Chapter 1: Probability Theory

MATH 667-01
Statistical Inference
University of Louisville
Textbook: Statistical Inference by Casella and Berger

August 25, 2015
Last corrected: 8/27/15

- *Definition*: The **sample space** is the set of all possible outcomes of a particular experiment.
- *Definition*: An **event** is any collection of possible outcomes of an experiment; that is, an event is a subset of the sample space.

Sets

- $A \subset B$ if $x \in A \Rightarrow x \in B$
- $A = B$ if and only if $A \subset B$ and $B \subset A$
- Union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$
- Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$
- Complementation: $A^c = \{x : x \notin A\}$
- Empty set: $\emptyset = \{\}$

- *Example*: Suppose we toss a fair coin twice and record H or T for each toss. Then the sample space can be represented by

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

Let A be the event that a head occurs on the first toss and B be the event that exactly one head occurs; that is,

$$A = \{HH, HT\} \text{ and } B = \{HT, TH\}.$$

The union of A and B is $A \cup B = \{HH, HT, TH\}$.
The intersection of A and B is $A \cap B = \{HT\}$.
The complement of A is $A^c = \{TH, TT\}$.
The complement of S is $S^c = \emptyset$.

- *Theorem*: Let A, B, and C be any events defined on sample space S. Then the following properties hold.
  - Commutativity: $$A \cup B = B \cup A$$ $$A \cap B = B \cap A$$
  - Associativity: $$A \cup (B \cup C) = (A \cup B) \cup C$$ $$A \cap (B \cap C) = (A \cap B) \cap C$$
  - Distributive Laws: $$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$ $$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
  - DeMorgan's Laws: $$(A \cup B)^c = A^c \cap B^c$$ $$(A \cap B)^c = A^c \cup B^c$$

- If Γ is an index set, then
  $\bigcup_{a \in \Gamma} A_a = \{x \in S : x \in A_a \text{ for some } a\}$ and
  $\bigcap_{a \in \Gamma} A_a = \{x \in S : x \in A_a \text{ for all } a\}$.
- *Definition*: 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** if $A_i \cap A_j = \emptyset$ for all $i \neq j$.
- *Definition*: If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$, then $A_1, A_2, \ldots$ forms a **partition** of S.

- *Example*: Let $S = [0,1)$, $A_i = [0, 2^{-i})$, $i = 0, 1, 2, \ldots$, and $\Gamma = \{0, 1, \ldots\}$. Then

$$\bigcup_{a \in \Gamma} A_a = S \text{ and } \bigcap_{a \in \Gamma} A_a = \{0\}.$$

Let $B_i = A_i \cap A_{i+1}^c = [2^{-i+1}, 2^{-i})$, $i = 0, 1, \ldots$. Then $B_0, B_1, B_2 \ldots$ are pairwise disjoint and

$$\bigcup_{a \in \Gamma} B_a = S \cap \{0\}^c \text{ and } \bigcap_{a \in \Gamma} B_a = \emptyset.$$

A partition of S is formed by $\{0\}, B_0, B_1, B_2, \ldots$.

- *Definition*: A collection $\mathcal{B}$ of subsets of S is called a **Borel field** (or σ-**algebra**) if
  1. $\emptyset \in \mathcal{B}$
  2. $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$
  3. $A_1, A_2, \ldots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.
- *Definition*: (**Axioms of Probability** or **Kolmogorov Axioms**) Given a sample space S and an associated Borel field $\mathcal{B}$, a **probability function** is a function P with domain $\mathcal{B}$ such that
  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
     $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

- *Theorem*: If P is a probability function and $A \in \mathcal{B}$, then
  a. $P(\emptyset) = 0$
  b. $P(A) \leq 1$
  c. $P(A^c) = 1 - P(A)$.
- *Theorem*: If P is a probability function and $A, B \in \mathcal{B}$, then
  a. $P(A^c \cap B) = P(B) - P(A \cap B)$
  b. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  c. If $A \subset B$, then $P(A) \leq P(B)$.
- *Theorem*: If P is a probability function, then
  a. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ for any partition $C_1, C_2, \ldots$ of S
  b. (**Boole's inequality**) $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$ for any sets $A_1, A_2, \ldots$.

- *Example*: Let $S = (0, 1]$, $A_i = (0, 2^{-i}]$, $i = 0, 1, 2, \ldots$, and let $\mathcal{B}$ be the smallest Borel field that contains A_i for $i = 0, 1, 2, \ldots$. For any $B \in \mathcal{B}$, let $P(B) =$ "length of B". For $i < j$, let $A_{i,j} = (2^{-j}, 2^{-i}]$. Show that $A_{i,j} \in \mathcal{B}$ and compute $P(A_{i,j})$.

- *Answer*:
  - $P(A_i) = 2^{-i}$
  - $A_i^c = (2^{-i}, 1]$ and $P(A_i^c) = 1 - P(A_i) = 1 - 2^{-i}$
  - $A_{i,j} = (A_i^c \cup A_j)^c \in \mathcal{B}$

  and

$$
\begin{aligned}
P(A_{i,j}) &= 1 - P(A_i^c \cup A_j) = 1 - P(A_i^c) - P(A_j) \\
&= 1 - (1 - 2^{-i}) - 2^{-j} = 2^{-i} - 2^{-j}.
\end{aligned}
$$

Counting

- *Theorem*: (Fundamental Theorem of Counting) 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.
- *Definition*: (**factorials**) If n is a positive integer, then $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$. By definition, $0! = 1$
- *Definition*: (**binomial coefficients**) For nonnegative integers n and $r \leq n$, $\binom{n}{r} = \dfrac{n!}{r!(n-r)!}$.

- *Example*: Suppose we want to create a 3-symbol code with one letter and two numerical digits (for example, R58 or 0G0).

  a. How many codes are possible?
  b. How many codes have different numbers?
  c. For how many codes is the left number less than the right number?
  d. For how many codes is the left number less than or equal to the right number?

- *Answers*:
  a. How many codes are possible? $3 \times 26 \times 10^2 = 7800$
  b. How many codes have different numbers? $3 \times 26 \times 10 \times 9 = 7020$
  c. For how many codes is the left number less than the right number? $3 \times 26 \times \binom{10}{2} = 3510$
  d. For how many codes is the left number less than or equal to the right number? $3 \times 26 \times \binom{11}{2} = 4290$

Equally likely outcomes

- If $S = \{s_1, \ldots, s_N\}$ is a finite sample space, then the outcomes are **equally likely** if $P(\{s_i\}) = \dfrac{1}{N}$ for $i = 1, \ldots, N$.
- For any event A in an equally likely sample space S, the probability of A is

$$P(A) = \frac{\text{number of elements of } A}{\text{number of elements of } S}.$$

- *Example*: Suppose we randomly select a 3-symbol code with one letter and two numerical digits (for example, R58 or 0G0) from a list of all possible codes. The probability that the code is a letter followed by two numbers which are not the same is

$$\frac{26 \times 10 \times 9}{3 \times 26 \times 10^2} = \frac{9}{3 \times 10} = \frac{3}{10}.$$

- *Definition*: If $A, B \in S$ and $P(B) > 0$, then

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

- **Bayes' Rule**: Let $A_1, A_2, \dots$ be a partition of the sample space S and $B \subset S$. If $P(B) > 0$ and $P(A_i) > 0$, then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j:P(A_j)>0} P(B|A_j)P(A_j)}.$$

- *Example*: Suppose that 0.1% of individuals in a population have a disease. If an individual has the disease, the probability that a particular symptom is present is 0.9. If an individual does not have the disease, the probability that the symptom is present is 0.01. What is the probability that an individual has the disease, given that the individual has the symptom? (*Answer* ≈ 0.0826)

- *Definition*: A and B are **statistically independent** if $P(A \cap B) = P(A)P(B)$
- *Theorem*: If A and B are independent, then the following pairs are also independent:
  a. A and B^c
  b. A^c and B
  c. A^c and B^c
- *Definition*: 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

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

- *Example*: In each round of a game, a fair coin is tossed 4 times. The game ends when the number of heads exceeds the number of tails during a round. What is the probability that the game lasts at least 6 rounds? (*Answer* ≈ 0.1536)

- *Definition*: A **random variable** is a function from the sample space to the real numbers.
- Random variables are usually denoted by uppercase letters. Realized values of random varaibles are usually denoted by lowercase letters.
- Finite probability space:
  - $S = \{s_1, \ldots, s_n\}$ with probability function P
  - X is a random variable from S to $\mathbb{R}$ with range $\mathcal{X} = \{x_1, \ldots, x_m\}$
  - P_X is the induced probability function of X
  - $P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$

- *Example*: Suppose we roll a pair of dice so that

$$
\begin{aligned}
S &= \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\
&(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\
&(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\
&(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\
&(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\
&(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}
\end{aligned}
$$

  and we let X be the total number of spots on the pair of dice. Then X is a set function that assigns a value to every point in the sample space. For example, $X((4,2)) = 6$. We can find the probability of events by inverting this function. For instance,

$$
\begin{aligned}
P_{X}(X=6) &\equiv P(\{\omega \in S : X(\omega) = 6\})\\
&= P(\{(1,5),(2,4),(3,3),(4,2),(5,1)\}) = \frac{5}{36}.
\end{aligned}
$$

- *Definition*: The **cumulative distribution function** (cdf) of X is $F_X(x) = P_X(X \leq x)$ for all x.
- *Theorem*: The function $F(x)$ is a cdf if and only if the following conditions hold:
  a. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
  b. $F(x)$ is nondecreasing in x
  c. $F(x)$ is right continuous; that is, for every x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

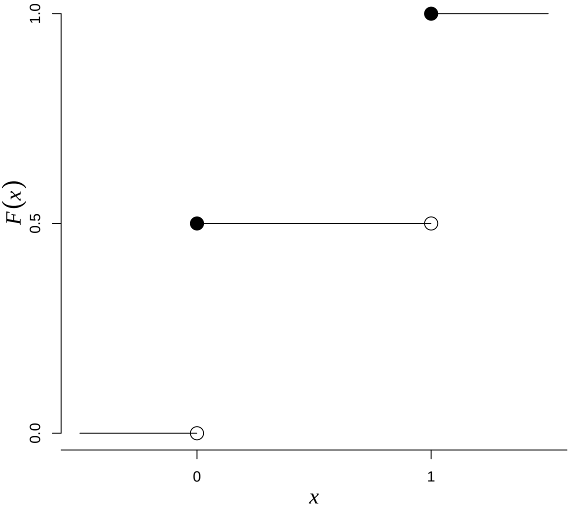
- *Definition*: A random variable is **continuous** if $F_X(x)$ is a continuous function of x. A random variable is **discrete** if $F_X(x)$ is a step function of x.
- *Definition*: Random variables X and Y are **identically distributed** if for every set A, $P(X \in A) = P(Y \in A)$.
- *Theorem*: X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for every x.

- *Definition*: The **probability mass function** (pmf) of a discrete random variable X is given by $f_X(x) = P(X = x)$ for all x.
- *Definition*: The **probability density function** (pdf) of a continuous random variable X is a function $f_X(x)$ that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$ for all x.
- For a continuous random variable X, $\dfrac{dF_X(x)}{dx} = f_X(x)$.

- *Theorem*: A function $f_X(x)$ is a pmf (or pdf) of a random variable X if and only if
  a. $f_X(x) \geq 0$ for all x
  b. $\sum_x f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$ (pdf)
- If X has cdf F or pmf/pdf f, we often write $X \sim F_X(x)$ or $X \sim f_X(x)$, respectively.
- If X and Y have the same distribution, then we can write $X \sim Y$.

- *Example*: Suppose we define X such that $P(X = 0) = .5$ and $P(X = 1) = .5$. Then the cdf of X is

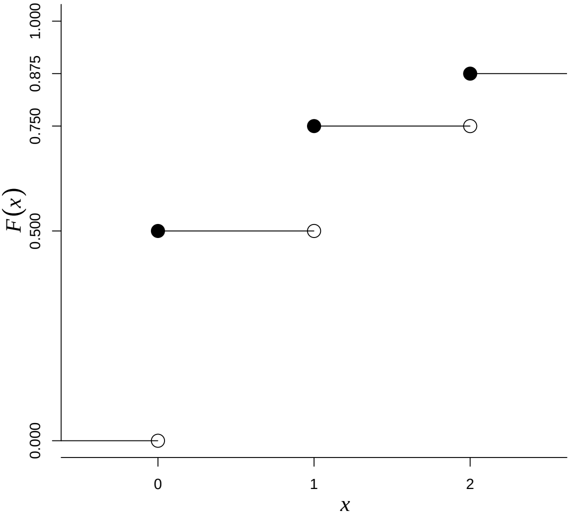
$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ .5 & \text{for } 0 \le x < 1 \\ 1 & \text{for } 1 \le x \end{cases}.$$


1.0
0.5
0.0
$F(x)$
0
1
x

- *Example*: Suppose we define X such that $P(X=0)=\frac{1}{2}$, $P(X=1)=\frac{1}{4}$, $P(X=2)=\frac{1}{8},\ldots,P(X=i)=\frac{1}{2^{i+1}},\ldots$. Then the cdf of X is

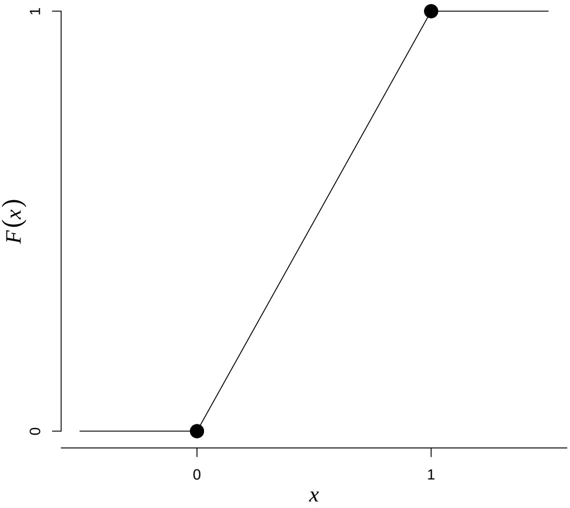
$$F(x)=\begin{cases} 0 & \text{for } x<0 \\ .5 & \text{for } 0\le x<1 \\ .75 & \text{for } 1\le x<2 \\ .875 & \text{for } 2\le x<3 \\ \vdots & \vdots \\ \frac{2^{i+1}-1}{2^{i+1}} & \text{for } i\le x<i+1 \\ \vdots & \vdots \end{cases}.$$

1.6. Density and Mass Functions



- *Example*: Suppose X is continuous with density $f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$. When $0 < x < 1$, $F(x) = \int_0^x 1 \, d\omega = [\omega]_0^x = x$. Thus, the cdf of X is

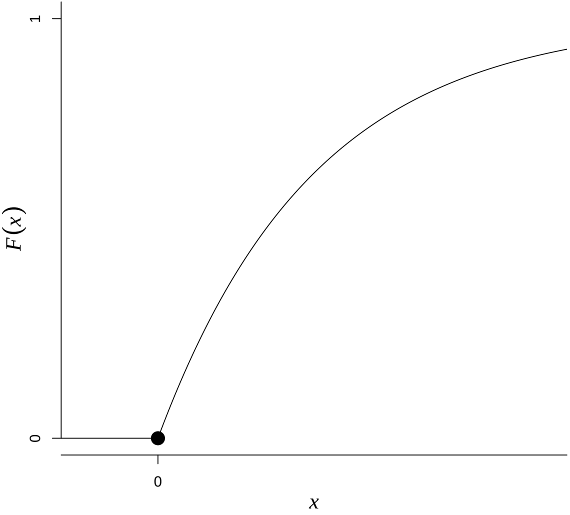
$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \le x < 1 \\ 1 & \text{for } 1 \le x \end{cases}.$$



1
0
$F(x)$
0
1
x

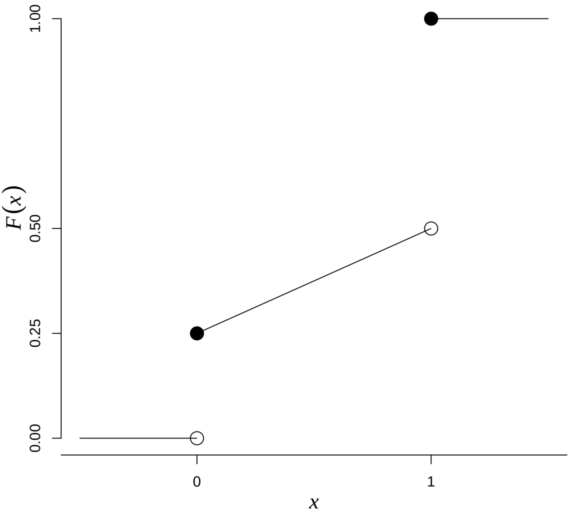
- *Example*: Suppose X is continuous with density $f(x) = e^{-x}, x > 0$. When $x > 0$, $F(x) = \int_0^x e^{-\omega}\, d\omega = [-e^{-\omega}]_0^x = 1 - e^{-x}$. Thus, the cdf of X is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-x} & \text{for } 0 \leq x \end{cases}.$$


1
0
$F(x)$
0
x

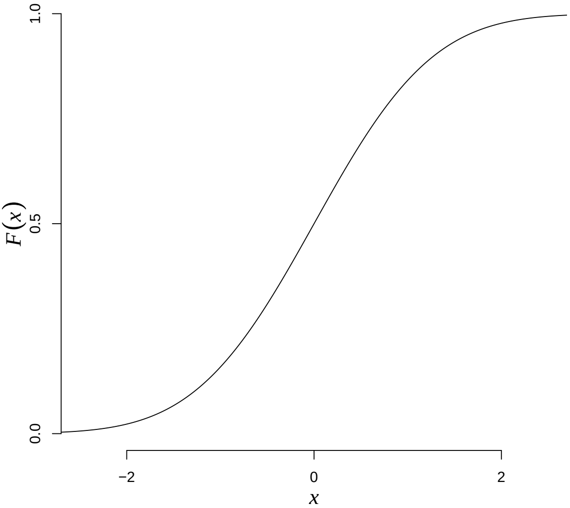
- *Example*: Consider the mixed distribution that has $P(X = 0) = \frac{1}{4}$, $P(X = 1) = \frac{1}{2}$, and $\frac{1}{4}$ of the probability is uniformly distributed over the interval $(0, 1)$. Then the cdf of X is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1+x}{4} & \text{for } 0 \leq x < 1 \\ 1 & \text{for } 1 \leq x \end{cases} .$$


1.00
0.50
0.25
0.00
$F(x)$
0
1
x

- *Example*: Suppose X is continuous with density $f(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Then, the cdf of X is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\omega^2/2} \, d\omega.$$


1.0
0.5
0.0
$F(x)$
−2
0
2
x