Aug 24st 2021 ※ Chapter 1.

Definition 1.1.1 The set, S, of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

Ex. flip a coin : $S=\{H,T\}$

Definition 1.1.2 An *event* is any collection of possible outcomes of an experiment, that is, any subset of S (including S itself).

Theorem 1.1.4 *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	$(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

Definition 1.1.5 Two events A and B are *disjoint* (or *mutually exclusive*) if $A \cap B = \emptyset$. The events $A_1, A_2, \ldots$ are *pairwise disjoint* (or *mutually exclusive*) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition 1.1.6 If $A_1, A_2, \ldots$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, \ldots$ forms a *partition* of S.

Definition 1.2.1 A collection of subsets of S is called a *sigma algebra* (or *Borel field*), denoted by $\mathcal{B}$, if it satisfies the following three properties:

a. $\emptyset \in \mathcal{B}$ (the empty set is an element of $\mathcal{B}$).

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

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

Ex. $\mathcal{B}=\{\emptyset, S\}$ trivial algebra

$\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}$. $|\mathcal{B}|=2^n$ (if $|S|=n$)

Example 1.2.3 (Sigma algebra–II) Let $S = (-\infty, \infty)$, the real line. Then $\mathcal{B}$ is chosen to contain all sets of the form

$$[a, b], \quad (a, b], \quad (a, b), \quad \text{and} \quad [a, b)$$

for all real numbers a and b. Also, 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. ‖

The empty set $\emptyset$ is a subset of any set. Thus, $\emptyset \subset S$. Property (a) states that this subset is always in a sigma algebra. Since $S = \emptyset^c$, properties (a) and (b) imply that S is always in $\mathcal{B}$ also. In addition, from DeMorgan's Laws it follows that $\mathcal{B}$ is closed under countable intersections. If $A_1, A_2, \ldots \in \mathcal{B}$, then $A_1^c, A_2^c, \ldots \in \mathcal{B}$ by property (b), and therefore $\cup_{i=1}^{\infty} A_i^c \in \mathcal{B}$. However, using DeMorgan's Law (as in Exercise 1.9), we have

$$\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c = \bigcap_{i=1}^{\infty} A_i. \tag{1.2.1}$$

Thus, again by property (b), $\cap_{i=1}^{\infty} A_i \in \mathcal{B}$. → closed under countable intersections

Definition 1.2.4 Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function P with domain $\mathcal{B}$ that satisfies

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$.
2. $P(S) = 1$.
3. If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

$P: \mathcal{B} \to [0,1]$

Prop1: $P(\emptyset) = 0$

Pf: Let $A_i = \emptyset$, then $\bigcup_{i=1}^{\infty} A_i = \emptyset$

$$P(\emptyset) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} P(\emptyset) \Rightarrow P(\emptyset) = 0.$$

2. *Axiom of Finite Additivity:* If $A \in \mathcal{B}$ and $B \in \mathcal{B}$ are disjoint, then

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

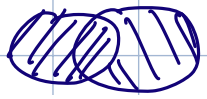
Pf: Let $A_1 = A$, $A_2 = B$, $A_i = \emptyset$, $\forall i \geq 3$

$$P(A \cup B) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = P(A) + P(B) + \underbrace{\sum_{i=3}^{\infty} P(\emptyset)}_{=0} = P(A) + P(B)$$

3. $P(A^c) = 1 - P(A)$.

Pf: $1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$

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



Pf: $P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B) - P(A \cap B)$

$$P(B) = P(\{B \cap A\} \cup \{B \cap A^c\}) = P(B \cap A) + P(B \cap A^c),$$

5. *If $A \subset B$, then $P(A) \leq P(B)$*

Pf: $0 \leq P(B \cap A^c) = P(B) - P(A)$

6/7. **Theorem 1.2.11** *If P is a probability function, then*

a. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ *for any partition* $C_1, C_2, \ldots$;

b. $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ *for any sets* $A_1, A_2, \ldots$. *(Boole's Inequality)*

(a) Pf: $$A = A \cap S = A \cap \left(\bigcup_{i=1}^{\infty} C_i\right) = \bigcup_{i=1}^{\infty} (A \cap C_i). \quad P\left(\bigcup_{i=1}^{\infty} (A \cap C_i)\right) = \sum_{i=1}^{\infty} P(A \cap C_i)$$

(b) Pf: To establish (b) we first construct a disjoint collection $A_1^*, A_2^*, \ldots$, with the property that $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$. We define A_i^* by

Need to check:
① $A_i^* \subseteq A_i$
② $\bigcup_{i=1}^{\infty} A_i^* = \bigcup_{i=1}^{\infty} A_i$
③ A_i^* pairwise disjoint

$$A_1^* = A_1, \quad A_i^* = A_i \setminus \left(\bigcup_{j=1}^{i-1} A_j\right), \quad i = 2, 3, \ldots,$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i^*\right) = \sum_{i=1}^{\infty} P(A_i^*) \leq \sum_{i=1}^{\infty} P(A_i)$$

Theorem 1.2.6 *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 $P(A)$ by*

$$P(A) = \sum_{\{i: s_i \in A\}} p_i.$$

(The sum over an empty set is defined to be 0.) Then P is a probability function on $\mathcal{B}$. This remains true if $S = \{s_1, s_2, \ldots\}$ is a countable set.

Proof: We will give the proof for finite S. For any $A \in \mathcal{B}$, $P(A) = \sum_{\{i: s_i \in A\}} p_i \geq 0$, because every $p_i \geq 0$. Thus, Axiom 1 is true. Now,

$$P(S) = \sum_{\{i: s_i \in S\}} p_i = \sum_{i=1}^{n} p_i = 1.$$

Thus, Axiom 2 is true. Let $A_1, \ldots, A_k$ denote pairwise disjoint events. ($\mathcal{B}$ contains only a finite number of sets, so we need consider only finite disjoint unions.) Then,

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{\{j: s_j \in \cup_{i=1}^{k} A_i\}} p_j \overset{A_i \text{ disjoint}}{=} \sum_{i=1}^{k} \sum_{\{j: s_j \in A_i\}} p_j = \sum_{i=1}^{k} P(A_i).$$

The first and third equalities are true by the definition of $P(A)$. The disjointedness of the A_is ensures that the second equality is true, because the same p_js appear exactly once on each side of the equality. Thus, Axiom 3 is true and Kolmogorov's Axioms are satisfied. □

Theorem 1.2.14 *If a job consists of k separate tasks, the ith of which can be done in n_i ways, $i = 1, \ldots, k$, then the entire job can be done in $n_1 \times n_2 \times \cdots \times n_k$ ways.*

Table 1.2.1. *Number of possible arrangements of size r from n objects*

	Without replacement	With replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

△ unordered + w.r.

X_i = #times object i was selected

$\{(X_1, X_2, \cdots, X_n) : X_i \in \mathbb{Z}^+, \sum_{i=1}^{n} X_i = r\}$

→ n non-negative integers sum to r, so $\binom{n+r-1}{r}$

Aug 26st 2021

Very easy. No need to review this example.

Example 1.2.18 (Poker) Consider choosing a five-card poker hand from a standard deck of 52 playing cards. Obviously, we are sampling without replacement from the deck. But to specify the possible outcomes (possible hands), we must decide whether we think of the hand as being dealt sequentially (ordered) or all at once (unordered). If we wish to calculate probabilities for events that depend on the order, such as the probability of an ace in the first two cards, then we must use the ordered outcomes. But if our events do not depend on the order, we can use the unordered outcomes. For this example we use the unordered outcomes, so the sample space consists of all the five-card hands that can be chosen from the 52-card deck. There are $\binom{52}{5} = 2{,}598{,}960$ possible hands. If the deck is well shuffled and the cards are randomly dealt, it is reasonable to assign probability $1/2{,}598{,}960$ to each possible hand.

We now calculate some probabilities by counting outcomes in events. What is the probability of having four aces? How many different hands are there with four aces? If we specify that four of the cards are aces, then there are 48 different ways of specifying the fifth card. Thus,

$$P(\text{four aces}) = \frac{48}{2{,}598{,}960},$$

less than 1 chance in 50,000. Only slightly more complicated counting, using Theorem 1.2.14, allows us to calculate the probability of having four of a kind. There are 13

ways to specify which denomination there will be four of. After we specify these four cards, there are 48 ways of specifying the fifth. Thus, the total number of hands with four of a kind is (13)(48) and

$$P(\text{four of a kind}) = \frac{(13)(48)}{2{,}598{,}960} = \frac{624}{2{,}598{,}960}.$$

To calculate the probability of exactly one pair (not two pairs, no three of a kind, etc.) we combine some of the counting techniques. The number of hands with exactly one pair is

$$13\binom{4}{2}\binom{12}{3}4^3 = 1{,}098{,}240. \tag{1.2.11}$$

Expression (1.2.11) comes from Theorem 1.2.14 because

13 = # of ways to specify the denomination for the pair,

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

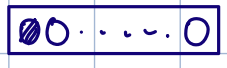
$\binom{12}{3}$ = # of ways of specifying the other three denominations,

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

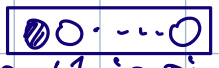
Thus,

$$P(\text{exactly one pair}) = \frac{1{,}098{,}240}{2{,}598{,}960}. \qquad \|$$

Ex. $\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$



object1 isn't picked $= \binom{n-1}{r}$



object1 is picked $= \binom{n-1}{r-1}$

the *inclusion-exclusion identity* says that

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P_1 - P_2 + P_3 - P_4 + \cdots \pm P_n.$$

$(-1)^{n+1}$

where

$$P_1 = \sum_{i=1}^{n} P(A_i)$$

$$P_2 = \sum_{1 \le i < j \le n} P(A_i \cap A_j)$$

$$P_3 = \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k)$$

$$\vdots$$

$$P_n = P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Pf: $P(A_1\cup\cdots\cup A_{n+1}) = P(A_1\cup\cdots\cup A_n) + P(A_{n+1}) - P((A_1\cup\cdots\cup A_n)\cap A_{n+1})$

$= P(A_1\cup\cdots\cup A_n) + P(A_{n+1}) - P((A_1\cap A_{n+1})\cup\cdots\cup(A_n\cap A_{n+1}))$

$= \sum_{i=1}^{n} P(A_i) - \sum_{1\le i_1<i_2\le n} P(A_{i_1}\cap A_{i_2}) + \cdots + (-1)^{n+1} P(A_1\cap\cdots\cap A_n)$

$+ P(A_{n+1}) - \sum_{i=1}^{n} P(A_i\cap A_{n+1}) + \sum_{1\le i_1<i_2\le n} P(A_{i_1}\cap A_{i_2}\cap A_{n+1})$

$- \cdots - (-1)^n \sum_{1\le i_1<\cdots<i_{n-1}\le n} P(A_{i_1}\cap\cdots\cap A_{i_{n-1}}\cap A_{n+1})$

$-(-1)^{n+1} P(A_1\cap A_2\cap\cdots\cap A_{n+1})$

$= P_1 - P_2 + P_3 - \cdots + (-1)^{n+2} P_{n+1}$

Ex. N men repick their hats randomly

- Calculate the prob. of <u>at least one man selects his own hat</u> → A.

A_i = ith man selects his own hat, $1\le i\le N$

$A = \bigcup_{i=1}^{N} A_i$

$P(A) = P(\bigcup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i) - \sum_{1\le i<j\le N} P(A_i\cap A_j) + \cdots -$

$|S| = N!$ $\quad P(A_i) = \frac{(N-1)!}{N!} = \frac{1}{N}$

$i\ne j$: $\quad P(A_i\cap A_j) = \frac{(N-2)!}{N!} = \frac{1}{N(N-1)}$

$\vdots$

$P(A_{i_1}\cap\cdots\cap A_{i_k}) = \frac{(N-k)!}{N!}$

$\therefore P(\bigcup_{i=1}^{N} A_i) = N\cdot\frac{1}{N} - \binom{N}{2}\frac{(N-2)!}{N!} + \binom{N}{3}\frac{(N-3)!}{N!} - \cdots + (-1)^{N+1}\binom{N}{N}\frac{(N-N)!}{N!}$

$= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{N+1}\frac{1}{N!}$

$\left(e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!}\right) \xrightarrow{\text{large } N} 1 - e^{-1} \approx 63\%$

Aug 31st, 2021

Definition 1.3.2 If A and B are events in S, and $P(B) > 0$, then the <u>*conditional probability of A given B*</u>, written $P(A|B)$, is

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

Theorem 1.3.5 (Bayes' Rule) *Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots$,*

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

Multiplicative Rule:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\cdots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

Example 1.3.6 (Coding) When coded messages are sent, there are sometimes errors in transmission. In particular, Morse code uses "dots" and "dashes," which are known to occur in the proportion of 3:4. This means that for any given symbol,

$$P(\text{dot sent}) = \frac{3}{7} \quad \text{and} \quad P(\text{dash sent}) = \frac{4}{7}.$$

Suppose there is interference on the transmission line, and with probability $\frac{1}{8}$ a dot is mistakenly received as a dash, and vice versa. If we receive a dot, can we be sure that a dot was sent? Using Bayes' Rule, we can write

$$P(\text{dot sent} \,|\, \text{dot received}) = P(\text{dot received} \,|\, \text{dot sent}) \frac{P(\text{dot sent})}{P(\text{dot received})}.$$

Now, from the information given, we know that $P(\text{dot sent}) = \frac{3}{7}$ and $P(\text{dot received} \,|\, \text{dot sent}) = \frac{7}{8}$. Furthermore, we can also write

$$\begin{aligned} P(\text{dot received}) &= P(\text{dot received} \cap \text{dot sent}) + P(\text{dot received} \cap \text{dash sent}) \\ &= P(\text{dot received} \,|\, \text{dot sent})P(\text{dot sent}) \\ &\quad + P(\text{dot received} \,|\, \text{dash sent})P(\text{dash sent}) \\ &= \frac{7}{8} \times \frac{3}{7} + \frac{1}{8} \times \frac{4}{7} = \frac{25}{56}. \end{aligned}$$

Combining these results, we have that the probability of correctly receiving a dot is

$$P(\text{dot sent} \,|\, \text{dot received}) = \frac{(7/8) \times (3/7)}{25/56} = \frac{21}{25} = 84\% \qquad \|$$

Definition 1.3.7 Two events, A and B, are *statistically independent* if

(1.3.8) $$P(A \cap B) = P(A)P(B).$$ $[P(A|B) = P(A)]$

Definition 1.3.12 A collection of events $A_1, \ldots, A_n$ are *mutually independent* if for any subcollection $A_{i_1}, \ldots, A_{i_k}$, we have

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

Theorem 1.3.9 *If A and B are independent events, then the following pairs are also independent:*

a. *A and B^c,*

b. *A^c and B,*

c. *A^c and B^c.*

Proof: We will prove only (a), ~~leaving the rest as Exercise 1.40~~. To prove (a) we must show that $P(A \cap B^c) = P(A)P(B^c)$. From Theorem 1.2.9a we have

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \qquad (A \text{ and } B \text{ are independent}) \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c). \end{aligned}$$

□

Definition 1.4.1 A *random variable* is a function from a sample space S into the real numbers.

$X: S \longrightarrow \mathbb{R}$

Ex. Toss a fair coin 3 times. $X :=$ # heads

$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$

$X:$ 3, 2, 2, 2, 1, 1, 1, 0

$P(X=2) = P(\{HHT, HTH, THH\}) = 3/8$

For any set $A \in \mathcal{X}$,

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

P_X: $\mathbb{R}$; P: S; $\{s \in S : X(s) \in A\}$: events in S; $X^{-1}(A)$: S

P_X is an induced prob. func on $\mathbb{R}$ (satisfies axioms of prob.)

Definition 1.5.1 The *cumulative distribution function* or *cdf* of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \le x), \quad \text{for all } x.$$

Example 1.5.2 (Tossing three coins) Consider the experiment of tossing three fair coins, and let X = number of heads observed. The cdf of X is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{8} & \text{if } 0 \le x < 1 \\ \frac{1}{2} & \text{if } 1 \le x < 2 \\ \frac{7}{8} & \text{if } 2 \le x < 3 \\ 1 & \text{if } 3 \le x < \infty. \end{cases} \tag{1.5.1}$$

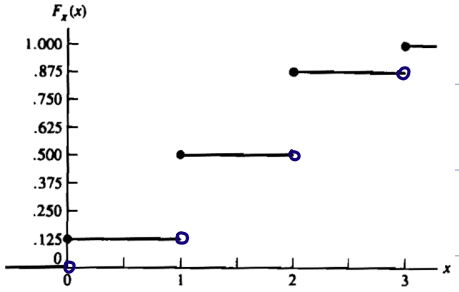


Figure 1.5.1. *Cdf of Example 1.5.2*

Theorem 1.5.3 *The function $F(x)$ is a cdf if and only if 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; that is, for every number* x_0, $\lim_{x\downarrow x_0} F(x) = F(x_0)$.

Axiom of continuity:

① $A_1, A_2, \cdots$ s.t. $A_i \subseteq A_{i+1}$; $i = 1, 2, \cdots$, then:

$P(A_n) \uparrow P(A)$; $A = \bigcup_{i=1}^{+\infty} A_i$

$\lim_{n\to+\infty} P(A_n) = \lim_{n\to+\infty} P(\bigcup_{i=1}^{n} A_i) \overset{?}{=} P(\lim_{n\to+\infty} \bigcup_{i=1}^{n} A_i) = P(\bigcup_{i=1}^{+\infty} A_i)$

② $A_1, A_2, \cdots$ s.t. $A_{i+1} \subseteq A_i$; $i = 1, 2, \cdots$, then:

$P(A_n) \downarrow P(A)$; $A = \bigcap_{i=1}^{+\infty} A_i$

Proof of 1.5.3:

(Necessity)

$\forall$ seq $x_n \to +\infty$ as $n \to +\infty$

a. $\lim_{x\to+\infty} F(x) = 1$: $A_n = \{X \le x_n\} \Rightarrow A_n \uparrow \bigcup_{i=1}^{+\infty} A_i = \{X < +\infty\} = S$

By the axiom of continuity: $F_X(x_n) = P(A_n) \uparrow P(S) = 1$

$-\lim_{x\to-\infty} F(x) = 0$: $\forall$ seq $x_n \to -\infty$ as $n \to +\infty$

$A_n = \{X \le x_n\} \Rightarrow A_n \downarrow \bigcap_{i=1}^{+\infty} A_i = \phi$

By the axiom of continuity: $F_X(x_n) = P(A_n) \downarrow P(\phi) = 0$

b. F_X is nondecreasing:

$x \le y \Rightarrow \{s: X(s) \le x\} \subseteq \{s: X(s) \le y\}$

$\Rightarrow P(\quad\downarrow\quad) \le P(\quad\downarrow\quad)$

$F_X(x) \qquad F_X(y)$

c. Right continuous: NTS: $\lim_{n\to+\infty} F_X(y_n) = F_X(x)$; $\forall y_n \downarrow x$

$A_n = \{X \le y_n\}$ $A_n \downarrow \bigcap_{i=1}^{+\infty} A_i = \{X \le x\}$

By the axiom of continuity: $F_X(y_n) = P(A_n) \downarrow P(\{X \le x\}) = F_X(x)$

- $F_X(\cdot)$ has left limit:

$y_n \uparrow x$ as $n \to +\infty$; $A_n = \{X \le y_n\}$ $A_n \uparrow \bigcup_{i=1}^{+\infty} A_i = \{X < x\}$

By the axiom of continuity: $F_X(y_n) = P(A_n) \uparrow P(\{X < x\})$

Example 1.5.4 (Tossing for a head) Suppose we do an experiment that consists of tossing a coin until a head appears. Let p = probability of a head on any given toss, and define a random variable X = number of tosses required to get a head. Then, for any $x = 1, 2, \ldots,$

$$P(X = x) = (1-p)^{x-1}p, \tag{1.5.2}$$

since we must get $x - 1$ tails followed by a head for the event to occur and all trials are independent. From (1.5.2) we calculate, for any positive integer x,

$$P(X \le x) = \sum_{i=1}^{x} P(X = i) = \sum_{i=1}^{x} (1-p)^{i-1}p. \tag{1.5.3}$$

The partial sum of the geometric series is

$$\sum_{k=1}^{n} t^{k-1} = \frac{1-t^n}{1-t}, \quad t \neq 1, \tag{1.5.4}$$

a fact that can be established by induction (see Exercise 1.50). Applying (1.5.4) to our probability, we find that the cdf of the random variable X is

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= \frac{1-(1-p)^x}{1-(1-p)}p \\ &= 1-(1-p)^x, \quad x = 1, 2, \ldots . \end{aligned}$$

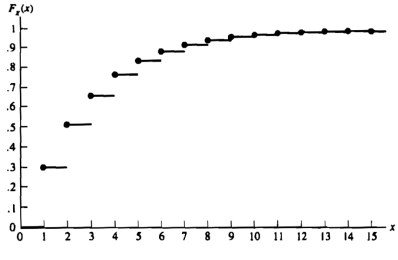


Figure 1.5.2. *Geometric cdf, p = .3*

Definition 1.5.7 A random variable X is *continuous* if $F_X(x)$ is a continuous function of x. A random variable X is *discrete* if $F_X(x)$ is a step function of x.

Definition 1.5.8 The random variables X and Y are *identically distributed* if, for every set $A \in \mathcal{B}^1$, $P(X \in A) = P(Y \in A)$.

Example 1.5.9 (Identically distributed random variables) Consider the experiment of tossing a fair coin three times as in Example 1.4.3. Define the random variables X and Y by

$$X = \text{number of heads observed} \quad \text{and} \quad Y = \text{number of tails observed.}$$

The distribution of X is given in Example 1.4.3, and it is easily verified that the distribution of Y is exactly the same. That is, for each $k = 0, 1, 2, 3$, we have $P(X = k) = P(Y = k)$. So X and Y are identically distributed. However, for no sample points do we have $X(s) = Y(s)$. ‖

Theorem 1.5.10 *The following two statements are equivalent:*

a. *The random variables X and Y are identically distributed.*

b. $F_X(x) = F_Y(x)$ *for every x.*

Definition 1.6.1 The *probability mass function* (*pmf*) of a discrete random variable X is given by

$$f_X(x) = P(X = x) \quad \text{for all } x.$$

Example 1.6.2 (Geometric probabilities) For the geometric distribution of Example 1.5.4, we have the pmf

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

Sep 2nd 2021

Definition 1.6.3 The *probability density function* or *pdf*, $f_X(x)$, of a continuous random variable X is the function that satisfies

NOT Unique

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \quad \text{for all } x. \tag{1.6.3}$$

Using the Fundamental Theorem of Calculus, if $f_X(x)$ is continuous, we have the further relationship

$$\frac{d}{dx}F_X(x) = f_X(x). \tag{1.6.2}$$

Theorem 1.6.5 *A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if*

a. $f_X(x) \geq 0$ *for* ~~all~~ x. a.s.

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

Chapter 2: Transformations and Expectations

$Y = g(X)$

Def: $\mathcal{X} = \{x : f_X(x) > 0\}$ support of r.v. X.

(2.1.7) $\mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}$ support of r.v. Y.

- If g is monotone:

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

$$\Rightarrow F_Y(y) = \int_{\{x \in \mathcal{X} \mid x \le g^{-1}(y)\}} f_X(x)dx = \int_{-\infty}^{g^{-1}(y)} f_X(x)dx = F_X(g^{-1}(y))$$

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

$$\Rightarrow F_Y(y) = \int_{g^{-1}(y)}^{+\infty} f_X(x)dx = 1 - F_X(g^{-1}(y))$$

Theorem 2.1.5 *Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined by (2.1.7). Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \dfrac{d}{dy} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases} \tag{2.1.10}$$

Example 2.1.4 (Uniform-exponential relationship–I) Suppose $X \sim f_X(x) = 1$ if $0 < x < 1$ and 0 otherwise, the *uniform*$(0,1)$ *distribution*. It is straightforward to check that $F_X(x) = x$, $0 < x < 1$. We now make the transformation $Y = g(X) = -\log X$. Since

$$\frac{d}{dx} g(x) = \frac{d}{dx}(-\log x) = \frac{-1}{x} < 0, \qquad \text{for} \quad 0 < x < 1,$$

$g(x)$ is a decreasing function. As X ranges between 0 and 1, $-\log x$ ranges between 0 and ∞, that is, $\mathcal{Y} = (0, \infty)$. For $y > 0$, $y = -\log x$ implies $x = e^{-y}$, so $g^{-1}(y) = e^{-y}$. Therefore, for $y > 0$,

$$F_Y(y) = 1 - F_X\left(g^{-1}(y)\right) = 1 - F_X(e^{-y}) = 1 - e^{-y}. \qquad (F_X(x) = x)$$

Of course, $F_Y(y) = 0$ for $y \le 0$. Note that it was necessary only to verify that $g(x) = -\log x$ is monotone on $(0,1)$, the support of X. ‖

Example 2.1.1 (Binomial transformation) A discrete random variable X has a *binomial distribution* if its pmf is of the form

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \tag{2.1.3}$$

where n is a positive integer and $0 \le p \le 1$. Values such as n and p that can be set to different values, producing different probability distributions, are called *parameters*. Consider the random variable $Y = g(X)$, where $g(x) = n - x$; that is, $Y = n - X$. Here $\mathcal{X} = \{0, 1, \ldots, n\}$ and $\mathcal{Y} = \{y : y = g(x),\ x \in \mathcal{X}\} = \{0, 1, \ldots, n\}$. For any $y \in \mathcal{Y}$, $n - x = g(x) = y$ if and only if $x = n - y$. Thus, $g^{-1}(y)$ is the single point $x = n - y$, and

$$\begin{aligned} f_Y(y) &= \sum_{x \in g^{-1}(y)} f_X(x) \\ &= f_X(n-y) \\ &= \binom{n}{n-y} p^{n-y} (1-p)^{n-(n-y)} \\ &= \binom{n}{y} (1-p)^y p^{n-y}. \qquad \left(\begin{array}{c}\text{Definition 1.2.17} \\ \text{implies } \binom{n}{y} = \binom{n}{n-y}\end{array}\right) \end{aligned}$$

Thus, we see that Y also has a binomial distribution, but with parameters n and $1 - p$. ‖

Example 2.1.7 (Square transformation) Suppose X is a continuous random variable. For $y > 0$, the cdf of $Y = X^2$ is

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}).$$

Because x is continuous, we can drop the equality from the left endpoint and obtain

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

The pdf of Y can now be obtained from the cdf by differentiation:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} [F_X(\sqrt{y}) - F_X(-\sqrt{y})] \\ &= \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}), \end{aligned}$$

where we use the chain rule to differentiate $F_X(\sqrt{y})$ and $F_X(-\sqrt{y})$. Therefore, the pdf is

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right). \quad (y>0) \tag{2.1.11}$$

Notice that the pdf of Y in (2.1.11) is expressed as the sum of two pieces, pieces that represent the intervals where $g(x) = x^2$ is monotone. In general, this will be the case. ‖

Theorem 2.1.8 *Let X have pdf $f_X(x)$, let $Y = g(X)$, and define the sample space $\mathcal{X}$ as in (2.1.7). Suppose there exists a partition, $A_0, A_1, \ldots, A_k$, of $\mathcal{X}$ such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i. Further, suppose there exist functions $g_1(x), \ldots, g_k(x)$, defined on $A_1, \ldots, A_k$, respectively, satisfying*

i. $g(x) = g_i(x)$, *for* $x \in A_i$,

ii. $g_i(x)$ *is monotone on* A_i,

iii. *the set* $\mathcal{Y} = \{y\colon\ y = g_i(x)$ *for some* $x \in A_i\}$ *is the same for each* $i = 1, \ldots, k$, *and*

iv. $g_i^{-1}(y)$ *has a continuous derivative on* $\mathcal{Y}$, *for each* $i = 1, \ldots, k$.

Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X\left(g_i^{-1}(y)\right) \left|\frac{d}{dy} g_i^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases}$$

Sep 7th, 2021

Theorem 2.1.10 (Probability integral transformation) *Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is, $P(Y \le y) = y, 0 < y < 1$.*

Before we prove this theorem, we will digress for a moment and look at F_X^{-1}, the inverse of the cdf F_X, in some detail. If F_X is strictly increasing, then F_X^{-1} is well defined by

$$F_X^{-1}(y) = x \quad \Leftrightarrow \quad F_X(x) = y. \tag{2.1.12}$$

However, if F_X is constant on some interval, then F_X^{-1} is not well defined by (2.1.12), as Figure 2.1.2 illustrates. Any x satisfying $x_1 \le x \le x_2$ satisfies $F_X(x) = y$.

This problem is avoided by defining $F_X^{-1}(y)$ for $0 < y < 1$ by

$$F_X^{-1}(y) = \inf\{x\colon\ F_X(x) \ge y\}, \tag{2.1.13}$$

a definition that agrees with (2.1.12) when F_X is nonconstant and provides an F_X^{-1} that is single-valued even when F_X is not strictly increasing. Using this definition, in Figure 2.1.2b, we have $F_X^{-1}(y) = x_1$. At the endpoints of the range of y, $F_X^{-1}(y)$ can also be defined. $F_X^{-1}(1) = \infty$ if $F_X(x) < 1$ for all x and, for any F_X, $F_X^{-1}(0) = -\infty$.

Proof of Theorem 2.1.10: For $Y = F_X(X)$ we have, for $0 < y < 1$,

$$\begin{aligned} P(Y \le y) &= P(F_X(X) \le y) \\ &= P(F_X^{-1}[F_X(X)] \le F_X^{-1}(y)) && (F_X^{-1} \text{ is increasing}) \\ &= P(X \le F_X^{-1}(y)) && (\text{see paragraph below}) \\ &= F_X\left(F_X^{-1}(y)\right) && (\text{definition of } F_X) \\ &= y. && (\text{continuity of } F_X) \end{aligned}$$

At the endpoints we have $P(Y \le y) = 1$ for $y \ge 1$ and $P(Y \le y) = 0$ for $y \le 0$, showing that Y has a uniform distribution.

The reasoning behind the equality

$$P\left(F_X^{-1}(F_X(X)) \le F_X^{-1}(y)\right) = P(X \le F_X^{-1}(y))$$

is somewhat subtle and deserves additional attention. If F_X is strictly increasing, then it is true that $F_X^{-1}(F_X(x)) = x$. (Refer to Figure 2.1.2a.) However, if F_X is flat, it may be that $F_X^{-1}(F_X(x)) \neq x$. Suppose F_X is as in Figure 2.1.2b and let $x \in [x_1, x_2]$. Then $F_X^{-1}(F_X(x)) = x_1$ for any x in this interval. Even in this case, though, the probability equality holds, since $P(X \le x) = P(X \le x_1)$ for any $x \in [x_1, x_2]$. The flat cdf denotes a region of 0 probability ($P(x_1 < X \le x) = F_X(x) - F_X(x_1) = 0$). □

Definition 2.2.1 The *expected value* or *mean* of a random variable $g(X)$, denoted by $\mathrm{E}\,g(X)$, is

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

provided that the integral or sum exists. If $\mathrm{E}|g(X)| = \infty$, we say that $\mathrm{E}\,g(X)$ does not exist. (Ross 1988 refers to this as the "law of the unconscious statistician." We

Example 2.2.3 (Binomial mean) If X has a *binomial distribution*, its pmf is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n,$$

where n is a positive integer, $0 \le p \le 1$, and for every fixed pair n and p the pmf sums to 1. The expected value of a binomial random variable is given by

$$\mathrm{E}\,X = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

(the $x = 0$ term is 0). Using the identity $x\binom{n}{x} = n\binom{n-1}{x-1}$, we have

$$\begin{aligned} \mathrm{E}\,X &= \sum_{x=1}^{n} n \binom{n-1}{x-1} p^x (1-p)^{n-x} \\ &= \sum_{y=0}^{n-1} n \binom{n-1}{y} p^{y+1} (1-p)^{n-(y+1)} \qquad \text{(substitute } y = x-1\text{)} \\ &= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} \\ &= np, \end{aligned}$$

since the last summation must be 1, being the sum over all possible values of a binomial$(n-1,\ p)$ pmf. ‖

Example 2.2.4 (Cauchy mean) A classic example of a random variable whose expected value does not exist is a *Cauchy random variable*, that is, one with pdf

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

It is straightforward to check that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$, but $\mathrm{E}|X| = \infty$. Write

$$\mathrm{E}|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi}\frac{1}{1+x^2}\,dx = \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\,dx.$$

For any positive number M,

$$\int_0^M \frac{x}{1+x^2}\,dx = \frac{\log(1+x^2)}{2}\bigg|_0^M = \frac{\log(1+M^2)}{2}.$$

Thus,

$$\mathrm{E}|X| = \lim_{M\to\infty} \frac{2}{\pi}\int_0^M \frac{x}{1+x^2}\,dx = \frac{1}{\pi}\lim_{M\to\infty}\log(1+M^2) = \infty$$

and $\mathrm{E}\,X$ does not exist. ‖

Theorem 2.2.5 *Let X be a random variable and let a, b, and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

a. $\mathrm{E}(ag_1(X) + bg_2(X) + c) = a\mathrm{E}\,g_1(X) + b\mathrm{E}\,g_2(X) + c$.

b. *If $g_1(x) \geq 0$ for all x, then $\mathrm{E}\,g_1(X) \geq 0$.*

c. *If $g_1(x) \geq g_2(x)$ for all x, then $\mathrm{E}\,g_1(X) \geq \mathrm{E}\,g_2(X)$.*

d. *If $a \leq g_1(x) \leq b$ for all x, then $a \leq \mathrm{E}\,g_1(X) \leq b$.*

Proof: We will give details for only the continuous case, the discrete case being similar. By definition,

$$\begin{aligned}
&\mathrm{E}(ag_1(X) + bg_2(X) + c)\\
&\quad = \int_{-\infty}^{\infty} (ag_1(x) + bg_2(x) + c) f_X(x)\,dx\\
&\quad = \int_{-\infty}^{\infty} ag_1(x) f_X(x)\,dx + \int_{-\infty}^{\infty} bg_2(x) f_X(x)\,dx + \int_{-\infty}^{\infty} cf_X(x)\,dx
\end{aligned}$$

by the additivity of the integral. Since a, b, and c are constants, they factor out of their respective integrals and we have

$$\begin{aligned}
&\mathrm{E}(ag_1(X) + bg_2(X) + c)\\
&\quad = a\int_{-\infty}^{\infty} g_1(x) f_X(x)\,dx + b\int_{-\infty}^{\infty} g_2(x) f_X(x)\,dx + c\int_{-\infty}^{\infty} f_X(x)\,dx\\
&\quad = a\mathrm{E}\,g_1(X) + b\mathrm{E}\,g_2(x) + c,
\end{aligned}$$

establishing (a). The other three properties are proved in a similar manner. □

Definition 2.3.1 For each integer n, the *nth moment of* X (or $F_X(x)$), μ'_n, is

$$\mu'_n = \mathrm{E}\, X^n.$$

The *nth central moment of* X, μ_n, is

$$\mu_n = \mathrm{E}(X - \mu)^n,$$

where $\mu = \mu'_1 = \mathrm{E}\, X$.

Definition 2.3.2 The *variance* of a random variable X is its second central moment, $\mathrm{Var}\, X = \mathrm{E}(X - \mathrm{E}\, X)^2$. The positive square root of $\mathrm{Var}\, X$ is the *standard deviation* of X.

$SD(X) = \sqrt{Var(X)}$

Theorem 2.3.4 *If X is a random variable with finite variance, then for any constants a and b,*

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

$SD(aX+b) = |a| SD(X)$

Proof: From the definition, we have

$$\begin{aligned}
\mathrm{Var}(aX + b) &= \mathrm{E}\left((aX + b) - \mathrm{E}(aX + b)\right)^2 \\
&= \mathrm{E}(aX - a\mathrm{E}\, X)^2 \qquad (\mathrm{E}(aX + b) = a\mathrm{E}\, X + b) \\
&= a^2 \mathrm{E}(X - \mathrm{E}\, X)^2 \\
&= a^2\, \mathrm{Var}\, X. \qquad \square
\end{aligned}$$

Another formula for Var X:

$$\begin{aligned}
\mathrm{Var}\, X = \mathrm{E}(X - \mathrm{E}\, X)^2 &= \mathrm{E}[X^2 - 2X\mathrm{E}\, X + (\mathrm{E}\, X)^2] \\
&= \mathrm{E}\, X^2 - 2(\mathrm{E}\, X)^2 + (\mathrm{E}\, X)^2 \\
&= \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2, \qquad \Rightarrow \ (EX)^2 \le EX^2
\end{aligned}$$

(special case of Jensen's Inequa)

Sep. 9th 2021

Example 2.3.5 (Binomial variance) Let $X \sim$ binomial(n, p), that is,

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

We have previously seen that $\mathrm{E}\, X = np$. To calculate $\mathrm{Var}\, X$ we first calculate $\mathrm{E}\, X^2$. We have

$$\mathrm{E}\, X^2 = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x}. \tag{2.3.2}$$

In order to sum this series, we must first manipulate the binomial coefficient in a manner similar to that used for $\mathrm{E}\, X$ (Example 2.2.3). We write

$$x^2 \binom{n}{x} = x \frac{n!}{(x-1)!(n-x)!} = xn \binom{n-1}{x-1}. \tag{2.3.3}$$

The summand in (2.3.2) corresponding to $x = 0$ is 0, and using (2.3.3), we have

$$\begin{aligned}
\mathrm{E}\,X^2 &= n\sum_{x=1}^{n} x\binom{n-1}{x-1}p^x(1-p)^{n-x} \\
&= n\sum_{y=0}^{n-1}(y+1)\binom{n-1}{y}p^{y+1}(1-p)^{n-1-y} \qquad \text{(setting } y=x-1\text{)} \\
&= np\sum_{y=0}^{n-1} y\binom{n-1}{y}p^y(1-p)^{n-1-y} + np\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y}.
\end{aligned}$$

Now it is easy to see that the first sum is equal to $(n-1)p$ (since it is the mean of a binomial$(n-1,p)$), while the second sum is equal to 1. Hence,

(2.3.4) $$\mathrm{E}\,X^2 = n(n-1)p^2 + np.$$

Using (2.3.1), we have

$$\mathrm{Var}\,X = n(n-1)p^2 + np - (np)^2 = -np^2 + np = np(1-p). \qquad \|$$

Definition 2.3.6 Let X be a random variable with cdf F_X. The *moment generating function (mgf) of* X (or F_X), denoted by $M_X(t)$, is

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

provided that the expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, $\mathrm{E}e^{tX}$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx \qquad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx} P(X = x) \qquad \text{if } X \text{ is discrete.}$$

Theorem. Suppose there exists a $t^*>0$ such that $M_X(t^*)<+\infty$, $M_X(-t^*)<+\infty$. Then,

$\forall t_0 \in (-\frac{t^*}{2}, \frac{t^*}{2})$: $\frac{d^n}{dt^n} M_X(t)\Big|_{t=t_0} = E(X^n \cdot e^{t_0 X})$, $n=1,2,3,\cdots$

In particular, when $t_0 = 0$:

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

Lemma (Dominated Convergence Theorem) (DCT):

Suppose $\frac{\partial}{\partial t} f(t,x)$ exists $\forall t \in (t_0-\delta, t_0+\delta)$ & $\forall x \in \mathbb{R}$

for some $\delta>0$. If:

① $\left|\frac{d}{dt}f(t,x)\right| \le g(x)$ (uniformly over $t\in(t_0-\delta, t_0+\delta)$)

② $\int_{\mathbb{R}} g(x)dx < +\infty$

then $\left(\frac{d}{dt}\int_{-\infty}^{+\infty} f(t,x)dx\right)\Big|_{t=t_0} = \int_{-\infty}^{+\infty}\left(\frac{d}{dt}f(t,x)\Big|_{t=t_0}\right)dx$

Pf: Start with $n=1$

$$\frac{d}{dt}M_X(t)\Big|_{t=t_0} = \frac{d}{dt}\left(\int_{\mathbb{R}} e^{tx}f_X(x)dx\right)\Big|_{t=t_0}$$

$$\overset{DCT}{=} \int_{-\infty}^{+\infty}\frac{d}{dt}e^{tx}f_X(x)\Big|_{t=t_0}dx$$

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

$$= E(X\cdot e^{tX})$$

Verify DCT conditions: $f(t,x)=e^{tx}f_X(x)$

$$\left|\frac{d}{dt}f(t,x)\right| = \left|xe^{tx}f_X(x)\right| = |x|\cdot e^{tx}f_X(x)$$

$$\le |x|\cdot e^{\frac{t^*}{2}|x|}f_X(x) \quad ;\forall t\in\left(-\tfrac{t^*}{2},\tfrac{t^*}{2}\right)$$

Now, since $t_0\in(-\frac{t^*}{2},\frac{t^*}{2})$, we can choose δ small enough such that $(t_0-\delta, t_0+\delta)\subseteq(-\frac{t^*}{2},\frac{t^*}{2})$

Since $|y|\le e^{|y|}$, $\forall y\in\mathbb{R}$, we have: $\left(\left|\frac{t^*}{2}x\right|\le e^{\frac{t^*}{2}|x|}\right)$

$$\left|\frac{d}{dt}f(t,x)\right| \le \frac{2}{t^*}e^{t^*|x|}f_X(x)$$

$$\le \frac{2}{t^*}\left(e^{t^*x}+e^{-t^*x}\right)f_X(x) =: g(x)$$

$$\int_{\mathbb{R}} g(x)dx = \frac{2}{t^*}\int_{\mathbb{R}}\left(e^{t^*x}+e^{-t^*x}\right)f_X(x)dx$$

$$= \frac{2}{t^*}\left(M_X(t^*)+M_X(-t^*)\right) < +\infty$$

Proceed the proof by induction:

Suppose: $\frac{d^n}{dt^n} M_X(t)\Big|_{t=t_0} = E(X^n e^{t_0 X})$; $\forall t_0 \in (-\frac{t^*}{2}, \frac{t^*}{2})$

Then: $\frac{d^{n+1}}{dt^{n+1}} M_X(t)\Big|_{t=t_0} = \frac{d}{dt} E(X^n e^{t_0 X})\Big|_{t=t_0}$

$= \left(\frac{d}{dt}\int_{\mathbb{R}} \underbrace{x^n e^{tx} f_X(x)}_{\triangleq f(t,x)} dx\right)\Big|_{t=t_0}$

Let δ be chosen such that $(t_0-\delta, t_0+\delta) \subseteq (-\frac{t^*}{2}, \frac{t^*}{2})$, then:

$\forall t \in (t_0-\delta, t_0+\delta)$: $\left|\frac{d}{dt} f(t,x)\right| = \left|\frac{d}{dt} x^n e^{tx} f_X(x)\right|$

$= |x|^{n+1} e^{tx} f_X(x) \le |x|^{n+1} \cdot e^{\frac{t^*}{2}|x|} f_X(x)$

$\left(|y|^{n+1} \le (n+1)! e^{|y|}\right)$ $\le \frac{2^{n+1}}{(t^*)^{n+1}} e^{t^*|x|} \cdot f_X(x)$

$\le \frac{2^{n+1}}{(t^*)^{n+1}} (e^{t^* x} + e^{-t^* x}) f_X(x) =: g(x)$

$\int_{\mathbb{R}} g(x)dx = \frac{2^{n+1}}{(t^*)^{n+1}} (M_X(t^*) + M_X(-t^*)) < +\infty$

Sep 14th. 2021

Theorem 2.3.11 *Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.*

a. *If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only if $\mathrm{E}X^r = \mathrm{E}Y^r$ for all integers $r = 0, 1, 2, \ldots$.*

b. *If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.*

Example 2.3.10 (Nonunique moments) Consider the two pdfs given by

$$f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}, \qquad 0 \le x < \infty,$$

$$f_2(x) = f_1(x)[1 + \sin(2\pi \log x)], \qquad 0 \le x < \infty.$$

(The pdf f_1 is a special case of a *lognormal pdf.*)

It can be shown that if $X_1 \sim f_1(x)$, then

$$\mathrm{E}X_1^r = e^{r^2/2}, \qquad r = 0, 1, \ldots,$$

so X_1 has all of its moments. Now suppose that $X_2 \sim f_2(x)$. We have

$$\begin{aligned} \mathrm{E}\, X_2^r &= \int_0^\infty x^r f_1(x)[1 + \sin(2\pi \log x)]\, dx \\ &= \mathrm{E}\, X_1^r + \int_0^\infty x^r f_1(x) \sin(2\pi \log x)\, dx. \end{aligned}$$

$= EX_1^r$

$= 0$ (see below)

However, the transformation $y = \log x - r$ shows that this last integral is that of an odd function over $(-\infty, \infty)$ and hence is equal to 0 for $r = 0, 1, \ldots$. Thus, even though X_1 and X_2 have distinct pdfs, they have the same moments for all r. The two pdfs are pictured in Figure 2.3.2.

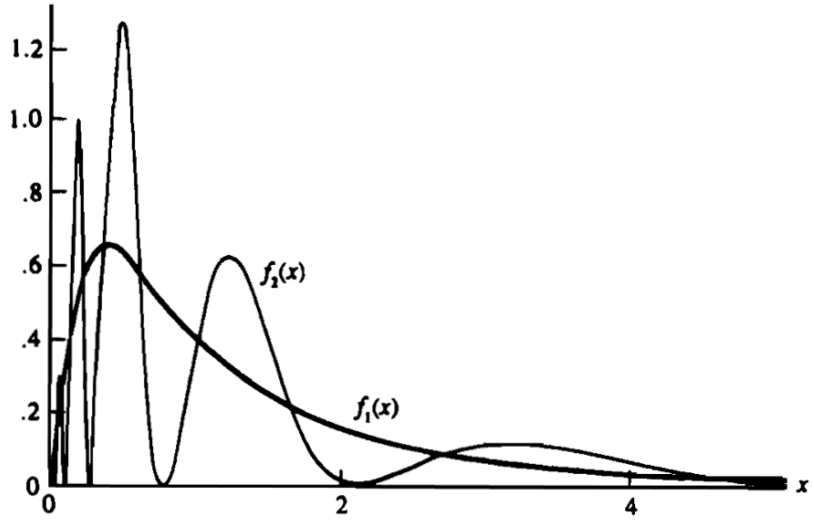


Figure 2.3.2. *Two pdfs with the same moments:* $f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}$ *and* $f_2(x) = f_1(x)[1 + \sin(2\pi \log x)]$

$$\int_0^\infty x^r f_1(x) \sin(2\pi \log x)\, dx \overset{y=\log x - r}{=} \int_{-\infty}^{+\infty} (e^{y+r})^r \frac{1}{\sqrt{2\pi}e^{y+r}} e^{-\frac{(y+r)^2}{2}} \cdot \sin(2\pi(y+r)) e^{y+r}\, dy$$

$= \sin(2\pi y)$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sin(2\pi y) \cdot e^{ry + r^2 - \frac{1}{2}y^2 - ry - \frac{1}{2}r^2}\, dy$$

$$= \frac{e^{\frac{r^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} \sin(2\pi y)\, dy \quad = 0$$

odd function & absolute integrable

Theorem 2.3.12 (Convergence of mgfs) *Suppose* $\{X_i, i = 1, 2, \ldots\}$ *is a sequence of random variables, each with mgf* $M_{X_i}(t)$. *Furthermore, suppose that*

$$\lim_{i \to \infty} M_{X_i}(t) = M_X(t), \qquad \text{for all } t \text{ in a neighborhood of } 0,$$

and $M_X(t)$ *is an mgf. Then there is a unique cdf* F_X *whose moments are determined by* $M_X(t)$ *and, for all* x *where* $F_X(x)$ *is continuous, we have*

$$\lim_{i \to \infty} F_{X_i}(x) = F_X(x).$$

That is, convergence, for $|t| < h$, *of mgfs to an mgf implies convergence of cdfs.*

Theorem 2.3.15 *For any constants a and b, the mgf of the random variable $aX+b$ is given by*

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

Proof: By definition,

$$\begin{aligned} M_{aX+b}(t) &= \mathrm{E}\left(e^{(aX+b)t}\right) \\ &= \mathrm{E}\left(e^{(aX)t}e^{bt}\right) && \text{(properties of exponentials)} \\ &= e^{bt}\mathrm{E}\left(e^{(at)X}\right) && (e^{bt} \text{ is constant}) \\ &= e^{bt}M_X(at), && \text{(definition of mgf)} \end{aligned}$$

proving the theorem. □

✳ Chapter 3: Common Families of Distributions

Discrete Distributions

① *Discrete Uniform Distribution*

A random variable X has a *discrete uniform $(1, N)$ distribution* if

$$P(X = x|N) = \frac{1}{N}, \quad x = 1, 2, \ldots, N, \tag{3.2.1}$$

We then have

$$\mathrm{E}X = \sum_{x=1}^{N} xP(X = x|N) = \sum_{x=1}^{N} x\frac{1}{N} = \frac{N+1}{2}$$

and

$$\mathrm{E}X^2 = \sum_{x=1}^{N} x^2\frac{1}{N} = \frac{(N+1)(2N+1)}{6},$$

and so

$$\begin{aligned} \mathrm{Var}\, X &= \mathrm{E}X^2 - (\mathrm{E}X)^2 \\ &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{(N+1)(N-1)}{12}. \end{aligned}$$

(2) Hypergeometric Distribution

Suppose we have a large urn filled with N balls that are identical in every way except that M are red and $N-M$ are green. We reach in, blindfolded, and select K balls at random (the K balls are taken all at once, a case of sampling without replacement). What is the probability that exactly x of the balls are red?

The total number of samples of size K that can be drawn from the N balls is $\binom{N}{K}$, as was discussed in Section 1.2.3. It is required that x of the balls be red, and this can be accomplished in $\binom{M}{x}$ ways, leaving $\binom{N-M}{K-x}$ ways of filling out the sample with $K-x$ green balls. Thus, if we let X denote the number of red balls in a sample of size K, then X has a *hypergeometric distribution* given by

$$P(X=x|N,M,K)=\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \quad x=0,1,\ldots,K. \tag{3.2.2}$$

Note that there is, implicit in (3.2.2), an additional assumption on the range of X. Binomial coefficients of the form $\binom{n}{r}$ have been defined only if $n \geq r$, and so the range of X is additionally restricted by the pair of inequalities

$$M \geq x \quad \text{and} \quad N-M \geq K-x,$$

which can be combined as

$$M-(N-K) \leq x \leq M.$$

In many cases K is small compared to M and N, so the range $0 \leq x \leq K$ will be contained in the above range and, hence, will be appropriate. The formula for the hypergeometric probability function is usually quite difficult to deal with. In fact, it is not even trivial to verify that

$$\sum_{x=0}^{K} P(X=x)=\sum_{x=0}^{K}\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}=1.$$

The hypergeometric distribution illustrates the fact that, statistically, dealing with finite populations (finite N) is a difficult task.

The mean of the hypergeometric distribution is given by

$$\mathrm{E}X=\sum_{x=0}^{K} x\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}=\sum_{x=1}^{K} x\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}. \qquad \text{(summand is 0 at } x=0\text{)}$$

To evaluate this expression, we use the identities (already encountered in Section 2.3)

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

$$\binom{N}{K}=\frac{N}{K}\binom{N-1}{K-1},$$

and obtain

$$EX = \sum_{x=1}^{K} \frac{M\binom{M-1}{x-1}\binom{N-M}{K-x}}{\frac{N}{K}\binom{N-1}{K-1}} = \frac{KM}{N}\sum_{x=1}^{K} \frac{\binom{M-1}{x-1}\binom{N-M}{K-x}}{\binom{N-1}{K-1}}.$$

We now can recognize the second sum above as the sum of the probabilities for another hypergeometric distribution based on parameter values $N-1$, $M-1$, and $K-1$. This can be seen clearly by defining $y = x-1$ and writing

$$\sum_{x=1}^{K} \frac{\binom{M-1}{x-1}\binom{N-M}{K-x}}{\binom{N-1}{K-1}} = \sum_{y=0}^{K-1} \frac{\binom{M-1}{y}\binom{(N-1)-(M-1)}{K-1-y}}{\binom{N-1}{K-1}}$$

$$= \sum_{y=0}^{K-1} P(Y = y|N-1, M-1, K-1) = 1,$$

where Y is a hypergeometric random variable with parameters $N-1$, $M-1$, and $K-1$. Therefore, for the hypergeometric distribution,

$$EX = \frac{KM}{N}.$$

A similar, but more lengthy, calculation will establish that

$$\operatorname{Var} X = \frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right).$$

$$E(X(X-1)) = \sum_{x=2}^{k} x(x-1)\frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}}$$

$$= \sum_{x=2}^{k} x(x-1)\frac{M(M-1)}{x(x-1)}\frac{\binom{M-2}{x-2}\binom{N-M}{k-x}}{\binom{N}{k}}$$

$$y = x-2 \quad = \frac{M(M-1)}{\binom{N}{k}}\sum_{y=0}^{k-2}\binom{M-2}{y}\binom{(N-2)-(M-2)}{k-2-y}$$

$$= \frac{M(M-1)}{\binom{N}{k}}\binom{N-2}{k-2} = \frac{M(M-1)k(k-1)}{N(N-1)}$$

$$\operatorname{Var}(X) = E(X(X-1)) + EX - (EX)^2 = \cdots = \frac{KM}{N}\left(1-\frac{M}{N}\right)\frac{N-K}{N-1}$$

③ *Binomial Distribution*

A random variable X has a *Bernoulli*(p) *distribution* if

(3.2.3)
$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p, \end{cases} \qquad 0 \le p \le 1.$$

$$EX = 1p + 0(1-p) = p,$$

$$\operatorname{Var} X = (1-p)^2 p + (0-p)^2(1-p) = p(1-p).$$

In a sequence of n identical, independent Bernoulli trials, each with success probability p, define the random variables $X_1, \ldots, X_n$ by

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

The random variable

$$Y = \sum_{i=1}^{n} X_i$$

has the binomial(n, p) distribution.

$$P(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, 2, \ldots, n,$$

$$\mathrm{E}X = np, \quad \mathrm{Var}\, X = np(1-p). \quad M_X(t) = [pe^t + (1-p)]^n.$$

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx} \cdot P(X=x)$$

$$= \sum_{x=0}^{n} \binom{n}{x} e^{tx} \cdot p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x}$$

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

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = n \cdot pe^t (pe^t + 1 - p)^{n-1} \Big|_{t=0} = np$$

Theorem: Suppose $X \sim B(n,p)$, $Y \sim B(m,p)$. Also, X and Y are independent. Then, $Z = X + Y \sim B(m+n, p)$

Pf: $P(Z=z) = \sum_{k=0}^{n} P(X=k, Y=z-k)$

$$= \sum_{k=0}^{n} P(X=k) P(Y=z-k)$$

$$= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \binom{m}{z-k} p^{z-k} (1-p)^{m-(z-k)}$$

$$= p^z (1-p)^{n+m-z} \sum_{k=0}^{z} \binom{n}{k} \binom{m}{z-k}$$

$$= \binom{n+m}{z} p^z (1-p)^{n+m-z}$$

(4) Poisson Distribution

A random variable X, taking values in the nonnegative integers, has a *Poisson(λ) distribution* if

(3.2.5) $$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

$$\begin{aligned}
EX &= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} \\
&= \lambda
\end{aligned}$$

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

$M_X(t) = \sum_{x=0}^{+\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!}$

$= e^{-\lambda}\sum_{x=0}^{+\infty}\frac{(\lambda e^t)^x}{x!}$

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

$Var(X) = \lambda$.

$X \sim B(n,p)$. Suppose $n \to +\infty$, $p \to 0$, $n\cdot p \to \lambda > 0$

Now: $P(X=x) = \binom{n}{x}p^x(1-p)^{n-x}$; $x = 0, 1, 2, \cdots$

$= \frac{1}{x!}\cdot\frac{n!}{(n-x)!}\cdot\frac{(np)^x}{n^x}\cdot\left(1-\frac{np}{n}\right)^{n-x}$

$= \frac{(np)^x}{x!}\cdot\frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x}\cdot\left(1-\frac{np}{n}\right)^n(1-p)^{-x}$

$\frac{(np)^x}{x!} \to \frac{\lambda^x}{x!}$; $\frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x} \to 1$; $\left(1-\frac{np}{n}\right)^n \to e^{-\lambda}$; $(1-p)^{-x} \to 1$

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

*Sep. 16th 2021

Example 3.2.4 (Waiting time) As an example of a waiting-for-occurrence application, consider a telephone operator who, on the average, handles five calls every 3 minutes. What is the probability that there will be no calls in the next minute? At least two calls?

If we let X = number of calls in a minute, then X has a Poisson distribution with $EX = \lambda = \frac{5}{3}$. So

$$\begin{aligned}
P(\text{no calls in the next minute}) &= P(X = 0) \\
&= \frac{e^{-5/3}\left(\frac{5}{3}\right)^0}{0!} \\
&= e^{-5/3} = .189;
\end{aligned}$$

$$\begin{aligned}
P(\text{at least two calls in the next minute}) &= P(X \geq 2) \\
&= 1 - P(X = 0) - P(X = 1) \\
&= 1 - .189 - \frac{e^{-5/3}\left(\frac{5}{3}\right)^1}{1!} \\
&= .496.
\end{aligned}$$

‖

Calculation of Poisson probabilities can be done rapidly by noting the following recursion relation:

$$P(X = x) = \frac{\lambda}{x} P(X = x - 1), \quad x = 1, 2, \ldots. \tag{3.2.6}$$

This relation is easily proved by writing out the pmf of the Poisson. Similar relations hold for other discrete distributions. For example, if $Y \sim \text{binomial}(n, p)$, then

$$P(Y = y) = \frac{(n - y + 1)}{y} \frac{p}{1 - p} P(Y = y - 1). \tag{3.2.7}$$

The recursion relations (3.2.6) and (3.2.7) can be used to establish the Poisson approximation to the binomial, which we have already seen in Section 2.3, where the approximation was justified using mgfs. Set $\lambda = np$ and, if p is small, we can write

$$\frac{n - y + 1}{y} \frac{p}{1 - p} = \frac{np - p(y - 1)}{y - py} \approx \frac{\lambda}{y}$$

since, for small p, the terms $p(y - 1)$ and py can be ignored. Therefore, to this level of approximation, (3.2.7) becomes

$$P(Y = y) = \frac{\lambda}{y} P(Y = y - 1), \tag{3.2.8}$$

which is the Poisson recursion relation. To complete the approximation, we need only establish that $P(X = 0) \approx P(Y = 0)$, since all other probabilities will follow from (3.2.8). Now

$$P(Y = 0) = (1 - p)^n = \left(1 - \frac{np}{n}\right)^n = \left(1 - \frac{\lambda}{n}\right)^n$$

upon setting $np = \lambda$. Recall from Section 2.3 that for fixed λ, $\lim_{n \to \infty} (1 - (\lambda/n))^n = e^{-\lambda}$, so for large n we have the approximation

$$P(Y = 0) = \left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda} = P(X = 0),$$

completing the Poisson approximation to the binomial.

The approximation is valid when n is large and p is small, which is exactly when it is most useful, freeing us from calculation of binomial coefficients and powers for large n.

$X \sim B(n,p) \Rightarrow M_X(t) = (pe^t + 1 - p)^n = (1 + \frac{1}{n}(e^t - 1)np)^n$

$\xrightarrow[\substack{p \to 0 \\ np \to \lambda}]{\text{as } n \to \infty} e^{\lambda(e^t - 1)}$

Example 3.2.5 (Poisson approximation) A typesetter, on the average, makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

If we assume that setting a word is a Bernoulli trial with success probability $p = \frac{1}{500}$ (notice that we are labeling an error as a "success") and that the trials are independent, then X = number of errors in five pages (1500 words) is binomial$(1500, \frac{1}{500})$. Thus

$$\begin{aligned} P(\text{no more than two errors}) &= P(X \le 2) \\ &= \sum_{x=0}^{2} \binom{1500}{x} \left(\frac{1}{500}\right)^x \left(\frac{499}{500}\right)^{1500-x} \\ &= .4230, \end{aligned}$$

which is a fairly cumbersome calculation. If we use the Poisson approximation with $\lambda = 1500(\frac{1}{500}) = 3$, we have

$$P(X \le 2) \approx e^{-3}\left(1 + 3 + \frac{3^2}{2}\right) = .4232. \quad \|$$

⑤ *Geometric Distribution*

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

which defines the pmf of a *geometric random variable* X with success probability p. X can be interpreted as the trial at which the first success occurs

The instructor's def of G(p): $P(X=x) = p(1-p)^x$, $x = 0, 1, 2, \cdots$

$$EX = \sum_{x=1}^{+\infty} x \cdot p \cdot (1-p)^x = p \sum_{x=1}^{+\infty} \left(\sum_{y=1}^{x} 1\right)(1-p)^x = p \sum_{x=1}^{+\infty} \sum_{y=1}^{x} (1-p)^x$$

$$= p \sum_{y=1}^{+\infty} \sum_{x=y}^{+\infty} (1-p)^x = p \cdot \sum_{y=1}^{+\infty} \frac{(1-p)^y}{1-(1-p)}$$

$$= \sum_{y=1}^{+\infty} (1-p)^y = \frac{1-p}{p}$$

$$M_X(t) = \sum_{x=0}^{+\infty} e^{tx} p(1-p)^x = p \sum_{x=0}^{+\infty} \left((1-p)e^t\right)^x$$

$$= \begin{cases} \frac{p}{1-(1-p)e^t}, & \text{if } (1-p)e^t < 1 \text{ or } t < -\log(1-p) \\ +\infty & \text{o.w.} \end{cases}$$

$$EX = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{p(1-p)e^t}{(1-(1-p)e^t)^2}\right|_{t=0} = \frac{p(1-p)}{p^2} = \frac{1-p}{p}$$

$$EX^2 = \left.\frac{d^2 M_X(t)}{dt^2}\right|_{t=0} = \left.\frac{p(1-p)e^t}{(1-(1-p)e^t)^2} + \frac{2p(1-p)^2 e^{2t}}{(1-(1-p)e^t)^3}\right|_{t=0}$$

$$= \frac{p(1-p)}{p^2} + \frac{2p(1-p)^2}{p^3} = \frac{1-p}{p} + \frac{2(1-p)^2}{p^2}$$

$$\mathrm{Var}(X) = EX^2 - (EX)^2 = \frac{1-p}{p^2}$$

The geometric distribution has an interesting property, known as the "memoryless" property. For integers $s > t$, it is the case that

$$P(X > s | X > t) = P(X > s - t); \tag{3.2.11}$$

Tue. Sep 21st. 2021

$$P(X \geq t+s \mid X \geq t) = P(X \geq s), \ \forall s,t \in \mathbb{N}_0$$

Pf: $P(X \geq x) = \sum_{t=x}^{+\infty} p(1-p)^t = p\frac{(1-p)^x}{1-(1-p)} = (1-p)^x$

$$P(X \geq t+s \mid X \geq t) = \frac{P(X \geq t+s,\ X \geq t)}{P(X \geq t)}$$

$$= \frac{P(X \geq t+s)}{P(X \geq t)}$$

$$= (1-p)^{t+s} / (1-p)^t$$

$$= (1-p)^s = P(X \geq s)$$

Thm. Suppose X is a r.v. whose range is inside integer numbers, i.e., $\mathrm{Range}(X) \subseteq \mathbb{Z}$, and $P(X \geq 0) > 0$. Now, if

$$P(X \geq t+s \mid X \geq t) = P(X \geq s) \ ; \forall s,t \geq 0,$$

then $X \sim \mathrm{Geometric}(P(X=0))$.

Pf: Use MP with $t=s=0$: $P(X \geq 0+0 \mid X \geq 0) = P(X \geq 0)$

$$\frac{P(X \geq 0)}{P(X \geq 0)} = 1 \quad \therefore P(X \geq 0) = 1$$

Use MP with $t=s=1$:

$$\frac{P(X \geq 2)}{P(X \geq 1)} = P(X \geq 2 \mid X \geq 1) = P(X \geq 1) \Rightarrow P(X \geq 2) = (P(X \geq 1))^2$$

By induction: $P(X\geq n)=(P(X\geq 1))^n \;;\forall n\geq 1$

$\left(\frac{P(X\geq n+1)}{P(X\geq 1)}=P(X\geq n+1|X\geq 1)=P(X\geq n)=(P(X\geq 1))^n \Rightarrow P(X\geq n+1)=(P(X\geq 1))^{n+1}\right)$

$$\begin{aligned}P(X=n)&=P(X\geq n)-P(X\geq n+1)\\&=(P(X\geq 1))^n-(P(X\geq 1))^{n+1}\\&=(P(X\geq 1))^n(1-P(X\geq 1))\\&=(1-P(X=0))^n P(X=0)\end{aligned}$$

⑥ ***Negative Binomial Distribution***

Repeat independently a Bernoulli(p)

Define X = #failure before the rth success

$X\sim$ Neg Binom(r,p)

Range$(X)=\{0,1,2,\dots\}$

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

$M_X(t)=\sum_{x=0}^{+\infty} e^{tx}\binom{x+r-1}{x}p^r(1-p)^x$

$=p^r\sum_{x=0}^{+\infty}\binom{x+r-1}{x}((1-p)e^t)^x$

negative binomial formula

$|a|<1$ & $r\in\mathbb{N}$

$(1-a)^{-r}=\sum_{x=0}^{+\infty}\binom{x+r-1}{x}a^x$

$=p^r(1-(1-p)e^t)^{-r}$, if $(1-p)e^t<1$ or $t<-\log(1-p)$

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0}=p^r(-r)(-(1-p)e^t)(1-(1-p)e^t)^{-r-1}\Big|_{t=0}$$

$$=rp^r(1-p)e^t(1-(1-p)e^t)^{-r-1}\Big|_{t=0}$$

$$=r\frac{1-p}{p}=EX$$

$$\left.\frac{d^2M_X(t)}{dt}\right|_{t=0}=rp^r(1-p)\left[e^t(1-(1-p)e^t)^{-r-1}+e^t(-r-1)(-(1-p)e^t)(1-(1-p)e^t)^{-r-2}\right]\Big|_{t=0}$$

$$=rp^r(1-p)\left(\frac{1}{p^{r+1}}+\frac{(r+1)(1-p)}{p^{r+2}}\right)$$

$$=\frac{r^2(1-p)^2}{p^2}+r\frac{1-p}{p^2}$$

$$\Rightarrow \mathrm{Var}(X)=EX^2-(EX)^2=r\frac{1-p}{p^2}$$

Proof of Negative Binomial Formula:

$f_m(z)=\sum_{k=0}^{+\infty}\binom{k+m}{m}z^k$

Note that $\binom{k+m}{m} = \binom{k+m-1}{m-1} + \binom{k+m-1}{m}$

$\Rightarrow f_m(z) = f_{m-1}(z) + z\sum_{k=1}^{+\infty}\binom{k-1+m}{m}z^{k-1}$

$= f_{m-1}(z) + z f_m(z)$

$|z|<1$

$\Rightarrow f_m(z) = \frac{1}{1-z} f_{m-1}(z)$

$f_0(z) = \sum_{k=0}^{+\infty}\binom{k+0}{0}z^k = \sum_{k=0}^{+\infty} z^k = \frac{1}{1-z}$

$\therefore f_m(z) = (1-z)^{-(m+1)}$ Negative Binomial Formula

Continuous Distributions

① *Uniform Distribution*

The continuous *uniform distribution* is defined by spreading mass uniformly over an interval $[a, b]$. Its pdf is given by

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

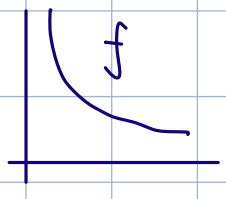
$$M_X(t) = E(e^{tX}) = \int_a^b \frac{e^{tx}}{b-a}dx = \begin{cases} \frac{e^{bt}-e^{at}}{t(b-a)} & ; t \neq 0 \\ 1 & ; t = 0 \end{cases}$$

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

$$\mathrm{Var}\,X = \int_a^b \frac{\left(x - \frac{b+a}{2}\right)^2}{b-a}\,dx = \frac{(b-a)^2}{12}.$$

② Exponential Distribution

$X \sim Exp(\theta), \theta > 0$, if $f_X(x) = \begin{cases} \frac{1}{\theta}e^{-\frac{x}{\theta}} & ; x > 0 \\ 0 & ; \text{o.w.} \end{cases}$



$P(X \in [a-h, a+h]) = e^{-\frac{a}{\theta}}(e^{\frac{h}{\theta}} - e^{-\frac{h}{\theta}})$

$EX = \int_0^{+\infty}\frac{x}{\theta}e^{-\frac{x}{\theta}}dx = \theta\Gamma(2) = \theta$

$EX^2 = \int_0^{+\infty}\frac{x^2}{\theta}e^{-\frac{x}{\theta}}dx = \theta^2\Gamma(3) = 2\theta^2$

$Var(X) = EX^2 - (EX)^2 = \theta^2$

$$M_X(t) = \int_0^{+\infty} e^{tx} \frac{1}{\theta} e^{-\frac{x}{\theta}} dx$$
$$= \frac{1}{\theta} \int_0^{+\infty} e^{(t-\frac{1}{\theta})x} dx$$
$$= \begin{cases} +\infty & ; t \geq \frac{1}{\theta} \\ \frac{1/\theta}{1/\theta - t} = \frac{1}{1-\theta t} & ; t < \frac{1}{\theta} \end{cases}$$

Thm: $X \sim Exp(\theta) \Rightarrow P(X > t+s \mid X > t) = P(X > s)$; $\forall t, s \in [0, +\infty)$

Pf: $P(X > u) = \int_u^{+\infty} \frac{1}{\theta} e^{-\frac{x}{\theta}} dx = (-e^{-\frac{x}{\theta}})\Big|_u^{+\infty} = e^{-\frac{u}{\theta}}$ $(u > 0)$

Now, $P(X > t+s \mid X > t) = P(X > t+s) / P(X > t)$
$$= e^{-(t+s)/\theta} / e^{-t/\theta}$$
$$= e^{-s/\theta} = P(X > s)$$

Thm: Suppose X is a continuous r.v. with $P(X > t) > 0$; $\forall t > 0$, and $P(X > s+t \mid X > t) = P(X > s)$; $\forall t, s \in [0, +\infty)$.

Then $X \sim Exp(\theta)$ for some $\theta > 0$.

Pf: $t = s = 0$: $P(X > 0 \mid X > 0) = P(X > 0) \Rightarrow P(X > 0) = 1 \Rightarrow Range(X) \in [0, +\infty)$

Define $G(x) = P(X > x)$; $\forall x \geq 0$ [$G(\cdot)$ is a continuous function]

$t = s = 1$: $G(2) = (G(1))^2$

By induction, $G(n) = (G(1))^n$; $\forall n \in \mathbb{N}$

Let $x = \frac{p}{q}$; $p, q \in \mathbb{N}$ using MP.

$$G(p) = G(\tfrac{p}{q} \cdot q) = (G(\tfrac{p}{q}))^q$$
$$\Rightarrow G(\tfrac{p}{q}) = (G(p))^{1/q} = ((G(1))^p)^{1/q} = (G(1))^{p/q}$$

Let $x \in [0, +\infty)$, then $\exists$ sequence of $\{x_n\}_{n \geq 1}$ of positive rational numbers such that $x_n \to x$ as $n \to \infty$. Now,

$$G(x) = G(\lim_{n\to\infty} x_n) \overset{\text{continuity of } G}{=} \lim_{n\to\infty} G(x_n)$$
$$= \lim_{n\to\infty} (G(1))^{x_n} = (G(1))^x = e^{x \log G(1)}$$
$$= e^{-(-\log G(1))x} \quad (\theta = \tfrac{-1}{\log G(1)})$$

$F_X(x) = 1 - G(x) = 1 - e^{-\frac{x}{\theta}}$; $\forall x > 0$ (matches the CDF of $Exp(\theta)$)

∴ Sep 23rd. 2021

③ *Gamma Distribution*

$[\Gamma(1,\beta) = \text{Exp}(\beta)]$

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

The parameter α is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter β is called the scale parameter, since most of its influence is on the spread of the distribution.

$$\begin{aligned} EX &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^\alpha e^{-x/\beta}\, dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \Gamma(\alpha+1)\beta^{\alpha+1} \quad \longleftarrow \quad \int_0^\infty x^{\alpha-1} e^{-x/\beta}\, dx = \Gamma(\alpha)\beta^\alpha, \\ &= \frac{\alpha\Gamma(\alpha)\beta}{\Gamma(\alpha)} \\ &= \alpha\beta. \end{aligned}$$

$$\text{Var}\, X = \alpha\beta^2.$$

$$M_X(t) = E(e^{tX}) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^{+\infty} e^{tx} x^{\alpha-1} e^{-x/\beta}\, dx$$

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

$$= \Gamma(\alpha)\left(\frac{\beta}{1-\beta t}\right)^\alpha \Big/ \Gamma(\alpha)\beta^\alpha = \begin{cases} \left(\frac{1}{1-\beta t}\right)^\alpha & ; \text{if } t < \frac{1}{\beta} \\ +\infty & ; \text{o.w.} \end{cases}$$

There are a number of important special cases of the gamma distribution. If we set $\alpha = p/2$, where p is an integer, and $\beta = 2$, then the gamma pdf becomes

$\Gamma(\frac{p}{2}, 2) = \chi_p^2$

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty, \tag{3.3.10}$$

which is the *chi squared pdf with p degrees of freedom*. The mean, variance, and mgf of the chi squared distribution can all be calculated by using the previously derived gamma formulas.

(4) *Normal Distribution*

The normal distribution has two parameters, usually denoted by μ and σ^2, which are its mean and variance. The pdf of the *normal distribution* with mean μ and variance σ^2 (usually denoted by $\mathrm{n}(\mu, \sigma^2)$) is given by

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty. \tag{3.3.13}$$

If $X \sim \mathrm{n}(\mu, \sigma^2)$, then the random variable $Z = (X-\mu)/\sigma$ has a $\mathrm{n}(0,1)$ distribution, also known as the *standard normal.* This is easily established by writing

$$\begin{aligned} P(Z \le z) &= P\left(\frac{X-\mu}{\sigma} \le z\right) \\ &= P(X \le z\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{z\sigma+\mu} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt, \qquad \left(\text{substitute } t = \frac{x-\mu}{\sigma}\right) \end{aligned}$$

showing that $P(Z \le z)$ is the standard normal cdf.

We have not yet established that (3.3.13) integrates to 1 over the whole real line. By applying the standardizing transformation, we need only to show that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 1.$$

Notice that the integrand above is symmetric around 0, implying that the integral over $(-\infty, 0)$ is equal to the integral over $(0, \infty)$. Thus, we reduce the problem to showing

$$\int_0^{\infty} e^{-z^2/2} dz = \frac{\sqrt{2\pi}}{2} = \sqrt{\frac{\pi}{2}}. \tag{3.3.14}$$

The function $e^{-z^2/2}$ does not have an antiderivative that can be written explicitly in terms of elementary functions (that is, in closed form), so we cannot perform the integration directly. In fact, this is an example of an integration that either you know how to do or else you can spend a very long time going nowhere. Since both sides of (3.3.14) are positive, the equality will hold if we establish that the squares are equal. Square the integral in (3.3.14) to obtain

$$\begin{aligned} \left(\int_0^{\infty} e^{-z^2/2} dz\right)^2 &= \left(\int_0^{\infty} e^{-t^2/2}\, dt\right)\left(\int_0^{\infty} e^{-u^2/2} du\right) \\ &= \int_0^{\infty}\int_0^{\infty} e^{-(t^2+u^2)/2}\, dt\, du. \end{aligned}$$

The integration variables are just dummy variables, so changing their names is allowed. Now, we convert to polar coordinates. Define

$$t = r\cos\theta \quad \text{and} \quad u = r\sin\theta.$$

Then $t^2 + u^2 = r^2$ and $dt\, du = r\, d\theta\, dr$ and the limits of integration become $0 < r < \infty$, $0 < \theta < \pi/2$ (the upper limit on θ is $\pi/2$ because t and u are restricted to be positive). We now have

$$\int_0^\infty \int_0^\infty e^{-(t^2+u^2)/2}\,dt\,du = \int_0^\infty \int_0^{\pi/2} re^{-r^2/2}\,d\theta\,dr$$
$$= \frac{\pi}{2}\int_0^\infty re^{-r^2/2}dr$$
$$= \frac{\pi}{2}\left[-e^{-r^2/2}\Big|_0^\infty\right]$$
$$= \frac{\pi}{2},$$

which establishes (3.3.14).

This integral is closely related to the gamma function; in fact, by making the substitution $w = \frac{1}{2}z^2$ in (3.3.14), we see that this integral is essentially $\Gamma(\frac{1}{2})$. If we are careful to get the constants correct, we will see that (3.3.14) implies

(3.3.15) $$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty w^{-1/2}e^{-w}dw = \sqrt{\pi}.$$

$\sqrt{\frac{\pi}{2}} = \int_0^{+\infty} e^{-\frac{z^2}{2}}dz \xrightarrow[dy=zdz]{z=\sqrt{2y}} \int_0^{+\infty} e^{-y}\frac{1}{\sqrt{2y}}dy = \frac{1}{\sqrt{2}}\Gamma(\frac{1}{2}) \Rightarrow \Gamma(\frac{1}{2}) = \sqrt{\pi}.$

$M_Z(t) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{tz} e^{-z^2/2}dz$

$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2-2tz+t^2-t^2)}dz$

$= e^{t^2/2}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}dz = e^{\frac{t^2}{2}}$

$M_X(t) = E(e^{tX}) = E(e^{t(\sigma Z+\mu)})$

$= e^{t\mu}E(e^{t\sigma Z}) = e^{t\mu}M_Z(t\sigma)$

$= e^{t\mu}e^{\frac{(t\sigma)^2}{2}} = e^{t\mu+t^2\sigma^2/2}$; $\forall t \in \mathbb{R}$

Sep 28th 2021

Example 3.3.2 (Normal approximation) Let $X \sim$ binomial$(25, .6)$. We can approximate X with a normal random variable, Y, with mean $\mu = 25(.6) = 15$ and standard deviation $\sigma = ((25)(.6)(.4))^{1/2} = 2.45$. Thus

$$P(X \le 13) \approx P(Y \le 13) = P\left(Z \le \frac{13-15}{2.45}\right) = P(Z \le -.82) = .206,$$

while the exact binomial calculation gives

$$P(X \le 13) = \sum_{x=0}^{13}\binom{25}{x}(.6)^x(.4)^{25-x} = .267,$$

showing that the normal approximation is good, but not terrific. The approximation can be greatly improved, however, by a "continuity correction." To see how this works, look at Figure 3.3.2, which shows the binomial$(25, .6)$ pmf and the n$(15, (2.45)^2)$ pdf. We have drawn the binomial pmf using bars of width 1, with height equal to the probability. Thus, the areas of the bars give the binomial probabilities. In the approximation, notice how the area of the approximating normal is smaller than the binomial area (the normal area is everything to the left of the line at 13, whereas the binomial area includes the entire bar at 13 up to 13.5). The continuity correction adds this area back by adding $\frac{1}{2}$ to the cutoff point. So instead of approximating $P(X \leq 13)$, we approximate the equivalent expression (because of the discreteness), $P(X \leq 13.5)$ and obtain

$$P(X \leq 13) = P(X \leq 13.5) \approx P(Y \leq 13.5) = P(Z \leq -.61) = .271,$$

a much better approximation. In general, the normal approximation with the continuity correction is far superior to the approximation without the continuity correction.

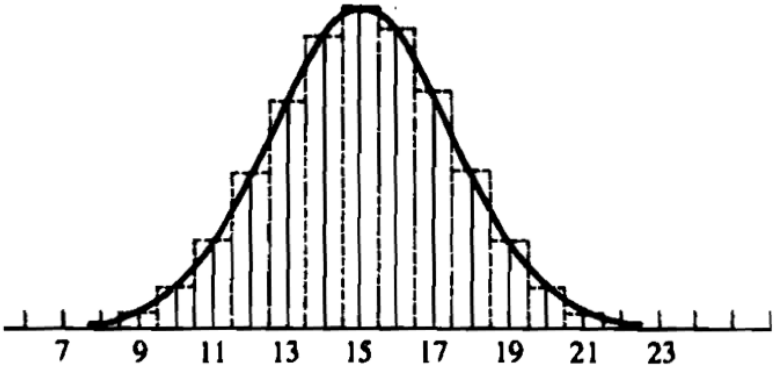


Figure 3.3.2. *Normal*$(15, (2.45)^2)$ *approximation to the binomial*$(25, .6)$

We also make the correction on the lower end. If $X \sim$ binomial(n, p) and $Y \sim$ n$(np, np(1-p))$, then we approximate

$$P(X \leq x) \approx P(Y \leq x + 1/2),$$
$$P(X \geq x) \approx P(Y \geq x - 1/2). \quad \|$$

(k) *Beta Distribution*

The beta family of distributions is a continuous family on $(0, 1)$ indexed by two parameters. The *beta*(α, β) *pdf* is

$$(3.3.16)\quad f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0,$$

where $B(\alpha, \beta)$ denotes the beta function,

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

The beta function is related to the gamma function through the following identity:

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{3.3.17}$$

$\alpha=1,\beta=1$: Beta(1,1) = Uniform(0,1)

$\alpha>1,\beta=1$: strictly increasing $f_X(x)$

$\alpha=1,\beta>1$: strictly decreasing $f_X(x)$

$\alpha<1,\beta<1$: U-shaped

$\alpha>1,\beta>1$: Unimodal

Calculation of moments of the beta distribution is quite easy, due to the particular form of the pdf. For $n>-\alpha$ we have

$$\begin{aligned}\mathrm{E}X^n&=\frac{1}{B(\alpha,\beta)}\int_0^1 x^n x^{\alpha-1}(1-x)^{\beta-1}\,dx\\&=\frac{1}{B(\alpha,\beta)}\int_0^1 x^{(\alpha+n)-1}(1-x)^{\beta-1}\,dx.\\&=\frac{B(\alpha+n,\beta)}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}.\end{aligned}$$

Using (3.3.3) and (3.3.18) with $n=1$ and $n=2$, we calculate the mean and variance of the beta(α,β) distribution as

$$\mathrm{E}X=\frac{\alpha}{\alpha+\beta}\quad\text{and}\quad\mathrm{Var}\,X=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

(6) *Cauchy Distribution*

The *Cauchy distribution* is a symmetric, bell-shaped distribution on $(-\infty,\infty)$ with pdf

$$f(x|\theta)=\frac{1}{\pi}\frac{1}{1+(x-\theta)^2},\quad -\infty<x<\infty,\quad -\infty<\theta<\infty. \tag{3.3.19}$$

Exponential Families

A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right). \tag{3.4.1}$$

Here $h(x) \geq 0$ and $t_1(x), \ldots, t_k(x)$ are real-valued functions of the observation x (they cannot depend on $\boldsymbol{\theta}$), and $c(\boldsymbol{\theta}) \geq 0$ and $w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})$ are real-valued functions of the possibly vector-valued parameter $\boldsymbol{\theta}$ (they cannot depend on x). Many common families introduced in the previous section are exponential families. These include the continuous families—normal, gamma, and beta, and the discrete families—binomial, Poisson, and negative binomial.

Example 3.4.1 (Binomial exponential family) Let n be a positive integer and consider the binomial(n, p) family with $0 < p < 1$. Then the pmf for this family, for $x = 0, \ldots, n$ and $0 < p < 1$, is

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} (1-p)^n \left(\frac{p}{1-p}\right)^x \\ &= \binom{n}{x} (1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right) x\right). \end{aligned}$$

Define

$$h(x) = \begin{cases} \binom{n}{x} & x = 0, \ldots, n \\ 0 & \text{otherwise,} \end{cases} \qquad c(p) = (1-p)^n, \quad 0 < p < 1,$$

$$w_1(p) = \log\left(\frac{p}{1-p}\right), \quad 0 < p < 1, \quad \text{and} \quad t_1(x) = x.$$

Then we have

$$f(x|p) = h(x)c(p) \exp[w_1(p)t_1(x)],$$

Theorem 3.4.2 *If X is a random variable with pdf or pmf of the form (3.4.1), then*

$$\mathrm{E}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial}{\partial \theta_j} \log c(\boldsymbol{\theta}); \tag{3.4.4}$$

$$\mathrm{Var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) - \mathrm{E}\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(X)\right). \tag{3.4.5}$$

Example 3.4.3 (Binomial mean and variance) From Example 3.4.1 we have

$$\frac{d}{dp}w_1(p) = \frac{d}{dp}\log\frac{p}{1-p} = \frac{1}{p(1-p)}$$
$$\frac{d}{dp}\log c(p) = \frac{d}{dp}n\log(1-p) = \frac{-n}{1-p}$$

and thus from Theorem 3.4.2 we have

$$\mathrm{E}\left(\frac{1}{p(1-p)}X\right) = \frac{n}{1-p}$$

and a bit of rearrangement yields $\mathrm{E}(X) = np$. The variance identity works in a similar manner. ‖

Example 3.4.4 (Normal exponential family) Let $f(x|\mu,\sigma^2)$ be the $\mathrm{n}(\mu,\sigma^2)$ family of pdfs, where $\boldsymbol{\theta} = (\mu,\sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$. Then

$$\begin{aligned} f(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\exp\left(-\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}\right). \end{aligned} \tag{3.4.6}$$

Define

$$h(x) = 1 \text{ for all } x;$$
$$c(\boldsymbol{\theta}) = c(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\mu^2}{2\sigma^2}\right), \quad -\infty < \mu < \infty,\ \sigma > 0;$$
$$w_1(\mu,\sigma) = \frac{1}{\sigma^2}, \quad \sigma > 0; \qquad w_2(\mu,\sigma) = \frac{\mu}{\sigma^2}, \ \sigma > 0;$$
$$t_1(x) = -x^2/2; \quad \text{and} \quad t_2(x) = x.$$

Then

$$f(x|\mu,\sigma^2) = h(x)c(\mu,\sigma)\exp[w_1(\mu,\sigma)t_1(x) + w_2(\mu,\sigma)t_2(x)],$$

From (3.4.1), since the factor $\exp(\cdot)$ is always positive, it can be seen that for any $\boldsymbol{\theta} \in \Theta$, that is, for any $\boldsymbol{\theta}$ for which $c(\boldsymbol{\theta}) > 0$, $\{x : f(x|\boldsymbol{\theta}) > 0\} = \{x : h(x) > 0\}$ and this set does not depend on $\boldsymbol{\theta}$. So, for example, the set of pdfs given by $f(x|\theta) = \theta^{-1}\exp(1-(x/\theta)), 0 < \theta < x < \infty$, is not an exponential family even though we can write $\theta^{-1}\exp(1-(x/\theta)) = h(x)c(\theta)\exp(w(\theta)t(x))$, where $h(x) = e^1$, $c(\theta) = \theta^{-1}$, $w(\theta) = \theta^{-1}$, and $t(x) = -x$. Writing the pdf with indicator functions makes this very clear. We have

$$f(x|\theta) = \theta^{-1}\exp\left(1-\left(\frac{x}{\theta}\right)\right)I_{[\theta,\infty)}(x).$$

The indicator function cannot be incorporated into any of the functions of (3.4.1) since it is not a function of x alone, not a function of θ alone, and cannot be expressed as an exponential. Thus, this is not an exponential family.

Definition 3.4.7 A *curved exponential family* is a family of densities of the form (3.4.1) for which the dimension of the vector $\boldsymbol{\theta}$ is equal to $d < k$. If $d = k$, the family is a *full exponential family.* (See also Miscellanea 3.8.3.)

Example 3.4.8 (A curved exponential family) The normal family of Example 3.4.4 is a full exponential family. However, if we assume that $\sigma^2 = \mu^2$, the family becomes curved. (Such a model might be used in the analysis of variance; see Exercises 11.1 and 11.2.) We then have

$$
\begin{aligned}
f(x|\mu) &= \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{(x-\mu)^2}{2\mu^2}\right) I_{(-\infty,+\infty)}(x) \\
&= \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2}\right) \exp\left(-\frac{x^2}{2\mu^2} + \frac{x}{\mu}\right) I_{(-\infty,+\infty)}(x)
\end{aligned}
\tag{3.4.9}
$$

For the normal family the full exponential family would have parameter space $(\mu, \sigma^2) = \Re \times (0, \infty)$, while the parameter space of the curved family $(\mu, \sigma^2) = (\mu, \mu^2)$ is a parabola. ‖

Example 3.4.9 (Normal approximations) In Chapter 5 we will see that if $X_1, \ldots, X_n$ is a sample from a Poisson(λ) population, then the distribution of $\bar{X} = \Sigma_i X_i/n$ is approximately

$$\bar{X} \sim \text{n}(\lambda, \lambda/n),$$

a curved exponential family.

The n($\lambda, \lambda/n$) approximation is justified by the Central Limit Theorem (Theorem 5.5.14). In fact, we might realize that most such CLT approximations will result in a curved normal family. We have seen the normal binomial approximation (Example 3.3.2): If $X_1, \ldots, X_n$ are iid Bernoulli(p), then

$$\bar{X} \sim \text{n}(p, p(1-p)/n),$$

approximately. For another illustration, see Example 5.5.16. ‖

Location & Scale Families

Theorem 3.5.1 *Let $f(x)$ be any pdf and let μ and $\sigma > 0$ be any given constants. Then the function*

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

is a pdf.

Pf: $g \geq 0$ a.s. and

$$\int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx = \int_{-\infty}^{\infty} f(y)\, dy \qquad \left(\text{substitute } y = \frac{x-\mu}{\sigma}\right)$$

$$= 1, \qquad (\text{since } f(y) \text{ is a pdf})$$

Definition 3.5.2 Let $f(x)$ be any pdf. Then the family of pdfs $f(x-\mu)$, indexed by the parameter μ, $-\infty < \mu < \infty$, is called the *location family with standard pdf* $f(x)$ and μ is called the *location parameter* for the family.

Example 3.5.3 (Exponential location family) Let $f(x) = e^{-x}$, $x \geq 0$, and $f(x) = 0$, $x < 0$. To form a location family we replace x with $x - \mu$ to obtain

$$f(x|\mu) = \begin{cases} e^{-(x-\mu)} & x-\mu \geq 0 \\ 0 & x-\mu < 0 \end{cases}$$

$$= \begin{cases} e^{-(x-\mu)} & x \geq \mu \\ 0 & x < \mu. \end{cases}$$

Graphs of $f(x|\mu)$ for various values of μ are shown in Figure 3.5.2. As in Figure 3.5.1, the graph has been shifted. Now the positive part of the graph starts at μ rather than at 0. If X measures time, then μ might be restricted to be nonnegative so that X will be positive with probability 1 for every value of μ. In this type of model, where μ denotes a bound on the range of X, μ is sometimes called a *threshold parameter*. ||

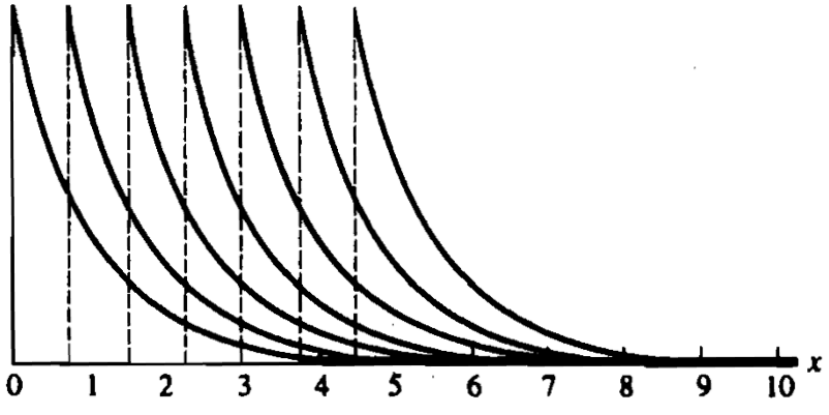


Figure 3.5.2. *Exponential location densities*

Definition 3.5.4 Let $f(x)$ be any pdf. Then for any $\sigma > 0$, the family of pdfs $(1/\sigma)f(x/\sigma)$, indexed by the parameter σ, is called the *scale family with standard pdf* $f(x)$ and σ is called the *scale parameter* of the family.

The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph.

Definition 3.5.5 Let $f(x)$ be any pdf. Then for any μ, $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs $(1/\sigma)f((x-\mu)/\sigma)$, indexed by the parameter (μ, σ), is called the *location–scale family with standard pdf* $f(x)$; μ is called the *location parameter* and σ is called the *scale parameter*.

Theorem 3.5.6 *Let $f(\cdot)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

Proof: To prove the "if" part, define $g(z) = \sigma z + \mu$. Then $X = g(Z)$, g is a monotone function, $g^{-1}(x) = (x-\mu)/\sigma$, and $\left|(d/dx)g^{-1}(x)\right| = 1/\sigma$. Thus by Theorem 2.1.5, the pdf of X is

$$f_X(x) = f_Z\left(g^{-1}(x)\right)\left|\frac{d}{dx}g^{-1}(x)\right| = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

To prove the "only if" part, define $g(x) = (x-\mu)/\sigma$ and let $Z = g(X)$. Theorem 2.1.5 again applies: $g^{-1}(z) = \sigma z + \mu$, $\left|(d/dz)g^{-1}(z)\right| = \sigma$, and the pdf of Z is

$$f_Z(z) = f_X\left(g^{-1}(z)\right)\left|\frac{d}{dz}g^{-1}(z)\right| = \frac{1}{\sigma}f\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)\sigma = f(z).$$

Also,

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{X-\mu}{\sigma}\right) + \mu = X.$$

□

Probability Inequalities

Theorem 3.6.1 (Chebychev's Inequality) *Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,*

$$P\left(g(X) \geq r\right) \leq \frac{\mathrm{E}g(X)}{r}.$$

Proof:

$$\begin{aligned}
\mathrm{E}g(X) &= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \\
&\geq \int_{\{x:g(x)\geq r\}} g(x) f_X(x)\, dx \qquad (g \text{ is nonnegative}) \\
&\geq r \int_{\{x:g(x)\geq r\}} f_X(x)\, dx \\
&= r P\left(g(X) \geq r\right). \qquad (\text{definition})
\end{aligned}$$

Rearranging now produces the desired inequality. □

Example 3.6.2 (Illustrating Chebychev) The most widespread use of Chebychev's Inequality involves means and variances. Let $g(x) = (x-\mu)^2/\sigma^2$, where $\mu = \mathrm{E}X$ and $\sigma^2 = \mathrm{Var}\, X$. For convenience write $r = t^2$. Then

$$P\left(\frac{(X-\mu)^2}{\sigma^2} \geq t^2\right) \leq \frac{1}{t^2}\mathrm{E}\frac{(X-\mu)^2}{\sigma^2} = \frac{1}{t^2}.$$

Doing some obvious algebra, we get the inequality

$$P(|X-\mu| \geq t\sigma) \leq \frac{1}{t^2} \qquad P\left(\left|\frac{X-\mu}{\sigma}\right| \geq t\right) \leq \frac{1}{t^2}$$

and its companion

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

which gives a universal bound on the deviation $|X-\mu|$ in terms of σ. For example, taking $t = 2$, we get

$$P(|X-\mu| \geq 2\sigma) \leq \frac{1}{2^2} = .25,$$

so there is at least a 75% chance that a random variable will be within 2σ of its mean (no matter what the distribution of X). ‖

Example 3.6.3 (A normal probability inequality) If Z is standard normal, then

$$P(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}, \quad \text{for all } t > 0. \tag{3.6.1}$$

Compare this with Chebychev's Inequality. For $t = 2$, Chebychev gives $P(|Z| \geq t) \leq .25$ but $\sqrt{(2/\pi)}e^{-2}/2 = .054$, a vast improvement.

To prove (3.6.1), write

$$
\begin{aligned}
P(Z \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \\
&\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2}\, dx \qquad \left(\begin{array}{c}\text{since } x/t > 1 \\ \text{for } x > t\end{array}\right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}
\end{aligned}
$$

and use the fact that $P(|Z| \geq t) = 2P(Z \geq t)$. A lower bound on $P(|Z| \geq t)$ can be established in a similar way (see Exercise 3.47). ‖

If $M_X(t)$ exists, then $P(X \geq a) \leq e^{-at} M_X(t)$,

Oct 7th 2021

Chapter 4: Multiple Random Variables

Definition 4.1.1 An *n-dimensional random vector* is a function from a sample space S into $\Re^n$, n-dimensional Euclidean space.

Definition 4.1.3 Let (X, Y) be a discrete bivariate random vector. Then the function $f(x, y)$ from $\Re^2$ into $\Re$ defined by $f(x, y) = P(X = x, Y = y)$ is called the *joint probability mass function* or *joint pmf* of (X, Y). If it is necessary to stress the fact that f is the joint pmf of the vector (X, Y) rather than some other vector, the notation $f_{X,Y}(x, y)$ will be used.

Expectations of functions of random vectors are computed just as with univariate random variables. Let $g(x, y)$ be a real-valued function defined for all possible values (x, y) of the discrete random vector (X, Y). Then $g(X, Y)$ is itself a random variable and its expected value $\mathrm{E}g(X, Y)$ is given by

$$\mathrm{E}g(X, Y) = \sum_{(x,y)\in\Re^2} g(x, y) f(x, y).$$

Theorem 4.1.6 *Let (X, Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then the marginal pmfs of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by*

$$f_X(x) = \sum_{y\in\Re} f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_{x\in\Re} f_{X,Y}(x, y).$$

Proof: We will prove the result for $f_X(x)$. The proof for $f_Y(y)$ is similar. For any $x \in \Re$, let $A_x = \{(x, y) : -\infty < y < \infty\}$. That is, A_x is the line in the plane with first coordinate equal to x. Then, for any $x \in \Re$,

$$\begin{aligned}
f_X(x) &= P(X = x) \\
&= P(X = x, -\infty < Y < \infty) && (P(-\infty < Y < \infty) = 1) \\
&= P((X, Y) \in A_x) && \text{(definition of } A_x) \\
&= \sum_{(x,y)\in A_x} f_{X,Y}(x, y) \\
&= \sum_{y\in\Re} f_{X,Y}(x, y). && \square
\end{aligned}$$

Oct 12th 2021

Definition 4.1.10 A function $f(x, y)$ from $\Re^2$ into $\Re$ is called a *joint probability density function* or *joint pdf of the continuous bivariate random vector* (X, Y) if, for every $A \subset \Re^2$,

$$P((X, Y) \in A) = \int_A \int f(x, y)\, dx\, dy.$$

the *expected value of* $g(X,Y)$ is defined to be

$$\mathrm{E}g(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy.$$

The *marginal probability density functions* of X and Y are also defined as in the discrete case with integrals replacing sums. The marginal pdfs may be used to compute probabilities or expectations that involve only X or Y. Specifically, the marginal pdfs of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy, \quad -\infty < x < \infty,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx, \quad -\infty < y < \infty. \tag{4.1.3}$$

The joint probability distribution of (X,Y) can be completely described with the *joint cdf* (cumulative distribution function) rather than with the joint pmf or joint pdf. The joint cdf is the function $F(x,y)$ defined by

$$F(x,y) = P(X \le x, Y \le y)$$

for all $(x,y) \in \Re^2$. The joint cdf is usually not very handy to use for a discrete random vector. But for a continuous bivariate random vector we have the important relationship, as in the univariate case,

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)\,dt\,ds.$$

From the bivariate Fundamental Theorem of Calculus, this implies that

$$\frac{\partial^2 F(x,y)}{\partial x\,\partial y} = f(x,y) \tag{4.1.4}$$

at continuity points of $f(x,y)$. This relationship is useful in situations where an expression for $F(x,y)$ can be found. The mixed partial derivative can be computed to find the joint pdf.

Definition 4.2.1 Let (X,Y) be a discrete bivariate random vector with joint pmf $f(x,y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X=x) = f_X(x) > 0$, the *conditional pmf of Y given that $X = x$* is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = P(Y = y|X = x) = \frac{f(x,y)}{f_X(x)}.$$

For any y such that $P(Y=y) = f_Y(y) > 0$, the *conditional pmf of X given that $Y = y$* is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = P(X = x|Y = y) = \frac{f(x,y)}{f_Y(y)}.$$

Definition 4.2.3 Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the *conditional pdf of Y given that $X = x$* is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $f_Y(y) > 0$, the *conditional pdf of X given that $Y = y$* is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

In addition to their usefulness for calculating probabilities, the conditional pdfs or pmfs can also be used to calculate expected values. Just remember that $f(y|x)$ as a function of y is a pdf or pmf and use it in the same way that we have previously used unconditional pdfs or pmfs. If $g(Y)$ is a function of Y, then the *conditional expected value of $g(Y)$ given that $X = x$* is denoted by $\mathrm{E}(g(Y)|x)$ and is given by

$$\mathrm{E}(g(Y)|x) = \sum_y g(y) f(y|x) \quad \text{and} \quad \mathrm{E}(g(Y)|x) = \int_{-\infty}^{\infty} g(y) f(y|x)\, dy$$

The variance of the probability distribution described by $f(y|x)$ is called the *conditional variance of Y given $X = x$*. Using the notation $\mathrm{Var}(Y|x)$ for this, we have, using the ordinary definition of variance,

$$\mathrm{Var}(Y|x) = \mathrm{E}(Y^2|x) - (\mathrm{E}(Y|x))^2.$$

Definition 4.2.5 Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called *independent random variables* if, for every $x \in \Re$ and $y \in \Re$,

$$f(x, y) = f_X(x) f_Y(y). \tag{4.2.1}$$

$X \perp\!\!\!\perp Y$

If X and Y are independent, the conditional pdf of Y given $X = x$ is

$$\begin{aligned} f(y|x) &= \frac{f(x, y)}{f_X(x)} && \text{(definition)} \\ &= \frac{f_X(x) f_Y(y)}{f_X(x)} && \text{(from (4.2.1))} \\ &= f_Y(y), \end{aligned}$$

regardless of the value of x. Thus, for any $A \subset \Re$ and $x \in \Re$, $P(Y \in A|x) = \int_A f(y|x)\, dy = \int_A f_Y(y)\, dy = P(Y \in A)$. The knowledge that $X = x$ gives us no additional information about Y.

≠ Oct 14th. 2021

Lemma 4.2.7 *Let (X,Y) be a bivariate random vector with joint pdf or pmf $f(x,y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \Re$ and $y \in \Re$,*

$$f(x,y) = g(x)h(y).$$

Proof: The "only if" part is proved by defining $g(x) = f_X(x)$ and $h(y) = f_Y(y)$ and using (4.2.1). To prove the "if" part for continuous random variables, suppose that $f(x,y) = g(x)h(y)$. Define

$$\int_{-\infty}^{\infty} g(x)\,dx = c \quad \text{and} \quad \int_{-\infty}^{\infty} h(y)\,dy = d,$$

where the constants c and d satisfy

$$\begin{aligned} cd &= \left(\int_{-\infty}^{\infty} g(x)\,dx\right)\left(\int_{-\infty}^{\infty} h(y)\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\,dx\,dy \qquad (4.2.2) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy \\ &= 1. \qquad (f(x,y) \text{ is a joint pdf}) \end{aligned}$$

Furthermore, the marginal pdfs are given by

(4.2.3)

$$f_X(x) = \int_{-\infty}^{\infty} g(x)h(y)\,dy = g(x)d \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} g(x)h(y)\,dx = h(y)c.$$

Thus, using (4.2.2) and (4.2.3), we have

$$f(x,y) = g(x)h(y) = g(x)h(y)cd = f_X(x)f_Y(y),$$

showing that X and Y are independent. Replacing integrals with sums proves the lemma for discrete random vectors. □

Theorem 4.2.10 *Let X and Y be independent random variables.*

a. *For any $A \subset \Re$ and $B \subset \Re$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.*

b. *Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then*

$$\mathrm{E}\left(g(X)h(Y)\right) = \left(\mathrm{E}g(X)\right)\left(Eh(Y)\right).$$

Proof: For continuous random variables, part (b) is proved by noting that

$$\mathrm{E}\left(g(X)h(Y)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,dx\,dy \qquad \text{(by (4.2.1))}$$
$$= \int_{-\infty}^{\infty} h(y)f_Y(y)\int_{-\infty}^{\infty} g(x)f_X(x)\,dx\,dy$$
$$= \left(\int_{-\infty}^{\infty} g(x)f_X(x)\,dx\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)\,dy\right)$$
$$= (\mathrm{E}g(X))\,(\mathrm{E}h(Y))\,.$$

The result for discrete random variables is proved by replacing integrals by sums. Part (a) can be proved by a series of steps similar to those above or by the following argument. Let $g(x)$ be the indicator function of the set A. Let $h(y)$ be the indicator function of the set B. Note that $g(x)h(y)$ is the indicator function of the set $C \subset \Re^2$ defined by $C = \{(x,y) : x \in A, y \in B\}$. Also note that for an indicator function such as $g(x), \mathrm{E}g(X) = P(X \in A)$. Thus using the expectation equality just proved, we have

$$P(X \in A, Y \in B) = P((X,Y) \in C) = \mathrm{E}\,(g(X)h(Y))$$
$$= (\mathrm{E}g(X))\,(\mathrm{E}h(Y)) = P(X \in A)P(Y \in B). \qquad \square$$

Example 4.2.11 (Expectations of independent variables) Let X and Y be independent exponential(1) random variables. From Theorem 4.2.10 we have

$$P(X \geq 4, Y < 3) = P(X \geq 4)P(Y < 3) = e^{-4}(1 - e^{-3}).$$

Letting $g(x) = x^2$ and $h(y) = y$, we see that

$$\mathrm{E}\big(X^2Y\big) = \big(\mathrm{E}X^2\big)\,(\mathrm{E}Y) = \big(\mathrm{Var}\,X + (\mathrm{E}X)^2\big)\,\mathrm{E}Y = (1 + 1^2)1 = 2. \qquad \|$$

Theorem 4.2.12 *Let X and Y be independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then the moment generating function of the random variable $Z = X + Y$ is given by*

$$M_Z(t) = M_X(t)M_Y(t).$$

Proof: Using the definition of the mgf and Theorem 4.2.10, we have

$$M_Z(t) = \mathrm{E}e^{tZ} = \mathrm{E}e^{t(X+Y)} = \mathrm{E}(e^{tX}e^{tY}) = (\mathrm{E}e^{tX})(\mathrm{E}e^{tY}) = M_X(t)M_Y(t). \qquad \square$$

Example 4.2.13 (Mgf of a sum of normal variables) Sometimes Theorem 4.2.12 can be used to easily derive the distribution of Z from knowledge of the distribution of X and Y. For example, let $X \sim \mathrm{n}(\mu, \sigma^2)$ and $Y \sim \mathrm{n}(\gamma, \tau^2)$ be independent normal random variables. From Exercise 2.33, the mgfs of X and Y are

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2) \quad \text{and} \quad M_Y(t) = \exp(\gamma t + \tau^2 t^2/2).$$

Thus, from Theorem 4.2.12, the mgf of $Z = X + Y$ is

$$M_Z(t) = M_X(t)M_Y(t) = \exp\left((\mu + \gamma)t + (\sigma^2 + \tau^2)t^2/2\right).$$

This is the mgf of a normal random variable with mean $\mu+\gamma$ and variance $\sigma^2+\tau^2$. This result is important enough to be stated as a theorem. ||

Theorem 4.2.14 *Let $X \sim n(\mu,\sigma^2)$ and $Y \sim n(\gamma,\tau^2)$ be independent normal random variables. Then the random variable $Z = X+Y$ has a $n(\mu+\gamma,\sigma^2+\tau^2)$ distribution.*

∴ Oct 19th 2021

Example 4.3.1 (Distribution of the sum of Poisson variables) Let X and Y be independent Poisson random variables with parameters θ and λ, respectively. Thus the joint pmf of (X,Y) is

$$f_{X,Y}(x,y) = \frac{\theta^x e^{-\theta}}{x!}\frac{\lambda^y e^{-\lambda}}{y!}, \quad x = 0,1,2,\ldots,\ y = 0,1,2,\ldots.$$

The set $\mathcal{A}$ is $\{(x,y) : x = 0,1,2,\ldots \text{ and } y = 0,1,2,\ldots\}$. Now define $U = X+Y$ and $V = Y$. That is, $g_1(x,y) = x+y$ and $g_2(x,y) = y$. We will describe the set $\mathcal{B}$, the set of possible (u,v) values. The possible values for v are the nonnegative integers. The variable $v = y$ and thus has the same set of possible values. For a given value of v, $u = x+y = x+v$ must be an integer greater than or equal to v since x is a nonnegative integer. The set of all possible (u,v) values is thus given by $\mathcal{B} = \{(u,v) : v = 0,1,2,\ldots \text{ and } u = v, v+1, v+2,\ldots\}$. For any $(u,v) \in \mathcal{B}$, the only (x,y) value satisfying $x+y = u$ and $y = v$ is $x = u-v$ and $y = v$. Thus, in this example, A_{uv} always consists of only the single point $(u-v,v)$. From (4.3.1) we thus obtain the joint pmf of (U,V) as

$$f_{U,V}(u,v) = f_{X,Y}(u-v,v) = \frac{\theta^{u-v}e^{-\theta}}{(u-v)!}\frac{\lambda^v e^{-\lambda}}{v!}, \quad \begin{array}{l} v = 0,1,2,\ldots, \\ u = v, v+1, v+2,\ldots. \end{array}$$

In this example it is interesting to compute the marginal pmf of U. For any fixed nonnegative integer u, $f_{U,V}(u,v) > 0$ only for $v = 0,1,\ldots,u$. This gives the set of v values to sum over to obtain the marginal pmf of U. It is

$$f_U(u) = \sum_{v=0}^{u} \frac{\theta^{u-v}e^{-\theta}}{(u-v)!}\frac{\lambda^v e^{-\lambda}}{v!} = e^{-(\theta+\lambda)}\sum_{v=0}^{u}\frac{\theta^{u-v}}{(u-v)!}\frac{\lambda^v}{v!}, \quad u = 0,1,2,\ldots.$$

This can be simplified by noting that, if we multiply and divide each term by $u!$, we can use the Binomial Theorem to obtain

$$f_U(u) = \frac{e^{-(\theta+\lambda)}}{u!}\sum_{v=0}^{u}\binom{u}{v}\lambda^v\theta^{u-v} = \frac{e^{-(\theta+\lambda)}}{u!}(\theta+\lambda)^u, \quad u = 0,1,2,\ldots.$$

This is the pmf of a Poisson random variable with parameter $\theta+\lambda$. This result is significant enough to be stated as a theorem. ||

Theorem 4.3.2 *If $X \sim \text{Poisson}(\theta)$ and $Y \sim \text{Poisson}(\lambda)$ and X and Y are independent, then $X+Y \sim \text{Poisson}(\theta+\lambda)$.*

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

If (X, Y) is a continuous random vector with joint pdf $f_{X,Y}(x, y)$, then the joint pdf of (U, V) can be expressed in terms of $f_{X,Y}(x, y)$ in a manner analogous to (2.1.8). As before, $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$ and $\mathcal{B} = \{(u, v) : u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in \mathcal{A}\}$. The joint pdf $f_{U,V}(u, v)$ will be positive on the set $\mathcal{B}$. For the simplest version of this result we assume that the transformation $u = g_1(x, y)$ and $v = g_2(x, y)$ defines a one-to-one transformation of $\mathcal{A}$ onto $\mathcal{B}$. The transformation is onto because of the definition of $\mathcal{B}$. We are assuming that for each $(u, v) \in \mathcal{B}$ there is only one $(x, y) \in \mathcal{A}$ such that $(u, v) = (g_1(x, y), g_2(x, y))$. For such a one-to-one, onto transformation, we can solve the equations $u = g_1(x, y)$ and $v = g_2(x, y)$ for x and y in terms of u and v. We will denote this inverse transformation by $x = h_1(u, v)$ and $y = h_2(u, v)$. The role played by a derivative in the univariate case is now played by a quantity called the *Jacobian of the transformation*. This function of (u, v), denoted by J, is the *determinant of a matrix* of partial derivatives. It is defined by

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

where

$$\frac{\partial x}{\partial u} = \frac{\partial h_1(u, v)}{\partial u}, \quad \frac{\partial x}{\partial v} = \frac{\partial h_1(u, v)}{\partial v}, \quad \frac{\partial y}{\partial u} = \frac{\partial h_2(u, v)}{\partial u}, \quad \text{and} \quad \frac{\partial y}{\partial v} = \frac{\partial h_2(u, v)}{\partial v}.$$

We assume that J is not identically 0 on $\mathcal{B}$. Then the joint pdf of (U, V) is 0 outside the set $\mathcal{B}$ and on the set $\mathcal{B}$ is given by

(4.3.2) $$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v),\ h_2(u, v))|J|,$$

Example 4.3.3 (Distribution of the product of beta variables) Let $X \sim \text{beta}(\alpha, \beta)$ and $Y \sim \text{beta}(\alpha + \beta, \gamma)$ be independent random variables. The joint pdf of (X, Y) is

$$f_{X,Y}(x, y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha + \beta)\Gamma(\gamma)} y^{\alpha+\beta-1}(1 - y)^{\gamma-1},$$
$$0 < x < 1, \quad 0 < y < 1.$$

Consider the transformation $U = XY$ and $V = X$. The set of possible values for V is $0 < v < 1$ since $V = X$. For a fixed value of $V = v$, U must be between 0 and v since $X = V = v$ and Y is between 0 and 1. Thus, this transformation maps the set

$\mathcal{A}$ onto the set $\mathcal{B} = \{(u, v) : 0 < u < v < 1\}$. For any $(u, v) \in \mathcal{B}$, the equations $u = xy$ and $v = x$ can be uniquely solved for $x = h_1(u, v) = v$ and $y = h_2(u, v) = u/v$. Note that if considered as a transformation defined on all of $\Re^2$, this transformation is not one-to-one. Any point $(0, y)$ is mapped into the point $(0, 0)$. But as a function defined only on $\mathcal{A}$, it is a one-to-one transformation onto $\mathcal{B}$. The Jacobian is given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}.$$

Thus, from (4.3.2) we obtain the joint pdf as

$$(4.3.3)\quad f_{U,V}(u,v) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} v^{\alpha-1}(1-v)^{\beta-1}\left(\frac{u}{v}\right)^{\alpha+\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\frac{1}{v}, \quad 0<u<v<1.$$

The marginal distribution of $V = X$ is, of course, a beta(α, β) distribution. But the distribution of U is also a beta distribution:

$$\begin{aligned} f_U(u) &= \int_u^1 f_{U,V}(u,v)\,dv \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}\int_u^1 \left(\frac{u}{v}-u\right)^{\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\left(\frac{u}{v^2}\right)dv. \end{aligned}$$

The expression (4.3.3) was used but some terms have been rearranged. Now make the univariate change of variable $y = (u/v - u)/(1-u)$ so that $dy = -u/[v^2(1-u)]dv$ to obtain

$$\begin{aligned} f_U(u) &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}\int_0^1 y^{\beta-1}(1-y)^{\gamma-1}\,dy \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}\frac{\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\beta+\gamma)} \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta+\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}, \quad 0<u<1. \end{aligned}$$

To obtain the second identity we recognized the integrand as the kernel of a beta pdf and used (3.3.17). Thus we see that the marginal distribution of U is beta$(\alpha, \beta+\gamma)$. ||

Ex. $f_{X_1,X_2}(x_1,x_2) = \lambda e^{-\lambda x_1}\cdot\lambda e^{-\lambda x_2}\cdot I_{(0,+\infty)}(x)\cdot I_{(0,+\infty)}(y)$

Define $(Y_1, Y_2) = (g_1(X_1,X_2), g_2(X_1,X_2)) = \left(X_1+X_2, \frac{X_1}{X_1+X_2}\right)$

$x_1 = h_1(y_1,y_2) = y_1 y_2$, $x_2 = h_2(y_1,y_2) = y_1 - y_1 y_2$

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ 1-y_2 & -y_1 \end{vmatrix} = -y_1 y_2 - y_1(1-y_2) = -y_1$$

$\Rightarrow f_{Y_1,Y_2}(y_1,y_2) = \lambda^2 e^{-\lambda y_1} \cdot |-y_1| = \lambda^2 y_1 e^{-\lambda y_1}, \quad y_1 > 0, \ 0 < y_2 < 1$

$\Rightarrow$ ① $Y_2 \sim \text{unif}(0,1)$ ② $Y_1 \sim \text{Gamma}(2, \frac{1}{\lambda})$ ③ $Y_1 \perp\!\!\!\perp Y_2$

Theorem 4.3.5 *Let X and Y be independent random variables. Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then the random variables $U = g(X)$ and $V = h(Y)$ are independent.*

Proof: We will prove the theorem assuming U and V are continuous random variables. For any $u \in \Re$ and $v \in \Re$, define

$$A_u = \{x : g(x) \le u\} \quad \text{and} \quad B_v = \{y : h(y) \le v\}.$$

Then the joint cdf of (U, V) is

$$\begin{aligned} F_{U,V}(u,v) &= P(U \le u, V \le v) && \text{(definition of cdf)} \\ &= P(X \in A_u, Y \in B_v) && \text{(definition of } U \text{ and } V) \\ &= P(X \in A_u) P(Y \in B_v). && \text{(Theorem 4.2.10)} \end{aligned}$$

The joint pdf of (U, V) is

$$\begin{aligned} f_{U,V}(u,v) &= \frac{\partial^2}{\partial u \partial v} F_{U,V}(u,v) && \text{(by (4.1.4))} \\ &= \left(\frac{d}{du} P(X \in A_u)\right)\left(\frac{d}{dv} P(Y \in B_v)\right), \end{aligned}$$

where, as the notation indicates, the first factor is a function only of u and the second factor is a function only of v. Hence, by Lemma 4.2.7, U and V are independent. □

Of course, in many situations, the transformation of interest is not one-to-one. Just as Theorem 2.1.8 generalized the univariate method to many-to-one functions, the same can be done here. As before, $\mathcal{A} = \{(x,y) : f_{X,Y}(x,y) > 0\}$. Suppose $A_0, A_1, \ldots, A_k$ form a partition of $\mathcal{A}$ with these properties. The set A_0, which may be empty, satisfies $P((X,Y) \in A_0) = 0$. The transformation $U = g_1(X,Y)$ and $V = g_2(X,Y)$ is a one-to-one transformation from A_i onto $\mathcal{B}$ for each $i = 1, 2, \ldots, k$. Then for each i, the inverse functions from $\mathcal{B}$ to A_i can be found. Denote the ith inverse by $x = h_{1i}(u,v)$ and $y = h_{2i}(u,v)$. This ith inverse gives, for $(u,v) \in \mathcal{B}$, the unique $(x,y) \in A_i$ such that $(u,v) = (g_1(x,y), g_2(x,y))$. Let J_i denote the Jacobian computed from the ith inverse. Then assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the joint pdf, $f_{U,V}(u,v)$:

$$f_{U,V}(u,v) = \sum_{i=1}^{k} f_{X,Y}(h_{1i}(u,v), h_{2i}(u,v)) |J_i|. \tag{4.3.6}$$

Example 4.3.6 (Distribution of the ratio of normal variables) Let X and Y be independent n$(0,1)$ random variables. Consider the transformation $U = X/Y$ and $V = |Y|$. (U and V can be defined to be any value, say $(1,1)$, if $Y = 0$ since $P(Y = 0) = 0$.) This transformation is not one-to-one since the points (x,y) and $(-x,-y)$ are both mapped into the same (u,v) point. But if we restrict consideration to either positive or negative values of y, then the transformation is one-to-one. In the above notation, let

$$A_1 = \{(x,y) : y > 0\}, \quad A_2 = \{(x,y) : y < 0\}, \quad \text{and} \quad A_0 = \{(x,y) : y = 0\}.$$

A_0, A_1, and A_2 form a partition of $\mathcal{A} = \Re^2$ and $P((X,Y) \in A_0) = P(Y = 0) = 0$. For either A_1 or A_2, if $(x,y) \in A_i$, $v = |y| > 0$, and for a fixed value of $v = |y|$, $u = x/y$ can be any real number since x can be any real number. Thus, $\mathcal{B} = \{(u,v) : v > 0\}$ is the image of both A_1 and A_2 under the transformation. Furthermore, the inverse transformations from $\mathcal{B}$ to A_1 and $\mathcal{B}$ to A_2 are given by $x = h_{11}(u,v) = uv$, $y = h_{21}(u,v) = v$, and $x = h_{12}(u,v) = -uv$, $y = h_{22}(u,v) = -v$. Note that the first inverse gives positive values of y and the second gives negative values of y. The Jacobians from the two inverses are $J_1 = J_2 = v$. Using

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-x^2/2} e^{-y^2/2},$$

from (4.3.6) we obtain

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi} e^{-(uv)^2/2} e^{-v^2/2} |v| + \frac{1}{2\pi} e^{-(-uv)^2/2} e^{-(-v)^2/2} |v| \\ &= \frac{v}{\pi} e^{-(u^2+1)v^2/2}, \quad -\infty < u < \infty, \quad 0 < v < \infty. \end{aligned}$$

From this the marginal pdf of U can be computed to be

$$\begin{aligned} f_U(u) &= \int_0^\infty \frac{v}{\pi} e^{-(u^2+1)v^2/2}\, dv \\ &= \frac{1}{2\pi} \int_0^\infty e^{-(u^2+1)z/2}\, dz \quad z = v^2) \qquad \text{(change of variable)} \\ &= \frac{1}{2\pi} \frac{2}{(u^2+1)} \qquad \begin{pmatrix} \text{integrand is kernel of} \\ \text{exponential } (\beta = 2/(u^2+1)) \text{ pdf} \end{pmatrix} \\ &= \frac{1}{\pi(u^2+1)}, \quad -\infty < u < \infty. \end{aligned}$$

So we see that the ratio of two independent standard normal random variables is a Cauchy random variable. (See Exercise 4.28 for more relationships between normal and Cauchy random variables.) ‖

Oct 21st 2021

4.4 Hierarchical Models and Mixture Distributions

Example 4.4.1 (Binomial-Poisson hierarchy) Perhaps the most classic hierarchical model is the following. An insect lays a large number of eggs, each surviving with probability p. On the average, how many eggs will survive?

The "large number" of eggs laid is a random variable, often taken to be Poisson(λ). Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials. Therefore, if we let X = number of survivors and Y = number of eggs laid, we have

$$X|Y \sim \text{binomial}(Y, p),$$
$$Y \sim \text{Poisson}(\lambda),$$

a hierarchical model. (Recall that we use notation such as $X|Y \sim \text{binomial}(Y, p)$ to mean that the conditional distribution of X given $Y = y$ is binomial(y, p).) ||

The advantage of the hierarchy is that complicated processes may be modeled by a sequence of relatively simple models placed in a hierarchy. Also, dealing with the hierarchy is no more difficult than dealing with conditional and marginal distributions.

Example 4.4.2 (Continuation of Example 4.4.1) The random variable of interest, X = number of survivors, has the distribution given by

$$\begin{aligned} P(X = x) &= \sum_{y=0}^{\infty} P(X = x, Y = y) \\ &= \sum_{y=0}^{\infty} P(X = x|Y = y)P(Y = y) && \left(\begin{array}{c}\text{definition of}\\ \text{conditional probability}\end{array}\right) \\ &= \sum_{y=x}^{\infty} \left[\binom{y}{x} p^x (1-p)^{y-x}\right] \left[\frac{e^{-\lambda}\lambda^y}{y!}\right], && \left(\begin{array}{c}\text{conditional probability}\\ \text{is 0 if } y < x\end{array}\right) \end{aligned}$$

since $X|Y = y$ is binomial(y, p) and Y is Poisson(λ). If we now simplify this last expression, canceling what we can and multiplying by λ^x/λ^x, we get

$$\begin{aligned} P(X = x) &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} (t = y - x) \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} && \left(\begin{array}{c}\text{sum is a kernel for}\\ \text{a Poisson distribution}\end{array}\right) \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p}, \end{aligned}$$

so $X \sim \text{Poisson}(\lambda p)$. Thus, any marginal inference on X is with respect to a Poisson(λp) distribution, with Y playing no part at all. Introducing Y in the hierarchy was mainly to aid our understanding of the model. There was an added bonus in that the parameter of the distribution of X is the product of two parameters, each relatively simple to understand.

The answer to the original question is now easy to compute:

$$\mathrm{E}X = \lambda p,$$

so, on the average, λp eggs will survive. If we were interested only in this mean and did not need the distribution, we could have used properties of conditional expectations. ||

Theorem 4.4.3 *If X and Y are any two random variables, then*

$$\mathrm{E}X = \mathrm{E}\left(\mathrm{E}(X|Y)\right), \tag{4.4.1}$$

E[g(x)] = E[E[g(x)|Y]]

provided that the expectations exist.

Proof: Let $f(x,y)$ denote the joint pdf of X and Y. By definition, we have

$$\mathrm{E}X = \int\int x f(x,y)\,dx\,dy = \int \left[\int x f(x|y)\,dx\right] f_Y(y)\,dy, \tag{4.4.2}$$

where $f(x|y)$ and $f_Y(y)$ are the conditional pdf of X given $Y = y$ and the marginal pdf of Y, respectively. But now notice that the inner integral in (4.4.2) is the conditional expectation $\mathrm{E}(X|y)$, and we have

$$\mathrm{E}X = \int \mathrm{E}(X|y) f_Y(y)\,dy = \mathrm{E}\left(\mathrm{E}(X|Y)\right),$$

as desired. Replace integrals by sums to prove the discrete case. □

We can now easily compute the expected number of survivors in Example 4.4.1. From Theorem 4.4.3 we have

$$\begin{aligned}
\mathrm{E}X &= \mathrm{E}\left(\mathrm{E}(X|Y)\right) \\
&= \mathrm{E}(pY) && (\text{since } X|Y \sim \text{binomial}(Y,p)) \\
&= p\lambda. && (\text{since } Y \sim \text{Poisson}(\lambda))
\end{aligned}$$

Example 4.4.5 (Generalization of Example 4.4.1) Consider a generalization of Example 4.4.1, where instead of one mother insect there are a large number of mothers and one mother is chosen at random. We are still interested in knowing the average number of survivors, but it is no longer clear that the number of eggs laid follows the same Poisson distribution for each mother. The following three-stage hierarchy may be more appropriate. Let X = number of survivors in a litter; then

$$\begin{aligned}
X|Y &\sim \text{binomial}(Y,p), \\
Y|\Lambda &\sim \text{Poisson}(\Lambda), \\
\Lambda &\sim \text{exponential}(\beta),
\end{aligned}$$

where the last stage of the hierarchy accounts for the variability across different mothers.

The mean of X can easily be calculated as

$$\begin{aligned}
\mathrm{E}X &= \mathrm{E}\left(\mathrm{E}(X|Y)\right) \\
&= \mathrm{E}(pY) && \text{(as before)} \\
&= \mathrm{E}\left(\mathrm{E}(pY|\Lambda)\right) \\
&= \mathrm{E}(p\Lambda) \\
&= p\beta, && \text{(exponential expectation)}
\end{aligned}$$

completing the calculation. ‖

Note that this three-stage model can also be thought of as a two-stage hierarchy by combining the last two stages. If $Y|\Lambda \sim \text{Poisson}(\Lambda)$ and $\Lambda \sim \text{exponential}(\beta)$, then

$$\begin{aligned}
P(Y=y) &= P(Y=y, 0<\Lambda<\infty) \\
&= \int_0^\infty f(y,\lambda)\,d\lambda \\
&= \int_0^\infty f(y|\lambda) f(\lambda)\,d\lambda \\
&= \int_0^\infty \left[\frac{e^{-\lambda}\lambda^y}{y!}\right]\frac{1}{\beta}e^{-\lambda/\beta}\,d\lambda \\
&= \frac{1}{\beta y!}\int_0^\infty \lambda^y e^{-\lambda(1+\beta^{-1})}\,d\lambda && \begin{pmatrix}\text{gamma}\\ \text{pdf kernel}\end{pmatrix} \\
&= \frac{1}{\beta y!}\Gamma(y+1)\left(\frac{1}{1+\beta^{-1}}\right)^{y+1} \\
&= \frac{1}{(1+\beta)}\left(\frac{1}{1+\beta^{-1}}\right)^{y}.
\end{aligned}$$

This expression for the pmf of Y is the form (3.2.10) of the negative binomial pmf. Therefore, our three-stage hierarchy in Example 4.4.5 is equivalent to the two-stage hierarchy

$$\begin{aligned}
X|Y &\sim \text{binomial}(Y,p), \\
Y &\sim \text{negative binomial}\left(p=\frac{1}{1+\beta}, r=1\right).
\end{aligned}$$

Theorem 4.4.7 (Conditional variance identity) *For any two random variables X and Y,*

$$\operatorname{Var} X = \operatorname{E}(\operatorname{Var}(X|Y)) + \operatorname{Var}(\operatorname{E}(X|Y)), \tag{4.4.4}$$

provided that the expectations exist.

Proof: By definition, we have

$$\operatorname{Var} X = \operatorname{E}\left([X - \operatorname{E}X]^2\right) = \operatorname{E}\left([X - \operatorname{E}(X|Y) + \operatorname{E}(X|Y) - \operatorname{E}X]^2\right),$$

where in the last step we have added and subtracted $\operatorname{E}(X|Y)$. Expanding the square in this last expectation now gives

$$\begin{aligned} \operatorname{Var} X = {} & \operatorname{E}\left([X - \operatorname{E}(X|Y)]^2\right) + \operatorname{E}\left([\operatorname{E}(X|Y) - \operatorname{E}X]^2\right) \\ & + 2\operatorname{E}\left([X - \operatorname{E}(X|Y)][\operatorname{E}(X|Y) - \operatorname{E}X]\right). \end{aligned} \tag{4.4.5}$$

The last term in this expression is equal to 0, however, which can easily be seen by iterating the expectation:

(4.4.6)

$$\operatorname{E}\left([X - \operatorname{E}(X|Y)][\operatorname{E}(X|Y) - \operatorname{E}X]\right) = \operatorname{E}\left(\operatorname{E}\left\{[X - \operatorname{E}(X|Y)][\operatorname{E}(X|Y) - \operatorname{E}X]|Y\right\}\right).$$

In the conditional distribution $X|Y$, X is the random variable. So in the expression

$$\operatorname{E}\left\{[X - \operatorname{E}(X|Y)][\operatorname{E}(X|Y) - \operatorname{E}X]|Y\right\},$$

$\operatorname{E}(X|Y)$ and $\operatorname{E}X$ are constants. Thus,

$$\begin{aligned} \operatorname{E}\left\{[X - \operatorname{E}(X|Y)][\operatorname{E}(X|Y) - \operatorname{E}X]|Y\right\} &= (\operatorname{E}(X|Y) - \operatorname{E}X)\left(\operatorname{E}\left\{[X - \operatorname{E}(X|Y)]|Y\right\}\right) \\ &= (\operatorname{E}(X|Y) - \operatorname{E}X)\left(\operatorname{E}(X|Y) - \operatorname{E}(X|Y)\right) \\ &= (\operatorname{E}(X|Y) - \operatorname{E}X)(0) \\ &= 0. \end{aligned}$$

Thus, from (4.4.6), we have that $\operatorname{E}((X - \operatorname{E}(X|Y))(\operatorname{E}(X|Y) - \operatorname{E}X)) = \operatorname{E}(0) = 0$. Referring back to equation (4.4.5), we see that

$$\begin{aligned} \operatorname{E}\left([X - \operatorname{E}(X|Y)]^2\right) &= \operatorname{E}\left(\operatorname{E}\left\{[X - \operatorname{E}(X|Y)]^2|Y\right\}\right) \\ &= \operatorname{E}(\operatorname{Var}(X|Y)) \end{aligned}$$

and

$$\operatorname{E}\left([\operatorname{E}(X|Y) - \operatorname{E}X]^2\right) = \operatorname{Var}(\operatorname{E}(X|Y)),$$

establishing (4.4.4). □

Example 4.4.6 (Beta-binomial hierarchy) One generalization of the binomial distribution is to allow the success probability to vary according to a distribution. A standard model for this situation is

$$X|P \sim \text{binomial}(P), \quad i = 1, \ldots, n,$$
$$P \sim \text{beta}(\alpha, \beta).$$

By iterating the expectation, we calculate the mean of X as

$$\text{E}X = \text{E}[\text{E}(X|P)] = \text{E}[nP] = n\frac{\alpha}{\alpha+\beta}. \qquad \|$$

Example 4.4.8 (Continuation of Example 4.4.6) To calculate the variance of X, we have from (4.4.4),

$$\text{Var}\, X = \text{Var}\left(\text{E}(X|P)\right) + \text{E}\left(\text{Var}(X|P)\right).$$

Now $\text{E}(X|P) = nP$, and since $P \sim \text{beta}(\alpha, \beta)$,

$$\text{Var}\left(\text{E}(X|P)\right) = \text{Var}(nP) = n^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Also, since $X|P$ is binomial(n, P), $\text{Var}(X|P) = nP(1-P)$. We then have

$$\text{E}\left[\text{Var}(X|P)\right] = n\text{E}\left[P(1-P)\right] = n\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p(1-p)p^{\alpha-1}(1-p)^{\beta-1}dp.$$

Notice that the integrand is the kernel of another beta pdf (with parameters $\alpha+1$ and $\beta+1$) so

$$\text{E}\left(\text{Var}(X|P)\right) = n\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left[\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}\right] = n\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}.$$

Adding together the two pieces and simplifying, we get

$$\text{Var}\, X = n\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}. \qquad \|$$

*Oct. 26th 2021

4.5 Covariance and Correlation

Throughout this section we will frequently be referring to the mean and variance of X and the mean and variance of Y. For these we will use the notation $EX = \mu_X$, $EY = \mu_Y$, $\text{Var}\, X = \sigma_X^2$, and $\text{Var}\, Y = \sigma_Y^2$. We will assume throughout that $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$.

Definition 4.5.1 The *covariance of X and Y* is the number defined by

$$\text{Cov}(X, Y) = \text{E}\left((X - \mu_X)(Y - \mu_Y)\right).$$

Definition 4.5.2 The *correlation of X and Y* is the number defined by

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

The value ρ_{XY} is also called the *correlation coefficient.*

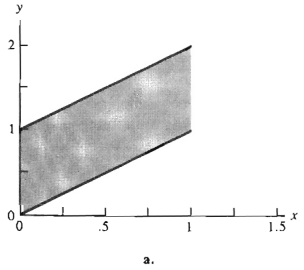
Theorem 4.5.3 *For any random variables X and Y,*

$$\text{Cov}(X, Y) = EXY - \mu_X \mu_Y.$$

Proof:
$$\begin{aligned}
\text{Cov}(X, Y) &= \text{E}\,(X - \mu_X)(Y - \mu_Y) \\
&= \text{E}\,(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) && \text{(expanding the product)} \\
&= EXY - \mu_X EY - \mu_Y EX + \mu_X \mu_Y && (\mu_X \text{ and } \mu_Y \text{ are constants}) \\
&= EXY - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\
&= EXY - \mu_X \mu_Y. && \square
\end{aligned}$$

Example 4.5.4 (Correlation–I) Let the joint pdf of (X, Y) be $f(x, y) = 1$, $0 < x < 1, x < y < x + 1$. See Figure 4.5.1. The marginal distribution of X is uniform$(0, 1)$ so $\mu_X = \frac{1}{2}$ and $\sigma_X^2 = \frac{1}{12}$. The marginal pdf of Y is $f_Y(y) = y$, $0 < y < 1$, and $f_Y(y) = 2 - y, 1 \le y < 2$, with $\mu_Y = 1$ and $\sigma_Y^2 = \frac{1}{6}$. We also have

$$\begin{aligned}
EXY &= \int_0^1 \int_x^{x+1} xy \, dy \, dx = \int_0^1 \frac{1}{2} x y^2 \Big|_x^{x+1} dx \\
&= \int_0^1 \left(x^2 + \frac{1}{2}x\right) dx = \frac{7}{12}.
\end{aligned}$$



a.

Using Theorem 4.5.3, we have $\text{Cov}(X, Y) = \frac{7}{12} - \left(\frac{1}{2}\right)(1) = \frac{1}{12}$. The correlation is

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1/12}{\sqrt{1/12}\sqrt{1/6}} = \frac{1}{\sqrt{2}}.$$

Theorem 4.5.5 *If X and Y are independent random variables, then* $\text{Cov}(X,Y) = 0$ *and* $\rho_{XY} = 0$.

Proof: Since X and Y are independent, from Theorem 4.2.10 we have $EXY = (EX)(EY)$. Thus

$$\text{Cov}(X,Y) = EXY - (EX)(EY) = (EX)(EY) - (EX)(EY) = 0$$

and

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{0}{\sigma_X \sigma_Y} = 0. \qquad \square$$

Thus, the values $\text{Cov}(X,Y) = \rho_{XY} = 0$ in some sense indicate that there is no relationship between X and Y. It is important to note, however, that Theorem 4.5.5 does *not* say that if $\text{Cov}(X,Y) = 0$, then X and Y are independent. For example, if $X \sim f(x-\theta)$, symmetric around 0 with $EX = \theta$, and Y is the indicator function $Y = I(|X-\theta| < 2)$, then X and Y are obviously not independent. However,

$$\text{E}(XY) = \int_{-\infty}^{\infty} xI(|x-\theta|<2) f(x-\theta)\,dx = \int_{-2}^{2} (t+\theta) f(t)\,dt = \theta \int_{-2}^{2} f(t)\,dt = EXEY,$$

where we used the fact that, by symmetry, $\int_{-2}^{2} t f(t)\,dt = 0$. So it is easy to find uncorrelated, dependent random variables.

Theorem 4.5.6 *If X and Y are any two random variables and a and b are any two constants, then*

$$\text{Var}(aX+bY) = a^2 \text{Var}\, X + b^2 \text{Var}\, Y + 2ab\,\text{Cov}(X,Y).$$

If X and Y are independent random variables, then

$$\text{Var}(aX+bY) = a^2 \text{Var}\, X + b^2 \text{Var}\, Y.$$

Proof: The mean of $aX+bY$ is $\text{E}(aX+bY) = a\text{E}X + b\text{E}Y = a\mu_X + b\mu_Y$. Thus

$$\begin{aligned}
\text{Var}(aX+bY) &= \text{E}\left((aX+bY) - (a\mu_X + b\mu_Y)\right)^2 \\
&= \text{E}\left(a(X-\mu_X) + b(Y-\mu_Y)\right)^2 \\
&= \text{E}\left(a^2(X-\mu_X)^2 + b^2(Y-\mu_Y)^2 + 2ab(X-\mu_X)(Y-\mu_Y)\right) \\
&= a^2\text{E}(X-\mu_X)^2 + b^2\text{E}(Y-\mu_Y)^2 + 2ab\text{E}(X-\mu_X)(Y-\mu_Y) \\
&= a^2 \text{Var}\, X + b^2 \text{Var}\, Y + 2ab\,\text{Cov}(X,Y).
\end{aligned}$$

If X and Y are independent, then, from Theorem 4.5.5, $\text{Cov}(X,Y) = 0$ and the second equality is immediate from the first. $\square$

Theorem 4.5.7 *For any random variables X and Y,*

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

b. $|\rho_{XY}| = 1$ *if and only if there exist numbers* $a \neq 0$ *and* b *such that* $P(Y = aX + b) = 1$. *If* $\rho_{XY} = 1$, *then* $a > 0$, *and if* $\rho_{XY} = -1$, *then* $a < 0$.

Proof: Consider the function $h(t)$ defined by

$$h(t) = \mathrm{E}\left((X - \mu_X)t + (Y - \mu_Y)\right)^2.$$

Expanding this expression, we obtain

$$\begin{aligned} h(t) &= t^2\mathrm{E}(X - \mu_X)^2 + 2t\mathrm{E}(X - \mu_X)(Y - \mu_Y) + \mathrm{E}(Y - \mu_Y)^2 \\ &= t^2\sigma_X^2 + 2t\mathrm{Cov}(X, Y) + \sigma_Y^2. \end{aligned}$$

This quadratic function of t is greater than or equal to 0 for all values of t since it is the expected value of a nonnegative random variable. Thus, this quadratic function can have at most one real root and thus must have a nonpositive discriminant. That is,

$$(2\mathrm{Cov}(X, Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0.$$

This is equivalent to

$$-\sigma_X\sigma_Y \leq \mathrm{Cov}(X, Y) \leq \sigma_X\sigma_Y.$$

Dividing by $\sigma_X\sigma_Y$ yields

$$-1 \leq \frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y} = \rho_{XY} \leq 1.$$

Also, $|\rho_{XY}| = 1$ if and only if the discriminant is equal to 0. That is, $|\rho_{XY}| = 1$ if and only if $h(t)$ has a single root. But since $((X - \mu_X)t + (Y - \mu_Y))^2 \geq 0$, the expected value $h(t) = \mathrm{E}((X - \mu_X)t + (Y - \mu_Y))^2 = 0$ if and only if

$$P\big([(X - \mu_X)t + (Y - \mu_Y)]^2 = 0\big) = 1.$$

This is equivalent to

$$P((X - \mu_X)t + (Y - \mu_Y) = 0) = 1.$$

This is $P(Y = aX + b) = 1$ with $a = -t$ and $b = \mu_X t + \mu_Y$, where t is the root of $h(t)$. Using the quadratic formula, we see that this root is $t = -\mathrm{Cov}(X, Y)/\sigma_X^2$. Thus $a = -t$ has the same sign as ρ_{XY}, proving the final assertion. □

If there is a line $y = ax + b$, with $a \neq 0$, such that the values of (X, Y) have a high probability of being near this line, then the correlation between X and Y will be near 1 or -1. But if no such line exists, the correlation will be near 0. This is an intuitive notion of the linear relationship that is being measured by correlation. This idea will be illustrated further in the next two examples.

Example 4.5.8 (Correlation–II) This example is similar to Example 4.5.4, but we develop it differently to illustrate other model building and computational techniques. Let X have a uniform$(0, 1)$ distribution and Z have a uniform$(0, \frac{1}{10})$ distribution. Suppose X and Z are independent. Let $Y = X + Z$ and consider the random vector (X, Y). The joint distribution of (X, Y) can be derived from the joint distribution of (X, Z) using the techniques of Section 4.3. The joint pdf of (X, Y) is

$$f(x, y) = 10, \quad 0 < x < 1, \quad x < y < x + \frac{1}{10}.$$

Rather than using the formal techniques of Section 4.3, we can justify this as follows. Given $X = x$, $Y = x + Z$. The conditional distribution of Z given $X = x$ is just uniform$(0, \frac{1}{10})$ since X and Z are independent. Thus x serves as a location parameter in the conditional distribution of Y given $X = x$, and this conditional distribution is just uniform$(x, x + \frac{1}{10})$. Multiplying this conditional pdf by the marginal pdf of X (uniform$(0, 1)$) yields the joint pdf above. This representation of $Y = X + Z$ makes the computation of the covariance and correlation easy. The expected values of X and Y are $EX = \frac{1}{2}$ and $EY = E(X + Z) = EX + EZ = \frac{1}{2} + \frac{1}{20} = \frac{11}{20}$, giving

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= EXY - (EX)(EY) \\
&= EX(X + Z) - (EX)(E(X + Z)) \\
&= EX^2 + EXZ - (EX)^2 - (EX)(EZ) \\
&= EX^2 - (EX)^2 + (EX)(EZ) - (EX)(EZ) \quad \begin{pmatrix}\text{independence of} \\ X \text{ and } Z\end{pmatrix} \\
&= \sigma_X^2 = \frac{1}{12}.
\end{aligned}$$

From Theorem 4.5.6, the variance of Y is $\sigma_Y^2 = \mathrm{Var}(X + Z) = \mathrm{Var}\, X + \mathrm{Var}\, Z = \frac{1}{12} + \frac{1}{1200}$. Thus

$$\rho_{XY} = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}}\sqrt{\frac{1}{12} + \frac{1}{1200}}} = \sqrt{\frac{100}{101}}.$$

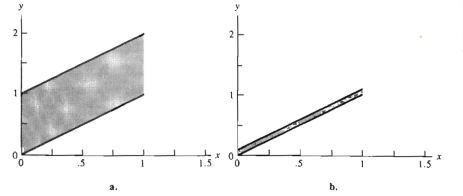


Figure 4.5.1. (a) *Region where $f(x, y) > 0$ for Example 4.5.4;* (b) *region where $f(x, y) > 0$ for Example 4.5.8*

This is much larger than the value of $\rho_{XY} = 1/\sqrt{2}$ obtained in Example 4.5.4. The sets on which $f(x, y)$ is positive for Example 4.5.4 and this example are illustrated in Figure 4.5.1. (Recall that this set is called the support of a distribution.) In each case, (X, Y) is a random point from the set. In both cases there is a linearly increasing relationship between X and Y, but the relationship is much stronger in Figure 4.5.1b. Another way to see this is by noting that in this example, the conditional distribution of Y given $X = x$ is uniform$(x, x + \frac{1}{10})$. In Example 4.5.4, the conditional distribution of Y given $X = x$ is uniform$(x, x + 1)$. The knowledge that $X = x$ gives us much more information about the value of Y in this model than in the one in Example 4.5.4. Hence the correlation is nearer to 1 in this example. ‖

The next example illustrates that there may be a strong relationship between X and Y, but if the relationship is not linear, the correlation may be small.

Example 4.5.9 (Correlation–III) In this example, let X have a uniform$(-1, 1)$ distribution and let Z have a uniform$(0, \frac{1}{10})$ distribution. Let X and Z be independent. Let $Y = X^2 + Z$ and consider the random vector (X, Y). As in Example 4.5.8, given $X = x$, $Y = x^2 + Z$ and the conditional distribution of Y given $X = x$ is uniform$(x^2, x^2 + \frac{1}{10})$. The joint pdf of X and Y, the product of this conditional pdf and the marginal pdf of X, is thus

$$f(x, y) = 5, \quad -1 < x < 1, \quad x^2 < y < x^2 + \frac{1}{10}.$$

The set on which $f(x, y) > 0$ is illustrated in Figure 4.5.2. There is a strong relationship between X and Y, as indicated by the conditional distribution of Y given $X = x$. But the relationship is not linear. The possible values of (X, Y) cluster around a parabola rather than a straight line. The correlation does not measure this nonlinear relationship. In fact, $\rho_{XY} = 0$. Since X has a uniform$(-1, 1)$ distribution, $EX = EX^3 = 0$, and since X and Z are independent, $EXZ = (EX)(EZ)$. Thus,

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathrm{E}(X(X^2 + Z)) - (EX)(\mathrm{E}(X^2 + Z)) \\ &= EX^3 + EXZ - 0\mathrm{E}(X^2 + Z) \\ &= 0 + (EX)(EZ) = 0(EZ) = 0, \end{aligned}$$

and $\rho_{XY} = \mathrm{Cov}(X, Y)/(\sigma_X \sigma_Y) = 0$.

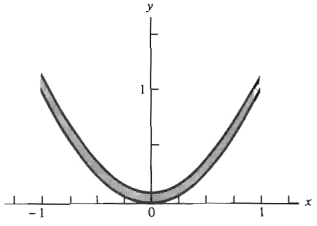


Figure 4.5.2. *Region where $f(x, y) > 0$ for Example 4.5.9*

Oct. 28th 2021

Ex: Let $X_1, \dots, X_n$ be independent random variables all with $\mathbb{E}X_i = \mu$, $\text{var}(X_i) = \sigma^2$.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. The random variable

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is called the sample variable, calculate ⓐ $\text{var}(\bar{X})$ ⓑ $\mathbb{E}(S^2)$

ⓐ $\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) \overset{\text{indep}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

$\mathbb{E}(\bar{X}) = \mu$

ⓑ $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2$

$= \sum_{i=1}^{n}(X_i-\mu)^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 - 2\sum_{i=1}^{n}(X_i-\mu)(\bar{X}-\mu)$

$\Rightarrow \mathbb{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right) = \sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^2 + \sum_{i=1}^{n}\mathbb{E}(\bar{X}-\mu)^2 - 2\sum_{i=1}^{n}\mathbb{E}(X_i-\mu)(\bar{X}-\mu)$

$= \sum_{i=1}^{n}\left(\sigma^2 + \frac{\sigma^2}{n} - 2\text{cov}(X_i, \bar{X})\right)$

$= \sum_{i=1}^{n}\left(\sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n}\right)$

$= (n-1)\sigma^2$

$\Rightarrow \mathbb{E}(S^2) = \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right) = \sigma^2$

$\text{cov}(X_i, \bar{X})$

$= \text{cov}\left(X_i, \sum_{j=1}^{n} X_j/n\right)$

$= \frac{1}{n}\sum_{j=1}^{n}\text{cov}(X_i, X_j)$

$= \frac{\sigma^2}{n}$

Ex A group of N people throw their hats into the center of the room. The hats are mixed, and each person randomly selects one

Let X = # of people who select their own hat

Compute $\mathbb{E}(X)$, $\text{var}(X)$

- Number the people from 1 to N. Let: $X_i = \begin{cases} 1 & \text{; if person } i \text{ choose own hat} \\ 0 & \text{; otherwise} \end{cases}$

So $X = X_1 + X_2 + \cdots + X_N$ $\quad P(X_i = 1) = \frac{1}{N} \Rightarrow \mathbb{E}(X_i) = 1\cdot\frac{1}{N} + 0\left(1-\frac{1}{N}\right) = \frac{1}{N}$

$\Rightarrow \mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N}\mathbb{E}(X_i) = N\cdot\frac{1}{N} = 1$

$\left[\text{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N}\text{Var}(X_i) + 2\sum_{i<j}\text{cov}(X_i, X_j)\right]$

NO.

Date

$X_i X_j = \begin{cases} 1 & \text{if persons } i \& j \text{ both choose their own hat} \\ 0 & \text{otherwise} \end{cases}$

So $P(X_i=1, X_j=1) = P(X_j=1 \mid X_i=1) P(X_i=1) = \frac{1}{N(N-1)}$

$\Rightarrow E(X_i, X_j) = \frac{1}{N(N-1)}$

$\Rightarrow Cov(X_i, X_j) = E(X_i X_j) - E(X_i) E(X_j) = \frac{1}{N(N-1)} - \frac{1}{N^2} = \frac{1}{N^2(N-1)}$

$var(X) = \underbrace{N \frac{N-1}{N^2}}_{N \cdot var(Bernoulli(\frac{1}{N}))} + 2\binom{N}{2} \frac{1}{N^2(N-1)} = \frac{N-1}{N} + \frac{1}{N} = 1$

- Let N be a r.v. (discrete) with values in the natural numbers $(1, 2, \dots)$ & $E(N) < +\infty$. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variable (independent & identically distributed) that are independent of N. Also, assume that $E(X_i) = E(X_1) < +\infty$. Define $Y = X_1 + X_2 + \cdots + X_N$.

What is $E(Y)$? $Var(Y)$? (assume $var(N) < +\infty$, $var(X_i) < +\infty$)

- $E(Y) = E(E(Y \mid N)) = E(E(\sum_{i=1}^{N} X_i \mid N=n)) = E(N \cdot E(X_1)) = E(X_1) E(N)$

$E(Y \mid N) = N E(X_1) \Rightarrow var(E(Y \mid N)) = Var(N E(X_1)) = (E(X_1))^2 var(N)$

$var(Y \mid N) = N Var(X_1) \Rightarrow E(Var(Y \mid N)) = Var(X_1) E(N)$

$\Rightarrow Var(Y) = E(Var(Y \mid N)) + Var(E(Y \mid N))$

$= E(N) Var(X) + (E(X_1))^2 Var(N)$

Def (bivariate Normal distribution):

Let $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y > 0$, and $-1 < \rho < 1$. The bivariate normal pdf with means μ_X & μ_Y, variances σ_X^2 & σ_Y^2, and covariance ρ, is the

Date

bivariate pdf given by

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)\right]$$

$x, y \in \mathbb{R}$

Properties

① Marginal distribution of X is Normal (μ_x, σ_x^2)

② Marginal distribution of Y is Normal (μ_Y, σ_y^2)

③ The correlation between X & Y is $\rho_{XY}=\rho$

④ $Y|X=x$ is also Normal $(\mu_y+\rho(\sigma_Y/\sigma_x)(X-\mu_x),\ \sigma_y^2(1-\rho^2))$

⑤ For any constants a & b, the distribution of $aX+bY$ is Normal with mean $=a\mu_x+b\mu_Y$, and var $=a^2\sigma_x^2+b^2\sigma_y^2+2ab\rho\sigma_x\sigma_y$

proof ① ② $f_X(x)=\int_{\mathbb{R}} f(x,y)\,dy=\int_{-\infty}^{+\infty}\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(w^2-2\rho wz+z^2)\right)\sigma_y\,dz$

Define $z=\frac{y-\mu_y}{\sigma_y}$

$dy=\sigma_y dz$ $w=\frac{x-\mu_x}{\sigma_x}$

$$=\frac{\exp\left(-\frac{w^2}{2(1-\rho^2)}\right)}{2\pi\sigma_x\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2(1-\rho^2)}(z^2-2\rho wz+\rho^2w^2-\rho^2w^2)\right)dz$$

$$=\frac{e^{-\frac{w^2}{2(1-\rho^2)}}e^{\frac{\rho^2w^2}{2(1-\rho^2)}}}{2\pi\sigma_x\sqrt{1-\rho^2}}\underbrace{\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2(1-\rho^2)}(z-\rho w)^2\right)dz}$$

↳ pdf with normal with mean ρw variance $1-\rho^2$

$$=\frac{e^{-w^2/2}}{2\pi\sigma_x\sqrt{1-\rho^2}}\sqrt{2\pi}\sqrt{1-\rho^2}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_x}e^{-w^2/2}=\frac{1}{\sqrt{2\pi}\sigma_x}e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}},\quad x\in\mathbb{R}$$

② is the same

④ $$f_{Y|X}(y|x)=\frac{\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)\right]}{\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2-(1-\rho^2)\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right)\right)$$

NO.

Date

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\rho^2\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)\right)$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2\sigma_y^2(1-\rho^2)}\left((y-\mu_y)-\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2\right)$$

$\Rightarrow Y|X\sim$ Normal $\left(\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x),\ \sigma_y^2(1-\rho^2)\right)$

※ Nov. 2nd 2021

Definition 4.5.10 Let $-\infty < \mu_X < \infty, -\infty < \mu_Y < \infty$, $0 < \sigma_X$, $0 < \sigma_Y$, and $-1 < \rho < 1$ be five real numbers. The *bivariate normal pdf with means* μ_X *and* μ_Y, *variances* σ_X^2 *and* σ_Y^2, *and correlation* ρ is the bivariate pdf given by

$$f(x,y) = \left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \times \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

a. The marginal distribution of X is $\mathrm{n}(\mu_X, \sigma_X^2)$.

b. The marginal distribution of Y is $\mathrm{n}(\mu_Y, \sigma_Y^2)$.

c. The correlation between X and Y is $\rho_{XY} = \rho$.

d. For any constants a and b, the distribution of $aX + bY$ is $\mathrm{n}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$.

We will leave the verification of properties (a), (b), and (d) as exercises (Exercise 4.45). Assuming (a) and (b) are true, we will prove (c). We have by definition

$$\begin{aligned}\rho_{XY} &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} \\ &= \frac{\mathrm{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} \\ &= \mathrm{E}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) f(x,y)\,dx\,dy.\end{aligned}$$

Make the change of variable

$$s = \left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) \quad \text{and} \quad t = \left(\frac{x-\mu_X}{\sigma_X}\right).$$

Then $x = \sigma_X t + \mu_X$, $y = (\sigma_Y s/t) + \mu_Y$, and the Jacobian of the transformation is $J = \sigma_X\sigma_Y/t$. With this change of variable, we obtain

$$\begin{aligned}\rho_{XY} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} sf\left(\sigma_X t + \mu_X, \frac{\sigma_Y s}{t} + \mu_Y\right)\left|\frac{\sigma_X\sigma_Y}{t}\right| ds\,dt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s\left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \\ &\quad \times \exp\left(-\frac{1}{2(1-\rho^2)}\left(t^2 - 2\rho s + \left(\frac{s}{t}\right)^2\right)\right)\frac{\sigma_X\sigma_Y}{|t|}\,ds\,dt.\end{aligned}$$

Noting that $|t| = \sqrt{t^2}$ and $t^2 - 2\rho s + \left(\frac{s}{t}\right)^2 = \left(\frac{s-\rho t^2}{t}\right)^2 + (1-\rho^2)t^2$, we can rewrite this as

$$\rho_{XY} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \left[\int_{-\infty}^{\infty} \frac{s}{\sqrt{2\pi}\sqrt{(1-\rho^2)t^2}} \exp\left(-\frac{(s-\rho t^2)^2}{2(1-\rho^2)t^2}\right) ds \right] dt.$$

The inner integral is ES, where S is a normal random variable with $ES = \rho t^2$ and $\operatorname{Var} S = (1-\rho^2)t^2$. Thus the inner integral is ρt^2. Hence we have

$$\rho_{XY} = \int_{-\infty}^{\infty} \frac{\rho t^2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt.$$

But this integral is ρET^2, where T is a $\mathrm{n}(0,1)$ random variable. Hence $ET^2 = 1$ and $\rho_{XY} = \rho$.

All the conditional distributions of Y given $X = x$ and of X given $Y = y$ are also normal distributions. Using the joint and marginal pdfs given above, it is straightforward to verify that the conditional distribution of Y given $X = x$ is

$$\mathrm{n}(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \sigma_Y^2(1-\rho^2)).$$

As ρ converges to 1 or -1, the conditional variance $\sigma_Y^2(1-\rho^2)$ converges to 0. Thus, the conditional distribution of Y given $X = x$ becomes more concentrated about the point $\mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X)$, and the joint probability distribution of (X, Y) becomes more concentrated about the line $y = \mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X)$. This illustrates again the point made earlier that a correlation near 1 or -1 means that there is a line $y = ax + b$ about which the values of (X, Y) cluster with high probability.

Note one important fact: All of the normal marginal and conditional pdfs are derived from the starting point of bivariate normality. The derivation does not go in the opposite direction. That is, marginal normality does not imply joint normality. See Exercise 4.47 for an illustration of this.

proof of (d):

$$M_{aX+bY}(t) = E(e^{t(aX+bY)}) = E(e^{taX+tbY})$$

$$= E(E(e^{taX+tbY}|X))$$

$$= E(e^{taX} \cdot E(e^{tbY}|X))$$

$$= E\left(e^{taX} \cdot e^{tb(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)) + \frac{t^2b^2(1-\rho^2)\sigma_Y^2}{2}}\right)$$

$$= e^{tb\mu_Y - tb\rho\frac{\sigma_Y}{\sigma_X}\mu_X + \frac{t^2b^2(1-\rho^2)\sigma_Y^2}{2}} E\left(e^{(ta + \frac{tb\rho\sigma_Y}{\sigma_X})X}\right)$$

$$= \exp\left[tb\mu_Y - \frac{tb\rho\sigma_Y}{\sigma_X}\mu_X + \frac{t^2b^2(1-\rho^2)\sigma_Y^2}{2} + ta\mu_X + \frac{tb\rho\sigma_Y}{\sigma_X}\mu_X + \frac{t^2a^2\sigma_X^2}{2} + t^2ab\rho\sigma_X\sigma_Y + \frac{t^2b^2\rho^2\sigma_Y^2}{2}\right]$$

$$= \exp\left[t(a\mu_X + b\mu_Y) + \frac{t^2(a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_X\sigma_Y)}{2}\right]$$

$$\Rightarrow aX + bY \sim N(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_X\sigma_Y)$$

4.6 Multivariate Distributions

The random vector $\mathbf{X} = (X_1, \ldots, X_n)$ has a sample space that is a subset of $\Re^n$. If $(X_1, \ldots, X_n)$ is a discrete random vector (the sample space is countable), then the *joint pmf of* $(X_1, \ldots, X_n)$ is the function defined by $f(\mathbf{x}) = f(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$ for each $(x_1, \ldots, x_n) \in \Re^n$. Then for any $A \subset \Re^n$,

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}). \tag{4.6.1}$$

If $(X_1, \ldots, X_n)$ is a continuous random vector, the *joint pdf of* $(X_1, \ldots, X_n)$ is a function $f(x_1, \ldots, x_n)$ that satisfies

$$P(\mathbf{X} \in A) = \int \cdots \int_A f(\mathbf{x}) d\mathbf{x} = \int \cdots \int_A f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n. \tag{4.6.2}$$

Let $g(\mathbf{x}) = g(x_1, \ldots, x_n)$ be a real-valued function defined on the sample space of $\mathbf{X}$. Then $g(\mathbf{X})$ is a random variable and the *expected value of* $g(\mathbf{X})$ is

$$\mathrm{E}g(\mathbf{X}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad \text{and} \quad \mathrm{E}g(\mathbf{X}) = \sum_{\mathbf{x} \in \Re^n} g(\mathbf{x}) f(\mathbf{x}) \tag{4.6.3}$$

in the continuous and discrete cases, respectively. These and other definitions are analogous to the bivariate definitions except that now the integrals or sums are over the appropriate subset of $\Re^n$ rather than $\Re^2$.

The *marginal pdf or pmf* of any subset of the coordinates of $(X_1, \ldots, X_n)$ can be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates. Thus, for example, the marginal distribution of $(X_1, \ldots, X_k)$, the first k coordinates of $(X_1, \ldots, X_n)$, is given by the pdf or pmf

$$f(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_{k+1} \cdots dx_n \tag{4.6.4}$$

or

$$f(x_1, \ldots, x_k) = \sum_{(x_{k+1}, \ldots, x_n) \in \Re^{n-k}} f(x_1, \ldots, x_n) \tag{4.6.5}$$

for every $(x_1, \ldots, x_k) \in \Re^k$. The *conditional pdf or pmf* of a subset of the coordinates of $(X_1, \ldots, X_n)$ given the values of the remaining coordinates is obtained by dividing the joint pdf or pmf by the marginal pdf or pmf of the remaining coordinates. Thus, for example, if $f(x_1, \ldots, x_k) > 0$, the conditional pdf or pmf of $(X_{k+1}, \ldots, X_n)$ given $X_1 = x_1, \ldots, X_k = x_k$ is the function of $(x_{k+1}, \ldots, x_n)$ defined by

$$f(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_k)}. \tag{4.6.6}$$

These ideas are illustrated in the following example.

Example 4.6.1 (Multivariate pdfs) Let $n = 4$ and

$$f(x_1, x_2, x_3, x_4) = \begin{cases} \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) & 0 < x_i < 1, i = 1, 2, 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

This nonnegative function is the joint pdf of a random vector (X_1, X_2, X_3, X_4) and it can be verified that

$$\begin{aligned} &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_1\, dx_2\, dx_3\, dx_4 \\ &\quad = \int_0^1\int_0^1\int_0^1\int_0^1 \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2)\, dx_1\, dx_2\, dx_3\, dx_4 \\ &\quad = 1. \end{aligned}$$

This joint pdf can be used to compute probabilities such as

$$\begin{aligned} &P\left(X_1 < \frac{1}{2}, X_2 < \frac{3}{4}, X_4 > \frac{1}{2}\right) \\ &\quad = \int_{\frac{1}{2}}^1\int_0^1\int_0^{\frac{3}{4}}\int_0^{\frac{1}{2}} \frac{3}{4}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2\right)\, dx_1\, dx_2\, dx_3\, dx_4. \end{aligned}$$

Note how the limits of integration restrict the integration to those values of (x_1, x_2, x_3, x_4) that are in the event in question and for which $f(x_1, x_2, x_3, x_4) > 0$. Each of the four terms, $\frac{3}{4}x_1^2$, $\frac{3}{4}x_2^2$, etc., can be integrated separately and the results summed. For example,

$$\int_{\frac{1}{2}}^1\int_0^1\int_0^{\frac{3}{4}}\int_0^{\frac{1}{2}} \frac{3}{4}x_1^2\, dx_1\, dx_2\, dx_3\, dx_4 = \frac{3}{256}.$$

The other three integrals are $\frac{7}{1024}$, $\frac{3}{64}$, and $\frac{21}{256}$. Thus

$$P\left(X_1 < \frac{1}{2}, X_2 < \frac{3}{4}, X_4 > \frac{1}{2}\right) = \frac{3}{256} + \frac{7}{1024} + \frac{3}{64} + \frac{21}{256} = \frac{151}{1024}.$$

Using (4.6.4), we can obtain the marginal pdf of (X_1, X_2) by integrating out the variables x_3 and x_4 to obtain

$$\begin{aligned} f(x_1, x_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_3\, dx_4 \\ &= \int_0^1\int_0^1 \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2)\, dx_3\, dx_4 = \frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2} \end{aligned}$$

for $0 < x_1 < 1$ and $0 < x_2 < 1$. Any probability or expected value that involves only X_1 and X_2 can be computed using this marginal pdf. For example,

$$\begin{aligned} \mathrm{E}X_1X_2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f(x_1, x_2)\, dx_1\, dx_2 \\ &= \int_0^1\int_0^1 x_1x_2\left(\frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2}\right) dx_1\, dx_2 \\ &= \int_0^1\int_0^1 \left(\frac{3}{4}x_1^3x_2 + \frac{3}{4}x_1x_2^3 + \frac{1}{2}x_1x_2\right) dx_1\, dx_2 \\ &= \int_0^1 \left(\frac{3}{16}x_2 + \frac{3}{8}x_2^3 + \frac{1}{4}x_2\right) dx_2 = \frac{3}{32} + \frac{3}{32} + \frac{1}{8} = \frac{5}{16}. \end{aligned}$$

For any (x_1, x_2) with $0 < x_1 < 1$ and $0 < x_2 < 1$, $f(x_1, x_2) > 0$ and the conditional pdf of (X_3, X_4) given $X_1 = x_1$ and $X_2 = x_2$ can be found using (4.6.6). For any such (x_1, x_2), $f(x_1, x_2, x_3, x_4) > 0$ if $0 < x_3 < 1$ and $0 < x_4 < 1$, and for these values of (x_3, x_4), the conditional pdf is

$$\begin{aligned} f(x_3, x_4|x_1, x_2) &= \frac{f(x_1, x_2, x_3, x_4)}{f(x_1, x_2)} \\ &= \frac{\frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2)}{\frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2}} \\ &= \frac{x_1^2 + x_2^2 + x_3^2 + x_4^2}{x_1^2 + x_2^2 + \frac{2}{3}}. \end{aligned}$$

For example, the conditional pdf of (X_3, X_4) given $X_1 = \frac{1}{3}$ and $X_2 = \frac{2}{3}$ is

$$f\left(x_3, x_4\middle|x_1 = \frac{1}{3}, x_2 = \frac{2}{3}\right) = \frac{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + x_3^2 + x_4^2}{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \frac{2}{3}} = \frac{5}{11} + \frac{9}{11}x_3^2 + \frac{9}{11}x_4^2.$$

This can be used to compute

$$\begin{aligned} P\left(X_3 > \frac{3}{4}, X_4 < \frac{1}{2}\middle|X_1 = \frac{1}{3}, X_2 = \frac{2}{3}\right) &= \int_0^{\frac{1}{2}}\int_{\frac{3}{4}}^1 \left(\frac{5}{11} + \frac{9}{11}x_3^2 + \frac{9}{11}x_4^2\right) dx_3\, dx_4 \\ &= \int_0^{\frac{1}{2}} \left(\frac{5}{44} + \frac{111}{704} + \frac{9}{44}x_4^2\right) dx_4 \\ &= \frac{5}{88} + \frac{111}{1408} + \frac{3}{352} = \frac{203}{1408}. \end{aligned}$$

Definition 4.6.2 Let n and m be positive integers and let $p_1, \ldots, p_n$ be numbers satisfying $0 \le p_i \le 1$, $i = 1, \ldots, n$, and $\sum_{i=1}^n p_i = 1$. Then the random vector $(X_1, \ldots, X_n)$ has a *multinomial distribution with m trials and cell probabilities* $p_1, \ldots, p_n$ if the joint pmf of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \frac{m!}{x_1! \cdot \cdots \cdot x_n!} p_1^{x_1} \cdot \cdots \cdot p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!}$$

on the set of $(x_1, \ldots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$.

Theorem 4.6.4 (Multinomial Theorem) *Let m and n be positive integers. Let $\mathcal{A}$ be the set of vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$. Then, for any real numbers $p_1, \ldots, p_n$,*

$$(p_1 + \cdots + p_n)^m = \sum_{\mathbf{x} \in \mathcal{A}} \frac{m!}{x_1! \cdot \cdots \cdot x_n!} p_1^{x_1} \cdot \cdots \cdot p_n^{x_n}.$$

Theorem 4.6.4 shows that a multinomial pmf sums to 1. The set $\mathcal{A}$ is the set of points with positive probability in Definition 4.6.2. The sum of the pmf over all those points is, by Theorem 4.6.4, $(p_1 + \cdots + p_n)^m = 1^m = 1$.

Now we consider some marginal and conditional distributions for the multinomial model. Consider a single coordinate X_i. If the occurrence of the ith outcome is labeled a "success" and anything else is labeled a "failure," then X_i is the count of the number of successes in m independent trials where the probability of a success is p_i on each trial. Thus X_i should have a binomial(m, p_i) distribution. To verify this the marginal distribution of X_i should be computed using (4.6.5). For example, consider the marginal pmf of X_n. For a fixed value of $x_n \in \{0, 1, \ldots, n\}$, to compute the marginal pmf $f(x_n)$, we must sum over all possible values of $(x_1, \ldots, x_{n-1})$. That is, we must sum over all $(x_1, \ldots, x_{n-1})$ such that the x_is are all nonnegative integers and $\sum_{i=1}^{n-1} x_i = m - x_n$. Denote this set by $\mathcal{B}$. Then

$$\begin{aligned}
f(x_n) &= \sum_{(x_1, \ldots, x_{n-1}) \in \mathcal{B}} \frac{m!}{x_1! \cdot \cdots \cdot x_n!} (p_1)^{x_1} \cdot \cdots \cdot (p_n)^{x_n} \\
&= \sum_{(x_1, \ldots, x_{n-1}) \in \mathcal{B}} \frac{m!}{x_1! \cdot \cdots \cdot x_n!} p_1^{x_1} \cdot \cdots \cdot p_n^{x_n} \frac{(m - x_n)!}{(m - x_n)!} \frac{(1 - p_n)^{m - x_n}}{(1 - p_n)^{m - x_n}} \\
&= \frac{m!}{x_n!(m - x_n)!} p_n^{x_n} (1 - p_n)^{m - x_n} \\
&\quad \times \sum_{(x_1, \ldots, x_{n-1}) \in \mathcal{B}} \frac{(m - x_n)!}{x_1! \cdot \cdots \cdot x_{n-1}!} \left(\frac{p_1}{1 - p_n}\right)^{x_1} \cdot \cdots \cdot \left(\frac{p_{n-1}}{1 - p_n}\right)^{x_{n-1}}.
\end{aligned}$$

$= 1$

But using the facts that $x_1 + \cdots + x_{n-1} = m - x_n$ and $p_1 + \cdots + p_{n-1} = 1 - p_n$ and Theorem 4.6.4, we see that the last summation is 1. Hence the marginal distribution of X_n is binomial(m, p_n). Similar arguments show that each of the other coordinates is marginally binomially distributed.

Given that $X_n = x_n$, there must have been $m - x_n$ trials that resulted in one of the first $n-1$ outcomes. The vector $(X_1, \ldots, X_{n-1})$ counts the number of these $m - x_n$ trials that are of each type. Thus it seems that given $X_n = x_n$, $(X_1, \ldots, X_{n-1})$ might have a multinomial distribution. This is true. From (4.6.6), the conditional pmf of $(X_1, \ldots, X_{n-1})$ given $X_n = x_n$ is

$$\begin{aligned} f(x_1, \ldots, x_{n-1}|x_n) &= \frac{f(x_1, \ldots, x_n)}{f(x_n)} \\ &= \frac{\frac{m!}{x_1! \cdots x_n!}(p_1)^{x_1} \cdots (p_n)^{x_n}}{\frac{m!}{x_n!(m-x_n)!}(p_n)^{x_n}(1-p_n)^{m-x_n}} \\ &= \frac{(m-x_n)!}{x_1! \cdots x_{n-1}!}\left(\frac{p_1}{1-p_n}\right)^{x_1} \cdots \left(\frac{p_{n-1}}{1-p_n}\right)^{x_{n-1}}. \end{aligned}$$

This is the pmf of a multinomial distribution with $m - x_n$ trials and cell probabilities $p_1/(1-p_n), \ldots, p_{n-1}/(1-p_n)$. In fact, the conditional distribution of any subset of the coordinates of $(X_1, \ldots, X_n)$ given the values of the rest of the coordinates is a multinomial distribution.

We see from the conditional distributions that the coordinates of the vector $(X_1, \ldots, X_n)$ are related. In particular, there must be some negative correlation. It turns out that all of the pairwise covariances are negative and are given by (Exercise 4.39)

$$\text{Cov}(X_i, X_j) = \text{E}[(X_i - p_i)(X_j - p_j)] = -mp_ip_j.$$

Definition 4.6.5 Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random vectors with joint pdf or pmf $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Let $f_{\mathbf{X}_i}(\mathbf{x}_i)$ denote the marginal pdf or pmf of $\mathbf{X}_i$. Then $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are called *mutually independent random vectors* if, for every $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$,

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = f_{\mathbf{X}_1}(\mathbf{x}_1) \cdots f_{\mathbf{X}_n}(\mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i).$$

If the X_is are all one-dimensional, then $X_1, \ldots, X_n$ are called *mutually independent random variables.*

Theorem 4.6.6 (Generalization of Theorem 4.2.10) *Let $X_1, \ldots, X_n$ be mutually independent random variables. Let $g_1, \ldots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of x_i, $i = 1, \ldots, n$. Then*

$$\text{E}(g_1(X_1) \cdots g_n(X_n)) = (\text{E}g_1(X_1)) \cdots (\text{E}g_n(X_n)).$$

☆Nov. 16th. 2021

Theorem 4.6.7 (Generalization of Theorem 4.2.12) *Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $Z = X_1 + \cdots + X_n$. Then the mgf of Z is*

$$M_Z(t) = M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t).$$

In particular, if $X_1, \ldots, X_n$ all have the same distribution with mgf $M_X(t)$, then

$$M_Z(t) = (M_X(t))^n.$$

Example 4.6.8 (Mgf of a sum of gamma variables) Suppose $X_1, \ldots, X_n$ are mutually independent random variables, and the distribution of X_i is gamma(α_i, β). From Example 2.3.8, the mgf of a gamma(α, β) distribution is $M(t) = (1 - \beta t)^{-\alpha}$. Thus, if $Z = X_1 + \cdots + X_n$, the mgf of Z is

$$M_Z(t) = M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t) = (1 - \beta t)^{-\alpha_1} \cdot \cdots \cdot (1 - \beta t)^{-\alpha_n} = (1 - \beta t)^{-(\alpha_1 + \cdots + \alpha_n)}.$$

This is the mgf of a gamma$(\alpha_1 + \cdots + \alpha_n, \beta)$ distribution. Thus, the sum of independent gamma random variables that have a common scale parameter β also has a gamma distribution. ‖

Corollary 4.6.9 *Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Let $Z = (a_1X_1 + b_1) + \cdots + (a_nX_n + b_n)$. Then the mgf of Z is*

$$M_Z(t) = (e^{t(\Sigma b_i)}) M_{X_1}(a_1 t) \cdot \cdots \cdot M_{X_n}(a_n t).$$

Proof: From the definition, the mgf of Z is

$$\begin{aligned} M_Z(t) &= \mathrm{E} e^{tZ} \\ &= \mathrm{E} e^{t\Sigma(a_i X_i + b_i)} \\ &= (e^{t(\Sigma b_i)}) \mathrm{E}(e^{t a_1 X_1} \cdot \cdots \cdot e^{t a_n X_n}) \qquad \left(\begin{array}{c}\text{properties of exponentials} \\ \text{and expectations}\end{array}\right) \\ &= (e^{t(\Sigma b_i)}) M_{X_1}(a_1 t) \cdot \cdots \cdot M_{X_n}(a_n t), \qquad \text{(Theorem 4.6.6)} \end{aligned}$$

as was to be shown. □

Corollary 4.6.10 *Let $X_1, \ldots, X_n$ be mutually independent random variables with $X_i \sim \mathrm{n}(\mu_i, \sigma_i^2)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Then*

$$Z = \sum_{i=1}^{n} (a_i X_i + b_i) \sim \mathrm{n}\left(\sum_{i=1}^{n} (a_i \mu_i + b_i), \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right).$$

Proof: Recall that the mgf of a $\mathrm{n}(\mu, \sigma^2)$ random variable is $M(t) = e^{\mu t + \sigma^2 t^2/2}$. Substituting into the expression in Corollary 4.6.9 yields

$$\begin{aligned} M_Z(t) &= (e^{t(\Sigma b_i)}) e^{\mu_1 a_1 t + \sigma_1^2 a_1^2 t^2/2} \cdot \cdots \cdot e^{\mu_n a_n t + \sigma_n^2 a_n^2 t^2/2} \\ &= e^{\left((\Sigma(a_i \mu_i + b_i))t + (\Sigma a_i^2 \sigma_i^2) t^2/2\right)}, \end{aligned}$$

the mgf of the indicated normal distribution. □

Theorem 4.6.11 (Generalization of Lemma 4.2.7) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random vectors. Then $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent random vectors if and only if there exist functions $g_i(\mathbf{x}_i), i = 1, \ldots, n$, such that the joint pdf or pmf of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ can be written as*

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = g_1(\mathbf{x}_1) \cdot \cdots \cdot g_n(\mathbf{x}_n).$$

Theorem 4.6.12 (Generalization of Theorem 4.3.5) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors. Let $g_i(\mathbf{x}_i)$ be a function only of $\mathbf{x}_i, i = 1, \ldots, n$. Then the random variables $U_i = g_i(\mathbf{X}_i), i = 1, \ldots, n$, are mutually independent.*

Let $(X_1, \ldots, X_n)$ be a random vector with pdf $f_{\mathbf{X}}(x_1, \ldots, x_n)$. Let $\mathcal{A} = \{\mathbf{x} : f_{\mathbf{X}}(\mathbf{x}) > 0\}$. Consider a new random vector $(U_1, \ldots, U_n)$, defined by $U_1 = g_1(X_1, \ldots, X_n), U_2 = g_2(X_1, \ldots, X_n), \ldots, U_n = g_n(X_1, \ldots, X_n)$. Suppose that $A_0, A_1, \ldots, A_k$ form a partition of $\mathcal{A}$ with these properties. The set A_0, which may be empty, satisfies $P((X_1, \ldots, X_n) \in A_0) = 0$. The transformation $(U_1, \ldots, U_n) = (g_1(\mathbf{X}), \ldots, g_n(\mathbf{X}))$ is a one-to-one transformation from A_i onto $\mathcal{B}$ for each $i = 1, 2, \ldots, k$. Then for each i, the inverse functions from $\mathcal{B}$ to A_i can be found. Denote the ith inverse by $x_1 = h_{1i}(u_1, \ldots, u_n), x_2 = h_{2i}(u_1, \ldots, u_n), \ldots, x_n = h_{ni}(u_1, \ldots, u_n)$. This ith inverse gives, for $(u_1, \ldots, u_n) \in \mathcal{B}$, the unique $(x_1, \ldots, x_n) \in A_i$ such that $(u_1, \ldots, u_n) = (g_1(x_1, \ldots, x_n), \ldots, g_n(x_1, \ldots, x_n))$. Let J_i denote the Jacobian computed from the ith inverse. That is,

$$J_i = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{1i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_n} \\ \frac{\partial h_{2i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_{ni}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_n} \end{vmatrix},$$

the determinant of an $n \times n$ matrix. Assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the joint pdf, $f_{\mathbf{U}}(u_1, \ldots, u_n)$, for $\mathbf{u} \in \mathcal{B}$:

$$f_{\mathbf{U}}(u_1, \ldots, u_n) = \sum_{i=1}^{k} f_{\mathbf{X}}(h_{1i}(u_1, \ldots, u_n), \ldots, h_{ni}(u_1, \ldots, u_n)) |J_i|. \tag{4.6.7}$$

Example 4.6.13 (Multivariate change of variables) Let (X_1, X_2, X_3, X_4) have joint pdf

$$f_{\mathbf{X}}(x_1, x_2, x_3, x_4) = 24 e^{-x_1 - x_2 - x_3 - x_4}, \quad 0 < x_1 < x_2 < x_3 < x_4 < \infty.$$

Consider the transformation

$$U_1 = X_1, \quad U_2 = X_2 - X_1, \quad U_3 = X_3 - X_2, \quad U_4 = X_4 - X_3.$$

This transformation maps the set $\mathcal{A}$ onto the set $\mathcal{B} = \{\mathbf{u}: 0 < u_i < \infty, i = 1, 2, 3, 4\}$. The transformation is one-to-one, so $k = 1$, and the inverse is

$$X_1 = U_1, \quad X_2 = U_1 + U_2, \quad X_3 = U_1 + U_2 + U_3, \quad X_4 = U_1 + U_2 + U_3 + U_4.$$

The Jacobian of the inverse is

$$J = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{vmatrix} = 1.$$

Since the matrix is triangular, the determinant is equal to the product of the diagonal elements. Thus, from (4.6.7) we obtain

$$\begin{aligned} f_{\mathbf{U}}(u_1, \ldots, u_4) &= 24e^{-u_1-(u_1+u_2)-(u_1+u_2+u_3)-(u_1+u_2+u_3+u_4)} \\ &= 24e^{-4u_1-3u_2-2\mu_3-u_4} \end{aligned}$$

on $\mathcal{B}$. From this the marginal pdfs of U_1, U_2, U_3, and U_4 can be calculated. It turns out that $f_U(u_i) = (5-i)e^{-(5-i)u_i}, 0 < u_i$; that is, $U_i \sim$ exponential$(1/(5-i))$. From Theorem 4.6.11 we see that U_1, U_2, U_3, and U_4 are mutually independent random variables. ‖

4.7 Inequalities

Lemma 4.7.1 *Let a and b be any positive numbers, and let p and q be any positive numbers (necessarily greater than 1) satisfying*

(4.7.1) $$\frac{1}{p} + \frac{1}{q} = 1. \qquad (p+q=pq)$$

Then

(4.7.2) $$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality if and only if $a^p = b^q$.

Proof: Fix b, and consider the function

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab.$$

To minimize $g(a)$, differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \Rightarrow a^{p-1} - b = 0 \Rightarrow b = a^{p-1}.$$

A check of the second derivative will establish that this is indeed a minimum. The value of the function at the minimum is

$$\frac{1}{p}a^p + \frac{1}{q}\left(a^{p-1}\right)^q - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p \quad \left(\begin{array}{c}(p-1)q = p \text{ follows} \\ \text{from (4.7.1)}\end{array}\right)$$
$$= 0. \qquad \text{(again from (4.7.1))}$$

Hence the minimum is 0 and (4.7.2) is established. Since the minimum is unique (why?), equality holds only if $a^{p-1} = b$, which is equivalent to $a^p = b^q$, again from (4.7.1). □

Theorem 4.7.2 (Hölder's Inequality) *Let X and Y be any two random variables, and let p and q satisfy (4.7.1). Then*

$\|XY\|_1 \le \|X\|_p \|Y\|_q$

$$|\mathrm{E}XY| \le \mathrm{E}|XY| \le (\mathrm{E}|X|^p)^{1/p}\,(\mathrm{E}|Y|^q)^{1/q}. \tag{4.7.3}$$

Proof: The first inequality follows from $-|XY| \le XY \le |XY|$ and Theorem 2.2.5. To prove the second inequality, define

$$a = \frac{|X|}{(\mathrm{E}|X|^p)^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{(\mathrm{E}|Y|^q)^{1/q}}.$$

Applying Lemma 4.7.1, we get

$$\frac{1}{p}\frac{|X|^p}{\mathrm{E}\,|X|^p} + \frac{1}{q}\frac{|Y|^q}{\mathrm{E}\,|Y|^q} \ge \frac{|XY|}{(\mathrm{E}\,|X|^p)^{1/p}\,(\mathrm{E}|Y|^q)^{1/q}}.$$

Now take expectations of both sides. The expectation of the left-hand side is 1, and rearrangement gives (4.7.3). □

Theorem 4.7.3 (Cauchy–Schwarz Inequality) *For any two random variables X and Y,*

$$|\mathrm{E}XY| \le \mathrm{E}|XY| \le (\mathrm{E}|X|^2)^{1/2}(\mathrm{E}|Y|^2)^{1/2}. \tag{4.7.4}$$

Example 4.7.4 (Covariance inequality) If X and Y have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively, we can apply the Cauchy–Schwarz Inequality to get

$$\mathrm{E}\,|(X-\mu_X)(Y-\mu_Y)| \le \left\{\mathrm{E}(X-\mu_X)^2\right\}^{1/2}\left\{\mathrm{E}(Y-\mu_Y)^2\right\}^{1/2}.$$

Squaring both sides and using statistical notation, we have

$$(\mathrm{Cov}(X,Y))^2 \le \sigma_X^2\sigma_Y^2.$$

Recalling the definition of the correlation coefficient, ρ, we have proved that $0 \le \rho^2 \le 1$. Furthermore, the condition for equality in Lemma 4.7.1 still carries over, and equality is attained here only if $X - \mu_X = c(Y - \mu_Y)$, for some constant c. That is, the correlation is ± 1 if and only if X and Y are linearly related. Compare the ease of this proof to the one used in Theorem 4.5.7, before we had the Cauchy–Schwarz Inequality. ‖

Some other special cases of Hölder's Inequality are often useful. If we set $Y \equiv 1$ in (4.7.3), we get

$$\mathrm{E}|X| \le \{\mathrm{E}(|X|^p)\}^{1/p}, \quad 1 < p < \infty. \tag{4.7.5}$$

For $1 < r < p$, if we replace $|X|$ by $|X|^r$ in (4.7.5), we obtain

$$\mathrm{E}|X|^r \le \{\mathrm{E}(|X|^{pr})\}^{1/p}.$$

Now write $s = pr$ (note that $s > r$) and rearrange terms to get

$$\{\mathrm{E}|X|^r\}^{1/r} \le \{\mathrm{E}|X|^s\}^{1/s}, \quad 1 < r < s < \infty, \tag{4.7.6}$$

$\|X\|_r \le \|X\|_s$ $(r<s)$

which is known as *Liapounov's Inequality.*

Theorem 4.7.5 (Minkowski's Inequality) *Let X and Y be any two random variables. Then for $1 \le p < \infty$,*

$$[\mathrm{E}|X+Y|^p]^{1/p} \le [\mathrm{E}|X|^p]^{1/p} + [\mathrm{E}|Y|^p]^{1/p}. \tag{4.7.7}$$

$\|X+Y\|_p \le \|X\|_p + \|Y\|_p$

Proof: Write

$$\begin{aligned} \mathrm{E}|X+Y|^p &= \mathrm{E}\left(|X+Y||X+Y|^{p-1}\right) \\ &\le \mathrm{E}\left(|X||X+Y|^{p-1}\right) + \mathrm{E}\left(|Y||X+Y|^{p-1}\right), \end{aligned} \tag{4.7.8}$$

where we have used the fact that $|X+Y| \le |X| + |Y|$ (the *triangle inequality*; see Exercise 4.64). Now apply Hölder's Inequality to each expectation on the right-hand side of (4.7.8) to get

$$\begin{aligned} \mathrm{E}(|X+Y|^p) \le &\left[\mathrm{E}(|X|^p)\right]^{1/p}\left[\mathrm{E}|X+Y|^{q(p-1)}\right]^{1/q} \\ &+ \left[\mathrm{E}(|Y|^p)\right]^{1/p}\left[\mathrm{E}|X+Y|^{q(p-1)}\right]^{1/q}, \end{aligned}$$

where q satisfies $1/p + 1/q = 1$. Now divide through by $\left[\mathrm{E}(|X+Y|^{q(p-1)})\right]^{1/q}$. Noting that $q(p-1) = p$ and $1 - 1/q = 1/p$, we obtain (4.7.7). □

*Nov. 18th. 2021

Theorem 4.7.7 (Jensen's Inequality) *For any random variable X, if $g(x)$ is a convex function, then*

$$\mathrm{E}g(X) \ge g(\mathrm{E}X).$$

Equality holds if and only if, for every line $a + bx$ that is tangent to $g(x)$ at $x = EX$, $P(g(X) = a + bX) = 1$.

Proof: To establish the inequality, let $l(x)$ be a tangent line to $g(x)$ at the point $g(EX)$. (Recall that EX is a constant.) Write $l(x) = a + bx$ for some a and b. The situation is illustrated in Figure 4.7.2.

Now, by the convexity of g we have $g(x) \ge a + bx$. Since expectations preserve inequalities,

$$
\begin{aligned}
Eg(X) &\geq E(a+bX) \\
&= a + bEX && \left(\begin{array}{c}\text{linearity of expectation,}\\ \text{Theorem 2.2.5}\end{array}\right) \\
&= l(EX) && (\text{definition of } l(x)) \\
&= g(EX), && (l \text{ is tangent at } EX)
\end{aligned}
$$

as was to be shown.

If $g(x)$ is linear, equality follows from properties of expectations (Theorem 2.2.5). For the "only if" part see Exercise 4.62. □

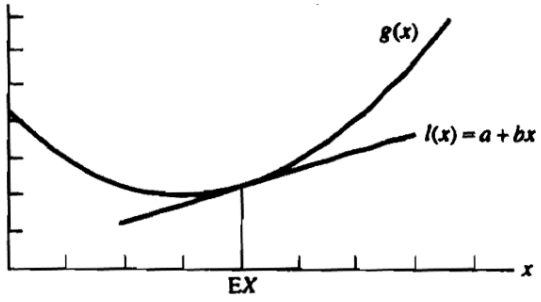


Figure 4.7.2. *Graphical illustration of Jensen's Inequality*

Example 4.7.8 (An inequality for means) Jensen's Inequality can be used to prove an inequality between three different kinds of means. If $a_1, \ldots, a_n$ are positive numbers, define

$$
\begin{aligned}
a_A &= \frac{1}{n}(a_1 + a_2 + \cdots + a_n), && \text{(arithmetic mean)} \\
a_G &= [a_1 a_2 \cdot \cdots \cdot a_n]^{1/n}, && \text{(geometric mean)} \\
a_H &= \frac{1}{\frac{1}{n}\left(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}\right)}. && \text{(harmonic mean)}
\end{aligned}
$$

An inequality relating these means is

$$
a_H \leq a_G \leq a_A.
$$

To apply Jensen's Inequality, let X be a random variable with range $a_1, \ldots, a_n$ and $P(X = a_i) = 1/n, i = 1, \ldots, n$. Since $\log x$ is a concave function, Jensen's Inequality shows that $E(\log X) \leq \log(EX)$; hence,

$$
\log a_G = \frac{1}{n}\sum_{i=1}^{n} \log a_i = E(\log X) \leq \log(EX) = \log\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) = \log a_A,
$$

so $a_G \leq a_A$. Now again use the fact that $\log x$ is concave to get

$$
\log \frac{1}{a_H} = \log\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_i}\right) = \log E\frac{1}{X} \geq E\left(\log \frac{1}{X}\right) = -E(\log X).
$$

Since $E(\log X) = \log a_G$, it then follows that $\log(1/a_H) \geq \log(1/a_G)$, or $a_G \geq a_H$. ‖

The next inequality merely exploits the definition of covariance, but sometimes proves to be useful. If X is a random variable with finite mean μ and $g(x)$ is a nondecreasing function, then

$$\mathrm{E}\left(g(X)(X-\mu)\right) \geq 0,$$

since

$$\begin{aligned}
&\mathrm{E}(g(X)(X-\mu)) \\
&\quad = \mathrm{E}\left(g(X)(X-\mu)I_{(-\infty,0)}(X-\mu)\right) + \mathrm{E}\left(g(X)(X-\mu)I_{[0,\infty)}(X-\mu)\right) \\
&\quad \geq \mathrm{E}\left(g(\mu)(X-\mu)I_{(-\infty,0)}(X-\mu)\right) \\
&\qquad + \mathrm{E}\left(g(\mu)(X-\mu)I_{[0,\infty)}(X-\mu)\right) \qquad \text{(since } g \text{ is nondecreasing)} \\
&\quad = g(\mu)\mathrm{E}(X-\mu) \\
&\quad = 0.
\end{aligned}$$

Theorem 4.7.9 (Covariance Inequality) *Let X be any random variable and $g(x)$ and $h(x)$ any functions such that $\mathrm{E}g(X)$, $\mathrm{E}h(X)$, and $\mathrm{E}(g(X)h(X))$ exist.*

a. *If $g(x)$ is a nondecreasing function and $h(x)$ is a nonincreasing function, then*

$$\mathrm{E}\left(g(X)h(X)\right) \leq \left(\mathrm{E}g(X)\right)\left(\mathrm{E}h(X)\right).$$

b. *If $g(x)$ and $h(x)$ are either both nondecreasing or both nonincreasing, then*

$$\mathrm{E}\left(g(X)h(X)\right) \geq \left(\mathrm{E}g(X)\right)\left(\mathrm{E}h(X)\right).$$

The intuition behind the inequality is easy. In case (a) there is negative correlation between g and h, while in case (b) there is positive correlation. The inequalities merely reflect this fact. The usefulness of the Covariance Inequality is that it allows us to bound an expectation without using higher-order moments.

Properties of a Random Sample

Definition 5.1.1 The random variables $X_1, \ldots, X_n$ are called a *random sample of size n from the population $f(x)$* if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called *independent and identically distributed random variables with pdf or pmf $f(x)$*. This is commonly abbreviated to iid random variables.

From Definition 4.6.5, the joint pdf or pmf of $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2)\cdots\cdots f(x_n) = \prod_{i=1}^{n} f(x_i). \tag{5.1.1}$$

Example 5.1.2 (Sample pdf–exponential) Let $X_1, \ldots, X_n$ be a random sample from an exponential(β) population. Specifically, $X_1, \ldots, X_n$ might correspond to the times until failure (measured in years) for n identical circuit boards that are put on test and used until they fail. The joint pdf of the sample is

$$f(x_1, \ldots, x_n|\beta) = \prod_{i=1}^{n} f(x_i|\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1+\cdots+x_n)/\beta}.$$

This pdf can be used to answer questions about the sample. For example, what is the probability that all the boards last more than 2 years? We can compute

$$\begin{aligned}
&P(X_1 > 2, \ldots, X_n > 2) \\
&\quad = \int_2^\infty \cdots \int_2^\infty \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_1 \cdots dx_n \\
&\quad = e^{-2/\beta} \int_2^\infty \cdots \int_2^\infty \prod_{i=2}^{n} \frac{1}{\beta} e^{-x_i/\beta}\, dx_2 \cdots dx_n \qquad \text{(integrate out } x_1\text{)} \\
&\quad \vdots \qquad\qquad \text{(integrate out the remaining } x_i\text{s successively)} \\
&\quad = (e^{-2/\beta})^n \\
&\quad = e^{-2n/\beta}.
\end{aligned}$$

Or:

$$\begin{aligned}
&P(X_1 > 2, \ldots, X_n > 2) \\
&\quad = P(X_1 > 2) \cdots P(X_n > 2) && \text{(independence)} \\
&\quad = [P(X_1 > 2)]^n && \text{(identical distributions)} \\
&\quad = (e^{-2/\beta})^n && \text{(exponential calculation)} \\
&\quad = e^{-2n/\beta}. && \|
\end{aligned}$$

Nov. 23rd 2021

Definition 5.2.1 Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a *statistic*. The probability distribution of a statistic Y is called the *sampling distribution of* Y.

Definition 5.2.2 The *sample mean* is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Definition 5.2.3 The *sample variance* is the statistic defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The *sample standard deviation* is the statistic defined by $S = \sqrt{S^2}$.

Theorem 5.2.4 *Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then*

a. $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$,

b. $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Proof: To prove part (a), add and subtract $\bar{x}$ to get

$$\begin{aligned}\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2. \qquad \text{(cross term is 0)}\end{aligned}$$

It is now clear that the right-hand side is minimized at $a = \bar{x}$. (Notice the similarity to Example 2.2.6 and Exercise 4.13.)

To prove part (b), take $a = 0$ in the above. □

Theorem 5.2.6 *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

a. $\mathrm{E}\bar{X} = \mu$,

b. $\mathrm{Var}\,\bar{X} = \dfrac{\sigma^2}{n}$,

c. $\mathrm{E}S^2 = \sigma^2$.

Proof: To prove (a), let $g(X_i) = X_i/n$, so $\mathrm{E}g(X_i) = \mu/n$. Then, by Lemma 5.2.5,

$$\mathrm{E}\bar{X} = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} n\mathrm{E}X_1 = \mu.$$

Similarly for (b), we have

$$\mathrm{Var}\ \bar{X} = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n\,\mathrm{Var}\ X_1 = \frac{\sigma^2}{n}.$$

For the sample variance, using Theorem 5.2.4, we have

$$\begin{aligned}
\mathrm{E}S^2 &= \mathrm{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\
&= \frac{1}{n-1}\left(n\mathrm{E}X_1^2 - n\mathrm{E}\bar{X}^2\right) \\
&= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \sigma^2,
\end{aligned}$$

establishing part (c) and proving the theorem. □

Theorem 5.2.9 *If X and Y are independent continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$, then the pdf of $Z = X + Y$ is*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z-w)\, dw. \tag{5.2.3}$$

Proof: Let $W = X$. The Jacobian of the transformation from (X, Y) to (Z, W) is 1. So using (4.3.2), we obtain the joint pdf of (Z, W) as

$$f_{Z,W}(z,w) = f_{X,Y}(w, z-w) = f_X(w) f_Y(z-w).$$

Integrating out w, we obtain the marginal pdf of Z as given in (5.2.3). □

Example 5.2.10 (Sum of Cauchy random variables) As an example of a situation where the mgf technique fails, consider sampling from a Cauchy distribution. We will eventually derive the distribution of $\bar{Z}$, the mean of $Z_1, \ldots, Z_n$, iid Cauchy$(0,1)$ observations. We start, however, with the distribution of the sum of two independent Cauchy random variables and apply formula (5.2.3).

Let U and V be independent Cauchy random variables, $U \sim$ Cauchy$(0, \sigma)$ and $V \sim$ Cauchy$(0, \tau)$; that is,

$$f_U(u) = \frac{1}{\pi\sigma}\frac{1}{1+(u/\sigma)^2}, \qquad f_V(v) = \frac{1}{\pi\tau}\frac{1}{1+(v/\tau)^2}, \qquad \begin{aligned} &-\infty < u < \infty, \\ &-\infty < v < \infty. \end{aligned}$$

Based on formula (5.2.3), the pdf of $Z = U + V$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\pi\sigma}\frac{1}{1+(w/\sigma)^2}\,\frac{1}{\pi\tau}\frac{1}{1+((z-w)/\tau)^2}\, dw, \quad -\infty < z < \infty. \tag{5.2.4}$$

This integral is somewhat involved but can be solved by a partial fraction decomposition and some careful antidifferentiation (see Exercise 5.7). The result is

$$f_Z(z) = \frac{1}{\pi(\sigma+\tau)} \frac{1}{1+(z/(\sigma+\tau))^2}, \quad -\infty < z < \infty. \tag{5.2.5}$$

$Z \sim \text{Cauchy}(0, \sigma+\tau)$

Thus, the sum of two independent Cauchy random variables is again a Cauchy, with the scale parameters adding. It therefore follows that if $Z_1, \ldots, Z_n$ are iid Cauchy$(0,1)$ random variables, then $\sum Z_i$ is Cauchy$(0,n)$ and also $\bar{Z}$ is Cauchy$(0,1)$! The sample mean has the same distribution as the individual observations. (See Example A.0.5 in Appendix A for a computer algebra version of this calculation.) ||

5.4 Order Statistics

Definition 5.4.1 The *order statistics* of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$.

The *sample range*, $R = X_{(n)} - X_{(1)}$, is the distance between the smallest and largest observations. It is a measure of the dispersion in the sample and should reflect the dispersion in the population.

The *sample median*, which we will denote by M, is a number such that approximately one-half of the observations are less than M and one-half are greater. In terms of the order statistics, M is defined by

$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \left(X_{(n/2)} + X_{(n/2+1)}\right)/2 & \text{if } n \text{ is even.} \end{cases} \tag{5.4.1}$$

mid-range $V = \frac{X_{(1)} + X_{(n)}}{2}$

Theorem 5.4.4 *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the pdf of $X_{(j)}$ is*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j}. \tag{5.4.4}$$

Proof: We first find the cdf of $X_{(j)}$ and then differentiate it to obtain the pdf. As in Theorem 5.4.3, let Y be a random variable that counts the number of $X_1, \ldots, X_n$ less than or equal to x. Then, defining a "success" as the event $\{X_j \le x\}$, we see that $Y \sim$ binomial$(n, F_X(x))$. (Note that we can write $P_i = F_X(x_i)$ in Theorem 5.4.3. Also, although $X_1, \ldots, X_n$ are continuous random variables, the counting variable Y is discrete.) Thus,

$$F_{X_{(j)}}(x) = P(Y \ge j) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1-F_X(x)]^{n-k},$$

and the pdf of $X_{(j)}$ is

$$
\begin{aligned}
f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\
&= \sum_{k=j}^{n} \binom{n}{k} \Big(k\,[F_X(x)]^{k-1}\,[1-F_X(x)]^{n-k} f_X(x) \\
&\quad - (n-k)[F_X(x)]^k [1-F_X(x)]^{n-k-1} f_X(x) \Big) && \text{(chain rule)} \\
&= \binom{n}{j} j f_X(x) [F_X(x)]^{j-1} [1-F_X(x)]^{n-j} \\
&\quad + \sum_{k=j+1}^{n} \binom{n}{k} k [F_X(x)]^{k-1} [1-F_X(x)]^{n-k} f_X(x) \\
&\quad - \sum_{k=j}^{n-1} \binom{n}{k} (n-k) [F_X(x)]^k [1-F_X(x)]^{n-k-1} f_X(x) && \binom{k=n \text{ term}}{\text{is 0}} \\
&= \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1-F_X(x)]^{n-j} \\
&\quad + \sum_{k=j}^{n-1} \binom{n}{k+1} (k+1)\,[F_X(x)]^k\,[1-F_X(x)]^{n-k-1} f_X(x) && \binom{\text{change}}{\text{dummy}} \\
&\quad - \sum_{k=j}^{n-1} \binom{n}{k} (n-k)\,[F_X(x)]^k\,[1-F_X(x)]^{n-k-1} f_X(x). && \text{(5.4.5)}
\end{aligned}
$$

Noting that

$$
\binom{n}{k+1}(k+1) = \frac{n!}{k!(n-k-1)!} = \binom{n}{k}(n-k), \tag{5.4.6}
$$

we see that the last two sums in (5.4.5) cancel. Thus, the pdf $f_{X_{(j)}}(x)$ is given by the expression in (5.4.4). □

Example 5.4.5 (Uniform order statistic pdf) Let $X_1, \ldots, X_n$ be iid uniform$(0,1)$, so $f_X(x) = 1$ for $x \in (0,1)$ and $F_X(x) = x$ for $x \in (0,1)$. Using (5.4.4), we see that the pdf of the jth order statistic is

$$
\begin{aligned}
f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j} \quad \text{for } x \in (0,1) \\
&= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1}.
\end{aligned}
$$

Thus, the jth order statistic from a uniform$(0,1)$ sample has a beta$(j, n-j+1)$ distribution. From this we can deduce that

$$
\mathrm{E}X_{(j)} = \frac{j}{n+1} \quad \text{and} \quad \mathrm{Var}\, X_{(j)} = \frac{j(n-j+1)}{(n+1)^2(n+2)}. \qquad \|
$$

Theorem 5.4.6 *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is*

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} \times [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j} \tag{5.4.7}$$

for $-\infty < u < v < \infty$.

The joint pdf of three or more order statistics could be derived using similar but even more involved arguments. Perhaps the other most useful pdf is $f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n)$, the joint pdf of all the order statistics, which is given by

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! f_X(x_1) \cdot \cdots \cdot f_X(x_n) & -\infty < x_1 < \cdots < x_n < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Example 5.4.7 (Distribution of the midrange and range) Let $X_1, \ldots, X_n$ be iid uniform$(0, a)$ and let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics. The range was earlier defined as $R = X_{(n)} - X_{(1)}$. The *midrange*, a measure of location like the sample median or the sample mean, is defined by $V = (X_{(1)} + X_{(n)})/2$. We will derive the joint pdf of R and V from the joint pdf of $X_{(1)}$ and $X_{(n)}$. From (5.4.7) we have that

$$\begin{aligned} f_{X_{(1)},X_{(n)}}(x_1, x_n) &= \frac{n(n-1)}{a^2} \left(\frac{x_n}{a} - \frac{x_1}{a} \right)^{n-2} \\ &= \frac{n(n-1)(x_n - x_1)^{n-2}}{a^n}, \quad 0 < x_1 < x_n < a. \end{aligned}$$

Solving for $X_{(1)}$ and $X_{(n)}$, we obtain $X_{(1)} = V - R/2$ and $X_{(n)} = V + R/2$. The Jacobian for this transformation is -1. The transformation from $(X_{(1)}, X_{(n)})$ to (R, V) maps $\{(x_1, x_n) : 0 < x_1 < x_n < a\}$ onto the set $\{(r, v) : 0 < r < a, r/2 < v < a - r/2\}$. To see this, note that obviously $0 < r < a$ and for a fixed value of r, v ranges from $r/2$ (corresponding to $x_1 = 0, x_n = r$) to $a - r/2$ (corresponding to $x_1 = a - r, x_n = a$). Thus, the joint pdf of (R, V) is

$$f_{R,V}(r, v) = \frac{n(n-1)r^{n-2}}{a^n}, \quad 0 < r < a, \quad r/2 < v < a - r/2.$$

The marginal pdf of R is thus

$$\begin{aligned} f_R(r) &= \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n} \, dv \\ &= \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a. \end{aligned} \tag{5.4.8}$$

If $a = 1$, we see that r has a beta$(n-1, 2)$ distribution. Or, for arbitrary a, it is easy to deduce from (5.4.8) that R/a has a beta distribution. Note that the constant a is a scale parameter.

The set where $f_{R,V}(r, v) > 0$ is shown in Figure 5.4.1, where we see that the range of integration of r depends on whether $v > a/2$ or $v \le a/2$. Thus, the marginal pdf of V is given by

$$f_V(v) = \int_0^{2v} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n(2v)^{n-1}}{a^n}, \quad 0 < v \le a/2,$$

and

$$f_V(v) = \int_0^{2(a-v)} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n[2(a-v)]^{n-1}}{a^n}, \quad a/2 < v \le a.$$

This pdf is symmetric about $a/2$ and has a peak at $a/2$. ‖

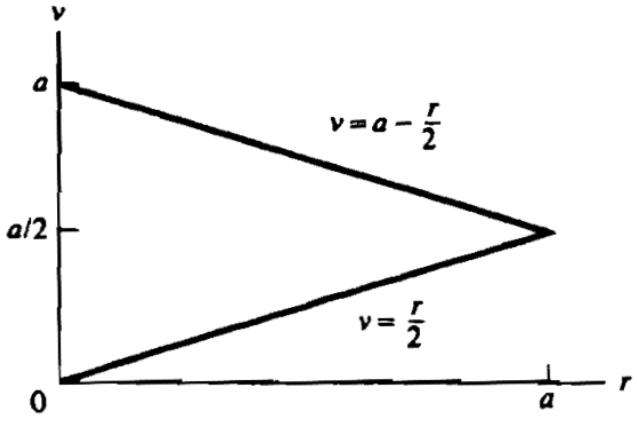


Figure 5.4.1. *Region on which $f_{R,V}(r, v) > 0$ for Example 5.4.7*

※Nov. 30th 2021

5.5 Convergence Concepts

Definition 5.5.1 A sequence of random variables, $X_1, X_2, \ldots$, *converges in probability* to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \ge \epsilon) = 0 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

Theorem 5.5.2 (Weak Law of Large Numbers) *Let $X_1, X_2, \ldots$ be iid random variables with $EX_i = \mu$ and Var $X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then, for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1;$$

that is, $\bar{X}_n$ converges in probability to μ. $\bar{X}_n \xrightarrow{P} \mu$

Proof: The proof is quite simple, being a straightforward application of Chebychev's Inequality. We have, for every $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{\mathrm{E}(\bar{X}_n - \mu)^2}{\epsilon^2} = \frac{\mathrm{Var}\ \bar{X}_n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Hence, $P(|\bar{X}_n - \mu| < \epsilon) = 1 - P(|\bar{X}_n - \mu| \geq \epsilon) \geq 1 - \sigma^2/(n\epsilon^2) \to 1$, as $n \to \infty$. □

Example 5.5.3 (Consistency of S^2) Suppose we have a sequence $X_1, X_2, \ldots$ of iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}\ X_i = \sigma^2 < \infty$. If we define

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

can we prove a WLLN for S_n^2? Using Chebychev's Inequality, we have

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathrm{E}(S_n^2 - \sigma^2)^2}{\epsilon^2} = \frac{\mathrm{Var}\ S_n^2}{\epsilon^2}$$

and thus, a sufficient condition that S_n^2 converges in probability to σ^2 is that $\mathrm{Var}\ S_n^2 \to 0$ as $n \to \infty$. ‖

Theorem 5.5.4 *Suppose that $X_1, X_2, \ldots$ converges in probability to a random variable X and that h is a continuous function. Then $h(X_1), h(X_2), \ldots$ converges in probability to $h(X)$.*

Example 5.5.5 (Consistency of S) If S_n^2 is a consistent estimator of σ^2, then by Theorem 5.5.4, the sample standard deviation $S_n = \sqrt{S_n^2} = h(S_n^2)$ is a consistent estimator of σ. Note that S_n is, in fact, a biased estimator of σ (see Exercise 5.11), but the bias disappears asymptotically. ‖

Definition 5.5.6 A sequence of random variables, $X_1, X_2, \ldots$, *converges almost surely* to a random variable X if, for every $\epsilon > 0$,

$$P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1.$$

Example 5.5.7 (Almost sure convergence) Let the sample space S be the closed interval $[0, 1]$ with the uniform probability distribution. Define random variables $X_n(s) = s + s^n$ and $X(s) = s$. For every $s \in [0, 1)$, $s^n \to 0$ as $n \to \infty$ and $X_n(s) \to s = X(s)$. However, $X_n(1) = 2$ for every n so $X_n(1)$ does not converge to $1 = X(1)$. But since the convergence occurs on the set $[0, 1)$ and $P([0, 1)) = 1$, X_n converges to X almost surely. ‖

Example 5.5.8 (Convergence in probability, not almost surely) In this example we describe a sequence that converges in probability, but not almost surely. Again, let the sample space S be the closed interval [0,1] with the uniform probability distribution. Define the sequence $X_1, X_2, \ldots$ as follows:

$$X_1(s) = s + I_{[0,1]}(s), \quad X_2(s) = s + I_{[0,\frac{1}{2}]}(s), \quad X_3(s) = s + I_{[\frac{1}{2},1]}(s),$$

$$X_4(s) = s + I_{[0,\frac{1}{3}]}(s), \quad X_5(s) = s + I_{[\frac{1}{3},\frac{2}{3}]}(s), \quad X_6(s) = s + I_{[\frac{2}{3},1]}(s),$$

etc. Let $X(s) = s$. It is straightforward to see that X_n converges to X in probability. As $n \to \infty, P(|X_n - X| \geq \epsilon)$ is equal to the probability of an interval of s values whose length is going to 0. However, X_n does not converge to X almost surely. Indeed, there is no value of $s \in S$ for which $X_n(s) \to s = X(s)$. For every s, the value $X_n(s)$ alternates between the values s and $s+1$ infinitely often. For example, if $s = \frac{3}{8}, X_1(s) = 1\frac{3}{8}, X_2(s) = 1\frac{3}{8}, X_3(s) = \frac{3}{8}, X_4(s) = \frac{3}{8}, X_5(s) = 1\frac{3}{8}, X_6(s) = \frac{3}{8}$, etc. No pointwise convergence occurs for this sequence. ‖

Theorem 5.5.9 (Strong Law of Large Numbers) *Let $X_1, X_2, \ldots$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 < \infty$, and define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then, for every $\epsilon > 0$,*

$$P(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon) = 1;$$

that is, $\bar{X}_n$ converges almost surely to μ.

Definition 5.5.10 A sequence of random variables, $X_1, X_2, \ldots,$ *converges in distribution* to a random variable X if

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

at all points x where $F_X(x)$ is continuous.

Example 5.5.11 (Maximum of uniforms) If $X_1, X_2, \ldots$ are iid uniform$(0,1)$ and $X_{(n)} = \max_{1\leq i\leq n} X_i$, let us examine if (and to where) $X_{(n)}$ converges in distribution.

As $n \to \infty$, we expect $X_{(n)}$ to get close to 1 and, as $X_{(n)}$ must necessarily be less than 1, we have for any $\varepsilon > 0$,

$$\begin{aligned} P(|X_{(n)} - 1| \geq \varepsilon) &= P(X_{(n)} \geq 1 + \varepsilon) + P(X_{(n)} \leq 1 - \varepsilon) \\ &= 0 + P(X_{(n)} \leq 1 - \varepsilon). \end{aligned}$$

Next using the fact that we have an *iid* sample, we can write

$$P(X_{(n)} \leq 1 - \varepsilon) = P(X_i \leq 1 - \varepsilon, i = 1, \ldots n) = (1 - \varepsilon)^n,$$

which goes to 0. So we have proved that $X_{(n)}$ converges to 1 in probability. However, if we take $\varepsilon = t/n$, we then have

$$P(X_{(n)} \leq 1 - t/n) = (1 - t/n)^n \to e^{-t},$$

which, upon rearranging, yields

$$P(n(1 - X_{(n)}) \leq t) \to 1 - e^{-t};$$

that is, the random variable $n(1 - X_{(n)})$ converges in distribution to an exponential(1) random variable. ‖

Theorem 5.5.14 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be a sequence of iid random variables whose mgfs exist in a neighborhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$, for some positive h). Let $\mathrm{E}X_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 > 0$. (Both μ and σ^2 are finite since the mgf exists.) Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x, $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy;$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution.

Proof of Theorem 5.5.14: We will show that, for $|t| < h$, the mgf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges to $e^{t^2/2}$, the mgf of a n$(0, 1)$ random variable.

Define $Y_i = (X_i - \mu)/\sigma$, and let $M_Y(t)$ denote the common mgf of the Y_is, which exists for $|t| < \sigma h$ and is given by Theorem 2.3.15. Since

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i, \tag{5.5.1}$$

we have, from the properties of mgfs (see Theorems 2.3.15 and 4.6.7),

$$\begin{aligned} M_{\sqrt{n}(\bar{X}_n-\mu)/\sigma}(t) &= M_{\Sigma_{i=1}^n Y_i/\sqrt{n}}(t) \\ &= M_{\Sigma_{i=1}^n Y_i}\left(\frac{t}{\sqrt{n}}\right) && \text{(Theorem 2.3.15)} \\ &= \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n. && \text{(Theorem 4.6.7)} \end{aligned} \tag{5.5.2}$$

We now expand $M_Y(t/\sqrt{n})$ in a Taylor series (power series) around 0. (See Definition 5.5.20.) We have

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} M_Y^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!}, \tag{5.5.3}$$

where $M_Y^{(k)}(0) = \left(d^k/dt^k\right) M_Y(t)|_{t=0}$. Since the mgfs exist for $|t| < h$, the power series expansion is valid if $t < \sqrt{n}\sigma h$.

Using the facts that $M_Y^{(0)} = 1$, $M_Y^{(1)} = 0$, and $M_Y^{(2)} = 1$ (by construction, the mean and variance of Y are 0 and 1), we have

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right), \tag{5.5.4}$$

where R_Y is the remainder term in the Taylor expansion,

$$R_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=3}^{\infty} M_Y^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!}.$$

An application of Taylor's Theorem (Theorem 5.5.21) shows that, for fixed $t \neq 0$, we have

$$\lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0.$$

Since t is fixed, we also have

$$\lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n\to\infty} nR_Y\left(\frac{t}{\sqrt{n}}\right) = 0, \tag{5.5.5}$$

and (5.5.5) is also true at $t = 0$ since $R_Y(0/\sqrt{n}) = 0$. Thus, for any fixed t, we can write

$$\begin{aligned}\lim_{n\to\infty}\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n &= \lim_{n\to\infty}\left[1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &= \lim_{n\to\infty}\left[1 + \frac{1}{n}\left(\frac{t^2}{2} + nR_Y\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n \\ &= e^{t^2/2}\end{aligned} \tag{5.5.6}$$

by an application of Lemma 2.3.14, where we set $a_n = (t^2/2) + nR_Y(t/\sqrt{n})$. (Note that (5.5.5) implies that $a_n \to t^2/2$ as $n \to \infty$.) Since $e^{t^2/2}$ is the mgf of the n(0, 1) distribution, the theorem is proved. □

Theorem 5.5.17 (Slutsky's Theorem) *If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then*

a. *$Y_nX_n \to aX$ in distribution.*

b. *$X_n + Y_n \to X + a$ in distribution.*

Example 5.5.18 (Normal approximation with estimated variance) Suppose that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \text{n}(0, 1),$$

but the value of σ is unknown. We have seen in Example 5.5.3 that, if $\lim_{n\to\infty} \text{Var } S_n^2 = 0$, then $S_n^2 \to \sigma^2$ in probability. By Exercise 5.32, $\sigma/S_n \to 1$ in probability. Hence, Slutsky's Theorem tells us

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \text{n}(0, 1). \qquad \|$$

5.5.4 The Delta Method

The previous section gives conditions under which a standardized random variable has a limit normal distribution. There are many times, however, when we are not specifically interested in the distribution of the random variable itself, but rather some function of the random variable.

Theorem 5.5.24 (Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathrm{n}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathrm{n}(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.} \tag{5.5.10}$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \text{ Remainder}, \tag{5.5.11}$$

where the remainder $\to 0$ as $Y_n \to \theta$. Since $Y_n \to \theta$ in probability it follows that the remainder $\to 0$ in probability. By applying Slutsky's Theorem (Theorem 5.5.17) to

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta),$$

the result now follows. See Exercise 5.43 for details. □

Ex. $X_1, X_2, \ldots$ i.i.d. $E(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2 > 0$

$\sqrt{n}(\bar{X} - \mu) \Rightarrow N(0, \sigma^2)$

Define $g(x) = \frac{1}{x}$

Delta method $\Rightarrow \sqrt{n}(\frac{1}{\bar{X}} - \frac{1}{\mu}) \Rightarrow N(0, \sigma^2 \cdot \frac{1}{\mu^4})$; $\mu \neq 0$

$[S_n^2 \xrightarrow{P} \sigma^2,\ \bar{X}_n \xrightarrow{P} \mu]$ $\quad \frac{\sqrt{n}(\frac{1}{\bar{X}} - \frac{1}{\mu})}{S_n \cdot (\frac{1}{\bar{X}})^2} \Rightarrow N(0,1)$; $\mu \neq 0$