*Last modified November 2, 2019*

# STA257: Probability and Statistics I

## University of Toronto — Fall 2019

Jeff Shen

## Contents

# 1 Probability and Counting

## 1.1 Introduction

**Experiments**: situations where outcome is random. e.g. flipping a coin is an experiment.
**Sample space**: set of all possible outcomes, denoted $S$ or $\Omega$. The number of elements in the sample space (cardinality) is denoted $|\Omega|$.
**Event**: a subset of a sample space.
**Outcome**: a particular element of a sample space. e.g. $s_1 \in \Omega$.

## 1.2 Set Theory

### 1.2.1 Definitions

- **union**: $A \cup B$. Elements in either $A$ or $B$.
- **intersection**: $A \cap B$. Elements in both $A$ and $B$.
- **complement** of A: $A^c$. Elements not in $A$.
- **empty set**: $\varnothing$. Set with no elements in it.
- $A$ and $B$ are **disjoint**: $A \cap B = \varnothing$. There are no elements in the intersection of $A$ and $B$.

### 1.2.2 Laws of Set Theory

1. **commutativity**: $A \cup B = B \cup A$, $A \cap B = B \cap A$
2. **associativity**: $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap B)$
3. **distributivity**: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

## 1.3 Probability Measures

**Probability measure**: a function which maps subsets of $\Omega$, which can be defined on any space, to real numbers $\mathbb{R}$.

### 1.3.1 Axioms of Probability Measures

- $P(\Omega) = 1$
- $\forall A \in \Omega, P(A) \geq 0$
- if $A_1, A_2, \dots A_n, \dots$ are mutually disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

### 1.3.2 Properties of Probability Measures

- $\forall A \in \Omega, P(A^c) = 1 - P(A)$

  Proof:

$$\begin{aligned} 1 &= P(\Omega) && \text{by Axiom 1} \\ &= P(A \cup A^c) && \text{by definition of complement} \\ &= P(A) + P(A^c) && \text{by Axiom 3 (since } A, A^c \text{ are disjoint)} \end{aligned}$$

  Rearrange this to see that $P(A^c) = 1 - P(A)$.

- $P(\varnothing) = 0$

  Proof:

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \varnothing) && \text{since } \Omega \cup \varnothing = \Omega \\ &= P(\Omega) + P(\varnothing) && \text{by Axiom 3 (since } \Omega, \varnothing \text{ are disjoint)} \end{aligned}$$

  So $P(\varnothing) = 0$.

- For $A, B \subseteq \Omega, A \subseteq B \implies P(A) \leq P(B)$

  Proof:

$$\begin{aligned} P(B) &= P(A \cup (B \cap A^c)) \\ &= P(A) + P(B \cap A^c) && \text{by Axiom 3 (since } A, A^c \text{ are disjoint)} \end{aligned}$$

  But note that $P(B \cap A^c) \geq 0$ by Axiom 2.

  Then $P(B) = P(A) + P(B \cap A^c) \geq P(A)$.

- For $A, B \subseteq \Omega, P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  Proof:

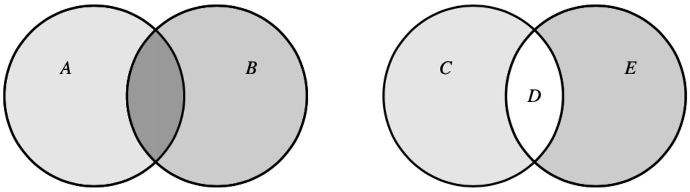
  Case 1: $A, B$ are disjoint. Then $A \cap B = \varnothing \implies P(A \cap B) = 0$.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) && \text{by Axiom 3 (since } A, B \text{ are disjoint)} \\ &= P(A) + P(B) + P(A \cap B) && \text{since we can add 0 wherever we want} \end{aligned}$$

  Case 2: $A, B$ not disjoint. Then $A \cap B \neq \varnothing$.

  Let $C = A \cap B^c$, $D = A \cap B$, $E = A^c \cap B$.

  Then $C, D, E$ are disjoint, and $A = C \cup D$, $B = D \cup E$, and $A \cup B = C \cup D \cup E$.



$$\begin{aligned} P(A) + P(B) - P(A \cap B) &= P(C \cup D) + P(D \cup E) - P(D) && \text{by how we defined } C, D, E \\ &= P(C) + P(D) + P(D) + P(E) - P(D) && \text{by Axiom 3 and disjointness of } C, D, E \\ &= P(C) + P(D) + P(E) \\ &= P(C \cup D \cup E) && \text{by Axiom 3 and disjointness of } C, D, E \\ &= P(A \cup B) \end{aligned}$$

## 1.4 Counting

**Multiplication principle**: if there are $m$ ways to do one thing, and $n$ ways to do another thing, then there are $mn$ ways to do both things.
**Permutation**: ordered arrangement of objects.

- Sampling **with replacement** means that duplicate item selection is allowed. (can pick the same object twice). For a set of size $n$ and a **sample size** (number of items selected) $r$, there are $n^r$ possible selections.
- Sampling **without replacement** means that each item is selected once at most. For a set of size $n$ and a sample size $r$, there are $n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}$ possible selections. In particular, there are $n(n-1)\dots(1) = n!$ ways to order $n$ elements.

**Combination**: arrangement of objects *without regard to order.* Think about this as the ways to select objects without replacement, divided by the ways that those objects can be ordered. For a set of size $n$ and a sample size $r$, we express the combination as follows:

$$\binom{n}{r} = \frac{n(n-1)\dots(n-r+1)}{r!} = \frac{n!}{(n-r)!\,r!}$$

**Binomial expansion**:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}$$

In particular, for $a = b = 1$,

$$(1+1)^n = 2^n = \sum_{k=0}^{n} \binom{n}{k} (1)^k (1)^{n-k} = \sum_{k=0}^{n} \binom{n}{k}$$

The number of ways to group $n$ objects into $r$ classes, with $n_i$ objects in the $i$th classes, where $1 < i < r$, is given by the following formula:

$$\binom{n}{n_1 n_2 \dots n_r} = \frac{n!}{n_1!\, n_2! \dots n_r!}$$

Proof: There are $\binom{n}{n_1}$ ways to select the first class, $\binom{n-n_1}{n_2}$ ways to select the second class, and so on. Repeat this for all $r$ classes, and then apply the multiplication rule. Then we have that

$$\begin{aligned}
\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r} &= \frac{n!}{n_1!\,(n-n_1)!}\,\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}\cdots\frac{(n-n_1-\cdots-n_{r-1})!}{n_r!\,0!} \\
&= \frac{n!}{n_1!\,\cancel{(n-n_1)!}}\,\frac{\cancel{(n-n_1)!}}{n_2!\,\cancel{(n-n_1-n_2)!}}\cdots\frac{\cancel{(n-n_1-\cdots-n_{r-1})!}}{n_r!\,0!} \\
&= \frac{n!}{n_1!\,n_2!\dots n_r!}
\end{aligned}$$

## 1.5 Conditional Probability and Independence

**Conditional probability**: probability that some event will occur given that another event has occured:

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

Two events are **independent** if knowing that one has occured gives no information about the likelihood of the other occuring. Events $A, B$ are independent if and only if any of the following hold:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

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

From the definition of conditional probability, we can derive **Bayes' Rule**, which describes the probability of an event given prior knowledge of conditions related to the event:

$$\begin{aligned}
&P(A|B)\,P(B) = P(A \cap B) = P(B \cap A) = P(B|A)\,P(A) \\
\implies &P(A|B)\,P(B) = P(B|A)\,P(A) \\
\implies &P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}
\end{aligned}$$

## 1.6 Law of Total Probability

Let $B_1, B_2, \ldots, B_n$ be sets satisfying the following conditions:

1. $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$ (the $B_i$'s are a **partition** of $\Omega$)
2. $\forall i, j \in [1, n] \subseteq \mathbb{N}, i \neq j \implies B_i \cap B_j = \varnothing$ (pairwise disjoint)
3. $\forall i \in [1, n] \subseteq \mathbb{N}, P(B_i) > 0$ (strictly positive probability),

we can conclude that for any event $A$, we have that

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\,P(B_i)$$

Proof: Since all $B_i$'s are pairwise disjoint, all $A \cap B_i$'s must also be pairwise disjoint. Then

$$\begin{aligned}
P(A) &= P(A \cap \Omega) \\
&= P\left(A \cap \left(\bigcup_{i=1}^{n} B_i\right)\right) && \text{by condition 1} \\
&= P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) \\
&= \sum_{i=1}^{n} P(A \cap B_i) && \text{by Axiom 3 of probability measures} \\
&= \sum_{i=1}^{n} P(A|B_i)\,P(B_i) && \text{by definition of conditional probability}
\end{aligned}$$

Under the same conditions, we can also conclude an alternate form of Bayes' Rule:

$$P(B_j|A) = \frac{P(A|B_j)\,P(B_j)}{\sum_{i=1}^{n} P(A|B_i)\,P(B_i)}$$

## 1.7 Equations

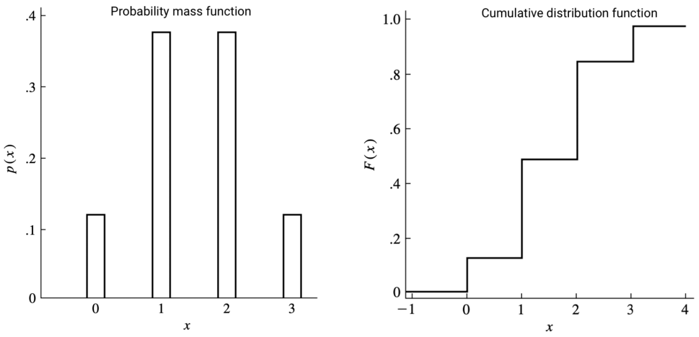
| | |
|---|---|
| Axioms | $P(\Omega) = 1$ |
| | $P(A) \geq 0$ |
| | for mutually disjoint sets, $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ |
| Properties | $P(A^c) = 1 - P(A)$ |
| | $P(\varnothing) = 0, P(\Omega) = 1$ |
| | $A \subseteq B \implies P(A) \leq P(B)$ |
| | $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ |
| Permutations | with replacement: $n^r$ |
| | without replacement: $\frac{n!}{(n-r)!}$ |
| Combinations | $\binom{n}{r} = \frac{n!}{(n-r)!\, r!}$ |
| Binomial Expansion | $(a+b)^n = \sum_{i=1}^{n} \binom{n}{k} a^k\, b^{n-k}$ |
| Grouping $n$ objects into $r$ classes | $\binom{n}{n_1 n_2 \ldots n_r} = \frac{n!}{n_1!\, n_2! \ldots n_r!}$ |
| Conditional probability | $P(A\|B) = \frac{P(A \cap B)}{P(B)}$ |
| Bayes' Rule | $P(A\|B) = \frac{P(B\|A)\, P(A)}{P(B)}$ |
| | $P(B_j\|A) = \frac{P(A\|B_j)\, P(B_j)}{\sum_{i=1}^{n} P(A\|B_i)\, P(B_i)}$ |
| Law of Total Probability | $P(A) = \sum_{i=1}^{n} P(A\|B_i)\, P(B_i)$ |

# 2 Random Variables

## 2.1 Discrete Random Variables

**Random variable**: function from $\Omega$ to $\mathbb{R}$.
**Discrete random variable**: random variable which can only take a finite number of values, or a countably infinite number of values.
**Probability mass function (PMF)**: describes probability properties of a random variable. For some discrete random variable $X$ taking on values $x_1, x_2, \ldots$, the probability mass function is some function $p$ such that $p(x_i) = P(X = x_i)$, and satisfies that $\sum_{i=1}^{\infty} p(x_i) = 1$.
**Cumulative distribution function (CDF)**: defined as some function $F$ such that $F(x) = P(X \leq x)$, where $-\infty < x < \infty$. It satisfies the following conditions:

- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- is right-continuous (can have jumps, but must be continuous when approaching values from the right side). Continuity implies right-continuity.
- is non-decreasing.



For random variables $X, Y$ taking on values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ respectively, $X$ and $Y$ are **independent** if $\forall i, j, P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$.

### 2.1.1 Bernoulli

$X \sim \text{Bern}(p)$. Was there a success?
$X$ takes on the values 1 and 0 with probabilities $p$ and $q = 1 - p$ respectively. For some event $A$, Bernoulli random variables are often used as an **indicator random variable**. That is, $I_A = 1$ if $A$ occurs, and $I_A = 0$ otherwise.

### 2.1.2 Binomial

$X \sim \text{Bin}(n, p)$. Something was done $n$ times. How many successes were there?
Binomial random variables can be thought of as the sum of $n$ **i.i.d (independent and identically distributed)** Bernoulli random variables. That is, $X = X_1 + X_2 + \cdots + X_n$, where $X_i \sim \text{Bern}(p)$.

### 2.1.3 Geometric

$X \sim \text{Geom}(p)$. Keep going until there is a success.
Constructed from a (countably) infinite sequence of Bernoulli trials. Count the number of trials it took until finally getting a success. Some people don't include the $k$th trial (the last trial, which is a success), but we will. So $k$ is the trial on which we have a success.

### 2.1.4 Negative Binomial

$X \sim \text{NB}(n, p)$. Keep going until there are $n$ successes.
A natural extension to the geometric distribution—can be thought of as the sum of $n$ i.i.d geometric random variables. That is, $X = X_1 + X_2 + \cdots + X_n$, where $X_i \sim \text{Geom}(p)$.

### 2.1.5 Hypergeometric

$X \sim \text{HG}(n, r, m)$. Total population of $n$ in which $r$ are marked/tagged. We randomly pick $m$ from the population without replacement. How many are tagged?

### 2.1.6 Poisson

$X \sim \text{Pois}(\lambda)$. Large number of trials, each will a small probability of success.
$\lambda$ is called the **rate parameter**. Think of dividing some interval into a large number of subintervals, such that the probability of an event occuring in each subinterval is small, but there are a large number of subintervals.
Assume that what happens in one subinterval is independent of what happens in the other subintervals, that the probability of some event occuring is the same for each subinterval, and that events do not occur simultaneously. Then this can be derived by setting up a binomial distribution in such a way that $n \to \infty$, $p \to 0$, and $np = \lambda$.

DERIVE POISSON FROM BINOMIAL

## 2.2 Continuous Random Variables

**Continuous random variables** can take on a continuum of values rather than just a finite (or countably infinite) number.
**Probability density function (pdf)**. Analogous to the pmf for discrete random variables. Satisfies the following conditions:

- $f(x) \geq 0$
- $f$ is piecewise continuous
- $\int_{-\infty}^{\infty} f(x)dx = 1$

For some $a, b$ such that $a < b$, the probability that $X$ is in the interval $(a, b)$ is given by the area under $f(x)$ from $a$ to $b$: $P(a < X < b) = \int_a^b f(x)dx$. Note that the probability of taking on a particular value is 0: $P(X = c) = \int_c^c f(x)dx = 0$. This implies that $P(a \leq X \leq b) = P(a < X < b)$.
For an (infinitessimally) small value of $dx$, the probability that $X$ is in the interval $(x, x + dx)$ is proportional to $f(x)$: $P(x \leq X \leq x + dx) = f(x)dx$.
The **cumulative distribution function (cdf)** can be expressed in terms of the pdf:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

If $f$ is continuous at $x$, then by the Fundamental Theorem of Calculus,

$$\begin{aligned} f(x) &= F'(x) \\ \implies P(a < X < b) &= \int_a^b f(x)dx = F(b) - F(a). \end{aligned}$$

If $F$ is strictly increasing on some interval $I$, with $F = 0$ to the left of $I$ and $F = 1$ to the right of $I$, then the **inverse** function $F^{-1}$ is well defined:

$$\begin{aligned} y &= F(x) \\ \implies x &= F^{-1}(y) \end{aligned}$$

The $p$**th quantile** of F is the value $x_p$ such that $x_p = F^{-1}(p)$. Note that the **percentile** is the same as the quantile, but expressed as a percentage from 0 to 100, whereas the quantile is expressed as a value between 0 and 1. Special quantiles are the **median** ($p = \frac{1}{2}$) and the **lower and upper quartiles** ($p = \frac{1}{4}, \frac{3}{4}$) respectively.

#### 2.2.1 Uniform

#### 2.2.2 Exponential

#### 2.2.3 Gamma

#### 2.2.4 Beta

#### 2.2.5 Uniform

#### 2.2.6 Standard Normal

#### 2.2.7 General Normal

### 2.3 Transformations of Random Variables

### 2.4 Distributions

pmf/pdf, cdf, ex, var, range of support for k

# 3 Expected Values

## 3.1 Mean and Variance

### 3.1.1 LOTUS

### 3.1.2 Inequalities

## 3.2 Moment Generating Functions

# 4 Joint Distributions

## 4.1 Joint and Marginal Distributions

### 4.1.1 Discrete

### 4.1.2 Continuous

## 4.2 Independence in Joint Distributions

## 4.3 Conditional Distributions

### 4.3.1 Discrete

### 4.3.2 Continuous

## 4.4 Functions of Joint Distributions

## 4.5 Order Statistics