ for each $y \in \mathcal{Y}$, and summing the appropriate probs. $\blacksquare$

Continuous RV's

If X is a cont. RV, g a nice (cont.) function,

then $Y = g(X)$ is a cont. RV.

So,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(x \in \mathcal{X} : g(x) \leq y) \\ &= \int_{\{x : g(x) \leq y\}} f_X(x)\,dx \end{aligned}$$

The set $\{x : g(x) \leq y\}$ may be difficult to identify so F_Y may be hard to derive in general

Prop. If g is (strictly) monotone, this gets simpler s.t. either

- g is increasing, i.e. $g(x) > g(y)$ if $x > y$
- g is decreasing, i.e. $g(x) < g(y)$ if $x > y$

recall: monotonic function values only move in one direction (increasing/decreasing) as the input increases

note: $g^{-1}(y) = \{x : g(x) = y\}$ is a singleton set.

Thm. Let X have the cdf $F_X(x)$,

$\mathcal{X} = \{x : f_X(x) > 0\}$ ← support of $\{f_x$

$\mathcal{Y} = g(\mathcal{X}) := \{y : y = F(x)$ for some $x \in \mathcal{X}\}$

a) If $g: \mathcal{X} \to \mathcal{Y}$ is (strictly) increasing, then $Y = g(X)$ is a RV taking values values in $\mathcal{Y}$ with the cdf

$$F_Y(y) = F_X(g^{-1}(y)) \quad \forall\, y \in \mathcal{Y}$$

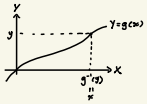
b) If $g: \mathcal{X} \to \mathcal{Y}$ is (strictly) decreasing, then $Y = g(X)$ is a RV in $\mathcal{Y}$ with cdf

$$F_Y(y) = 1 - F_X(g^{-1}(y)) \quad \forall\, y \in \mathcal{Y}$$

Proof (a). If $g\uparrow$ (strictly increasing), where $Y = g(X)$, then

$\{x : g(x) \le y\} = \{x \le g^{-1}(y)\}$

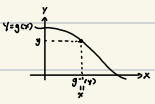
$\Rightarrow F_Y(y) = \int_{\{x \le g^{-1}(y)\}} f_X(x)\,dx = F_X(g^{-1}(y))$. ∎



Proof (b). If $g\downarrow$ (strictly decreasing), where $Y = g(X)$, then

$\{x : g(x) \le y\} = \{x \ge g^{-1}(y)\}$

$\Rightarrow F_Y(y) = \int_{\{x \ge g^{-1}(y)\}} f_X(x)\,dx = 1 - F_X(g^{-1}(y))$. ∎



corollary: we can deduce the pdf from cdf:

from (a), $f_Y(y) = \frac{d}{dy} F_Y(y)$

$= f_X(g^{-1}(y)) \underbrace{\frac{d}{dy} g^{-1}(y)}_{>0 \text{ since } g\uparrow \Rightarrow g^{-1}\uparrow}$.

and from (b),

$f_Y(y) = -f_X(g^{-1}(y)) \underbrace{\frac{d}{dy} g^{-1}(y)}_{<0 \text{ since } g\downarrow \Rightarrow g^{-1}\downarrow}$.

combining the two cases to obtain

change-of-var (formula): $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$

Thm (Probability Integral Transform). Let X - continuous RV w/ cdf $F_X(x)$.

(i) Let $Y = F_X(x) = g(x)$

Then

$Y \sim \text{Uniform}(0,1)$

Remarks: • doesn't hold for discrete RV's (F_X would be a step function)

• F_X might not be strictly increasing from 0 to 1

Recall (infimum): largest number that is less than or equal to every element in the set

Let $A \subset \mathbb{R}$.

Then $\inf A = \begin{cases} \text{largest } a \in \mathbb{R} \cup \{-\infty\} \text{ s.t. } a \le x \ \forall\, x \in A. \\ +\infty \text{ if } A \text{ doesn't exist} \end{cases}$

e.g. $\inf([a,b]) = a$

and $\inf((a,b]) = a \ne \min(a,b]$ which DNE

Inverse Function: $F_X^{-1}(y) = \inf\{x : F_X(x) \ge y\} \ \forall\, y \in (0,1)$

where • $F_X^{-1}(1) := +\infty$ only if $F_X(x) < 1 \ \forall\, x$.

• $F_X^{-1}(0) := -\infty$

Thm. Let $F_X(x)$ be a cdf of a RV.

(ii) Let $Y \sim \text{Uniform}(0,1)$

and set $Z = F_X^{-1}(Y) = \inf\{x : F_X(x) \ge y\} \ \forall\, y \in (0,1)$

Then Z has cdf F_X, i.e. $F_{X^{-1}(Y)} = F_X(x)$

Remarks • useful for generating RV's from Uniform RV's

• No restriction to continuity; holds generally (for real-valued RV's)

2.2: Expectation

Def: The expectation of a RV $g(X)$ is

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

(if)

Remarks: • Also known as 'expected values', 'average' of a RV or of the prob. dist. of the RV $g(X)$.

• Expectation is associated w/ a distribution

Q: what is the expectation of X?

Prop. If $X \in \mathbb{R}$, then by letting $g(X) := x$:

$$E(X) = \begin{cases} \int x f(x)\,dx, & X\text{-cont.} \\ \sum_{x \in \mathcal{X}} x \cdot f_X(x), & X\text{-discrete} \end{cases}$$

(def)

Note: If the domain of X is not a subset of $\mathbb{R}$ (Euclidean space), then $(\star\star)$ may be invalid but a notion of $E(X)$ may still be designed via the expectations of $\{E(g(X))$ in $\mathbb{C}\}$

Remarks: • a proper expectation exists only when the positive and negative part of integral are finite $\left(\int |x| f_X(x)\,dx < \infty\right)$
• $E(X)$ is undefined if $\int x f_X(x)\,dx$ is not absolutely convergent

Prop.
Let X-RV with dist. P_X.

Let $X_1, \dots, X_n$ be n mutually independent RV's that have identical distributions as X, i.e. $X_1, \dots, X_n$ is an n-iid sample of P_X

Def (Empirical Distribution): of $\{x_1, \dots, x_n\}$ is a prob. dist., denoted by P_n, s.t.
if $Y \sim P_n$ then $y \in \mathcal{Y} = \{x_1, \dots, x_n\}$ and
$P(Y = x_i) = \frac{1}{n}$ for $i = 1, \dots, n$

Remarks: • Thus, Y is discrete (regardless of X) and

$$E(Y) = \sum_{y \in \mathcal{Y}} P(Y=y) \cdot y = \frac{1}{n}(x_1 + \cdots + x_n) \neq E(X) = \int x f_X(x)\,dx$$

• Colloquially, this $E(Y)$ is also called the "average" of the (data) sample $x_1, \dots, x_n$.

Linearity of Expectation

Thm. Let X, Y : real-valued RV's for which the expectations exist.

(w) $\boxed{E(aX + bY + c) = a E(X) + b E(Y) + c}$ for $a, b, c \in \mathbb{R}$

i) If $X \geq 0$ almost surely, i.e. $P(X \geq 0) = 1$, then $E(X) \geq 0$

ii) If $X \geq Y$ almost surely, i.e. $P(X \geq Y) = 1$, then $E(X) \geq E(Y)$

iii) If $P(X \in [a, b]) = 1$, then $E(X) \in [a, b]$

2.3: Moments

Def: For each $n \in \mathbb{N}$, the n-th moment of X, or $F_X(x)$ is $\boxed{m_n' := E(X^n)}$

The n-th central moment of X is $\boxed{m_n := E(X - E(X))^n}$

Remarks: • Let $Y = X - E(X)$. Then $E(Y) = E[X - E(X)] = E(X) - [E(X)]^2$

• Let $n = 2$. The second central moment is the variance, i.e.

$$\boxed{\begin{aligned} \mathrm{Var}(X) &:= E(X - E(X))^2 \\ SD(X) &:= \sqrt{\mathrm{Var}(X)} \end{aligned}}$$

• $E(X)$ captures the location (center) of X dist.
• $\mathrm{Var}(X)$ and/or $SD(X)$ captures the spread

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

Thm If X is a RV w/ finite variance, then for any $a, b \in \mathbb{R}$:

$$\boxed{\mathrm{Var}(aX + b) = a^2 \mathrm{var}(X)}$$

Recall: $E(aX + b) = a E(X) + b$

Remark: Let $X = X_1 + \cdots + X_n$ where $X_1, \dots, X_n$ are mutually independent.
Suppose X_i, $i = 1, \dots, n$ are identically distributed.
Then, we will later show (more formally) that

$$V(X) = V(X_1) + \cdots + V(X_n) = n \cdot \mathrm{var}(X_i)$$

whereas $E(X) = E[X_1 + \cdots + X_n] = E(X_1) + \cdots + E(X_n)$ by (previously) defined L.O.E.

Def: Let X-RV w/ cdf F_X. The moment-generating function (mgf) is $\boxed{M_X(t) := E(e^{tX})}$
provided that the expectation exists in some neighborhood of $t = 0$ (i.e. $t \in$ some $(-h, h)$ for $h > 0$).

Remarks: • $M_X(t) = \begin{cases} \int e^{tx} f_X(x)\,dx & , X \text{ cont. RV} \\ \sum_x e^{tx} f_X(x) & , X \text{ discrete RV} \end{cases}$

• the mgf is/can be used to "characterize" the dist. (of the RV)
• mgf uniquely determines all moments

Thm. $\forall n \in \mathbb{N}$, $\boxed{E(X^n) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}}$

Remarks: - we also write the RHS as $M_X^{(n)}(0)$
(n-th derivative of the mgf evaluated @ $t=0$ gives the n-th moment)
- gives meaning of "moment generating"

2.3: Moments (cont.)

RV X is defined by cdf $\boxed{F_X(x)}$. harder to estimate

mgf $M_X(t)$ —generate→ all moments . easier to estimate

Q1: Does the mgf uniquely determine the cdf?

Q2: Does the set of moments $\{E(X^n)\}_n^\infty$ uniquely determine the cdf?

Thm. Let $F_X(x)$ and $F_Y(y)$ be two cdf's, all of whose moments exist.

i) If X and Y have bounded support, i.e. $\forall x: F_X(x) > 0 \Rightarrow x \in [a,b]$, then

$\boxed{F_X(u) = F_Y(u) \ \forall u \iff E(X^r) = E(Y^r) \ \forall r = 0, 1, 2, \dots}$

ii) If $M_X(t)$ and $M_Y(t)$ exist, and

$M_X(t) = M_Y(t)$ for all t in some neighborhood of 0,

then $F_X(u) = F_Y(u) \ \forall u$

Remark: the moments do not capture all information abt the dist. in the unbounded support scenarios

-motivation lies in the necessity to approximate RV's

Def: $X_i \to X$ in distribution ($X_i \xrightarrow{d} X$)

if $F_{X_i}(x) \to F_X(x)$ at all continuity points of F_X

- implicit in the below thm.

Thm (Convergence of MGF's leads to convergence of CDF).

Suppose $X_1, X_2, \dots$ is a sequence of random variables, each w/ mgf $M_{X_i}(t)$.

Suppose

$\lim_{i \to \infty} M_{X_i}(t) \to M_X(t)$

for all t in a neighborhood of 0, i.e.

$\exists h > 0$ s.t. $\forall t \in (-h, h)$, $M_{X_i}(t) = M_X(t)$,

and $M_X(t)$ a (valid) mgf.

Then $X_i \to X$ in distribution

where X is a RV w/ mgf $M_X(t)$

i.e. $F_{X_i}(x) \to F_X(x)$ at all points x where cdf F_X is continuous

Remark (Poisson Approx.): By the convergence thm. of mgf,

if $X_n \sim \text{Binomial}(n, p)$, $\lambda = np$. Then,

$F_{X_n}(x) \to F_Y(x)$ for $Y \sim \text{Poisson}(\lambda)$

as $n \to \infty$ $\forall x$ where F_Y is continuous.

($\Leftarrow M_{X_n}(t) \to M_Y(t) \ \forall t$

- For Poisson, F_Y is a step function w/ continuity at $n \in \mathbb{N}$;
for $n \in \mathbb{N}$, $F_Y(y) = F_Y(y + 1/2)$ continuous at $y = 1/2$.
$\Rightarrow P(X_n \le x) \to P(Y \le x) \ \forall x \notin \mathbb{N}$ (in fact, here $\forall x \in \mathbb{R}$)

- Still holds if $X_n \sim \text{Binomial}(n, p_n)$ where $n \to \infty$, $p_n \to 0$ s.t. $np_n \to \lambda$

Thm. $\forall a, b \in \mathbb{R}$,

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

2.4: TOOLS

Thm (Leibniz's Rule). If

- $f(x, \theta)$, $a(\theta)$, $b(\theta)$ are differentiable wrt θ
- $\frac{\partial f}{\partial \theta}(x, \theta)$ is continuous on $x \in \underbrace{(x_0, x_1)}_{\text{can be weakened}}$

Then for $a(\theta), b(\theta) \in (x_0, x_1)$:

$$\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x,\theta)\, dx = f(b(\theta), \theta) \frac{d}{d\theta} b(\theta) - f(a(\theta), \theta) \frac{d}{d\theta} a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x, \theta)\, dx$$

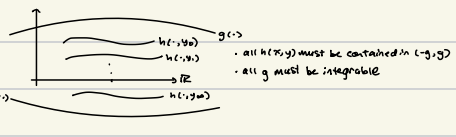
Corollary: If $a(\theta)=a$, $b(\theta)=b$, then

$$\frac{d}{d\theta}\int_a^b f(x,\theta)\,dx = \int_a^b \frac{\partial f}{\partial \theta}(x,\theta)\,dx$$

- To strengthen the above result(s) for infinite domain of the integral, we need a new tool.

Thm. Suppose
- $h(x,y)$ is continuous at $y=y_0$ for each fixed x
- there is a function $g(x)$ s.t.

envelope condition:
$$\begin{cases} |h(x,y)| \le g(x) \ \forall\ x,y \\ \int_{-\infty}^{\infty} g(x)\,dx < \infty \end{cases}$$



- all $h(x,y)$ must be contained in $(-g,g)$
- all g must be integrable

Then,
$$\lim_{y\to y_0}\int_{-\infty}^{\infty} h(x,y)\,dx = \int_{-\infty}^{\infty} \lim_{y\to y_0} h(x,y)\,dx$$

Remarks: - g is called the dominating or envelope function of f
- known as (a version/consequence of) Lebesgue's Dominated Convergence Thm (DCT)
- think of h as a collection of functions in x $\{h(x,y)\}_{y\in\mathbb{R}}$

Thm: Suppose
- For each x, $f(x,\theta)$ is differentiable wrt θ, at $\theta\in(\Theta)$
- For each $\theta\in(\Theta)$, there is a function $g(\cdot,\theta)$ and $\delta_\theta>0$ s.t.

$$\begin{cases} \left|\frac{\partial}{\partial\theta}f(x,\theta)\right|_{\theta=\theta'} \le g(x,\theta')\ \forall\ \theta'\in(\theta-\delta_\theta,\theta+\delta_\theta) \\ \int_{-\infty}^{\infty} g(x,\theta)\,dx < \infty \text{ for each } \theta\in(\Theta) \end{cases}$$

Then
$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)\,dx = \int_{-\infty}^{\infty}\frac{d}{d\theta}f(x,\theta)\,dx$$
holds for each $\theta\in(\Theta)$

Remarks: - Leibniz's rule holds under this weaker envelope condition for $f(x,\theta)$.
- If we need only to differentiate at $\theta=\theta_0$, then it is sufficient that the envelope condition be satisfied for a neighborhood of θ_0 i.e. $\forall\ \theta\in(\theta_0-\delta_0,\theta+\delta_0)$ for some $\delta_0>0$, and only a dominating function $g(x,\theta_0)$ (for fixed $\theta=\theta_0$) is required
- This thm is a direct consequence of Lebesgue's Dominated Convergence Thm (DCT)

Property: Let $X\sim\text{Exp}(\lambda)$; $f(x)=\frac{1}{\lambda}e^{-x/\lambda}$, $\lambda\in[0,\infty)$.

Moments: $E(X^n)=\int_0^\infty \frac{1}{\lambda}x^n e^{-x/\lambda}dx$, $n=1,2,\dots$

This gives a recursive relation, i.e.

$$E(X^{n+1}) = \lambda E(X^n) + \lambda^2\frac{d}{d\lambda}E(X^n)$$; (similar identity holds for broad family of dist.'s)

Def (Pointwise convergence): $\lim_{n\to\infty}\underbrace{\sum_{x=0}^{n} h(x,\theta)}_{S_n(\theta)} =: \underbrace{\sum_{x=0}^{\infty} h(x,\theta)}_{S(\theta)}$

Recall: Let $A\subset\mathbb{R}$. Then

$\sup A$ = Smallest $s\in\mathbb{R}$ s.t. $s\ge a\ \forall\ a\in A$, e.g.s.

: if $A=[u,v]$, then $\sup A=v$

: if $A=(u,v)$, then $\sup A=v$.

Def (Uniform convergence): $\sup_{\theta\in(\Theta)} |S_n(\theta)-S(\theta)| \to 0$ as $n\to\infty$

Suppose $\sum_{x=0}^{\infty} h(x,\theta)$ exists (i.e. converges pointwise) $\forall\ \theta\in(a,b)$

Moreover, assume

i) $\frac{\partial}{\partial\theta}h(x,\theta)$ is continuous in θ for each x

2) $\sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(x,\theta)$ converges uniformly for all θ in a closed subinterval of (a,b)

Then
$$\frac{d}{d\theta}\sum_{x=0}^{\infty}h(x,\theta) = \sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(x,\theta)$$

Remark: this is a consequence of Lebesgue's DCT

Thm. Suppose $\sum_{x=0}^{\infty} h(x,\theta)$ exists (i.e. converges pointwise) $\forall\ \theta\in[a,b]$

Moreover, assume

i) $h(x,\theta)$ is continuous in θ for each fixed x

2) $\sum_{x=0}^{\infty}h(x,\theta)$ converges uniformly on $[a,b]$

Then,
$$\int_a^b \sum_{x=0}^{\infty} h(x,\theta)\,d\theta = \sum_{x=0}^{\infty}\int_a^b h(x,\theta)\,d\theta$$

3.1: Discrete Distributions

Recall: RV X is discrete if the range of X (i.e. $\mathcal{X}$) is countable

Def (Uniform RV): $X \sim \text{Uniform}(1, N)$ if

$$P(X = x) = \frac{1}{N}, \quad x = 1, \dots, N.$$

we sometimes write

$$P(X = x \mid N) = \frac{1}{N}$$

Remarks: $E(X) = \frac{N+1}{2}$, $V(X) = \frac{(N+1)(N-1)}{12} = \frac{(N-1)^2}{12}$

Def (Hypergeometric Dist.):

Experiment: N balls, M red, $N-M$ green

Pick K balls uniformly at random (w/o replacement)

Let X: number of reds.

Then,

$$P(X = x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \quad x = 0, 1, \dots, K$$

Remarks:

- counting argument implies $\sum_{x=0}^{K} \binom{M}{x}\binom{N-M}{K-x} = \binom{N}{K}$
- $E(X) = \frac{KM}{N}$, $V(X) = \frac{KM}{N} \frac{(N-M)(N-K)}{N(N-1)}$

Def (Bernoulli Dist.): Let $X \sim \text{Bernoulli}(p)$, $0 \le p \le 1$

$$p(x) := \begin{cases} P(X=1 \mid p) = p \\ P(X=0 \mid p) = 1-p \end{cases}$$

Remarks: $E(X) = p$, $V(X) = p(1-p)$

Def (Binomial Dist.): Perform n-independent-and-identically-distributed (iid) Bernoulli trials.

Let Y: number of successes (i.e. 1's).

Then $Y \sim \text{Binomial}(n, p)$ and

$$P(Y = y \mid n, p) = \binom{n}{y} p^y (1-p)^{n-y}$$

Remarks: $E(Y) = np$, $V(Y) = np(1-p)$, $M_Y(t) = (pe^t + 1 - p)^n$

Def (Poisson Dist.): Let $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$.

$$P(X = x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, \dots$$

Remarks:

- $e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^x}{x!} + \cdots$
- arises in modelling the following (count) experiments:
  - number of buses arriving with a period of time
  - number of stars observed in a skymap
  - number of incidents found in a segment of a network
- $E(X) = \lambda = V(X)$, $M_X(t) = \exp(\lambda(e^t - 1))$.

Prop (Poisson Approx.) (of the Binomial Dist.):

If $X_n \sim \text{Binomial}(n, p_n)$ and $np_n \to \lambda$ as $n \to \infty$

then $X_n \xrightarrow{d} Y$ where $Y \sim \text{Poisson}(\lambda)$.

Def (Negative Binomial Dist.):

Experiment: count the number of independent Bernoulli(p) trials until obtaining the r-th success

$X \sim \text{NegBinom}(r, p)$, $0 \le p \le 1$, $r \in \mathbb{N}$.

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \dots$$

Corollary: Let $Y = X - r$ represent the number of failures that occur before the r-th success in a sequence of independent Bernoulli(p) trials

Then,

$$P(Y = y \mid p, r) = P(X = y+r \mid p, r) = \binom{y+r-1}{r-1} p^r (1-p)^y, \quad y = 0, 1, \dots$$

$$\overset{(*)}{=} (-1)^y \binom{-r}{y} p^r (1-p)^y$$

Remark: $(*)$ motivates the "negative binomial" label

- Here, $n = y+r-1 \Rightarrow y = n-r+1$
- $E(Y) = \frac{r(1-p)}{p}$, $\text{var}(Y) = \frac{r(1-p)}{p^2}$

Prop: Let $\mu = E(Y) = \frac{r(1-p)}{p}$. Then

$$Var(Y) = \frac{1}{r}\mu^2 + \mu$$

Remark: colloquially the 'quadratic relation'

Prop (Poisson Approx.): If $r \to \infty$, $p \to 1$ such that $r(1-p) \to \lambda$,

then $\begin{cases} E(Y) \to \lambda \\ V(Y) \to \lambda \end{cases}$

Moreover, $Y \xrightarrow{d} \text{Poisson}(\lambda)$

Def (Geometric Dist.): Let $X \sim \text{Geometric}(p)$, $p \in (0,1)$.

$$P(X = x|p) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

Remarks: this is a special case of $\text{NegBinom}(p, r=1)$;

thus $\begin{cases} E(X) = (1-p)/p \\ V(X) = (1-p)/p^2 \end{cases}$

$\Rightarrow X$: # trials until the first success

Prop ('Memoryless Property'): If $s > t$, then

$$P(X > s \mid X > t) = P(X > s - t)$$

Remarks: Given $X > t$ (no success in the first t trials), the dist. of additional waiting time until the first success does not depend on t, i.e.

past failures do not change the tail prob's for the future

3-2a: Continuous Distributions

- continuous distributions put prob. mass on cont. spaces
- Formally, a continuous RV is one whose cdf is a continuous function

Def (Uniform Dist.): Let $X \sim \text{Uniform}([a,b])$, $a < b$.

$$f(x|a,b) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

Remarks: $E(X) = \frac{b+a}{2}$, $V(X) = \frac{(b-a)^2}{12}$

- uniform RV can either be continuous or discrete
  - depends on continuous range or a finite, discrete set of values

Def (Gamma Dist.): Let $X \sim \text{Gamma}(\alpha, \beta)$ for $\alpha > 0$: shape, $\beta > 0$: scale

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

corollary (Gamma Function):

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt, \quad \alpha > 0$$

$$\Gamma(\alpha)\beta^\alpha = \int_0^\infty x^{\alpha-1} e^{-x/\beta}\, dx$$

$$\Gamma(\alpha+1) = \Gamma(\alpha) \cdot \alpha$$

$$\Gamma(n) = (n-1)\Gamma(n-1) = \cdots = (n-1)!$$

Remark: sometimes we use $\beta = 1/\lambda$ instead as scale parameter

$$E(X) = \alpha\beta, \; V(X) = \alpha\beta^2$$

$$\frac{X}{\beta} \sim \text{Gamma}(\alpha, 1)$$

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, \quad t < \frac{1}{\beta}$$

Prop: $\forall\, x > 0$, $P(\lambda \le x) = P(Y \ge \alpha)$ for $\begin{cases} X \sim \text{Gamma}(\alpha, \beta) \\ Y \sim \text{Poisson}(x|\beta) \end{cases}$

$$\Leftrightarrow \int_0^x \text{pdf}_{\text{Gam}}(t)\, dt = \sum_{y=\alpha}^{\infty} \text{pmf}_{\text{Poi}}(y)$$

Prop (Special cases):

a. Let $X \sim \text{Gamma}(\alpha, \beta)$: $\begin{cases} \alpha = p/2, \; p\text{-integer} \\ \beta = 2 \end{cases}$

Then $f(x|p) = \frac{1}{\Gamma(p/2)\, 2^{p/2}} x^{p/2-1} e^{-x/2}, \; x > 0 \sim \chi^2_{df=p}$

Remark: we will later see that, for $Z_i \sim N(0,1)$:

$$\sum_{i=1}^{p} Z_i^2 \sim \chi^2_p$$

b. Let $X \sim \text{Gamma}(\alpha,\beta)$ $\begin{cases} \alpha=1 \\ \beta>0 \end{cases}$ and note $\Gamma(1)=1$

Then $f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}$, $x>0$, i.e. the exponential pdf.

Note that the exponential dist. (like the geometric dist.) has the memoryless property, i.e.

$P(X>s \mid X>t) = P(X>s-t)$ for $s>t\geq 0$

c. If $X \sim \text{Exp}(\beta)$, $\beta>0$, letting $Y = X^{1/\gamma}$, $\gamma>0$, then

$F_Y(y) = P(Y \leq y) = P(X \leq y^\gamma) = 1 - e^{-y^\gamma/\beta}$

where $\frac{d}{dy} e^{-y^\gamma/\beta} = e^{-y^\gamma/\beta} \cdot \frac{d}{dy} e^{-y^\gamma/\beta}$

$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{\gamma}{\beta} y^{\gamma-1} e^{-y^\gamma/\beta}$, $0<y<\infty$

where $Y \sim \text{Weibull}(\gamma,\beta)$

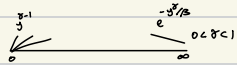
Recall: Let $X \sim \text{Exp}(\beta)$, $f(x) = \frac{1}{\beta} e^{-x/\beta}$, $x \geq 0$

$F(x) = \int_0^x \frac{1}{\beta} e^{-x/\beta} dx = \frac{1}{\beta}\left[-\beta e^{-x/\beta}\right]_0^x$

$= -e^{-x/\beta}\Big|_0^x$

$= 1 - e^{-x/\beta}$, $x>0$.

Remark: Weibull dist. useful for handling extreme/rare events



3.2b: Continuous Dist. (cont.)

Def (Normal Dist.): $X \sim \text{Normal}(\mu,\sigma^2)$, $\mu \in \mathbb{R}$, $\sigma>0$

$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$, $x \in \mathbb{R}$

Remark: Parameters μ: mean, σ^2: variance

Prop (identities):

$\frac{1}{\sqrt{2\pi\sigma^2}} \int e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1 = \int f(x) dx$ (1)

$\frac{1}{\sqrt{2\pi\sigma^2}} \int x\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \mu = \int x f(x) dx$ (2)

$\frac{1}{\sqrt{2\pi\sigma^2}} \int x^2 e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \mu^2 + \sigma^2 = E^2(x) + V(x)$ (3) ← why? reformulating $V(x) = E(x^2) - E^2(x)$

$\forall\, \mu \in \mathbb{R}$, $\sigma>0$.

Def (standardization): = subtracting the mean and rescaling by the STD

- does not always keep the RV to remain in the family of distributions)

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

Standard normal RV: $Z := \frac{X-\mu}{\sigma} \sim N(0,1)$

$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z^2)$, $z \in \mathbb{R}$

$F_Z(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$

Remark: $\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi} \Rightarrow \frac{1}{\sqrt{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{\pi}$

$\Leftrightarrow \int_0^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{\frac{\pi}{2}} \Rightarrow \sqrt{2} \int_0^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{\pi}$, i.e. a definition for π

- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (*)

Prop: If $X \sim N(0,1)$,

then $Y = X^2 \sim \chi^2_1$ s.t.

$f(y) = \frac{1}{\sqrt{2\pi y}} e^{-y^2/2}$

$= \frac{1}{\Gamma(1/2) 2^{1/2}} y^{-1/2} e^{-y^2/2}$ by (*)

i.e. a pdf for Gamma($\alpha=\frac{1}{2}$, $\beta=2$)

Remark: More generally, Let $X \sim \chi^2_p$. Then $f(x) = \frac{1}{\Gamma(p/2)\, 2^{p/2}} x^{p/2-1} e^{-x/2}$, $x>0$, $p \in \mathbb{N}$

where $\chi^2_p \overset{d}{=} \text{Gamma}(p/2, 2)$

Prop (Normal Approx.): Let $X \sim \text{Binomial}(n,p)$ s.t. $\begin{cases} E(X) = np \\ V(X) = np(1-p) \end{cases}$

Let $\begin{cases} np \to \infty \\ np(1-p) \to \infty \end{cases}$. Then $X \approx \text{Normal}(np, np(1-p))$

Remark: We will later show via CLT that

$\frac{1}{\sqrt{n}}(X-np) \xrightarrow{d} \text{Normal}(0, p_\infty(1-p_\infty))$ if $\begin{cases} n \to \infty \\ p \to p_\infty \end{cases}$

Corollary (continuity correction): Let $X \sim Bin(n,p)$ and $Y \sim N(np, np(1-p))$.

Then,

$$P(X=k) \approx P(k-0.5 < Y < k+0.5)$$
$$P(X \le k) \approx P(Y \le k+0.5) \approx P(X < k)$$
$$P(X \ge k) \approx P(Y \ge k-0.5) \approx P(X > k)$$

Remark:
- cont. dist., e.g. Normal, assign points zero prob; must look at interval around point to capture any (approximate) mass
- w/o correction, we tend to underestimate the tails

3.2 c: Continuous Dist. (cont.)

Def (Beta Dist.): Let $X \sim Beta(\alpha, \beta)$, $\alpha, \beta > 0$

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } x \in (0,1)$$

n-th moments:
- $E(X^n) = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+n)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n)}$
- $E(X) = \frac{\alpha}{\alpha+\beta}$
- $V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Remarks: **Beta Function:** $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}$

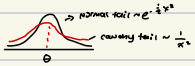
Fact: $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \Rightarrow Beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

Beta pdfs for different (α,β):
- $\alpha=\beta=1 \Rightarrow Beta(1,1) \equiv Uniform(0,1)$
- $\alpha>1, \beta>1 \Rightarrow$ Beta is unimodal
- $\alpha<1, \beta<1 \Rightarrow$ Beta is
- $\alpha<\beta$ or
- $\alpha>\beta$ or

Def (Cauchy Dist.): Let $X \sim Cauchy(\theta)$, $\theta \in \mathbb{R}$.

Then $f(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$, $x \in \mathbb{R}$

Remarks:
- cauchy pdf also bell-shaped:



- θ is the median (location) of X

Recall: $\begin{cases} E|X| = \infty \\ E(X) \text{ DNE} \end{cases}$

Prop. Let $X, Y \overset{iid}{\sim} N(0,1)$. Then $\frac{X}{Y} \sim Cauchy(0)$ Proof (see later).

Def (Log-Normal Dist.): If $Y \sim N(\mu, \sigma^2)$, then $X = e^Y \sim LogNormal$, i.e.

$Y = \log(X) \sim N(\mu, \sigma^2)$ s.t.

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

3.3: Exponential Families

Def: A family of pdf or pmf is called an **exponential family** if it has the form

$$f(x|\theta) = c(\theta) h(x) \exp\left\{\sum_{i=1}^{k} w_i(\theta) t_i(x)\right\}, \quad x \in \mathcal{X}$$
$$= c(\theta) h(x) \exp\langle w(\theta), t(x)\rangle \quad (*)$$

Remarks: where $w(\theta) = (w_1(\theta), \dots, w_k(\theta))$
$t(x) = (t_1(x), \dots, t_k(x))$
↳ vector of **sufficient statistics** (coming later)

Here,
- $h(x) \ge 0$
- $t(x) = (t_1(x), \dots, t_k(x))$ depends only on x (data)
- $w(\theta) = (w_1(\theta), \dots, w_k(\theta))$ depends only on θ (parameters)

clearly, $c(\theta)$ is the reciprocal of the normalizing constant, i.e.

$$\int f(x|\theta) = 1 \Rightarrow \frac{1}{c(\theta)} = \int_{x \in \mathcal{X}} h(x) e^{\langle w(\theta), t(x)\rangle} dx \quad \text{(or sum if discrete)}$$

Ex (Binomial): Let $X \sim \text{Bin}(n,p)$; $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0,1,\dots,n$, $\theta = p$

$= \cdots$

$= \underbrace{(1-p)^n}_{c(\theta)} \underbrace{\binom{n}{x}}_{h(x)} \exp\left\{ \sum_{i=1}^{1} \underbrace{\log\frac{p}{1-p}}_{w_i(\theta)} \cdot x \right\}$

Ex (Normal): Let $X \sim N(\mu, \sigma^2)$; $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right)$, $x \in \mathbb{R}$, $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$

$= \cdots$

$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\mu^2/2\sigma^2} \exp\left\{ \frac{-1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x \right\}$.

Remarks:
- other distributions in the exponential family: Poisson, Geometric, NegBinom, Gamma, Exponential, Log-normal
- Dist's not in exponential families: cauchy, Mixtures

Thm: If X is a RV w/ pdf/pmf in the exponential form $c(\theta)$, then

1) $E\left\{ \sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X) \right\} = \frac{\partial}{\partial \theta_j} \log(1/c(\theta))$, $j = 1,\dots,k$

2) $\text{var}\left(\left\{ \sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X) \right\} \right) = \frac{\partial^2}{\partial \theta_j^2} \log(1/c(\theta)) - E\left\{ \sum_{i=1}^{k} \frac{\partial^2 w_i(\theta)}{\partial \theta_j^2} t_i(X) \right\}$ ← check

Remark (high-level): Differentiating the logarithm of the normalizing constant w.r.t. a parameter results in suitable expectations of sufficient statistics

- Differentiating w.r.t. the natural parameter corresponding to the sufficient statistic of interest
- when we differentiate w.r.t. a natural parameter $w_i(\theta)$, we're targeting the expectation of $t_i(X)$, i.e. expectation of a sufficient statistic

Corollary (Simplicity): Suppose

$w(\theta) = \theta = (\theta_1, \theta_2, \dots, \theta_k) \in \mathbb{R}^k$

$t(x) = (t_1(x), t_2(x), \dots, t_k(x)) \in \mathbb{R}^k$

then $f(x|\theta) = c(\theta) h(x) \exp\langle \theta, t(x) \rangle$ and

$\mu_j := E(t_j(X)) = \frac{\partial}{\partial \theta_j} \log(1/c(\theta))$

$\sigma_j^2 := \text{Var}(t_j(X)) = \frac{\partial^2}{\partial \theta_j^2} \log(1/c(\theta))$

3.4: Location and Scale Families

Fact: If $f(x)$ is a valid pdf on $\mathbb{R}$, then $\forall \mu \in \mathbb{R}$, $\sigma > 0$, the function

$g(x|\mu,\sigma) := \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma} \right)$, $x \in \mathbb{R}$

is a valid pdf on $\mathbb{R}$

Def: Let $f(x)$ be any pdf (on $\mathbb{R}$). The family of pdf $\left\{ g(x|\mu,\sigma) := \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma} \right) \,\middle|\, \mu \in \mathbb{R}, \sigma > 0 \right\}$ is called a location-scale family of distributions for $\begin{cases} \mu: \text{location param.} \\ \sigma: \text{scale param.} \end{cases}$

Ex: Let $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x^2 \right)$ s.t. $X \sim N(0,1)$.

Then, we obtain the L-S family, i.e.

$\left\{ g(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{1}{\sigma}(x-\mu) \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right) \,\middle|\, \mu \in \mathbb{R}, \sigma > 0 \right\}$.

Special cases: 1) $\left\{ g(x|\mu) := f(x-\mu) \,|\, \mu \in \mathbb{R} \right\}$ is location family

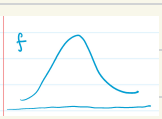
2) $\left\{ g(x|\sigma) := \frac{1}{\sigma} g\left(\frac{x}{\sigma} \right) \,|\, \sigma > 0 \right\}$ is scale family

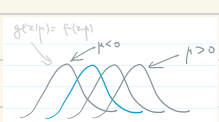
Prop: (why not $f(x+\mu)$?)

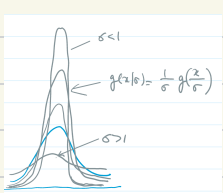
If $X \sim f$, then $X + \mu \sim g(x) = f(x-\mu)$

If $X \sim f$, then $\sigma X \sim g(x) = \frac{1}{\sigma} f\left(\frac{x}{\sigma} \right)$

If $X \sim f$, then $\sigma X + \mu \sim g(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma} \right)$







Thm: Suppose Y is a RV w/ pdf $f(y)$ and $E(Y), V(Y)$ exist.

If X is a RV w/ pdf $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma} \right)$, then $\begin{cases} E(X) = \sigma E(Y) + \mu \\ V(X) = \sigma^2 V(Y) \end{cases}$

3.5: Inequalities and Identities

- when we can't calculate probs, it is important to estimate them by inequalities to obtain bounds

Thm (Chebychev's Inequality): Let X be a RV, $g(x) \geq 0 \ \forall x$

Then, $\boxed{P(g(X) \geq r) \leq \frac{E(g(X))}{r} \quad \forall r > 0}$

$\Rightarrow \boxed{Eg(X) \geq r \cdot P(g(X) \geq r)}$

Corollary: Let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$ where $\begin{cases} \mu = E(X) \\ \sigma^2 = V(X) \end{cases}$

Then, $\underline{P(|X-\mu| \geq \sigma\sqrt{r}) \leq \frac{1}{r}}$

Alternatively, letting $r = t^2$:

$\underline{P(|X-\mu| \geq t\sigma) \leq \frac{1}{t^2}}$ and $\underline{P(|X-\mu| < t\sigma) \geq 1 - \frac{1}{t^2}}$

Remarks: $t=2$: $P(|X-\mu| \geq 2\sigma) \leq 1/4 = 25\%$

$t=3$: $P(|X-\mu| \geq 3\sigma) \leq 1/9 = 11.1\%$

$t=4$: $P(|X-\mu| \geq 4\sigma) \leq 1/16 = 6.25\%$

Thm. (tighter inequality for normal tails)

If $Z \sim N(0,1)$, then for $t > 0$:

$$\boxed{P(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}}$$

Remark: If $t=2$, $P(|Z| \geq 2) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-2}}{2} = 0.054 < 1/4$ (from above thm/corr)

Corollary: In general, $\sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t} << \frac{1}{t^2}$

(exponentially small; polynomially true)

Identities

Poisson (identity): If $X \sim \text{Poisson}(\lambda)$, then $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$

So, $\begin{cases} P(X=0) = e^{-\lambda} \\ P(X=x+1) = e^{-\lambda}\frac{\lambda^{x+1}}{(x+1)!} \\ \qquad\qquad = P(X=x)\frac{\lambda}{x+1} \end{cases}$

note: Recursion-like identities (like this one) may be useful in various situations that require such computations

Gamma (identity): If $X_{\alpha,\beta} \sim \text{Gamma}(\alpha,\beta)$ with $\begin{cases} f(x|\alpha,\beta) \\ \alpha > 1, \ \beta > 0 \end{cases}$

Then $\forall\ a, b$

$$\boxed{P(X_{\alpha,\beta} \in (a,b)) = \beta(F(b|\alpha,\beta) - F(b|\alpha,\beta)) + P(X_{\alpha-1,\beta} \in (a,b))}$$

Remark: If $\alpha \in \mathbb{N}$, the above identity allows us to recurse to Gamma$(\alpha-1,\beta)$; Gamma$(\alpha-2,\beta)$, ... , and so on to Gamma$(1,\beta) = \text{Exp}(\beta)$.

Stein's Identity for Normal RV: If $X \sim N(\theta, \sigma^2)$,

g is a differentiable function s.t. $E|g'(x)| < \infty$

Then

$$\boxed{E[g(x)(x-\theta)] = \sigma^2 E(g'(x))}$$

Remark: see applications in Practice Problems

Hwang's Identity (for discrete variables)

Let g: function with $|Eg(x)| < \infty$ and $|g(-1)| < \infty$.

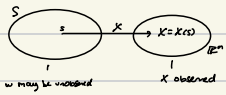
If $X \sim \text{Poisson}(\lambda)$, then $\underline{E(\lambda g(x)) = E(X g(x-1))}$

If $X \sim \text{Negbinom}(r,p)$, then $\underline{E((1-p)g(x)) = E\left[\frac{x}{r+x-1} g(x-1)\right]}$

Ch. 4: Multiple Random Variables

4.1: Joint and marginal distributions

Def: An n-dimensional random vector is a function from a sample space into space $\mathbb{R}^n$ (may change to other spaces)



note: If $n=2$, $X = (X_1, X_2) \in \mathbb{R}^2$ is called a bivariate vector

Def: Let (X,Y) be a discrete bivariate vector. Then the function from $\mathbb{R}^2 \to \mathbb{R}$:

$$f(x,y) := P(X=x, Y=y)$$

is the joint prob. mass function (pmf) on (X,Y)

Remark: Denote by $f_{XY}(x,y)$ or $f_{X,Y}(x,y)$

where $f_{XY}(x,y)$ not necessarily $f_{YX}(x,y)$

corollary: For $A \subset \mathbb{R}^2$, by AOP, we have that $P((X,Y) \in A) = \sum_{(x,y) \in A} f(x,y)$

Prop: Let $g(x,y)$ be a function from $\mathbb{R}^2 \to \mathbb{R}$. Then $g(X,Y)$ is a real-valued RV:

$$Eg(X,Y) := \sum_{(x,y) \in \mathbb{R}^2} g(x,y) f_{XY}(x,y)$$

Proposition (linearity of expectation): If g_1, g_2 are real-valued functions on $\mathbb{R}^2$; $a, b \in \mathbb{R}$. Then

$$E[a g_1(X,Y) + b g_2(X,Y)] = a E g_1(X,Y) + b E g_2(X,Y)$$

Thm: Let (X,Y) be a discrete bivariate RV w/ joint pmf $f_{XY}(x,y)$.

Then, X and Y are discrete RV's w/ the following pmf's:

marginal pmf's:

$$\begin{cases} f_X(x) = \sum_y f_{XY}(x,y) \\ f_Y(y) = \sum_x f_{XY}(x,y) \end{cases}$$

Remarks: · f_X and f_Y are called the marginal pmf's of X and Y

· the distributions of X and Y are the marginal distributions (w.r.t. dist. of (X,Y))

· the dist. of random vector (X,Y) is the joint dist. of X and Y

corollary: Joint dist/pmf completely determines its marginal dist/pmf

But marginal dist/pmf's do not determine the joint dist.

- Continuous Bivariate RVs are described via joint prob. density functions

Def: a function $f(x,y)$ from $\mathbb{R}^2 \to \mathbb{R}$ is called a joint pdf of the continuous bivariate random vector (X,Y) if, for every $A \subseteq \mathbb{R}^2$

$$P((X,Y) \in A) = \iint_{(x,y) \in A} f_{XY}(x,y)\,dx\,dy$$

Remarks: · $f(x,y) \geq 0 \ \forall (x,y) \in \mathbb{R}^2$

· $\iint_{\mathbb{R}^2} f(x,y)\,dx\,dy = 1$

Def: If $g(x,y)$ is a real-valued function, then $g(X,Y)$ is a random variable w/ the expectation

$$E[g(X,Y)] := \iint_{\mathbb{R}^2} g(x,y) f_{XY}(x,y)\,dx\,dy$$

marginal pdfs:

$$f_X(x) = \int_{\mathbb{R}} f_{XY}(x,y)\,dy$$

$$f_Y(y) = \int_{\mathbb{R}} f_{XY}(x,y)\,dx$$

Joint cdf := $F_{XY}(x,y) := P(X \le x, Y \le y)$

$$= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(s,t)\,dt\,ds$$

Corollary: $F_{XYZ}(x,y,z) := P(X \le x, Y \le y, Z \le z)$

$$= \int_{-\infty}^{x}\int_{-\infty}^{y}\int_{-\infty}^{z} f_{XYZ}(u,v,w)\,du\,dv\,dw$$

Remarks:
- $0 \le F_{XY}(x,y) \le 1 \quad \forall\ (x,y) \in \mathbb{R}^2$
- $F_{XY}(x,\cdot)\uparrow \quad \forall\ x \in \mathbb{R}$
- $F_{XY}(\cdot,y)\uparrow \quad \forall\ y \in \mathbb{R}$
- Fundamental Thm. of calculus (bivariate case) gives

$\frac{d^2 F_{XY}(x,y)}{dx\,dy} = f_{XY}(x,y)$ at continuous points (x,y) of function f_{XY}

4.2a: conditional distributions

- Given a bivariate RV $(X,Y) \in \mathbb{R}^2$, we are interested in, say

$P(Y \in B \mid X \in A)$ for $A, B \subset \mathbb{R}$.

Prop. Let (X,Y) be discrete.

$$P(Y \in B \mid X \in A) = \frac{P(X \in A \cap Y \in B)}{P(X \in A)}$$

where $P(X \in A \cap Y \in B) = \sum_{x \in A}\sum_{y \in B} f_{XY}(x,y)$

and $P(X \in A) = \sum_{x \in A}\sum_{y} f_{XY}(x,y) = \sum_{x \in A} f_X(x)$

so, $P(Y \in B \mid X \in A) = \sum_{x \in A}\sum_{y \in B} f_{XY}(x,y) \,/\, \sum_{x \in A} f_X(x)$

$$P(Y \in B \mid X \in A) = \sum_{y \in B} \frac{\sum_{x \in A} f_{XY}(x,y)}{\sum_{x \in A} f_X(x)}$$

Corollary: Let $A = \{x\}$, $B = \{y\}$. Then $P(Y=y \mid X=x) = \frac{f_{XY}(x,y)}{f_X(x)}$

Def: Let (X,Y) be a discrete random vector w/ joint pmf $f_{XY}(x,y)$. Then, for any x s.t. $f_X(x) = P(X=x) > 0$, the **conditional pmf** is:

$$f(y|x) = P(Y=y \mid X=x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\sum_y f_{XY}(x,y)}$$

Remarks:
- $f(y|x) \ge 0 \quad \forall\ y$
- $\sum_y f(y|x) = \sum_y \frac{f_{XY}(x,y)}{\sum_y f_{XY}(x,y)} = \frac{\sum_y f_{XY}(x,y)}{\sum_y f_{XY}(x,y)} = 1.$
- Denote $f_{Y|X}(y|x)$

Note: By definition, the marginal pmf (f_X, f_Y) and the conditional pmf $(f_{Y|X}, f_{X|Y})$ are completely determined by the joint pmf (f_{XY})

Corollary: $f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$

$$\Rightarrow f_{XY}(x,y) = f_X(x)\, f_{Y|X}(y|x)$$

similarly,

$$f_{XY}(x,y) = f_Y(y)\, f_{X|Y}(x|y)$$

wherever $f_X(x) > 0$ and $f_Y(y) > 0$

Remark: the marginal pmf and conditional pmf are combined to completely determine the joint pmf

Def (conditional pdf): Consider the continuous bivariate random vector $(X,Y) \in \mathbb{R}^2$. Then

$$P(Y \in B \mid X = x) := \lim_{\varepsilon \to 0} P(Y \in B \mid X \in (x-\varepsilon, x+\varepsilon)) = \int_B \frac{f_{XY}(x,y)}{f_X(x)}\,dy$$

assuming $f_{XY}(x,y)$ continuous at (x,y)

Def (conditional pdf): Let (X,Y) be a bivariate random vector w/ joint pdf $f_{XY}(x,y)$. Then, for any x s.t. $f_X(x) > 0$,

$$f_{Y|X}(y|x) := \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int f_{XY}(x,y)\,dy}$$

Remarks:
- $f_{Y|X}(x,y) \ge 0 \quad \forall\ y, x \quad (f_X(x) > 0)$
- $\int f_{Y|X}(y|x)\,dy = 1$
- Identities hold: $f_{XY}(x,y) = \begin{cases} f_X(x) f_{Y|X}(y|x) & \text{when } f_X(x) > 0 \\ f_Y(y) f_{X|Y}(x|y) & \text{when } f_Y(y) > 0 \end{cases}$

- By varying x, we actually have a collection of distributions for RV Y: $\{f_{Y|X}(y|x)\}_{x \in \mathbb{R}}$

Def (Cond. Expectation): Given $(X,Y) \sim f_{X,Y}$, $g: \mathbb{R} \to \mathbb{R}$ a function

Then, $\boxed{E[g(Y)|X=x] := E[g(Y)|X=x] := \int g(y) f_{Y|X}(y|x)\,dy}$

Remarks: two equivalent ways to define a joint dist. for a bivariate random vector (X,Y):

a) joint pdf: $f_{X,Y}(x,y)\ \forall\ x,y$

b) $f_{X,Y}(x,y) := \begin{cases} f_X(x) f_{Y|X}(y|x), & f_X(x) > 0 \\ 0, & \text{otherwise} \end{cases}$

where $f_X(x) = \int_y f_{X,Y}(x,y)\,dy$

note: (b) often easier in applications

4.3.6: Independence

Def: Let (X,Y) be a bivariate random vector, with joint pmf/pdf $f_{X,Y}(x,y)$ and marginal pdf/pmf $f_X(x)$ and $f_Y(y)$.

Then X and Y are independent RVs if, $\forall\ x,y \in \mathbb{R}$

$$\boxed{f_{X,Y}(x,y) = f_X(x) f_Y(y)}$$

remarks: 1) If $X \perp\!\!\!\perp Y$ then $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y)$

(does not depend on x)

Moreover, $\forall\ A, B \subset \mathbb{R}$,

$$P(Y \in B \mid X \in A) = \frac{\int_B \int_A f_{X,Y}(x,y)\,dx\,dy}{\int_A f_X(x)\,dx} = \frac{\int_B \int_A f_X(x) f_Y(y)\,dx\,dy}{\int_A f_X(x)\,dx} = \frac{\left(\int_B f_Y(y)\,dy\right)\left(\int_A f_X(x)\,dx\right)}{\int_A f_X(x)\,dx} = \int_B f_Y(y)\,dy = P(Y \in B)$$

Hence the event $\{Y \in B\}$ is independent of $\{X \in A\}\ \forall\ A, B$

Remark: to verify independence, need to check the above identity for all $x,y \in \mathbb{R}$ or all A,B

to show non-independence, need to identify a pair (x,y) or (A,B) where identity not satisfied.

Lemma: Let $(X,Y) \sim f_{X,Y}$. Then $X \perp\!\!\!\perp Y$ i.f.f. $\exists$ functions $g(x)$ and $h(y)$ s.t.

$$f_{X,Y}(x,y) = g(x) h(y)\ \forall\ x,y \in \mathbb{R}$$

Thm. If $X \perp\!\!\!\perp Y$, then

1) $\forall\ A \subset \mathbb{R}, B \subset \mathbb{R}$

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$$

2) $\forall$ funcs $g(x)$ - only of x

$h(y)$ - only of y

$$E[g(X)h(Y)] = E g(X)\, E h(Y)$$

Thm: If $X \perp\!\!\!\perp Y$, w/ mgf M_X and M_Y, then

RV $Z := X+Y$ has the mgf

$$M_Z(t) = M_X(t) \cdot M_Y(t)\ \forall\ t$$

P. $M_Z(t) = E e^{tZ} = E e^{t(X+Y)}$

$= E e^{tX} e^{tY}$

$= E e^{tX} E e^{tY}$, $X \perp\!\!\!\perp Y$

$= M_X(t) M_Y(t)$. ∎

Thm. The sum of two independent normal RV's is again normal

4.3: Bivariate Transformation

Let (X,Y) be a bivariate random vector

$g: \mathbb{R}^2 \to \mathbb{R}^2$

$g(x,y) := (g_1(x,y), g_2(x,y)) \in \mathbb{R}^2$

Then $(U,V) := g(X,Y)$ is a bivariate random vector

$\forall\ A \in \mathbb{R}^2$, $P((U,V) \in A) = P((X,Y) \in g^{-1}(A))$

where $g^{-1}(A) = \{(x,y) \mid g(x,y) \in A\}$

Discrete case: if (X,Y) is discrete, then so is (U,V)

$$\boxed{f_{U,V}(u,v) = \sum_{\substack{x,y: \\ g_1(x,y)=u \\ g_2(x,y)=v}} f_{X,Y}(x,y)}$$

Continuous case: change-of-var formula.

Let (X,Y) - continuous bivariate vector
$(X,Y) \sim f_{XY}$

Let $g: \mathbb{R}^2 \to \mathbb{R}^2$ be a one-to-one mapping
i.e. $g^{-1}(\{u,v\})$ has at most one element

Define $(U,V) = g(X,Y)$; what is $f_{UV}(u,v)$?

Let $A = \{(x,y) : f_{XY}(x,y) > 0\}$, support of f_{XY}

$B = g(A) := \{g(x,y) : (x,y) \in A\}$

Write $\begin{pmatrix} u \\ v \end{pmatrix} := g(x,y) := \begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix}$

which has the inverse function

$$\begin{pmatrix} x \\ y \end{pmatrix} = h(u,v) := g^{-1}(u,v) := \begin{pmatrix} h_1(u,v) \\ h_2(u,v) \end{pmatrix}$$

Then the pdf $f_{UV}(u,v)$ is given by the

change-of-var formula: $f_{UV}(u,v) = f_{XY}(h_1(u,v), h_2(u,v)) \cdot |J|$

where J: determinant of the Jacobian matrix:

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix} = \frac{\partial h_1}{\partial u}\frac{\partial h_2}{\partial v} - \frac{\partial h_1}{\partial v}\frac{\partial h_2}{\partial u}$$

When the transformation $(x,y) \to g(x,y) \in \mathbb{R}^2$ is many-to-one.

we may partition the support of f_{XY}

$$A = \{(x,y) : f_{XY}(x,y) > 0\}$$

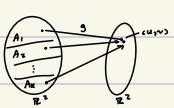
into disjoint subsets

$A = A_1 \cup \cdots \cup A_K$

s.t. g is one-to-one from each A_i to $g(A_i)$, $i = 1, \ldots, K$

Prop. Let $h_i = (h_{i1}, h_{i2})$ be the inverse function of g restricted to domain $A_i \to g(A_i)$ and J_i the corresponding Jacobian. Then

$$f_{UV}(u,v) = \sum_{i=1}^{K} f_{XY}(h_{i1}(u,v), h_{i2}(u,v)) \cdot |J_i|$$



4.4: Mixture & Hierarchical Models

- composite families of dist.'s in terms of simpler models of dist.'s, which are constructed in a hierarchical way

Thm. If X and Y are any RV's, then

$$EX = E[E(X|Y)]$$

provided that the expectations exist.

("law of iterated expectation")

Proof. Suppose $(X,Y) \sim F(x,y)$ - continuous setting. Then

$$\begin{aligned} EX &= \iint x \, f_{XY}(x,y)\,dx\,dy \\ &= \iint x \, f_{X|Y}(x|y)\, f_Y(y)\,dx\,dy \\ &= \int \left(\int x \, f_{X|Y}(x|y)\,dx \right) f_Y(y)\,dy \\ &= \int E(X|Y=y)\, f_Y(y)\,dy \\ &= E[E(X|Y)]. \end{aligned}$$

Def: A RV X is said to have a **mixture dist.** if the dist. of X depends on a quantity that is also random, e.g. (**Poisson-Binomial mixture**)

$$\left.\begin{aligned} X|Y &\sim \text{Binomial}(Y,p) \\ Y &\sim \text{Poisson}(\lambda) \end{aligned}\right\} \begin{array}{l} \text{mixture of Binomial distributions} \\ \text{mixing distribution given by Poisson} \end{array}$$

Adding more "models" to obtain arbitrarily complex distributions:

Prop.

Suppose $Y \in \{1, \ldots, K\}$, $P(Y=i) = p_i$, $i = 1, \ldots, K$

and $X|Y=i \sim \text{Normal}(\mu_i, \sigma_i^2)$

Then

$$\begin{aligned} f_X(x) &= \sum_{i=1}^{K} p_i \, N(x|\mu_i, \sigma_i^2) \\ &= \sum_{i=1}^{K} p_i \cdot \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(\frac{-1}{2\sigma_i^2}(x-\mu_i)^2\right) \end{aligned}$$

(a mixture of K normal components)

If a prob. dist. for X may be obtained by multiple stages of cond. dist.'s, then we obtain a **hierarchical model**, e.g.

$$\text{let} \begin{cases} X|Y \sim \text{Binomial}(Y,p) & (\#\text{ surviving eggs}) \\ Y|\lambda \sim \text{Poisson}(\lambda) & (\#\text{ eggs laid}) \\ \lambda \sim \text{Exponential}(\beta) \end{cases}$$

where the randomness of λ captures the variation across the (insect) mothers

Thm (Iterated Variance Formula): For any two RV's X, Y,

$$\operatorname{var} X = E[\operatorname{var}(X|Y)] + \operatorname{Var}[E(X|Y)]$$

P. Recall $V(X) = E(X^2) - E^2(X)$

$$\Rightarrow V(X|Y) = E(X^2|Y) - E^2(X|Y) \quad (a)$$

where $E(E(X|Y)) = E(X)$

$$\Rightarrow E(\operatorname{var}(X|Y)) = EX^2 - E(E(X|Y))^2 \quad (1)$$

$$\operatorname{var}(E(X|Y)) = E(E(X|Y))^2 - \underbrace{(E\,E(X|Y))^2}_{(EX)^2} \quad (2)$$

TWTS $RHS = (1) + (2) = (a)$

4.3: covariance and correlation

Let (X,Y) be a bivariate vector ($(X,Y) \in \mathbb{R}^2$)

e.g. $(X,Y) = (\text{weight, waist})$

want to make statement like

"if $X\uparrow$ then Y tends $\uparrow$ and vice versa"

covariance: = cov of RV's X and Y

$$\operatorname{cov}(X,Y) = E[(X - EX)(Y - EY)]$$

Remarks: - IF $X = Y$, then $\operatorname{cov}(X,X) \equiv \operatorname{var}(X) = E[X - EX]^2$

- IF $EX = EY = 0$ then $\operatorname{cov}(X,Y) = E(XY)$

correlation: =

$$\operatorname{corr}(X,Y) := \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}}$$

i.e. $\operatorname{corr}(X,Y)$ is covariance of standardized versions of X and Y

Thm: $\operatorname{cov}(X,Y) = E(XY) - (EX)(EY)$

(i) IF $X \perp\!\!\!\perp Y$ then $\operatorname{cov}(X,Y) = \operatorname{corr}(X,Y) = 0$

P(i). $\operatorname{cov}(X,Y) = E[(X - EX)(Y - EY)]$

$= E(XY - X(EY) - Y(EX) + (EX)(EY))$

i.e. E

$= E(XY) - (EX)(EY) - (EY)(EX) + (EX)(EY)$

$= E(XY) - (EX)(EY)$. ∎

Remarks: · generalizes $\operatorname{var} X = E(X^2) - E^2(X)$

· conversely, $\operatorname{cov}(X,Y) = 0$ does not imply $X \perp\!\!\!\perp Y$

(unless (X,Y) is bivariate normal - see later)

Thm: $\forall\, a, b \in \mathbb{R}$,

$$V(aX+bY) = a^2 \operatorname{var}(X) + b^2 \operatorname{var}(Y) + 2ab \operatorname{cov}(X,Y)$$

Cauchy-Schwartz Inequality: $E[XY] \le \sqrt{E(X^2)}\sqrt{E(Y^2)}$

$\Leftrightarrow$

$$[E(XY)]^2 \le E(X^2)E(Y^2)$$

Letting $U = X - \mu_X$, $V = Y - \mu_Y$,

$$[E(UV)]^2 \le E(U^2)E(V^2)$$

$$[\operatorname{cov}(X,Y)]^2 \le \operatorname{var}(X)\operatorname{var}(Y)$$

$$\frac{[\operatorname{cov}(X,Y)]^2}{V(X)V(Y)} \le 1$$

$$\left|\frac{\operatorname{cov}(X,Y)}{\sqrt{V_X}\sqrt{V_Y}}\right| \le 1$$

$$|\operatorname{corr}(X,Y)| \le 1$$

$E[XY] \le \sqrt{E(X^2)}\sqrt{E(Y^2)}$

Equality holds i.f.f. $Y = cX$ a.s., i.e.

$$P(Y = cX) = 1 \quad , c: \text{constant} \in \mathbb{R}$$

Letting $U = X - EX$, $V = Y - EY$,

equality holds i.f.f. $P(U = aV) = 1$, a: constant $\in \mathbb{R}$, i.e.

$$Y - EY = a(X - EX) \quad \text{a.s.}$$

$$\Rightarrow \operatorname{cov}(X,Y) = E[(X-EX)(Y-EY)] = E[(X-EX)a(X-EX)] = a\,E[(X-EX)^2] = a\operatorname{var}(X)$$

$$\operatorname{var}(Y) = E[(Y-EY)^2] = E[a^2(X-EX)^2] = a^2\operatorname{var}(X)$$

$$\operatorname{corr}(X,Y) = \frac{a\operatorname{var}(X)}{\sqrt{\operatorname{var}(X)}\sqrt{a^2\operatorname{var}(X)}} = \frac{a}{|a|} = \begin{cases} +1, & a > 0 \\ -1, & a < 0 \end{cases}$$

$$\Rightarrow |\operatorname{corr}(X,Y)| = 1$$

Thm. Suppose $\operatorname{var} X \neq 0$, $\operatorname{var} Y \neq 0$.

(i) If $\operatorname{corr}(X,Y) = 1$, then there are constants $b > 0$ and $c \in \mathbb{R}$ s.t.

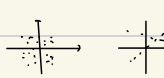
$$Y = bX + c \text{ w.p. } 1$$

(ii) If $\operatorname{corr}(X,Y) = -1$, then

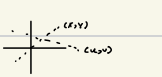
$$Y = bX + c \text{ w.p. } 1$$

for some constants $b < 0$, c.

Remarks:



i.f. $X \perp\!\!\!\perp Y$, then $\operatorname{corr}(X,Y) = 0$



$\operatorname{corr}(X,Y) = 1 \Rightarrow Y = bX + c, b > 0$

$\operatorname{corr}(X,Y) = -1 \Rightarrow Y = bX + c, b < 0$

Remark: $\text{corr}(X,Y)=0$ does not imply $X \perp\!\!\!\perp Y$.

· possible that X and Y are strongly dependent on one another (suggests nonlinear relationship)

Bivariate Normal Distribution

Def: $Z=(X,Y)\sim N(\mu,\Sigma)$ $\Big|$ $\mu\in\mathbb{R}^2$, Σ positive definite, symmetric $\in\mathbb{R}^{2\times 2}$

$$f_{XY}(x,y) := f_Z(z) := \frac{1}{((2\pi)^2|\Sigma|)^{1/2}} \exp\left\{ -\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right\} \quad (*)$$

Remark: $\mu=(\mu_1,\mu_2)^T$, $\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$

μ: mean vector, Σ: covariance matrix of Z: $\begin{cases} EZ=\mu \\ \text{cov}Z=\Sigma\end{cases}$

$EZ=\mu$ means $E\begin{pmatrix}X\\Y\end{pmatrix}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \Leftrightarrow \begin{cases} EX=\mu_1\\ EY=\mu_2\end{cases}$

$\text{cov}Z=\Sigma$ means $E[Z-EZ][Z-EZ]^T=\Sigma$

$$\Leftrightarrow E\begin{bmatrix}X-EX\\ Y-EY\end{bmatrix}\begin{bmatrix}X-EX\\ Y-EY\end{bmatrix}^T=\Sigma$$

$$\Leftrightarrow E\begin{bmatrix}(X-EX)^2 & (X-EX)(Y-EY)\\ (Y-EY)(X-EX) & (Y-EY)^2\end{bmatrix}=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

$$\Leftrightarrow \begin{cases}\Sigma_{11}=\text{var}X\\ \Sigma_{22}=\text{var}Y\\ \Sigma_{12}=\Sigma_{21}=\text{cov}(X,Y)\end{cases} \quad \text{where } \text{cov}Z=E[(Z-EZ)(Z-EZ)^T]$$

a remarkable property of the normal dist.

Due to the identities:

if $f_{XY}(x,y)=f_Z(z)=\frac{1}{\sqrt{(2\pi)^2|\Sigma|}}\exp\left\{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right\}$

then:

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f_{XY}(x,y)\,dx\,dy=1$$
$$\iint x f_{XY}(x,y)\,dx\,dy=\mu_1$$
$$\iint y f_{XY}(x,y)\,dx\,dy=\mu_2$$
$$\iint (z-\mu)(z-\mu)^T f_Z(z)\,dz=\Sigma$$

Remark: follows from identities for (univariate) normal dist.

Translating (*) into univariate forms (don't need to remember):

$$f_{XY}(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y(1-\rho_{XY}^2)^{1/2}}\times\exp\left\{\frac{-1}{2(1-\rho_{XY}^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2-2\rho_{XY}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

Thm. If $\text{cov}(X,Y)=0$ and $(X,Y)\sim$ bivariate normal then $X\perp\!\!\!\perp Y$

P. $\text{cov}(X,Y)=0 \Rightarrow \rho_{XY}=0$. From (*), $f_{XY}(x,y)$ factorizes into a product of $g(x)h(y)$, so $X\perp\!\!\!\perp Y$. ∎

Facts: If $Z=(X,Y)^T\sim N(\mu,\Sigma)$ $\Big|$ $\mu=(\mu_1,\mu_2)^T$, $\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$

then the marginal of X and Y are normal too:

$X\sim N(\mu_1,\Sigma_{11})$

$Y\sim N(\mu_2,\Sigma_{22})$

and $\text{cov}(X,Y)=\Sigma_{12}=\Sigma_{21}$

the conditional distributions are also normal:

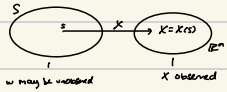
$$Y|_{X=x}\sim N(E(Y|X=x),\text{var}(Y|X=x))$$

where $E(Y|X=x)=\mu_2+\Sigma_{21}\Sigma_{11}^{-1}(x-\mu_1)$

$\text{var}(Y|X=x)=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

4.6: Multivariate distributions

Def: A random vector denoted $X=(X_1,\dots,X_n)$ is a function that maps elements of a sample space S into $\mathbb{R}^n$



$$X(s)=(X_1(s),\dots,X_n(s))\in\mathbb{R}^n$$

Now, we may specify probabilities, for $A\subset\mathbb{R}^n$:

$$\boxed{P(X\in A) := P(\{s\in S : X(s)\in A\})}$$

If X takes countably many values, then we say X is a discrete RV, which is associated w/ a

joint probability mass function (pmf): $=$ $\boxed{f(x)=f(x_1,\dots,x_n)=P(X_1=x_1,\dots,X_n=x_n)}$

Hence for $A\subset\mathbb{R}^n$, $P(X\in A)=\sum_{x\in A} P(X=x)=\sum_{x\in A} f(x)$

Similarly, pdf: $=$ $\boxed{f(x)=f(x_1,\dots,x_n)}$

where $P(X=(X_1,\dots,X_n)\in A)=\int_A f(x)\,dx=\iint\cdots\int_{(x_1,\dots,x_n)\in A} f(x_1,\dots,x_n)\,dx_1\dots dx_n$

Expectation: $=$ $\boxed{Eg(X)=\begin{cases}\int_{\mathbb{R}^n} g(x)f(x)dx & X\text{-cont.}\\ \sum_{x\in\mathbb{R}^n} g(x)f(x) & X\text{-discrete}\end{cases}}$

marginal dist.: $=$ $\boxed{f_{X_1,\dots,X_k}(x_1,\dots,x_k)=\begin{cases}\int_{\mathbb{R}}\cdots\int_{\mathbb{R}} f_{X_1,\dots,X_n}(x_1,\dots,x_n)\,dx_{k+1}\dots dx_n\\ \sum_{x_{k+1}}\dots\sum_{x_n} f_{X_1,\dots,X_n}(x_1,\dots,x_n)\end{cases}}$

Ex: $(X_1,\dots,X_4)\sim f_X(x_1,\dots,x_4)$

$$f_{X_1X_4}(x_1,x_4)=\int_{\mathbb{R}}\int_{\mathbb{R}} f_{X_1X_2X_3X_4}(x_1,x_2,x_3,x_4)\,dx_2dx_3$$

conditional dist.: $=$ $\boxed{f_{X_{k+1},\dots,X_n|X_1,\dots,X_k}(x_{k+1},\dots,x_n|x_1,\dots,x_k)=\frac{f_{X_1,\dots,X_n}(x_1,\dots,x_n)}{f_{X_1,\dots,X_k}(x_1,\dots,x_k)}}$

Remarks: · admits factorization: $f_{X_1,\dots,X_n}(x_1,\dots,x_n)=f_{X_1,\dots,X_k}(x_1,\dots,x_k)\,f_{(X_{k+1},\dots,X_n)|(X_1,\dots,X_k)}(x_{k+1},\dots,x_n|x_1,\dots,x_k)$

chain formula: $f_{X_1,\dots,X_n}(x_1,\dots,x_n)=f_{X_1}(x_1)\,f_{X_2|X_1}(x_2|x_1)\,f_{X_3|X_1,X_2}(x_3|x_1,x_2)\cdots f_{X_n|X_1,\dots,X_{n-1}}(x_n|x_1,\dots,x_{n-1})$

eg. $f_{XYZ}(x,y,z)=f_X(x)\,f_{Y|X}(y|x)=f_X(x)\,f_{Y|X}(y|x)\,f_{Z|Y,X}(z|y,x)$

Prop. Suppose $\begin{cases}X,Y\text{-discrete}\\ Z,T\text{-continuous}\end{cases}$ and let $f_{XYZT}(x,y,z,t)=\underbrace{f_{XY}(x,y)}_{\text{pmf}}\underbrace{f_{ZT|XY}(z,t|x,y)}_{\text{conditional pdf}}$

$$=\underbrace{f_{ZT}(z,t)}_{\text{pdf}}\underbrace{f_{XY|ZT}(x,y|z,t)}_{\text{conditional pmf}}$$

Then, $\forall A,B,C,D\subset\mathbb{R}$,

$$P(X\in A,Y\in B,Z\in C,T\in D)=\sum_{x\in A}\sum_{y\in B}\int_D\int_C f_{XYZT}(x,y,z,t)\,dz\,dt$$

Extn (multinomial dist.): $\begin{cases}n\text{ independent trials}\\ k\text{ possible outcomes w/}\\ \text{"cell probs" } (p_1,\dots,p_k),\ \sum_{i=1}^k p_i=1\end{cases}$

Let $(X_1,\dots,X_k)$: counts of each of k outcomes

$X_1+\dots+X_k=n$

Now, joint pmf for $x_1, \dots, x_n \in \mathbb{N}^n$, $\sum_{i=1}^n x_i = m$

$$f(x_1, \dots, x_n) = \binom{m}{x_1, \dots, x_n} p_1^{x_1} \cdots p_n^{x_n} = \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}$$

and 0 otherwise

Remarks: By multinomial formula

$$\sum_{\substack{x_1, \dots, x_n \\ x_1 + \cdots + x_n = m}} \binom{m}{x_1, \dots, x_n} p_1^{x_1} \cdots p_n^{x_n} = (p_1 + \cdots + p_n)^m = 1$$

- marginal distribution: $f_{x_n}(x_n) = \frac{m!}{x_n!(m-x_n)!} p_n^{x_n} (1-p_n)^{m-x_n}$

$\Rightarrow X_n \sim \text{Binomial}(m, p_n)$

- conditional distribution: $f_{x_1, \dots, x_{n-1} | x_n}(x_1, \dots, x_{n-1} | x_n) = \frac{(m-x_n)!}{x_1! \cdots x_{n-1}!} \left(\frac{p_1}{1-p_n}\right)^{x_1} \cdots \left(\frac{p_{n-1}}{1-p_n}\right)^{x_{n-1}}$

$\Rightarrow X_1, \dots, X_{n-1} | X_n = x_n \sim \text{Multinomial}(m - x_n \text{ trials}, \text{cell probs } \left(\frac{p_1}{1-p_n}, \dots, \frac{p_{n-1}}{1-p_n}\right))$

Notice: kind of invariance property by marginalization and conditioning;
this holds for quite a few families of distributions (often in exponential family)
for continuous distributions, one such family is the multivariate normal

Recall (independence): X and Y independent if $\forall x, y \in \mathbb{R}$, $F_{XY}(x,y) = F_X(x) F_Y(y)$

Def: Let $(X_1, \dots, X_n)$ be a random vector w/ joint pdf/pmf $f(x_1, \dots, x_n)$.
Let $f_{X_i}(x_i)$ be the marginal pdf of X_i.
Then $X_1, \dots, X_n$ are mutually independent RV's if $\forall (x_1, \dots, x_n)$

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

Remark: each X_i may be itself a random vector

Many properties established for bivariate random vector can be extended naturally to the multivariate setting

Thm. If $X_1, \dots, X_n$ are (mutually independent) RV's,
let $g_1(x_1), \dots, g_n(x_n)$ be real-valued functions on the domain of $X_1, \dots, X_n$ respectively. Then

$$E[g_1(x_1) \cdots g_n(x_n)] = E[g_1(x_1)] \cdots E[g_n(x_n)]$$

Thm. If $X_1, \dots, X_n$ - mutually independent RV's. Let $Z = X_1 + \cdots + X_n$
then $M_Z(t) = M_{X_1}(t) \cdots M_{X_n}(t)$

corollary: in particular, if $X_1, \dots, X_n$ - i.i.d, then $M_Z(t) = (M_{X_1}(t))^n$

Properties

1) If $X_i \sim \text{Gamma}(\alpha_i, \beta)$, X_i - independent.
then $X_1 + \cdots + X_n \sim \text{Gamma}(\alpha_1 + \cdots + \alpha_n, \beta)$

2) If $X_i \overset{\text{ind}}{\sim} N(\mu_i, \sigma_i^2)$, let $a_i, b_i \in \mathbb{R}$.
then $Z := \sum_{i=1}^n (a_i X_i + b_i) \sim N\left(\sum_{i=1}^n a_i \mu_i + b_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$
$= a^T X + b^T \mathbb{1}$

Lemma: Suppose $(X_1, \dots, X_n) \sim f_{X_1, \dots, X_n}$.
$X_1, \dots, X_n$ are mutually independent r.v.'s.

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = g_1(x_1) \cdots g_n(x_n)$$

for some functions $g_1, \dots, g_n$

Proof. Similar to bivariate case

Thm. If $X_1, \dots, X_n$ are mutually independent,
let $g_i(x_i)$: function of x_i, $i = 1, \dots, n$
then $g_1(x_1), \dots, g_n(x_n)$ are also mutually independent

Proof. Follows directly from the factorization of the joint pdf/pmf

Note: the change-of-variable formula can be extended to the multivariate case too (See textbook)

Chapter 5: Properties of a Random Sample

5.1: i.i.d samples

Def: Suppose $X_1, \dots, X_n \overset{iid}{\sim} f$ (pmf or pdf).
Then we say $(X_1, \dots, X_n)$ is a random sample from the population with pmf/pdf $f(x)$.

Remarks: (other variations):
- n-sample of f
- n-iid sample of f
- n-sample of size n

Def: Let $(X_1, \dots, X_n)$ be a n-sample from a population.
Let $T(x_1, \dots, x_n)$ be a real-valued function.
Then $Y = T(X_1, \dots, X_n)$ is called a statistic
i.e. a function of a random sample

Remarks: Statistic is a RV (r.v.)
- telling us something about an underlying population w/ pmf/pdf f.
- does so only on random samples (data)

Modern viewpoint: $X_1, \dots, X_n \to$ [T] $\to Y = T(X_1, \dots, X_n)$
algorithm/application

Examples: 1) $\bar{X} := \frac{1}{n}(X_1 + \cdots + X_n)$; sample mean

2) $S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

3) $X_{(1)} := \min\{X_1, \dots, X_n\}$
$X_{(2)} := \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}\}$
$\vdots$
$X_{(k)} := \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(k-1)}\}$
i.e. the k^{th} smallest number of the sample

order statistics of the n-sample $:= (X_{(1)}, \dots, X_{(n)})$

Thm. Let $(X_1, \dots, X_n)$ be an n-iid sample from a population with | mean μ / variance $\sigma^2 < \infty$

Then $\begin{cases} E\bar{X} = \mu \\ \operatorname{Var}\bar{X} = \sigma^2/n \\ ES^2 = \sigma^2 \end{cases}$

Remark: provides statistical justification for using $\bar{X}$ and s^2 as estimates of μ and σ^2 respectively; $\bar{X}$ and s^2 are unbiased estimates

Distribution of $\bar{X}$

two main methods: 1) method of moments
2) change-of-var formula (convolution formula)

1) $M_{\bar{X}}(t) = Ee^{t\bar{X}}$
$= Ee^{\frac{t}{n}(X_1 + \cdots + X_n)}$
$= Ee^{\frac{t}{n}X_1} \cdots e^{\frac{t}{n}X_n}$
$= (M_X(t/n))^n$

Remark: If $M_{\bar{X}}$ can be recognized as MGF of a known family, then we can find the dist. of $\bar{X}$ easily, e.g.

if $X_1, \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$
then $\bar{X} \sim N(\mu, \sigma^2/n)$

2) Convolution Formula

Thm. If $X \sim f_X$, $Y \sim f_Y$,
then $Z := X + Y$ has the pdf which is the convolution of f_X and f_Y:

$$f_Z(z) = f_X(x) * f_Y(z) := \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$$

Proof. Use the change-of-var formula for the mapping $(X, Y) \to (X+Y, X)$

Remark: useful for deriving pdf/pmf when the transformation $T(X_1, \dots, X_n)$ does not belong to a known/well-recognized family (e.g. exponential, location-scale, etc.)

5.2: useful classical facts

Thm. IF $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$,

let $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

then

a) $\bar{X}$ and S^2 are independent RV's

b) $\bar{X} \sim N(\mu, \sigma^2/n)$

c) $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

Proof (b). seen in previous section

Recall: χ^2_p w/ p dof:

$$f(x) = \frac{1}{\Gamma(p/2)\,2^{p/2}}\, x^{p/2-1} e^{-x/2}, \quad x \geq 0$$

and $\chi^2_p \equiv \text{Gamma}(p/2, 2)$

Other statistics and their distributions

IF $X_1, X_2, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ - unknown

then $\bar{x}$ is a statistic that tells us about μ

Q: how so? we know $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

this gives us a way to quantify the uncertainty about μ, using statistic $\bar{X}$ as an estimate

Q: what if σ is unknown too?

we may want to consider $\frac{\bar{X}-\mu}{S/\sqrt{n}}$

since S is a statistic that can be obtained from the sample

To be useful, we need to know the distribution of

$$T := \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

Thm. $T \sim t_{n-1}$: Student's t dist. w/ n-1 dof

t_p has the pdf

$$f_T(t) = \frac{\Gamma((p+1)/2)}{\Gamma(p/2)} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1+\frac{t^2}{p})^{(p+1)/2}} \quad -\infty < t < \infty$$

Remarks: • T is the ratio of two independent random variables as

$$\bar{X}-\mu \sim \text{Normal}, \quad (n-1)S^2 \sim \sqrt{\chi^2_{n-1}}$$

• IF $n=2$, then $T \sim$ Ratio of 2 ind. Normal RV's $\Rightarrow T \sim$ Cauchy

5.3: convergence concepts

Given a sequence of RV's $X_1, \dots, X_n$, we want to study various notions of convergence to a RV X:

i) $X_n \xrightarrow{P} X$ convergence in probability

ii) $X_n \xrightarrow{a.s.} X$ convergence almost surely/ with prob 1

iii) $X_n \xrightarrow{d} X$ convergence in dist.

we already encountered: $X_n \xrightarrow{d} X$ IF $F_{X_n}(x) \longrightarrow F_X(x)$ at all points where F_X is continuous

Def: $X_n \xrightarrow{P} X$ IF

$$\forall\, \varepsilon > 0,\ P(|X_n - X| \geq \varepsilon) \to 0 \text{ as } n \to \infty$$

Notice: $\forall\, \varepsilon > 0, P(|X_n - X| \geq \varepsilon) \to 0$ as $n \to \infty$

$= P(\omega \in \{|X_n(\omega) - X(\omega)| \geq \varepsilon\}) \to 0$ as $n \to \infty$

where S generally depends on n

Thm (weak law of large numbers - WLLN). Let $X_1, \dots, X_n \overset{iid}{\sim}$ RV's with $\mathbb{E}X_i = \mu$, $\mathrm{Var}X_i = \sigma^2 < \infty$.

Define $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$

Then $\bar{X}_n \xrightarrow{p} \mu$

Def: $X_n \xrightarrow{a.s.} X$ (w/ prob. 1) if

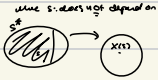
$$P\left(\lim_{n\to\infty} |X_n - X| = 0\right) = 1 \equiv P\left(\lim_{n\to\infty} X_n = X\right) = 1 \quad (*)$$

Remark: Equivalently (!): $\forall \varepsilon > 0$, $P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1$

Notice: $X_n \xrightarrow{a.s.} X$ if

$P(\exists S : X_n(s) \to X(s) \ \forall s) = 1$

where S does not depend on ω



Here, $X_n(s) \to X(s)$ $\forall s \in S^* \Rightarrow P(S^*) = 1$

Prop: If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{p} X$.

If $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{d} X$

Thm (strong law of large numbers - SLLN). Let $X_1, \dots, X_n \overset{iid}{\sim}$ RV's with $\mathbb{E}X_i = \mu$, $\mathbb{E}|X_i| < \infty$.

Define $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$

Then $\bar{X}_n \xrightarrow{a.s.} \mu$

Proof: similar in spirit but technically more involved than WLLN

Note: Condition $\mathbb{E}|X_i| < \infty$ is very mild

Thm (CLT). If $X_1, X_2, \dots \overset{iid}{\sim}$ RV w/ MGF exist in a neighborhood of 0.

Let $\mu = \mathbb{E}X_i$, $\sigma^2 = \mathrm{Var}X_i > 0$ and $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$, $Z \sim N(0,1)$

Then $\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} Z$

Remarks:
- perhaps most celebrated thm. in prob.
- widely applicable; only $\mathrm{Var}X_i < \infty$ is required

Corollary: Let $Y_i = (X_i - \mu)/\sigma$. Then $\mathbb{E}Y_i = 0$, $\mathbb{E}Y_i^2 = 1$

and we usually write $\frac{1}{\sqrt{n}}(Y_1 + \cdots + Y_n) \xrightarrow{d} N(0,1)$.

P. $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) = \frac{\sqrt{n}}{\sigma}\left(\frac{X_1 + \cdots + X_n - n\mu}{n}\right) = \frac{1}{\sqrt{n}}(Y_1 + \cdots + Y_n) \xrightarrow{d} N(0,1)$.

- convergence (a.s., in prob., in dist.) preserves in nice transformations:

Lemma (1): If $X_n \xrightarrow{a.s.} X$, $Y_n \xrightarrow{a.s.} b$: constant

then, $aX_n + Y_n \xrightarrow{a.s.} aX + b$

Lemma (2): If $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} b$: constant

then, $aX_n + Y_n \xrightarrow{p} aX + b$

Lemma (3): If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} b$: const.

then $aX_n + Y_n \xrightarrow{d} aX + b$

Remarks:
- L1 and L2 are almost immediate from definition
- L3 is known as Slutsky's lemma/thm.; P. based on characterization: $X_n \xrightarrow{d} X$ iff for all continuous and bounded function $f(x)$, $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ as $n \to \infty$ (higher-dim obs. as opposed to cdf - not covered)