MAT 525

Mathematical Statistics

Lecture Notes [1]

Spring 2025

Wei Li

Department of Mathematics

Syracuse University

[1] The notes outline the materials covered in MAT 525. Detailed examples and discussions are provided in the class.

1 Review

1.1 Probability

A random experiment is any mechanism that produces outcomes which are not predictable with certainty in advance.

Definition 1.1.1. (**Sample Space**). The set S, of all possible outcomes of a particular random experiment is called the sample space of the experiment.

Definition 1.1.2. (**Event**). An event is any subset of S, including S itself.

Theorem 1.1.1. For any three events A, B and C:

(1) Commutativity: $A \cup B = B \cup A, A \cap B = B \cap A$

(2) Associativity: $A \cup (B \cup C) = (A \cup B) \cup C, A \cap (B \cap C) = (A \cap B) \cap C$

(3) Distribution Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C), A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(4) DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c, (A \cap B)^c = A^c \cup B^c$

Definition 1.1.3. (**Disjoint/Mutually Exclusive**). Two events A and B are disjoint (mutually exclusive) if $A \cap B = \emptyset$. The events $A_1, A_2, \ldots$ are disjoint (pairwise disjoint or mutually exclusive) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition 1.1.4. (**Partition**). If $A_1, A_2, \ldots,$ are disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, \ldots$ forms a partition of S.

Definition 1.1.5. (**Probability**). Associated with each event A in the sample space S is a probability $\mathrm{P}(A)$. Here P is a function defined on subsets of S and taking values between $[0, 1]$ and further are required to have the following three properties:

(P1) $\mathrm{P}(A) \geq 0$ for any event A,

(P2) $\mathrm{P}(S) = 1$,

(P3) For every infinite sequence of disjoint events $A_1, A_2, \ldots$,

$$\mathrm{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathrm{P}(A_i). \tag{1}$$

A **probability measure** (or simply probability) on a sample space S is the function P that satisfies above properties (P1-P3).

Remark 1.1.1. * Ideally, P is defined on the collection of all subsets of S. However, it is generally impossible to define such P function so that (P1-P3) are all satisfied. In general, P is defined only on a collection $\mathcal{F}$ of subsets of S. The collection $\mathcal{F}$ is called a **sigma algebra** (or **Borel sigma field**). It is required that the sets in $\mathcal{F}$ satisfy three conditions: (1) The sample space S is in $\mathcal{F}$;

(2) If A is in $\mathcal{F}$, then A^c must also be in $\mathcal{F}$;
(3) If $A_1, A_2, \ldots$ is a countable collection of sets in $\mathcal{F}$, then $\cup_{i=1}^{\infty} A_i$ is also in $\mathcal{F}$. The sets in $\mathcal{F}$ are called **ble sets**. The triple $(\mathrm{P}, \mathcal{F}, S)$ is called a **probability space**. For practical purpose, we shall assume all the events of interest are measurable and therefore the probabilities for these events are all well-defined. Simply write (P, S) as a **probability space**.

Remark 1.1.2. Some properties of probability:

- If $A_1, \ldots, A_n$ are disjoint, then $\mathrm{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathrm{P}(A_i)$.
- $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.
- $A \subset B$ implies $\mathrm{P}(A) \leq \mathrm{P}(B)$.
- $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$ (additive rule).
- Inclusion-exclusion formula:

$$\mathrm{P}(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mathrm{P}(A_i) - \sum_{i<j} \mathrm{P}(A_i \cap A_j) + \sum_{i<j<k} \mathrm{P}(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} \mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_n)$$

- (Bonferroni) For events $A_1, \ldots, A_n$,
  $\mathrm{P}(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n \mathrm{P}(A_i)$,
  $\mathrm{P}(\bigcap_{i=1}^n A_i) \geq 1 - \sum_{i=1}^n \mathrm{P}(A_i^c) = \sum_{i=1}^n \mathrm{P}(A_i) - n + 1$.

Definition 1.1.6. (**Conditional Probability**). The conditional probability of the event A given that the event B has occurred is denoted by $\mathrm{P}(A|B)$. If $\mathrm{P}(B) > 0$, then

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

Note $\mathrm{P}(\cdot|B)$ is also a *probability measure* that satisfies the properties (P1-P3) in Definition 1.1.5.

The conditional probability asks "if we know that an outcome is in the set B, what is the probability that the event is also in A", "what proportion of time that B happens, does A also happen?"

Theorem 1.1.2. (**Multiplication Rule for Conditional Probabilities**). Suppose that $A_1, A_2, \ldots, A_n$ are events such that $\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$. Then

$$\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathrm{P}(A_1)\mathrm{P}(A_2|A_1)\mathrm{P}(A_3|A_1 \cap A_2) \cdots \mathrm{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Suppose that $A_1, A_2, \ldots, A_n, B$ are events such that $\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_{n-1}|B) > 0$ and $\mathrm{P}(B) > 0$. Then

$$\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n|B) = \mathrm{P}(A_1|B)\mathrm{P}(A_2|A_1 \cap B)\mathrm{P}(A_3|A_1 \cap A_2 \cap B) \cdots \mathrm{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap B).$$

Theorem 1.1.3. (**Law of Total Probability**). Suppose that the events $B_1, B_2, \dots, B_k$ form a *partition* of the space S and $\mathrm{P}(B_j) > 0$ for $j = 1, \dots, k$. Then for every event A in S,

$$\mathrm{P}(A) = \sum_{j=1}^{k} \mathrm{P}(B_j)\mathrm{P}(A|B_j).$$

Also, if $\mathrm{P}(C) > 0$,

$$\mathrm{P}(A|C) = \sum_{j=1}^{k} \mathrm{P}(B_j|C)\mathrm{P}(A|B_j \cap C).$$

Definition 1.1.7. (**Independence**). Two events A and B are independent if

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

It is denoted by $A \perp\!\!\!\perp B$.

Suppose $\mathrm{P}(A) > 0$ and $\mathrm{P}(B) > 0$, then A and B are independent if and only if $\mathrm{P}(A|B) = \mathrm{P}(A)$ if and only if $\mathrm{P}(B|A) = \mathrm{P}(B)$. Therefore, $A \perp\!\!\!\perp B$ says that knowing the occurrence or non-occurence of A does not affect your belief about $P(B)$ or the knowledge of it does not help in predicting B, and vice versa.

Theorem 1.1.4. (**Independence of Complements**). If two events A and B are independent, then A and B^c are also independent.

Definition 1.1.8. (**Conditional Independence**). Two events A_1 and A_2 are *conditionally independent given* B if

$$\mathrm{P}(A_1 \cap A_2|B) = \mathrm{P}(A_1|B)\mathrm{P}(A_2|B).$$

It is denoted by $(A_1 \perp\!\!\!\perp A_2)|B$.

Definition 1.1.9. (**(Mutually) Independent Events**). The k events $A_1, \dots, A_k$ are (mutually) independent if for every subset $A_{i_1}, \dots, A_{i_j}$ of j of these events ($j = 2, 3, \dots, k$),

$$\mathrm{P}(A_{i_1} \cap \cdots \cap A_{i_j}) = \mathrm{P}(A_{i_1}) \cdots \mathrm{P}(A_{i_j}).$$

The k events $A_1, \dots, A_k$ are **(mutually) independent given an event** B if for every subset $A_{i_1}, \dots, A_{i_j}$ of j of these events ($j = 2, 3, \dots, k$),

$$\mathrm{P}(A_{i_1} \cap \cdots \cap A_{i_j}|B) = \mathrm{P}(A_{i_1}|B) \cdots \mathrm{P}(A_{i_j}|B).$$

Remark 1.1.3. (Mutually) independent events are necessarily pairwise independent.

Theorem 1.1.5. Suppose A_1, A_2, and B are events such that $\mathrm{P}(A_1 \cap B) > 0$ and $\mathrm{P}(B) > 0$. Then A_1 and A_2 are conditionally independent given B if and only if $\mathrm{P}(A_2|A_1 \cap B) = \mathrm{P}(A_2|B)$.

Remark 1.1.4. If $\mathrm{P}(A) > 0$ and $\mathrm{P}(B) > 0$, then A and B cannot be simultaneously disjoint and independent.

Theorem 1.1.6. * (**Relation Between Disjoint and Independence**). Let $A_1, \ldots, A_n$ ($n > 1$) be disjoint (mutually exclusive) events. These events are also mutually independent if and only if all the events except possibly one of them has probability 0.

Theorem 1.1.7. (**Baye's Theorem**). Let the events $B_1, \ldots, B_k$ form a *partition* of the space S such that $\mathrm{P}(B_j) > 0$ for $j = 1, \ldots, k$ and let A be an event such that $\mathrm{P}(A) > 0$. Then for $i = 1, \ldots, k$,

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(B_i)\mathrm{P}(A|B_i)}{\sum_{j=1}^{k} \mathrm{P}(B_j)\mathrm{P}(A|B_j)}. \tag{3}$$

Note that the denominator is just by applying the law of total probability for $P(A)$.

Theorem 1.1.8. (**Baye's Theorem (Conditional Version)**). Let the events $B_1, \ldots, B_k$ form a *partition* of the space S such that $\mathrm{P}(B_j) > 0$ for $j = 1, \ldots, k$ and let A, C be two events such that $\mathrm{P}(A \cap C) > 0$. Then for $i = 1, \ldots, k$,

$$\mathrm{P}(B_i|A \cap C) = \frac{\mathrm{P}(B_i|C)\mathrm{P}(A|B_i \cap C)}{\sum_{j=1}^{k} \mathrm{P}(B_j|C)\mathrm{P}(A|B_j \cap C)}. \tag{4}$$

Remark 1.1.5. When the partition consists of countably many of sets $B_1, B_2, \ldots$, above Baye's theorems continue to hold with the summation replaced by $\sum_{j=1}^{\infty}$.

1.2 Combinatorics

Theorem 1.2.1. (**Theorem of Counting**) If a random experiment consists of k separate tasks, the i^{th} of which can be done in n_i ways, $i = 1, 2, \cdots, k$, then the entire experiment can be done in $\prod_{i=1}^{k} n_i$ ways.

Theorem 1.2.2. (**Permutations** – ordered without replacement). A permutation is an arrangement of objects in a particular order. The number of *distinct orderings* of k items selected *without replacement* from a collection of different n objects ($0 \le k \le n$) is $P_{n,k} = \frac{n!}{(n-k)!}$, which reads "*the number of k permutation out of n*".

Theorem 1.2.3. (**Combinations** – unordered without replacement). A combination is unordered group of objects. The number of *distinct subsets* of size k that can be chosen from a set of n different object is $C_{n,k} = \frac{n!}{k!(n-k)!}$. We use $\binom{n}{k}$ to denote $C_{n,k}$, which reads "*the number of k combination out of n, or n choose k*".

Theorem 1.2.4. (**Binomial theorem**). For all numbers x and y and each positive integer n,

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Theorem 1.2.5. For all n and all $k = 0, 1, \ldots, n$,

$$\binom{n}{k} = \binom{n}{n-k}.$$

Theorem 1.2.6. * (**Multinomial Coefficient**). The number of ways to partition a set of n items into k distinguishable subsets of sizes $n_1, \ldots, n_k$ where $n_1 + \cdots + n_k = n$ is

$$\binom{n}{n_1, \ldots, n_k} := \frac{n!}{n_1! n_2! \cdots n_k!}$$

Theorem 1.2.7. * (**Multinomial Theorem**). For all real numbers $x_1, \ldots, x_k$ and each positive integer n,

$$(x_1 + \cdots + x_k)^n = \sum \binom{n}{n_1, \ldots, n_k} x_1^{n_1} \cdots x_k^{n_k}$$

where summation extends over all possible combinations of nonnegative integers $n_1, \ldots, n_k$ such that $n_1 + \cdots + n_k = n$.

Remark 1.2.1. (ordered, with replacement) The total number of arrangements of k objects selected from n different objects with replacement is n^k.

Remark 1.2.2. (unordered, with replacement) The total number of distinct subsets of k objects selected from n different objects with replacement is $\binom{n+k-1}{k}$.

1.3 Random Variables

Definition 1.3.1. (**Random Variable (Vector)**). Let S be the sample space for an experiment. Let $X : S \mapsto \mathcal{X}$, where $\mathcal{X}$ is a subset of $\mathbb{R}^d$ for some $d = 1, 2, \ldots$, denote a random variable (when $d = 1$) or random vector (when $d > 1$). The set $\mathcal{X}$ is called the *range* of X or the *sample space* of X.

Remark 1.3.1. Throughout the notes, when orientation matters, vectors will always be taken as **column vectors**. Thus when $d > 1$, a $\mathbb{R}^d$-valued random vector X can be written as $X = (X_1, \ldots, X_d)^T$, where superscript T denotes the transpose. When the context is clear and the orientation does not matter, we may simply treat a multi-dimensional x as a vector by writing $x = (x_1, \ldots, x_d)$. We shall use the terminology **random variable** for any real-valued X.

Definition 1.3.2. (**Probability Induced by a Random Variable (Vector)**). Suppose we have a probability space (P, S). Let $X : S \mapsto \mathcal{X}$ be a random variable (or random vector). The probability that the value of X will belong to some subset $C \subset \mathbb{R}^d$ (such that $\{s : X(s) = C\}$ is an event) is given by

$$\mathrm{P}_X(C) := \mathrm{P}(X \in C) = \mathrm{P}(\{s \in S : X(s) \in C\})$$

Here P_X define a probability function on the subsets of $\mathcal{X}$ and must satisfy conditions (P1-P3). The (probability) distribution of X is described by the function P_X, hence by the collection of all the probabilities $P_X(C) = P(X \in C)$ for all the sets C so that $\{s : X(s) \in C\}$ is an event .

Remark 1.3.2. * Following above definition, in fact, $\{s : X(s) \in C\}$ is an event (thus measurable) for almost all kinds of subsets $C \subset \mathbb{R}^d$ that most readers will be able to imagine. To avoid technicalities, we shall assume $\{s : X(s) \in C\}$ is an event for any subset $C \subset \mathbb{R}^d$. We shall not concern with measurability thereafter.

Theorem 1.3.1. For real-valued X (that is, $\mathcal{X} \subset \mathbb{R}$), the distribution of X is fully described by the (cumulative) distribution function (c.d.f.)

$$F(x) := P_X((-\infty, x]) = P(X \leq x), \qquad -\infty < x < \infty.$$

The c.d.f. F has the following properties:

(1) $\lim_{x\to\infty} F(x) = 1, \lim_{x\to-\infty} F(x) = 0$.

(2) If $x_1 < x_2$, then $F(x_1) \leq F(x_2)$ (monotonic function).

(3) $\lim_{h\to 0^+} F(x + h) = F(x)$ (right continuous function).

(4) $\lim_{h\to 0^+} F(x - h) := F(x-) = F(x) - P(X = x) = P(X < x)$.

Theorem 1.3.2. If two random variables X and Y have the same c.d.f., then X and Y have the same probability distribution.

Theorem 1.3.3. * The c.d.f. F of a random variable X can have at most countably many discontinuity points.

Definition 1.3.3. (**Discrete Random Variable**). Let X be a random variable with c.d.f. F. It is a called discrete random variable if X can take only a finite number of different values $x_1, \dots, x_k$, or, at most countably many different values $x_1, x_2, \dots$. Equivalently, its c.d.f. F has at most countably many jumps (thus a step function). Indeed,

$$F(x) = \sum_{j:x_j \leq x} P(X = x_j)$$

where $P(X = x_j)$ is the size of the jump of F at x_j. We let $f(x_j) = P(X = x_j)$. The function f is called the **probability mass function** (p.m.f.). The set $\mathcal{X}_0 := \{x : f(x) > 0\}$ is called the **support** of X.

Definition 1.3.4. (**Continuous Random Variable**). Let X be a random variable with c.d.f. F. It is called a continuous random variable if there exists a nonnegative function $f : \mathbb{R} \mapsto \mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt, \qquad -\infty < x < \infty.$$

This function f is called **probability density function** (p.d.f.). If $f(x)$ is continuous at x, then $f(x) = F'(x)$. Since f is defined on the $\mathbb{R}$, it is customary to consider f to be $f(x)\mathbb{1}\{x \in \mathcal{X}_0\}$, i.e., the non-trivial part of p.d.f f is defined only on the support $\mathcal{X}_0 = \{x : f(x) > 0\}$.

Remark 1.3.3.

(1) For a discrete random variable X, the p.m.f. f must satisfy the following two requirements: $f(x) \geq 0$ for all x; and if the sequence $x_1, x_2, \ldots$ include all the possible values of X, then $\sum_{i=1}^{\infty} f(x_i) = 1$.

(2) For a continuous random variable X, the p.d.f. f must satisfy the following two requirements: $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)dx = 1$.

(3) For a continuous random variable X, $\mathrm{P}_X(x) = \mathrm{P}(X = x) = 0$ at any individual value x. Therefore, $\mathrm{P}(a < X < b) = \mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X < b)$. In particular, $F(x)$ is continuous.

(4) * The density function of a continuous random variable is not uniquely defined. If $f_1(x) = f_2(x)$ for almost all x and $F_1(x) = \int_{-\infty}^{x} f_1(t)dt$ and $F_2(x) = \int_{-\infty}^{x} f_2(t)dt$, then $F_1(x) = F_2(x)$. In this case, f_1 and f_2 are both density functions for X. In general, one can adopt the density function that is continuous if such one exists.

(5) * There exists some c.d.f. F that is continuous but cannot be defined as $F(x) = \int_{-\infty}^{x} f(t)dt$, i.e., F does not come from any density function. Therefore, the associated random variable X is neither discrete nor continuous. We shall not consider this type of random variable.

(6) * There exists some random variable that has a mixed distribution of discrete and continuous distributions. For example, $X = c\mathbb{1}(X^* \leq c) + X^*\mathbb{1}(X^* > c)$ for some continuous random variable X^* and some constant c.

Theorem 1.3.4. For a discrete random variable X with p.m.f. f, the probability of each subset C of the real line can be determined through

$$\mathrm{P}(X \in C) = \sum_{x \in C} f(x).$$

Theorem 1.3.5. For a continuous random variable X with p.d.f. f, the probability of each bounded closed interval $[a, b]$ can be determined through

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

More generally, for a subset $C \subset \mathbb{R}$,

$$\mathrm{P}(X \in C) = \int_C f(x)dx.$$

Definition 1.3.5. (**Empirical CDF**) . Suppose a random variable X is repeated observed, whose values are $x_1, \ldots, x_n$ (a *random sample*), the empirical cumulative distribution function (ECDF) for the sample is

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{x_i \leq x\right\}, \text{ for } x \in \mathbb{R}$$

In other words, the ECDF is the proportion of the sample that is less than or equal to x.

The corresponding p.d.f for ECDF is given heuristically as

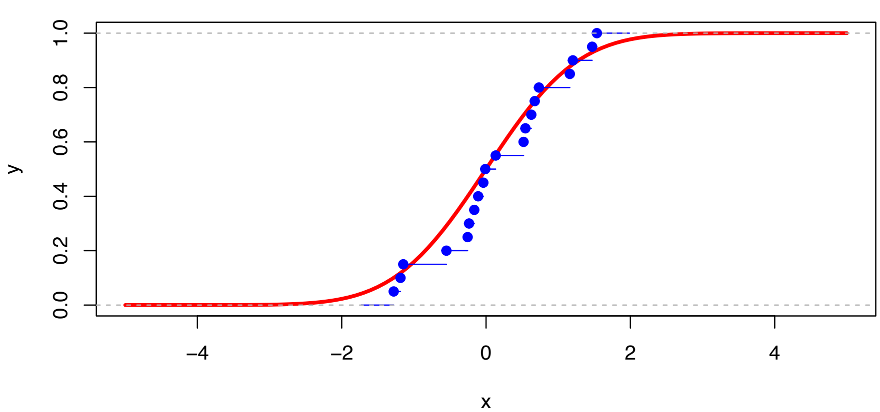
$$\begin{aligned}\hat{f}(x) &= \lim_{h\to 0}\frac{\hat{F}(x+h)-\hat{F}(x-h)}{2h}\\ &= \frac{1}{n}\lim_{h\to 0}\frac{\sum_{i=1}^{n}\mathbb{1}\left\{x-h < x_i \leq x+h\right\}}{2h}.\end{aligned}$$

Remark 1.3.4. The empirical CDF turns out to be a good approximation to the actual CDF of the random variable X, as sample size n grows:

$$\hat{F}(x) \to F(x), \quad \text{for all } x \in \mathbb{R}.$$

Precise probabilistic meaning of this approximation will be dealt with later.

The following is an example where red curve is the CDF of $N(0,1)$, and blue curve is the ECDF from a random sample of $N(0,1)$ of size 20.



1.4 Expectation and Moments

Definition 1.4.1. (**Expectation or Mean**). Let X be a random variable with c.d.f. F. The expectation of X denoted by $\mathrm{E}(X)$ is given by

$$\mu_X := \mathrm{E}(X) = \int_{-\infty}^{\infty} x dF(x),$$

provided the value exists (possibly $\pm\infty$).

(1)Here, $dF(x)$ is heuristically interpreted as $f(x)dx$ if X is continuous and has p.d.f. f. In this case $\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$.

(2) If X is discrete and has p.m.f. f, then $\mathrm{E}(X) = \sum_{x\in\mathcal{X}} xf(x)$.

(3) In both cases, the expectation exists and is finite if $E(|X|) < \infty$.

Remark 1.4.1. Interpretation of expectation $\mathrm{E}(X)$. (1) a weighted average of the possible values of X, i.e., the center of the distribution; (2) a "long run" average (e.g. betting a coin); (3) population mean (e.g. expected value of U.S. household income).

If X can be repeatedly observed and the observed values are $x_1, \dots, x_n$ (with a large sample size), then one might expect that using the ECDF $\hat{F}$ in replace of the true F would return a good approximation to $E(X)$: indeed,

$$\int x d_{\hat{F}(x)} = \int x\hat{f}(x)dx \approx \frac{\sum_{i=1}^{n} x_i}{n} := \bar{x}$$

giving us the well-known heuristic estimation by sample mean

$$E(X) \approx \bar{x}.$$

Definition 1.4.2. (**Expectation of a Function of Random Variables**). Let X be a random variable, possibly a random vector, with p.m.f. or p.d.f f. Let g denote a real-valued function defined on $\mathcal{X}$. Then

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

provided the integral or the sum exists. Special case: $g(X) = X$.

Remark 1.4.2. Let X be a random variable, possibly a random vector and a, b, and c be constants. Then for any real-valued function g_1 and g_2 whose expectations exists,

(1) $\mathrm{E}(ag_1(X) + bg_2(X) + c) = a\mathrm{E}g_1(X) + b\mathrm{E}g(X_2) + c$

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

Definition 1.4.3. (**Moments**). For each random variable X and every positive integer k, the expectation $\mathrm{E}(X^k)$ is called the k**th moment** of X. It is said that the kth moment *exists* if and only if $\mathrm{E}(|X|^k) < \infty$ (note: $\mathrm{E}(|X|^k)$ possibly equals to ∞). The expectation $\mathrm{E}[(X-\mu_X)^k]$ is called the k**th central moment**.

If the distribution of X is symmetric about zero, then all finite odd moments are zero (if the moments exist).

Definition 1.4.4. (**Variance and Standard Deviation**). The variance of a random variable X is given by

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

The positive square root of $\mathrm{Var}(X)$ is the **standard deviation** of X. We denote the standard deviation by σ_X, hence $\mathrm{Var}(X) = \sigma_X^2$.

If X can be repeatedly observed and the observed values are $x_1, \ldots, x_n$ (with a large sample size), then

$$\mathrm{Var}(X) \approx \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \quad \sigma_X \approx \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Remark 1.4.3. (1) The variance $\mathrm{Var}(X)$ measures the spread of the distribution of X. (2) The variance of any constant is 0. (3) The variance $\mathrm{Var}(X)$ is measured in the squared units of X; while the standard deviation σ_X is measured in the same unit as X.

Definition 1.4.5. (**Covariance and Correlation**). Let X and Y be random variables having finite means and finite variances σ_X^2 and σ_Y^2. The covariance of X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

The correlation of X and Y, denoted by $\rho(X, Y)$, is defined as

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

If the pair of (X,Y) can be repeatedly observed and the observed values are $(x_1, y_1), \ldots, (x_n, y_n)$ (with a large sample size), then

$$\mathrm{Cov}(X, Y) \approx \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Remark 1.4.4. (1) Covariance and correlation measures only *linear* relationship.

(2) If $Y = aX + b$ for some $a \neq 0$, then $|\rho(X, Y)| = 1$.

(3) If two random variables X and Y are uncorrelated if and only if $E(XY) = E(X)E(Y)$.

(4) (**independence**) Let X and Y be two random variables. We say X and Y are *independent*, if their joint p.d.f or p.m.f factors into products of individual p.d.f or p.m.f. If two random variables X and Y are independent, then X and Y are uncorrelated. The converse is not true (i.e., being uncorrelated does not imply independence). See Theorem 1.5.4 for more.

Remark 1.4.5. (**Properties of Covariance**). Let X, Y be two random variables. Then

(1) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

(2) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

(3) $\mathrm{Cov}(a+bX, c+dY) = bd\mathrm{Cov}(X,Y)$, for any constants a, b, c, d.

(4) $\mathrm{Cov}(aX+bY, cZ) = ac\mathrm{Cov}(X,Z) + bc\mathrm{Cov}(Y,Z)$, for any constants a, b, c.

Theorem 1.4.1. If $X_1, \ldots, X_d$ are *independent* or *(mutually) uncorrelated* random variables with finite variances and $c_1, \ldots, c_d$ are constants, then

$$\mathrm{Var}(a_1X_1 + \cdots + a_dX_d) = a_1^2\mathrm{Var}(X_1) + \cdots + a_d^2\mathrm{Var}(X_d).$$

Theorem 1.4.2. If $X_1, \ldots, X_d$ are random variables with finite variances, then

$$\mathrm{Var}\Big(\sum_{i=1}^{d} X_i\Big) = \sum_{i=1}^{d} \mathrm{Var}(X_i) + 2\sum_{i<j}\sum \mathrm{Cov}(X_i, X_j)$$

Suppose in addition, $X_1, \ldots, X_d$ are (mutually) uncorrelated, then

$$\mathrm{Var}\Big(\sum_{i=1}^{d} X_i\Big) = \sum_{i=1}^{d} \mathrm{Var}(X_i).$$

An important case when you have two random variables X and Y:

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

If in addition, X and Y are uncorrelated, then

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

Definition 1.4.6. (**Moment Generating Functions**). Let X be a random variable. The moment generating function (m.g.f.) of X, denoted by $M_X(t)$, is

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

provided that the expectation exists and is finite for $|t| < \delta$ for some number $\delta > 0$.

Theorem 1.4.3. Let X be a random variable having a m.g.f. $M_X(t)$ for $|t| < \delta$ for some $\delta > 0$. Then for each integer $n > 0$, $\mathrm{E}(X^n)$ exists and is finite, and

$$M_X(t) = \sum_{n=0}^{\infty} t^n \mathrm{E}(X^n)/n!, \quad |t| < \delta.$$

Also, the nth moment of X is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$:

$$\frac{d^n}{dt^n} M_X(t)\Big|_{t=0} = \mathrm{E}(X^n). \tag{5}$$

Theorem 1.4.4. Suppose X has m.g.f. M_X. For any constants a and b, the m.g.f. of the random variable $aX + b$ is given by

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

Theorem 1.4.5. Let X and Y be two random variable whose m.g.f. exist. X and Y have the same distribution if and only if there exists some $\delta > 0$ such that $M_X(t) = M_Y(t)$ for $|t| < \delta$.

Remark 1.4.6. A moment generating function uniquely determines the probability distribution of a random variable.

Previous discussion focuses on random variables. We turn to **random vectors** now. For a d-dimensional random vector $X = (X_1, \ldots, X_d)^T$, the expectation of X can be similarly defined as

$$\mathrm{E}(X) = [E(X_1), \ldots, E(X_d)]^T.$$

Remark 1.4.7. (**Properties of Expectation of Random Vectors**). Let X and Y be a d_X-dimensional and d_Y dimensional random vectors respectively, A be a $m \times d_X$ matrix of constants, B be a $m \times d_Y$ matrix of constants and b be a m-dimensional constant vector. Then

(1) $\mathrm{E}(X) = (\mathrm{E}(X^T))^T$

(2) $\mathrm{E}(AX + BY + c) = A\mathrm{E}(X) + B\mathrm{E}(Y) + c$

If X is a $m \times n$ random matrix, then we can similarly define its expectation as

$$\mathrm{E}(X) = \begin{bmatrix} \mathrm{E}(X_{1,1}) & \ldots & \mathrm{E}(X_{1,n}) \\ \vdots & \ddots & \vdots \\ \mathrm{E}(X_{m,1}) & \ldots & \mathrm{E}(X_{m,n}) \end{bmatrix}.$$

Remark 1.4.8. (**Properties of Expectation of Random Matrix**). Suppose A, B, C and D are non-random matrices and X and Y are random matrices. Assume that all matrix dimensions are compatible. Then

$$\mathrm{E}(AXB + CYD) = A\mathrm{E}(X)B + C\mathrm{E}(Y)D.$$

Definition 1.4.7. (**Covariance between Random Vectors**). Let $X = (X_1, \ldots, X_m)^T, Y = (Y_1, \ldots, Y_n)^T$ be two random vectors. The covariance of X and Y is given by

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)^T] = \mathrm{E}(XY^T) - \mathrm{E}(X)\mathrm{E}(Y)^T.$$

It can also be written as

$$\text{Cov}(X,Y) = \begin{bmatrix} \text{Cov}(X_1,Y_1) & \text{Cov}(X_1,Y_2) & \dots & \text{Cov}(X_1,Y_n) \\ \text{Cov}(X_2,Y_1) & \text{Cov}(X_2,Y_2) & \dots & \text{Cov}(X_2,Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_m,Y_1) & \text{Cov}(X_m,Y_2) & \dots & \text{Cov}(X_m,Y_n) \end{bmatrix}.$$

Remark 1.4.9. (**Properties of Covariance between Random Vectors**). Let X, Y and Z be three random vectors of dimensions d_X, d_Y, d_Z respectively. Let A_1, A_2 and A_3 be matrices of constants of dimensions $m \times d_X$,$m \times d_Y$ and $n \times d_Z$ respectively. Let c_1 and c_2 be vectors of constants of dimensions m and n respectively. Then covariance operation satisfies

$$\text{Cov}(A_1X + A_2Y + c_1, A_3Z + c_2) = A_1\text{Cov}(X,Z)A_3^T + A_2\text{Cov}(Y,Z)A_3^T$$
$$\text{Cov}(X,Y) = [\text{Cov}(Y,X)]^T$$

In particular:

(1) $\text{Cov}(A_1X, A_3Z) = A_1\text{Cov}(X,Z)A_3^T$.

(2) $\text{Cov}(X + c, Z) = \text{Cov}(X,Z)$, for any constant vector c.

(3) $\text{Cov}(X + Y, Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$.

The covariance matrix of a random vector is the outcome of a special case of covariance operation defined in the following.

Definition 1.4.8. (**Covariance Matrices of Random Vectors**). Let $X = (X_1, \dots, X_d)^T$ be a d-dimensional random vector. The covariance matrix of X is given by

$$\text{Cov}(X,X) = \text{E}[(X - \text{E}X)(X - \text{E}X)^T] = \text{E}(XX^T) - \text{E}(X)\text{E}(X)^T.$$

The covariance matrix of X can also be written as

$$\text{Var}(X) := \text{Cov}(X,X) = \begin{bmatrix} \text{Cov}(X_1,X_1) & \text{Cov}(X_1,X_2) & \dots & \text{Cov}(X_1,X_d) \\ \text{Cov}(X_2,X_1) & \text{Cov}(X_2,X_2) & \dots & \text{Cov}(X_2,X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_d,X_1) & \text{Cov}(X_d,X_2) & \dots & \text{Cov}(X_d,X_d) \end{bmatrix}.$$

Var(X) is commonly written as Σ_X.

Remark 1.4.10. (**Properties of Covariance Matrix**). Any $d \times d$ covariance matrix Σ satisfies

(1) Symmetric: $\Sigma = \Sigma^T$.

(2) Positive semidefinite: for any $d \times 1$ vector $a \neq 0$, $a^T\Sigma a \geq 0$.

Remark 1.4.11. (**Properties of Covariance Matrix**). Let X and Y be two dimensional random vectors of dimensions d_X and d_Y. Let A, B be matrices of dimensions $d \times d_X$ and $d \times d_Y$ respectively. Let c be a d dimensional vector of constants.

(1) $\mathrm{Var}(AX + c) = A\mathrm{Var}(X)A^T$.

(2) $\mathrm{Var}(AX + BY) = A\mathrm{Var}(X)A^T + B\mathrm{Var}(Y)B^T + 2A\mathrm{Cov}(X, Y)B^T$.

Definition 1.4.9. (**Moment Generating Functions for Random Vectors**). Let X be a d-dimensional random vector. The moment generating functions of X, denoted by $M_X(t)$, is

$$M_X(t) := \mathrm{E}(e^{t^T X}), \quad t \in \mathbb{R}^d$$

provided that the expectation exists and is finite for all t such that $||t|| < \delta$ for some number $\delta > 0$.

1.5 Joint Distributions and Marginal Distributions

Definition 1.5.1. (**Distribution for Random Vectors**). Let $X : S \mapsto \mathcal{X}$, where $\mathcal{X}$ is a subset of $\mathbb{R}^d$. The c.d.f. is define as the function $F : \mathbb{R}^d \mapsto [0, 1]$ given by

$$F(x) := F(x_1, \ldots, x_d) = \mathrm{P}(X \in (-\infty, x_1] \times (-\infty, x_2] \times \cdots (-\infty, x_d])$$

If X is written in terms of component random variables, i.e., $X = (X_1, \ldots, X_d)$, then

$$F(x) = \mathrm{P}(X_1 \le x_1, \ldots, X_d \le x_d).$$

The c.d.f. F is also called the *joint* c.d.f. of the random variables $X_1, \ldots, X_d$.

If the distribution function F of X can be written as

$$F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(t_1, \ldots, t_d) dt_1 \cdots d_{t_d} \tag{6}$$

for some function f on $\mathbb{R}^d$, then X is said to have a continuous distribution with density f. In particular,

$$f(x_1, \ldots, x_d) = \frac{\partial^d F}{\partial x_1 \ldots \partial x_d}.$$

Remark 1.5.1. For $X = (X_1, \ldots, X_d)$, whose range space $\mathcal{X}$ is a subset of $\mathbb{R}^d$. If the range $\mathcal{X}$ is countable set, then X is said to have a discrete distribution with p.m.f. given by $f(x) := f(x_1, \ldots, x_d) = \mathrm{P}(X_1 = x_2, \ldots, X_d = x_d)$. In this case, the distribution function F of X can be similarly written as Eqn.(6) , but with integration replaced by summations.

Definition 1.5.2. (**Marginal Distribution**). Consider a collection of random variables $X_1, \ldots, X_d$, where $X_i : S \mapsto \mathcal{X}_i \subset \mathbb{R}$ for $i = 1, \ldots, d$. For each $i = 1, \ldots, d$, the marginal distribution of X_i is given by

$$\mathrm{P}(X_i \in C_i) = \mathrm{P}(X_i \in C_i, \text{ and } X_j \in \mathcal{X}_j \quad \forall j \neq i), \quad C_i \subset \mathcal{X}_i$$

Definition 1.5.3. (**Marginal c.d.f.**). Consider a collection of random variables $X_1, \ldots, X_d$, where $X_i : S \mapsto \mathcal{X}_i \subset \mathbb{R}$ for $i = 1, \ldots, d$. The marginal c.d.f of X_1 is given by

$$F_1(x_1) := \mathrm{P}(X_1 \leq x_1) = \mathrm{P}(X_1 \leq x_1, X_2 < \infty, X_3 < \infty, \ldots, X_d < \infty)$$

The marginal c.d.f. for X_i for $i = 2, \ldots, d$ are similarly defined.

Theorem 1.5.1. (**Marginal p.m.f. for Discrete Random Variables**). Consider a collection of discrete random variables $X_1, \ldots, X_d$ with joint p.m.f. $f(x_1, \ldots, x_d)$, where $X_i : S \mapsto \mathcal{X}_i \subset \mathbb{R}$ for $i = 1, \ldots, d$. The marginal p.m.f. of X_1 is given by

$$f_1(x_1) := \mathrm{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} \cdots \sum_{x_d \in \mathcal{X}_d} f(x_1, x_2, x_3, \ldots, x_d)$$

The marginal p.m.f. for X_i for $i = 2, \ldots, d$ are similarly defined.

Theorem 1.5.2. (**Marginal p.d.f. for Continuous Random Variables**). Consider a collection of continuous random variables $X_1, \ldots, X_d$ with joint p.d.f. $f(x_1, \ldots, x_d)$, where $X_i : S \mapsto \mathcal{X}_i \subset \mathbb{R}$ for $i = 1, \ldots, d$. The marginal p.d.f. of X_1 is given by

$$f_1(x_1) := \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, x_3, \ldots, x_d) dx_2 \cdots dx_d$$

The marginal p.d.f. for X_i for $i = 2, \ldots, d$ are similarly defined.

Definition 1.5.4. (**Independence of a Collection of Random Variables**). Consider a collection of random variables $X_1, \ldots, X_d$ where $X_i : S \mapsto \mathcal{X}_i \subset \mathbb{R}$ for $i = 1, \ldots, d$. We say $X_1, \ldots, X_d$ are independent if for *any* collection of sets $C_1, \ldots, C_d$, $C_i \subset \mathcal{X}_i, i = 1, \ldots, d$, the events $X_1 \in C_1, \ldots, X_d \in C_d$ are independent, that is,

$$\mathrm{P}(X_1 \in C_1, \ldots, X_d \in C_d) = \mathrm{P}(X_1 \in C_1) \cdots \mathrm{P}(X_d \in C_d)$$

Definition 1.5.5. (**Independence of Random Vectors**). Consider two random vectors $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$. We say that the two random vectors are independent if for any sets $C_1, \ldots, C_m, \tilde{C}_1, \ldots, \tilde{C}_n$, the events $\{X_1 \in C_1, \ldots, X_m \in C_m\}$ and $\{Y_1 \in \tilde{C}_1, \ldots, Y_n \in \tilde{C}_n\}$ are independent, that is,

$$\mathrm{P}(X_i \in C_i, Y_j \in \tilde{C}_j, i = 1, \ldots, m, j = 1, \ldots, n) = \mathrm{P}(X_i \in C_i, i = 1, \ldots, m)\mathrm{P}(Y_j \in \tilde{C}_j, j = 1, \ldots, n)$$

Theorem 1.5.3. (1) If the collection of random variables $X_1, \dots, X_d$ are independent, then for any real-valued functions $g_1, \dots, g_d$ where $g_j : \mathcal{X}_j \mapsto \mathbb{R}, j = 1, \dots, d$, the collection of random variables $g_1(X_1), \dots, g_d(X_d)$ are also independent.

(2) If the collection of random variables $X_1, \dots, X_d$ are independent, for some m such that $1 \leq m < d$, then for any real valued function $g_1 : \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \mapsto \mathbb{R}$ and $g_2 : \mathcal{X}_{m+1} \times \cdots \times \mathcal{X}_d \mapsto \mathbb{R}$, it holds that

$$g_1(X_1, \dots, X_m) \perp\!\!\!\perp g_2(X_{m+1}, \dots, X_d).$$

Theorem 1.5.4. (**Characterization of Independence**) Consider a collection of random variables $X_1, \dots, X_d$. For each $i = 1, \dots, d$, let $\mathcal{X}_i$ and F_i denote the range and marginal c.d.f. for X_i respectively .

(1) $X_1, \dots, X_d$ are independent if and only if $F(x_1, \dots, x_d) = F_1(x_1) \cdots F_d(x_d)$ for all $x_1, \dots, x_d$.

(2) $X_1, \dots, X_d$ are independent if and only if for *any* sequence of bounded real-valued functions $g_1, \dots, g_d$ where $g_j : \mathcal{X}_i \mapsto \mathbb{R}, j = 1, \dots, d$,

$$\mathrm{E}[g_1(X_1)g_2(X_2) \cdots g_d(X_d)] = \mathrm{E}[g_1(X_1)] \cdots \mathrm{E}[g_d(X_d)].$$

(3) Suppose $X_1, \dots, X_d$ has joint p.m.f. given by f. For $1 \leq i \leq d$, let f_i denote the marginal p.m.f. for X_i. Then $X_1, \dots, X_d$ are independent if and only if $f(x_1, \dots, x_d) = f_1(x_1) \cdots f_d(x_d)$ for all $x_1, \dots, x_d$.

(4) Suppose $X_1, \dots, X_d$ has joint p.d.f. given by f. For $1 \leq i \leq d$, let f_i denote the marginal p.d.f. for X_i. Then $X_1, \dots, X_d$ are independent if and only if $f(x_1, \dots, x_d) = f_1(x_1) \cdots f_d(x_d)$ for (almost) all $x_1, \dots, x_d$.

(5) Suppose $X_1, \dots, X_d$ has joint p.d.f. or p.m.f. given by f. Then $X_1, \dots, X_d$ are independent if and only if there exist functions $h_1, \dots, h_d$ such that for *every* $x_i \in \mathbb{R}, i = 1, \dots, d$,

$$f(x_1, \dots, x_d) = h_1(x_1) \cdots h_d(x_d).$$

Theorem 1.5.5. Suppose $X_1, \dots, X_d$ are independent random variables. For each $i = 1, \dots, d$, let M_{X_i} denote the m.g.f. of X_i. Let $Y = \sum_{i=1}^d X_i$. Then for every t such that $M_{X_i}(t)$ is finite for $i = 1, \dots, d$,

$$M_Y(t) = \prod_{i=1}^d M_{X_i}(t).$$

1.6 Conditional Distributions

Definition 1.6.1. (**Conditional Distribution/p.m.f.**). Let X and Y have a **discrete** joint distribution p.m.f. f. Let f_Y denotes the marginal p.m.f. of Y. For each y such that $f_Y(y) > 0$, define

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

Then $f_{X|Y}$ is called the *conditional p.m.f.* of X given Y. For values of y such that $f_Y(y) = 0$, we are free to define $f_{X|Y}(x|y)$ so long as $f_{X|Y}(x|y)$ is a p.m.f. of X. Just like conditional probabilities of events, $f(\cdot|y)$ is a valid p.m.f. for a fixed y.

The discrete distribution whose p.m.f. is $f_{X|Y}(\cdot|y)$ is called the conditional distribution of X given $Y = y$. That is, the conditional distribution of X given $Y = y$ is given by

$$\mathrm{P}(X \in C|Y = y) = \sum_{x \in C} f_{X|Y}(x|y).$$

The conditional c.d.f. of a random variable X given $Y = y$ is given by

$$F_{X|Y}(t|y) = \sum_{x \in (-\infty, t]} f_{X|Y}(x|y).$$

Definition 1.6.2. (**Conditional Distribution/p.d.f.**). Let X and Y have a joint **continuous** distribution p.d.f. f. Let f_Y denotes the marginal p.d.f. of Y. For each y such that $f_Y(y) > 0$, define

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

Then $f_{X|Y}$ is called the conditional p.d.f. of X given Y. For values of y such that $f_Y(y) = 0$, we are free to define $f_{X|Y}(x|y)$ so long as $f_{X|Y}(x|y)$ is a p.d.f. of X. Just like conditional probabilities of events, $f(\cdot|y)$ is a valid p.d.f. for a fixed y.

The continuous distribution whose p.d.f. is $f_{X|Y}(\cdot|y)$ is called the conditional distribution of X given $Y = y$. That is, the conditional distribution of X given $Y = y$ is given by

$$\mathrm{P}(X \in C|Y = y) = \int_C f_{X|Y}(x|y)dx.$$

The conditional c.d.f. of a random variable X given $Y = y$ is given by

$$F_{X|Y}(t|y) = \int_{-\infty}^{t} f_{X|Y}(x|y)dx.$$

Remark 1.6.1. When the context is clear, we shall use a shorthand notation $f(x|y)$ for the conditional p.m.f. or p.d.f. $f_{X|Y}(x|y)$, and use $F(x|y)$ for the conditional c.d.f. $F_{X|Y}(x|y)$.

Theorem 1.6.1. (**Independence through Conditional Distributions**). Suppose that X and Y are two random variables having a joint p.m.f. or p.d.f. f, and, marginal p.m.f or p.d.f. f_X and f_Y respectively. Then X and Y are independent if and only if for *every* value of y such that $f_Y(y) > 0$ and *every* value of x,

$$f_{X|Y}(x|y) = f_X(x).$$

A similar statement holds with the roles of X and Y switched.

Definition 1.6.3. (**Conditionally independent Random Variables**). Let Y be a random vector with joint p.d.f. or p.m.f. f_Y. Several random variables $X_1, \ldots, X_d$ are **conditionally independent given** Y, if for all y such that $f_Y(y) > 0$, we have

$$f_{X|Y}(x|y) = f_{X|Y}(x_1, \ldots, x_d|y) = \prod_{i=1}^{d} f_i(x_i|y)$$

where $f(x|y)$ is the conditional (multivariate) p.d.f. or p.m.f. of X given $Y = y$ and $f_i(x_i|y)$ the conditional (univariate) p.d.f. or p.m.f. of X_i given $Y = y$

Theorem 1.6.2. (**Bayes' Theorem for Random Variables**). If $f_Y(y)$ is the marginal p.m.f. or p.d.f. of a random variable Y and $f(x|y)$ is the conditional p.m.f. or p.d.f. of X given $Y = y$, then

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)}.$$

Definition 1.6.4. (**Conditional Expectation**). Let (X, Y) be random variables with range $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R} \times \mathbb{R}$ and let $F(\cdot|y)$ denote the conditional c.d.f. of X given $Y = y$. Let $g : \mathcal{X} \mapsto \mathbb{R}$ such that $\mathrm{E}(|g(X)|) < \infty$. The conditional expectation of X given by $Y = y$ is given by

$$\mathrm{E}(g(X)|Y = y) = \int_{-\infty}^{\infty} g(x) dF(x|y),$$

provided the value exists (possibly $\pm\infty$).

(1)Here, $dF(x|y)$ is heuristically interpreted as $f(x|y)dx$ if X has a continuous conditional distribution given $Y = y$ with conditional p.d.f. $f(x|y)$. In this case $\mathrm{E}(g(X)|Y = y) = \int_{-\infty}^{\infty} g(x) f(x|y) dx$.

(2) If X has a discrete conditional distribution given $Y = y$ with conditional p.m.f. $f(x|y)$, then $\mathrm{E}(g(X)|Y = y) = \sum_{x \in \mathcal{X}} g(x) f(x|y)$.

(3) Note that in both cases, $\mathrm{E}(g(X)|Y = y)$ are functions of y. These functions may be computed before Y is observed. Therefore, it is legitimate to consider conditional expectation $\mathrm{E}(g(X)|Y)$ as a random variable (a function of Y) whose value becomes $\mathrm{E}(g(X)|Y = y)$ when $Y = y$ is observed.

Remark 1.6.2. Let X, Y be two independent random variables. Then $E(X|Y) = E(X)$.

Definition 1.6.5. (**Conditional Variance**). * Let X, Y be two random variables. The variance of the conditional distribution of X given $Y = y$ is given by

$$\mathrm{Var}(X|Y = y) = E\{[X - E(X|Y = y)]^2 | Y = y\} = E[X^2|Y = y] - [E(X|Y = y)]^2.$$

In short, $\mathrm{Var}(X|Y) = E[X^2|Y] - (E(X|Y))^2$.

Theorem 1.6.3. * Let X, Y be two random variables.

(1) (Law of iterated expectation) $\mathrm{E}[\mathrm{E}(g(X)|Y)] = \mathrm{E}[g(X)]$ for any $g : \mathcal{X} \mapsto \mathbb{R}$ such that $\mathrm{E}(|g(X)|) < \infty$.

(2) For any function r of X and Y, $\mathrm{E}(r(X,Y)|X = x) = \mathrm{E}(r(x,Y)|X = x)$.

(3) For any functions g_1 of X and g_2 of Y, $\mathrm{E}(g_1(X)g_2(Y)|X) = g_1(X)\mathrm{E}(g_2(Y)|X)$.

(4) (Law of total probability for variance) $\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathrm{E}(X|Y)]$.

1.7 Functions of Random Variables

Definition 1.7.1. (**Monotonic Functions**). Consider a real-valued function f. If for $x_1 < x_2$, $f(x_1) < f(x_2)$. Then we call f an **increasing function**. If for $x_1 < x_2$, $f(x_1) \leq f(x_2)$. Then we call f a **weakly increasing function**. The decreasing and weakly decreasing functions are similarly defined.

Given a random variable $X : S \mapsto \mathcal{X}$ with known distribution, consider another random variable $Y = g(X)$ for some known function g. Let $\mathcal{Y} = g(\mathcal{X})$ be the range of Y. The probability distribution of Y can be computed as, for any set $C \subset \mathcal{Y}$,

$$\mathrm{P}(Y \in C) = \mathrm{P}(g(X) \in C) = \mathrm{P}(\{x \in \mathcal{X} : g(x) \in C\})$$

If X is discrete random variable (thus Y is also discrete), then the p.m.f. of Y is

$$f_Y(y) = \mathrm{P}(Y = y) = \sum_{x \in \mathcal{X}: g(x) = y} f_X(x).$$

If X is continuous random variable with p.d.f. f_X (thus Y is also continuous), then the c.d.f. of Y is

$$F_Y(y) = \mathrm{P}(Y \leq y) = \int_{\{x \in \mathcal{X}: g(x) \leq y\}} f_X(x)dx.$$

Suppose in addition, that g is one-to-one (i.e. injective) function. We have the following results.

Theorem 1.7.1. Suppose X is a random variable with p.m.f. or p.d.f. f_X. Consider $Y = g(X)$ for $g : \mathbb{R} \mapsto \mathbb{R}$. Let $\mathcal{X}_0 := \{x : f_X(x) > 0\}$ and $\mathcal{Y}_0 := g(\mathcal{X}_0)$. Suppose g is a *one-to-one* and continuously differentiable function on $\mathcal{X}_0$. Let g^{-1} denotes the inverse of g.

(1) If g is an increasing function on $\mathcal{X}_0$, then $F_Y(y) = F_X(g^{-1}(y))$ for all $y \in \mathcal{Y}_0$.

(2) If g is a decreasing function on $\mathcal{X}_0$, then $F_Y(y) = 1 - F_X(g^{-1}(y)-)$ for all $y \in \mathcal{Y}_0$.

(3) If X is discrete, then Y has p.m.f. given by $f_Y(y) = f_X(g^{-1}(y))$ for all $y \in \mathcal{Y}_0$.

(4) If X is continuous and in addition $|g'(x)| > 0$ on $\mathcal{X}_0$, then Y has p.d.f. given by

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| \text{ for all } y \in \mathcal{Y}_0.$$

In (4), note also that by the Inverse Function Theorem,

$$\frac{d}{dy}g^{-1}(y) = \Big(\frac{d}{dx}g(x)\Big|_{x=g^{-1}(y)}\Big)^{-1}.$$

We also have a *multivariate* version of above theorem. To state the theorem, suppose $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ differentiable, where g takes the form of a vector of d functions:

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_d(x) \end{bmatrix}.$$

Recall the **Jacobian matrix** of g is the $d \times d$ matrix with (i, j)th element given by $\partial g_i/\partial x_j$, where g_i denotes the ith component of the function g. This matrix is denoted by $\partial g/\partial x$:

$$\frac{\partial g}{\partial x} = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_d} \\ \vdots & & \vdots \\ \frac{\partial g_d}{\partial x_1} & \cdots & \frac{\partial g_d}{\partial x_d} \end{bmatrix}.$$

The **Jacobian** of g at some point x is defined to be $|\det[\partial g/\partial x]|$, i.e., the absolute value of the determinant of the Jacobian matrix evaluated at x.

Theorem 1.7.2. Suppose X is a d-dimensional random vector with p.m.f. or p.d.f. f_X. Consider $Y = g(X)$ for $g : \mathbb{R}^d \mapsto \mathbb{R}^d$. Let $\mathcal{X}_0 := \{x : f_X(x) > 0\}$ and $\mathcal{Y}_0 := g(\mathcal{X}_0)$. Suppose g is a one-to-one and continuously differentiable function on $\mathcal{X}_0$.

(1) If X is discrete, then Y has p.m.f. given by $f_Y(y) = f_X(g^{-1}(y))$ for all $y \in \mathcal{Y}_0$.

(2) If X is continuous and in addition $|\det[\partial g/\partial x]|$ is nonzero on $\mathcal{X}_0$, then Y has p.d.f. given by

$$f_Y(y) = f_X(g^{-1}(y))\Big|\det\Big[\frac{\partial g^{-1}(y)}{\partial y}\Big]\Big| \text{ for all } y \in \mathcal{Y}_0.$$

In (2),

$$\frac{\partial g^{-1}(y)}{\partial y} = \Big(\frac{\partial g(x)}{\partial x}\Big)^{-1}\Big|_{x=g^{-1}(y)}.$$

Note also

$$\frac{\partial g^{-1}(y)}{\partial y}\Big|_{y=g(x)} = \Big(\frac{\partial g(x)}{\partial x}\Big)^{-1}.$$

If the function g is *not one-to-one*, we cannot apply above theorem. However, if the set $\mathcal{X}_0 = \{x : f_X(x) > 0\}$ can be partitioned into subsets such that g is one-to-one on each subset. We can use the following theorem.

Theorem 1.7.3. * Suppose X is a d-dimensional continuous random vector with p.d.f. f_X. Consider $Y = g(X)$ for $g : \mathbb{R}^d \mapsto \mathbb{R}^d$. Let $\mathcal{X}_0 := \{x : f_X(x) > 0\}$. Let $\mathcal{X}_1, \ldots, \mathcal{X}_m$ denote the disjoint open subsets of $\mathcal{X}_0$ such that $\mathrm{P}(X \in \cup_{i=1}^m \mathcal{X}_i) = 1$. Let $\mathcal{Y}_i := g(\mathcal{X}_i)$. Suppose g

is a function on $\mathcal{X}_0$ and let $g^{(i)}$ denotes the restriction of g to $\mathcal{X}_i, i = 1, .2, \ldots, m$. Assume that for each $i = 1, \ldots, m$, $g^{(i)}$ is one-to-one and continuously differentiable with inverse $h^{(i)}$ and the Jacobian of $g^{(i)}$ is nonzero on $\mathcal{X}_i$. Then Y is a continuous random vector with p.d.f. given by

$$f_Y(y) = \sum_{i=1}^{m} f_X(h^{(i)}(y)) \left| \det \left[\frac{\partial h^{(i)}(y)}{\partial y} \right] \right| \mathbb{1}\{y \in \mathcal{Y}_i\}, \text{ for } y \in g(\cup_{i=1}^{m} \mathcal{X}_i).$$

Remark 1.7.1. * Consider a case where X is a d-dimensional random vector and $Y = g_0(X)$ for some $g_0 : \mathbb{R}^d \mapsto \mathbb{R}^q$ where $q < d$. Above theorems cannot be applied directly. One possible solution is to construct a function g_1 such that $g = (g_0, g_1)$ satisfies the conditions of above theorems, i.e., $g : \mathbb{R}^d \mapsto \mathbb{R}^d$. We then can find the density of $g_0(X)$ by marginalizingthe out density of $g_1(X)$.

Theorem 1.7.4. (**Linear transformation**). Suppose X is a d-dimensional continuous random vector with p.d.f. f_X. Consider $Y = AX + \mu$, where A is $d \times d$ nonsingular matrix and μ is a d-dimensional vector. Let $\mathcal{Y}$ denote the range space of Y. Then Y is a continuous random vector with p.d.f. given by

$$f_Y(y) = \frac{1}{|\det A|} f_X(A^{-1}(y - \mu)) \text{ for } y \in \mathcal{Y}.$$

Remark 1.7.2. In the discussion above, in order to correctly apply Theorem 1.7.1, 1.7.2, 1.7.3 and 1.7.4, it is critical to keep track of the support of the original random variable (vector) X and the support of the transformed random variable (vector) Y.

1.8 Quantiles and Quantile Functions

Definition 1.8.1. (**Quantiles**). Let X be a random variable with c.d.f. F. For a given value $p \in (0, 1)$, the p **quantile** (or 100p **percentile**) of the distribution of X is defined to be

$$\inf\{x : F(x) \geq p\},$$

that is, the smallest value of x such that $F(x) \geq p$.

The **quantile function** of the distribution of X is the function $Q : (0, 1) \mapsto \mathbb{R}$ given by

$$Q(t) := \inf\{x : F(x) \geq t\}.$$

If the c.d.f. F is one-to-one (i.e. injective), then Q is simply F^{-1}, the inverse function of F. Throughout the note, even when a c.d.f. F is not one-to-one, we shall still use F^{-1} to denote the quantile function associated with F, that is,

$$F^{-1}(t) := \inf\{x : F(x) \geq t\}.$$

Remark 1.8.1. * (1) If F is one-to-one, i.e., injective, then $F^{-1} \circ F(x) = x$ and $F \circ F^{-1}(x) = x$; (2) For every $0 < p < 1$, and $x \in \mathbb{R}$, $F \circ F^{-1}(p) \geq p$ with equality if and only p is in the range of F; (3) $p \leq F(x)$ if and only if $F^{-1}(p) \leq x$;˙

Theorem 1.8.1. (**Probability Integral Transformation**). If X is a random variable with *continuous* c.d.f. F. Let $Y = F(X)$. The distribution of Y is the uniform distribution on $[0, 1]$.

Remark 1.8.2. The above theorem can be used to inspect if X follows some hypothesized distribution F_0 given a random sample $\{X_1, \ldots, X_n\}$. This is done by using $\{Y_1, \ldots, Y_n\}$ where $Y_i = F_0(X_i)$, and the empirical CDF of Y_i

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \{Y_i \leq y\}$$

to verify if $\hat{F}_Y(y)$ approximately follows a 45-degree line on $(0, 1)$ interval.

Corollary 1.8.1. (**Quantile Transformation**). If Y has the uniform distribution on $[0, 1]$, and let F be *any* c.d.f. with quantile function F^{-1}. Then $X = F^{-1}(Y)$ has c.d.f. F.

Remark 1.8.3. For all common continuous random variables, such as normal, t and F random variables, their c.d.f. F is continuous and one-to-one. Therefore, for every $0 < p < 1$, the p quantile of X is simply given by the value x_p such that $F(x_p) = p$, that is,

$$\mathrm{P}(X \leq x_p) = p.$$

In statistical inference, for some small value α assuming the role of $1 - p$, it is a convention to define the so-called **upper-α quantile** of X to be the value x_α such that

$$\mathrm{P}(X \geq x_\alpha) = \alpha.$$

Often α is chosen to be $0.01, 0.05$ or 0.1 in the problems of constructing confidence intervals and hypothesis testing.

For example, for normal distribution, the upper 0.5%, 2.5% and 5% quantiles are respectively $x_{0.5\%} = 2.58$, $x_{2.5\%} = 1.96$ and $x_{5\%} = 1.64$.

Normal probabilities and quantiles:

	$P(N(0,1) > x)$	$P(\lvert N(0,1)\rvert > x)$
$x = 0.00$	0.50	1.00
$x = 1.00$	0.16	0.32
$x = 1.64$	0.050	0.100
$x = 1.96$	0.025	0.050
$x = 2.00$	0.023	0.046
$x = 2.33$	0.010	0.020
$x = 2.58$	0.005	0.010

1.9 Some Common Discrete Distributions

1.9.1 Binomial Distributions

Definition 1.9.1. (**Bernoulli Trials and Binomial Distributions**). A sequence of trials, where

- each trial results in a "success" or a "failure",
- the trials are independent,
- the probability of "success," denoted by p, $0 < p < 1$, is the same on every trial.

Let X denote the number of successes out of n Bernoulli trials. Then X has a **Binomial distribution** with parameter n and p. We write $X \sim$Bin(n, p). Its p.m.f. is given by

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \cdots, n$$

The mean and variance are $\mathrm{E}(X) = np, \mathrm{Var}(X) = np(1-p)$. The m.g.f. is $M_X(t) = (1 - p + pe^t)^n$.

Example 1.9.1. Each of the following situations could be conceptualized as a binomial experiment. Does each of them satisfy with the Bernoulli assumptions?

- We flip a fair coin 10 times and let X denote the number of tails in 10 flips. Here, $X \sim Bin(n = 10, p = 0.5)$.
- In rural Kenya, the prevalence rate for HIV is estimated to be around 8 percent. Let X denote the number of HIV infecteds in a sample of 740 individuals. Here, $X \sim Bin(n = 740, p = 0.08)$.
- Parts produced by a certain company do not meet specifications (i.e., are defective) with probability 0.001. Let X denote the number of defective parts in a package of 40. Then, $X \sim Bin(n = 40, p = 0.001)$.

Example 1.9.2. At an automobile plant, 15 percent of all paint batches sent to the lab for chemical analysis do not conform to specifications. Assume the Bernoulli's trials assumptions hold the batches. Now suppose a shipment of 50 batches is examined. What is the probability at last 40 batches are conforming?

1.9.2 Geometric Distributions

Suppose that the Bernoulli trials are continually observed. Let X denotes the trial at which the first success occurs. Then X has a geometric distribution with parameter p. We write $X \sim$Gem(p). Its p.m.f. is given by

$$f_X(x) = p(1-p)^{x-1} \text{ for } x = 1, 2, \cdots.$$

The mean and variance are $\mathrm{E}(X) = \frac{1}{p}, \mathrm{Var}(X) = \frac{1-p}{p^2}$. The m.g.f. is $M_X(t) = \frac{pe^t}{1-(1-p)e^t}, t < -\log(1-p)$.

Note that $X \sim$Gem(p) has the memoryless property: for all nonnegative integers s and t, $\mathrm{P}(X \geq s+t | X \geq t) = \mathrm{P}(X \geq s)$. In words, given that X is at least t, the probability that X is at least $s+t$ is the same as if we were to look at X unconditionally being at least s.

Example 1.9.3. At an automobile plant, 15 percent of all paint batches sent to the lab for chemical analysis do not conform to specifications. Assume the Bernoulli's trials assumptions hold. Suppose a shipment is sent for analysis. What is the probability that the third batch is found to be the first non-conforming batch?

1.9.3 Negative Binomial Distributions*

Suppose that the Bernoulli trials are continually observed. Let X denotes the number of trials to observe the r-th success. Then X has a negative binomial distribution with parameter r and p. We write $X \sim$NB(r, p). Its p.m.f. is given by

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

Its mean and variance are given by $E(X) = \frac{r}{p}, \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$. The mgf $M_X(t) = (\frac{pe^t}{1-(1-p)e^t})^r, t < -\log(1-p)$. When $r = 1$, the NB$(1, p)$ reduces to the geometric distribution Gem(p).

Example 1.9.4. At an automobile plant, 15 percent of all paint batches sent to the lab for chemical analysis do not conform to specifications. Assume the Bernoulli's trials assumptions hold. Suppose a shipment is sent for analysis. What is the probability that no more than three non-conforming batches will be observed among the first 30 batches sent to the lab?

1.9.4 Poisson Distributions

Let the *number of occurrences* in a given continuous interval of time or space be counted. A *Poisson process* enjoys the following properties (as time flows):

1. The number of occurrences of certain event in non-overlapping intervals are independent random variables.

2. The probability of an occurrence in a sufficiently short interval is proportional to the length of the interval.

3. The probability of 2 or more occurrences in a sufficiently short interval is zero.

Suppose that a process satisfies the above three conditions, and let X denote the number of occurrences in an interval of length one. Our goal is to find an expression for $f_X(x) = P(X = x)$, the p.m.f. of X, the probability that x such events occur.

Envision partitioning the unit interval $[0, 1]$ into n subintervals, each of size $1/n$. Now, if n is sufficiently large (i.e., much larger than x), there can be at most one event occur in each of these subintervals. There we can approximate the probability that $X = x$ events (occurences) occur in this unit interval by finding the probability that exactly *one* event (occurrence) occurs in exactly x of these subintervals.

- By Property (2), we know that the probability of one event in any one subinterval is proportional to the subinterval's length, say λ/n, where $\lambda > 0$ is the proportionality constant.

- By Property (3), the probability of more than one occurrence in any subinterval is zero (for n large).

- Consider the occurrence/non-occurrence of an event in each subinterval as a Bernoulli trial. Then, by Property (1), we have a sequence of n Bernoulli trials, each with probability of "success" $p = \lambda/n$. Thus, a binomial (approximate) calculation gives

$$\mathrm{P}(X = x) \approx \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1 - \frac{\lambda}{n}\right)^{n-x}.$$

  To improve the approximation for $P(X = x)$, we let n get large without bound. One can show that $\lim_{n\to\infty} P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$.

We say that X follows a Poisson distribution with parameter λ, with p.m.f.

$$f(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, \ldots; 0 \le \lambda < \infty$$

The mean and variance are $\mathrm{E}(X) = \lambda, \mathrm{Var}(X) = \lambda$. The m.g.f. is $M_X(t) = e^{\lambda(e^t-1)}$. The parameter λ can be viewed as the average number of occurrences in the unit interval.

Example 1.9.5. In a certain region in the northeast U.S., the number of severe weather per year X is assumed to follow a Poisson distribution with mean $\lambda = 1.5$. What is the probability there are four or more severe weather events in a given year?

1.10 Some Common Continuous Distributions

1.10.1 Uniform Distributions

A random variable X is said to have a uniform distribution from a to b ($a < b$) if its p.d.f. is given by

$$f(x|a, b) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x).$$

We denote $X \sim U(a, b)$. Note the use of indicator function to keep track of the support.

Its mean and variance are given by $\mathrm{E}(X) = \frac{a+b}{2}; \mathrm{Var}(X) = \frac{(b-a)^2}{12}$. Its m.g.f. is $M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$. Note its c.d.f. is given by

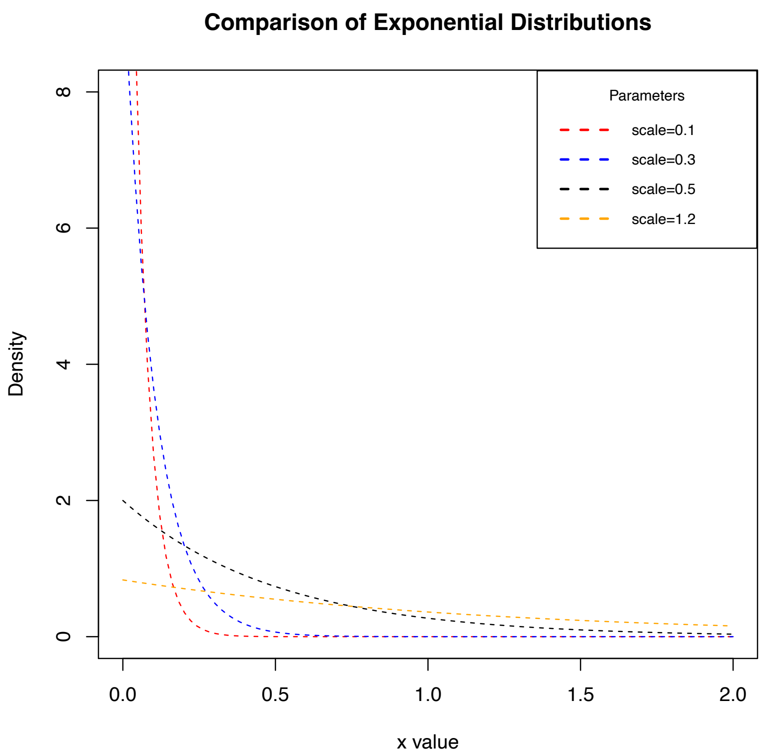
$$F_X(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \ge b. \end{cases}$$

1.10.2 Exponential Distributions

$$f(x|\beta) = \frac{1}{\beta} e^{-x/\beta} \mathbb{1}_{[0,\infty)}(x), \quad \beta > 0 \text{ is scale parameter.}$$

We denote $X \sim \text{Exp}(\beta)$. Its c.d.f. is $F(x) = 1 - e^{-x/\beta}$ for $x > 0$.

Its mean and variance are given by $\text{E}(X) = \beta, \text{Var}(X) = \beta^2$. The m.g.f. is $M_X(t) = \frac{1}{1-\beta t}, t < \beta^{-1}$.



Exponential distribution has memoryless property. Suppose that $X \sim \text{Exp}(\beta)$, and let r and s be positive constants. Then

$$P(X > r + s \mid X > r) = P(X > s).$$

Example 1.10.1. "Time to event" studies are common in medical applications. One recent study involved patients with leg ulcers. Some treatment was applied to the infected area. Let X denote the time (in days) until the leg ulcer was completely healed. Suppose X has an exponential distribution with mean $\beta = 190$. Find the probability the ulcer takes longer than 100 days to heal.

1.10.3 Gamma Distributions

Definition 1.10.1. (**Gamma Function**). The gamma function is a real function of t, defined by

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy,$$

for all $t > 0$.

The gamma function satisfies the properties (i) $\Gamma(1) = 1$; (ii) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$; (iii) recursive relationship

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1),$$

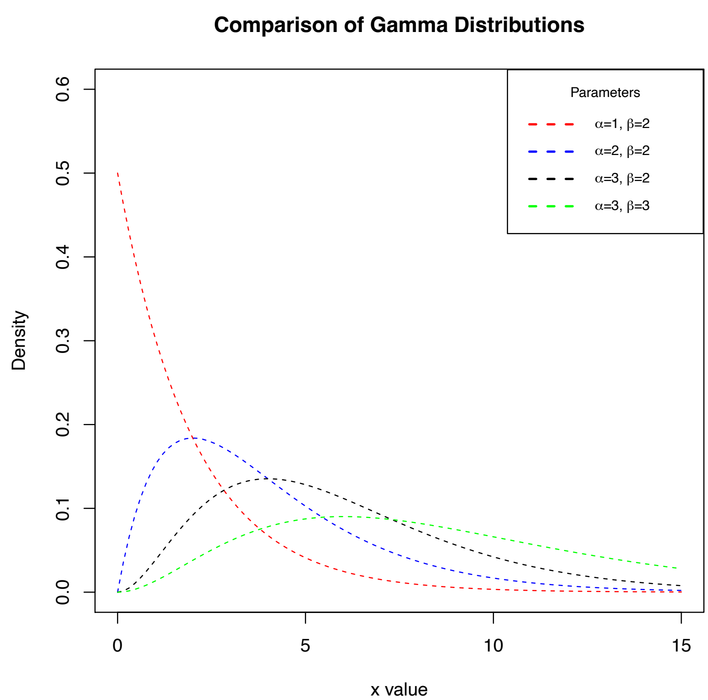
for $\alpha > 1$. From this fact, we can deduce that if α is an integer, then

$$\Gamma(\alpha) = (\alpha - 1)!$$

A random variable X is said to have a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \mathbb{1}(x > 0)$$

Shorthand notation is $X \sim$ gamma(α, β). The family of gamma distributions is indexed by two parameters: $\alpha =$ the shape parameter, $\beta =$ the scale parameter. The following picture shows this family is very flexible family of probability densities. Mean and variances are given by $\mathrm{E}(X) = \alpha\beta; \mathrm{Var}(X) = \alpha\beta^2$. Its m.g.f. is $M_X(t) = (\frac{1}{1-\beta t})^\alpha, t < \beta^{-1}$. Note that gamma$(\alpha, \beta)$ is the exponential distribution with scale β.



Theorem 1.10.1. If the random variables $X_1, \ldots, X_k$ are independent and if each $X_i \sim$ gamma$(\alpha_\mathrm{i}, \beta)$, then the sum $X_1 + \cdots + X_k \sim$ gamma$(\sum_{\mathrm{i}=1}^{\mathrm{k}} \alpha_\mathrm{i}, \beta)$

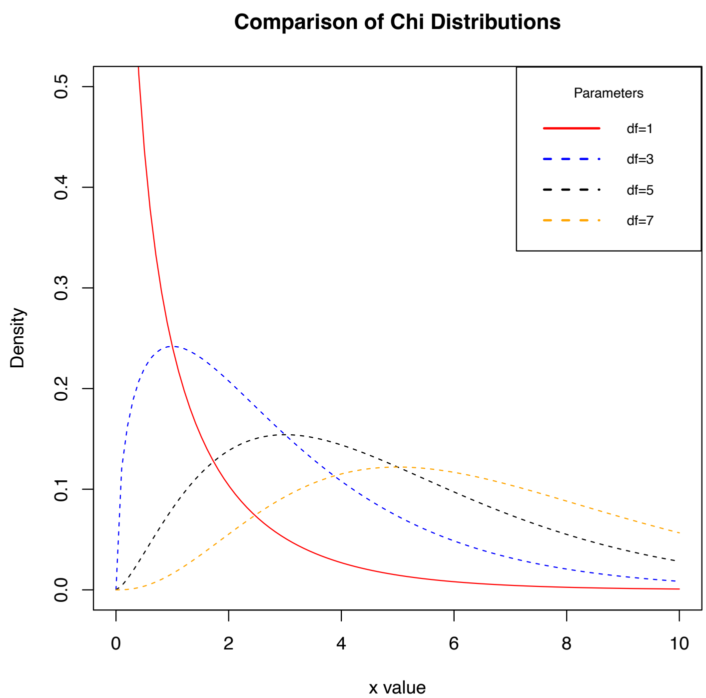
1.10.4 Chi-squared Distributions

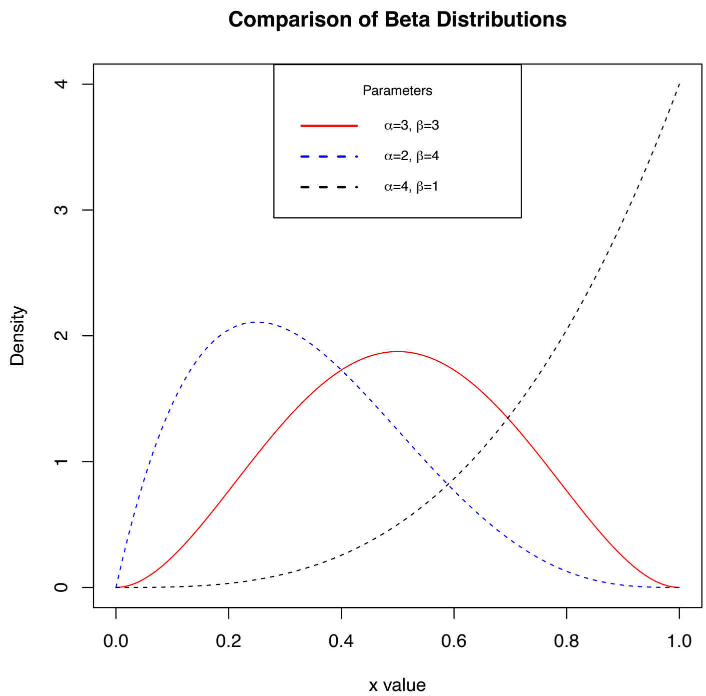
Let p be a positive integer. In the gamma(α, β) family, when $\alpha = p/2,\ \ \beta = 2$, we call the resulting distribution a χ^2 distribution with p degrees of freedom. We write $X \sim \chi^2(p)$. The p.d.f. is given by

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2} \mathbb{1}_{[0,\infty)}(x)$$

Mean and variance are $\mathrm{E}(X) = p, \mathrm{Var}(X) = 2p$. Its m.g.f. is $M_X(t) = (\frac{1}{1-2t})^{p/2}, t < 1/2$.

Theorem 1.10.2. If the random variables $X_1, \ldots, X_k$ are independent and if each $X_i \sim \chi^2(p_i)$, then the sum $X_1 + \cdots + X_k \sim \chi^2(\sum_{i=1}^k p_i)$





1.10.5 Beta Distributions

A random variable X is said to have a beta distribution with parameters $\alpha > 0$ and $\beta > 0$ if its p.d.f. is given by

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{[0,1]}(x).$$

The constant $B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is called the beta function. The family of beta distributions is useful for modeling densities supported on the unit interval $(0, 1)$. Its mean and variance are given by $\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}$, $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

1.10.6 Normal Distributions

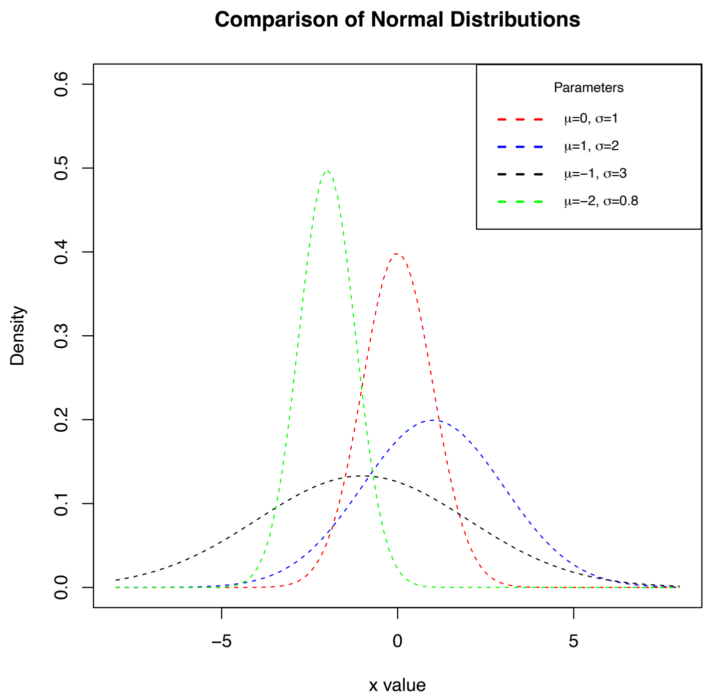
A random variable X is said to have a normal distribution if its p.d.f. is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathbb{1}(-\infty < x < \infty).$$

Shorthand notation is $X \sim N(\mu, \sigma^2)$. There are two parameters in the normal distribution: the mean $\mathrm{E}(X) = \mu$ and the variance $\mathrm{Var}(X) = \sigma^2$. Its m.g.f. is $M_X(t) = e^{\mu t+(\sigma^2 t^2)/2}$.

We also let $\Phi(t)$ denote the c.d.f. of the standard normal random variable:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds.$$



Note that $\Phi(t)$ does not have a closed form solution. However, probability table exists that tabulates its values at different t. A software can readily compute its value.

Facts:

(a) The $N(\mu, \sigma^2)$ p.d.f. is symmetric about μ; that is, for any $a \in R$, $f(\mu - a) = f(\mu + a)$.

(b) The $N(\mu, \sigma^2)$ p.d.f. has points of inflection (i.e., where curvature changes) located at $x = \mu \pm \sigma$.

(c) $\lim_{x \to \pm\infty} f_X(x) = 0$.

Theorem 1.10.3. If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

Corollary 1.10.1. If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0, 1)$.

Theorem 1.10.4. If $X \sim N(0, 1)$, then $X^2 \sim \chi^2(1)$.

1.10.7 The t distribution

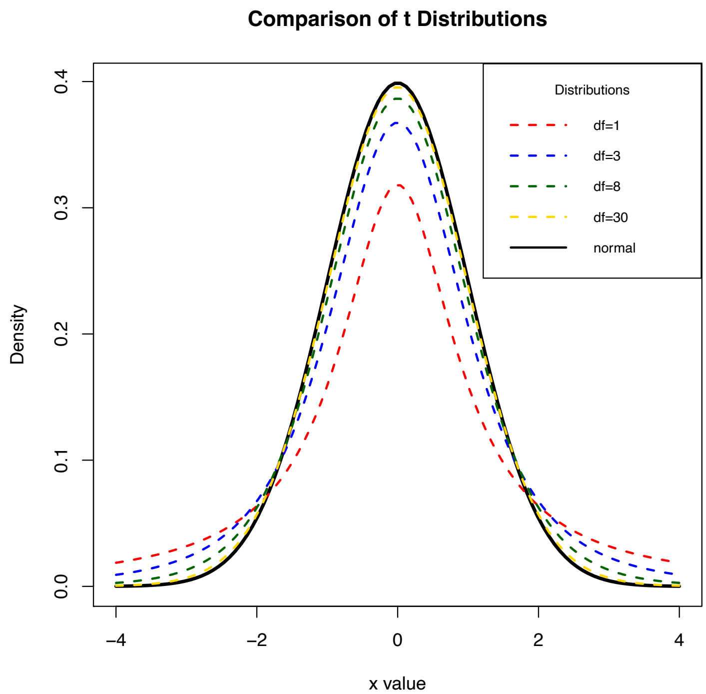
Suppose that $Z \sim N(0, 1)$ and that $W \sim \chi^2(\nu)$; Z and W are independent. Then the random variable

$$X = \frac{Z}{\sqrt{W/\nu}}$$

has a t distribution with ν degrees of freedom. In notation, $X \sim t(\nu)$. The p.d.f. of X is given by

$$f(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+\frac{x^2}{\nu})^{(\nu+1)/2}}, -\infty < x < \infty, \nu = 1, \dots$$

Its mean and variance are given by $\mathrm{E}(X) = 0, \nu > 1; \mathrm{Var}(X) = \frac{\nu}{\nu-2}, \nu > 2$. The m.g.f. does not exist.



Facts:

(a) t distribution is continuous and symmetric about 0.

(b) As $\nu \to \infty$, $t(\nu) \to N(0,1)$. That is, when ν becomes larger, the $t(\nu)$ and $N(0,1)$ distributions look more alike.

(c) Compared with $N(0,1)$, the t distribution is less peak and has more mass in both tails.

1.10.8 The F distribution

Suppose that $W_1 \sim \chi^2(\nu_1)$ and that $W_2 \sim \chi^2(\nu_2)$, $\nu_1, \nu_2 = 1, 2, \dots$; W_1 and W_2 are independent. Then the random variable

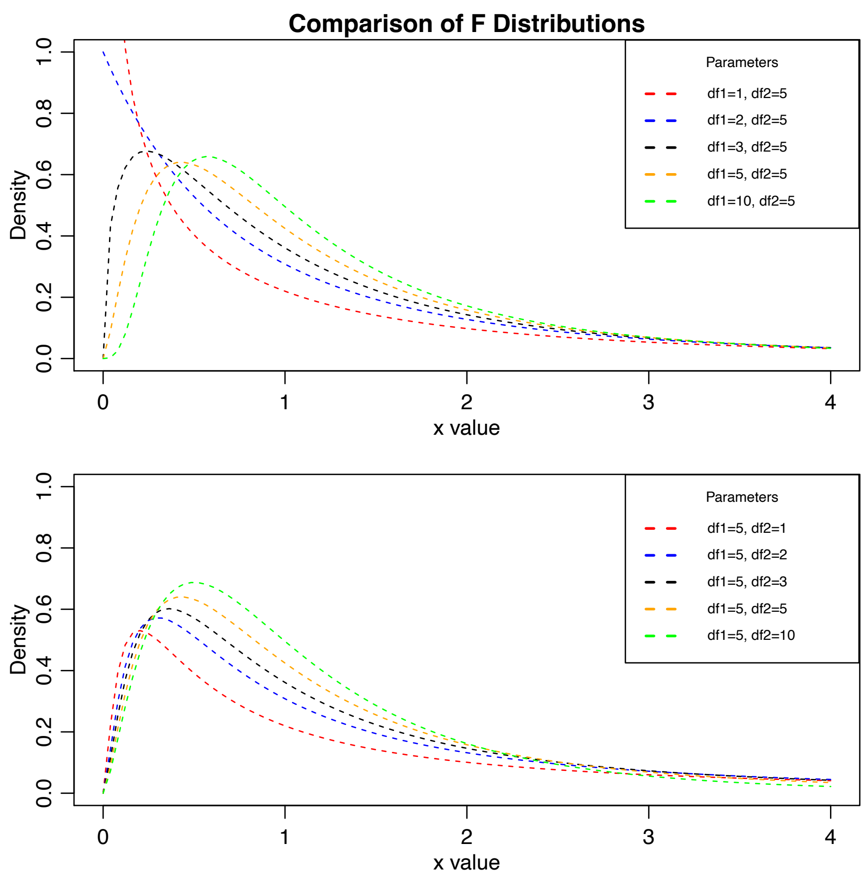
$$X = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has a F distribution with ν_1 (numerator) and ν_2 (denominator) degrees of freedom. In notation, $X \sim F(\nu_1, \nu_2)$.

The p.d.f. of X is given by

$$f(x|\nu_1, \nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} (\frac{\nu_1}{\nu_2})^{\nu_1/2} \frac{x^{(\nu_1-\nu_2)/2}}{(1+\frac{\nu_1}{\nu_2}x)^{(\nu_1+\nu_2)/2}} \mathbb{1}_{[0,\infty)}(x).$$

Its mean and variance are given by $\mathrm{E}(X) = \frac{\nu_2}{\nu_2-2}, \nu_2 > 2; \mathrm{Var}(X) = 2\frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)}(\frac{\nu_2}{\nu_2-2})^2, \nu_2 > 4$. The m.g.f. does not exist.



Facts:

(a) F distribution is continuous and skewed to the right.

(b) If $X \sim F(\nu_1, \nu_2)$, then $1/X \sim F(\nu_2, \nu_1)$.

(c) If $X \sim t(\nu)$, then $X^2 \sim F(1, \nu)$.

1.10.9 Multivariate normal distribution

Let X denote a random vector in $\mathbb{R}^k$. For a $k \times 1$ vector μ, and a positive definite matrix $\Sigma > 0$. The random vector X is said to have a multivariate normal distribution with mean

μ and variance Σ if its p.d.f. is given by

$$f(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}(\det\Sigma)^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)).$$

Its mean and variance are given by $\mathrm{E}(X) = \mu; \mathrm{Var}(X) = \Sigma$. Its m.g.f. is given by $M_X(t) = \exp(\mu^T t + \frac{1}{2}t^T\Sigma t)$. In notation, we write $X \sim N(\mu,\Sigma)$.

Special case 1: bivariate normal without correlation

To understand this definition, let's consider the case $k=2$, say two X_1, X_2 independent normal random variables, where μ_1, σ_1^2 are mean and variance of X_1, and μ_2, σ_2^2 are mean and variance of X_2, whose joint density is simply

$$f(x_2,x_2) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\big\{-\frac{1}{2}\big[\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\big]\big\}$$

Obviously in this case, $X = (X_1,X_2)^T \sim N(\mu,\Sigma)$, where $\mu = (\mu_1,\mu_2)^T$, and

$$\Sigma = \begin{pmatrix}\mathrm{Cov}(X_1,X_1) & \mathrm{Cov}(X_1,X_2)\\ \mathrm{Cov}(X_2,X_1) & \mathrm{Cov}(X_2,X_2)\end{pmatrix} = \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}.$$

Special case 2: bivariate normal with correlation

More generally, to construct a pair of bivariate normal $\{X_1,X_2\}$ such that X_1 and X_2 are correlated and each is a normal random variable: let

$$X_1 = a_1 + b_1Z_1 + c_1Z_2,$$
$$X_2 = a_2 + b_2Z_1 + c_2Z_2,$$

where Z_1, Z_2 are independent standard normal random variables. Note that $\mu_1 = E(X_1) = a_1, \mu_2 = E(X_2) = a_2$, and $\sigma_1^2 = \mathrm{Var}(X_1) = b_1^2 + c_1^2$, and $\sigma_2^2 = \mathrm{Var}(X_2) = b_2^2 + c_2^2$, and the covariance is $\sigma_{12} = \mathrm{Cov}(X_1,X_2) = b_1b_2 + c_1c_2$. Using the linear transformation formula Theorem 1.7.4, we can derive the joint density of (X_1,X_2) which is given by

$$f(x) = f(x_1,x_2) = \frac{1}{2\pi}\det(\Sigma)^{-1/2}\exp\big\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\big\},$$

where $\mu = (\mu_1,\mu_2)^T$ and

$$\Sigma = \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}.$$

Due to the unique structure of this joint density, one key observation is that if X_1 and X_2 are uncorrelated, i.e., $\sigma_{12} = 0$, then $f(x) = f_{X_1}(x_1)f_{X_2}(x_2)$ holds, i.e., X_1 and X_2 are independent. Recall, in general, uncorrelation does not imply independence.

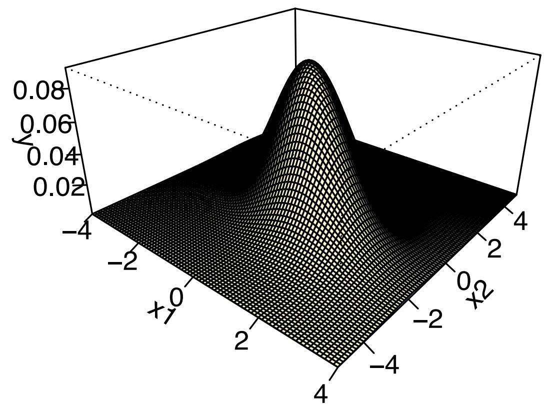


Figure 1: An example of bivariate normal, with $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 2$, $\sigma_{12} = -1$.

General case:

For $k \geq 2$, we call a k-dimensional random vector Z the *multivariate standard normal* random vector, if Z is a k-dimensional random vector that has the multivariate normal density with the mean vector $\mu = (0, \ldots, 0)^T$, and covariance matrix $\Sigma = I$ the identity matrix. One can show that if we generate k *independent* standard normal *random variables* $Z_1, \ldots, Z_k$, and put them as a vector $Z = (Z_1, \ldots, Z_k)^T$, then $Z \sim N(0, I)$ which is the **standard multivariate normal distribution**.

Indeed,

$$
\begin{aligned}
f(z_1, \ldots, z_k) &= f(z_1) f(z_2) \cdots f(z_k) \\
&= \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \\
&= \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{k} z_i^2\right) \\
&= \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{z^T z}{2}\right)
\end{aligned}
$$

How about a multivariate normal distribution with some non-standard covariance matrix? This can be achieved through the linear transformation: given a *multivariate standard normal* random vector $Z \sim N(0, I)$, where I is k by k identity matrix. Let $X = AZ + \mu$ for some nonsingular $k \times k$ matrix A and a vector of constant μ, then $X \sim N(\mu, AA^T)$. This can be verified by the Theorem 1.7.4. Here, the AA^T will be the covariance matrix Σ.

Some important properties of the multivariate normal distribution :

(1)*

$$\text{If } X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right],$$

then $X_1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}X_2 \perp\!\!\!\perp X_2$ and $X_1|X_2 \sim N(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$, where $\mu_{1\cdot 2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ and $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

(2) Bivariate normal (an alternative construction). Let Z_1, Z_2 are independent normal. For some given values $\mu_X, \mu_Y, \sigma_X > 0$, $\sigma_Y > 0$, $0 < \rho < 1$, define $X = \sigma_X Z_1 + \mu_X, Y = \sigma_Y(\rho Z_1 + (1-\rho^2)^{1/2}Z_2) + \mu_Y$. Then (X, Y) has bivariate normal distribution whose density is given by

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2(1-\rho^2)}\Big[\Big(\frac{x-\mu_X}{\sigma_X}\Big)^2 + \Big(\frac{y-\mu_Y}{\sigma_Y}\Big)^2 - 2\rho\Big(\frac{x-\mu_X}{\sigma_X}\Big)\Big(\frac{y-\mu_Y}{\sigma_Y}.\Big)\Big]\Big\}$$

In particular, one can see ρ in fact is the correlation coefficient between X and Y. Also, $f(x,y) = f_{X|Y}(x|y)f_Y(y), f_{X|Y} \sim N(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y), \sigma_X^2(1-\rho^2))$ and $f_Y \sim N(\mu_Y, \sigma_Y^2)$. Similarly, $f(x,y) = f_{Y|X}(y|x)f_X(x), f_{Y|X} \sim N(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2))$ and $f_X \sim N(\mu_X, \sigma_X^2)$.

(3) Assuming $X \sim N(\mu, \Sigma)$ where $\Sigma > 0$, then if $Y = AX + b$ for some matrix A and a vector of constant b, then $Y \sim N(A\mu + b, A\Sigma A^T)$. For $A\Sigma A^T > 0$, it must be that A is full row rank.

(4) If $X_1, \ldots, X_k$ are independent normal random variables, then any linear combination of $X_1, \ldots, X_k$ (i.e., any a^TX for nonzero constant vector a) is also a normal random variable.

(5) $X = (X_1, \ldots, X_k)^T$ is multivariate normal if and only if a^TX is normal for all nonzero constant vectors a.

(6) If $X = (X_1, \ldots, X_k)^T$ is multivariate normal, then each component X_i is also normal. The converse is not true in general unless $X_1, \ldots, X_k$ are also independent.

(7) If a random vector X has multivariate normal distribution, then two or more of its components that are (pairwise) uncorrelated are also (mutually) independent. As a consequence, pairwise independence of the components will imply the (mutual) independence of the components.

1.10.10 Some numerical commands

The corresponding Python commands uses the scipy.stats module.
For a given value x, compute $P(X = x)$ (p.m.f or p.d.f.):

Distribution	R	Python
$\mathrm{Bin}(n, p_0)$	$\mathrm{dbinom(x, n, p_0)}$	$\mathrm{scipy.stats.binom.pmf(x, n, p_0)}$
$\mathrm{Pois}(\lambda)$	$\mathrm{dpois(x, \lambda)}$	$\mathrm{scipy.stats.poisson.pmf(x, \lambda)}$
$N(0,1)$	$\mathrm{dnorm(x)}$	$\mathrm{scipy.stats.norm.pdf(x, loc = 0, scale = 1)}$
χ^2_r	$\mathrm{dchisq(x, r)}$	$\mathrm{scipy.stats.chi2.pdf(x, df = r)}$
t_r	$\mathrm{dt(x, r)}$	$\mathrm{scipy.stats.t.pdf(x, df = r)}$
$F_{r,k}$	$\mathrm{df(x, r, k)}$	$\mathrm{scipy.stats.f.pdf(x, dfn = r, dfd = k)}$

For a given value q, compute $P(X \le q)$:

Distribution	R	Python
$\mathrm{Bin}(n, p_0)$	$\mathrm{pbinom(q, n, p_0)}$	$\mathrm{scipy.stats.binom.cdf(q, n, p_0)}$
$\mathrm{Pois}(\lambda)$	$\mathrm{ppois(q, \lambda)}$	$\mathrm{scipy.stats.poisson.cdf(q, \lambda)}$
$N(0,1)$	$\mathrm{pnorm(q)}$	$\mathrm{scipy.stats.norm.cdf(q, loc = 0, scale = 1)}$
χ^2_r	$\mathrm{pchisq(q, r)}$	$\mathrm{scipy.stats.chi2.cdf(q, df = r)}$
t_r	$\mathrm{pt(q, r)}$	$\mathrm{scipy.stats.t.cdf(q, df = r)}$
$F_{r,k}$	$\mathrm{pf(q, r, k)}$	$\mathrm{scipy.stats.f.cdf(q, dfn = r, dfd = k)}$

For a given probability p, compute the x such that $P(X \le x) = p$:

Distribution	R	Python
$\mathrm{Bin}(n, p_0)$	$\mathrm{qbinom(p, n, p_0)}$	$\mathrm{scipy.stats.binom.ppf(p, n, p_0)}$
$\mathrm{Pois}(\lambda)$	$\mathrm{qpois(p, \lambda)}$	$\mathrm{scipy.stats.poisson.ppf(p, \lambda)}$
$N(0,1)$	$\mathrm{qnorm(p)}$	$\mathrm{scipy.stats.norm.ppf(p, loc = 0, scale = 1)}$
χ^2_r	$\mathrm{qchisq(p, r)}$	$\mathrm{scipy.stats.chi2.ppf(p, df = r)}$
t_r	$\mathrm{qt(p, r)}$	$\mathrm{scipy.stats.t.ppf(p, df = r)}$
$F_{r,k}$	$\mathrm{qf(p, r, k)}$	$\mathrm{scipy.stats.f.ppf(p, dfn = r, dfd = k)}$

1.10.11 Truncated and Censored Distributions

to add...

6 Appendix: List of Distributions

Some Common Discrete Distributions

For r any real number and n a nonnegative integer, define $\binom{-r}{n} = \frac{(-r)!}{(-r-n)!n!}$ which is understood as $\frac{(-r)(-r-1)\cdots(-r-n+1)}{n!} = \frac{r(r+1)\cdots(r+n-1)}{n!}(-1)^n$. Thus $\binom{-r}{n} = (-1)^n\binom{r+n-1}{n}$. Also, it holds that $(1+x)^{-r} = \sum_{n=0}^{\infty}\binom{-r}{n}x^n$ for $|x| < 1$.

Bernoulli

- p.m.f. $\mathrm{P}(X = x|p) = p^x(1-p)^{1-x}, x = 0, 1; 0 \leq p \leq 1$
- $\mathrm{E}(X) = p, \mathrm{Var}(X) = p(1-p)$
- m.g.f. $M_X(t) = 1 - p + pe^t$; c.f. $\psi_X(t) = 1 - p + pe^{it}$

Binomial(n, p)

- p.m.f. $\mathrm{P}(X = x|n, p) = \binom{n}{x}p^x(1-p)^{n-x}, x = 0, \ldots, n; 0 \leq p \leq 1$
- $\mathrm{E}(X) = np, \mathrm{Var}(X) = np(1-p)$
- m.g.f. $M_X(t) = (1 - p + pe^t)^n$; c.f. $\psi_X(t) = (1 - p + pe^{it})^n$

Discrete uniform

- p.m.f. $\mathrm{P}(X = x|N) = \frac{1}{N}, x = 1, \ldots, N; N = 1, 2, \ldots$
- $\mathrm{E}(X) = \frac{N+1}{2}, \mathrm{Var}(X) = \frac{(N+1)(N-1)}{12}$
- m.g.f. $M_X(t) = \frac{1}{N}\sum_{i=1}^{N} e^{it}$

Geometric(p)
Let X denote the number of trial to observe the first success in a sequence of independent Bernoulli(p) trials.

- p.m.f. $\mathrm{P}(X = x|p) = p(1-p)^{x-1}, x = 1, 2, \ldots; 0 \leq p \leq 1$
- $\mathrm{E}(X) = \frac{1}{p}, \mathrm{Var}(X) = \frac{1-p}{p^2}$
- m.g.f. $M_X(t) = \frac{pe^t}{1-(1-p)e^t}, t < -\log(1-p)$
- notes: (1) $Y = X - 1$ is negative binomial$(1, p)$. (2) Memoryless: $P(X > s|X > t) = P(X > s - t)$.

Hypergeometric
The population consists of N items, M of which are classified as successes. Let X denote the number of successes in the K random draws without replacement.

- p.m.f. $\mathrm{P}(X = x|N, M, K) = \binom{M}{x}\binom{N-M}{K-x}/\binom{N}{K}, x = 0, \ldots, M \wedge K;\ M - (N - K) \leq x \leq M \wedge K;\ N, M, K \geq 0$

- $\mathrm{E}(X) = \frac{KM}{N}, \mathrm{Var}(X) = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$
- notes: (1) If $K \ll M$ and N, the range $x = 0, 1, \ldots, K$ will be appropriate. (2) The identity holds that $\sum_{x=0}^{M\wedge K} \binom{M}{x}\binom{N-M}{K-x} = \binom{N}{K}$.

Negative binomial(r, p)
Let Y denote the number of failures before the r-th success in a sequence of independent Bernoulli(p) trials.

- p.m.f. $\mathrm{P}(Y = y|r, p) = \binom{r+y-1}{y}p^r(1-p)^y; y = 0, 1, \ldots; 0 \le p \le 1$
- $\mathrm{E}(Y) = \frac{r(1-p)}{p}, \mathrm{Var}(Y) = \frac{r(1-p)}{p^2}$
- m.g.f. $M_Y(t) = (\frac{p}{1-(1-p)e^t})^r, t < -\log(1-p)$
- notes: (1) The identity holds $\binom{r+y-1}{y}(-1)^y = \binom{-r}{y}$ for any real r. (2) Let $X = Y + r$ be the number of trials to observe the r-th success. $P(X = x|r, p) = \binom{x-1}{r-1}p^r(1-p)^{x-r}, x = r, r+1, \ldots$, $E(X) = \frac{r}{p}, \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$. The mgf $M_X(t) = (\frac{pe^t}{1-(1-p)e^t})^r, t < -\log(1-p)$.

Poisson(λ)

- p.m.f. $\mathrm{P}(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, \ldots; 0 \le \lambda < \infty$
- $\mathrm{E}(X) = \lambda, \mathrm{Var}(X) = \lambda$
- m.g.f. $M_X(t) = e^{\lambda(e^t-1)}$

Some Common Continuous Distributions

Some facts
The gamma function is defined as $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ for some $\alpha > 0$. It holds $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(1) = 1, \Gamma(\frac{1}{2}) = \sqrt{\pi}$. The identity holds $\int_0^\infty x^{\alpha-1}e^{-x/\beta}dx = \beta^\alpha\Gamma(\alpha)$. For any positive integer n, $\Gamma(n) = (n-1)!$ and $\Gamma(\frac{2n+1}{2}) = \frac{(2n)!}{2^{2n}n!}\sqrt{\pi}$.

Beta(α, β)

- p.d.f. $f(x|\alpha, \beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{[0,1]}(x); \alpha > 0, \beta > 0$.
- c.d.f. $F(x) = \frac{1}{B(\alpha,\beta)}\int_0^x u^{\alpha-1}(1-u)^{\beta-1}du = \frac{B_x(\alpha,\beta)}{B(\alpha,\beta)}$, where $B_x(\alpha, \beta) = \frac{x^\alpha}{\alpha}G(\alpha, 1-\beta; \alpha+1, x)$. Here $G(a, b; r, x)$ is the Gauss hypergeomeric function defined as $\sum_{k=0}^\infty \frac{(a)_k(b)_k}{(r)_k}\frac{x^k}{k!}$ (converges for $|x| < 1$), $(z)_k$ is ascending factorial $z(z+1)\cdots(z+k-1)$.
- $\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}, \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- m.g.f. $M_X(t) = 1 + \sum_{k=1}^\infty(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r})\frac{t^k}{k!}$, $\mathrm{E}(X^n) = \frac{B(\alpha+n,\beta)}{B(\alpha+\beta)}$

- notes: (1) $B(\alpha,\beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. (2) If $X \sim$Beta(α,β), then $1-X \sim$Beta(β,α). (3) If $X \sim$Beta$(\alpha,1)$, then $-\log(X) \sim$ exponential dist with scale parameter equal to α^{-1}.

Cauchy(θ,σ)

- p.d.f. $f(x|\theta,\sigma) = \frac{1}{\pi\sigma}\frac{1}{1+(\frac{x-\theta}{\sigma})^2}; -\infty < x < \infty, -\infty < \theta < \infty, \sigma > 0$ is scale parameter.
- $\mathrm{E}(X), \mathrm{Var}(X)$ do not exist.
- m.g.f. $M_X(t)$ does not exist.
- notes: (1) special case of t-distribution, where degrees of freedom=1. (2) If X and Y are independent $N(0,1)$, X/Y is Cauchy.

Chi squared(p)

- p.d.f. $f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}}x^{(p/2)-1}e^{-x/2}\mathbb{1}_{[0,\infty)}(x); p = 1, 2, \ldots$
- $\mathrm{E}(X) = p, \mathrm{Var}(X) = 2p$
- m.g.f. $M_X(t) = (\frac{1}{1-2t})^{p/2}, t < 1/2$
- notes: (1) Special cases of the gamma distribution. (2) Let Z be standard normal random variable, then Z^2 has the chi squared distribution χ^2 with degree of freedom 1. (2) If the random variables $Z_1, \ldots, Z_d$ are i.i.d. standard normal, then $\sum_{i=1}^d Z_i^2$ has the χ^2 distribution with d degrees of freedom.

Double exponential (Laplace)(μ,σ)

- p.d.f. $f(x|\mu,\sigma) = \frac{1}{2\sigma}e^{-|x-\mu|/\sigma}; -\infty < x < \infty, -\infty < \mu < \infty; \sigma > 0$
- c.d.f. $F(x|\mu,\sigma) = \frac{1}{2}e^{\frac{x-\mu}{\sigma}}$ for $x \le \mu$; $1 - \frac{1}{2}e^{-\frac{(x-\mu)}{\sigma}}$ for $x \ge \mu$.
- $\mathrm{E}(X) = \mu, \mathrm{Var}(X) = 2\sigma^2$
- m.g.f. $M_X(t) = \frac{e^{\mu t}}{1-(\sigma t)^2}, |t| < \sigma^{-1}$
- notes: (1) The double exponential distribution is a symmetric distribution with much fatter tails than the normal but still retains all of its moments. (2) It is not bell-shaped, and has a peak (non-differentiability) at the point $x = \mu$. (3) $|X - \mu|$ is exponential distribution(σ).

Exponential(β)

- p.d.f. $f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}\mathbb{1}_{[0,\infty)}(x); \beta > 0$ is scale parameter.
- c.d.f. $F(x) = 1 - e^{-x/\beta}$
- $\mathrm{E}(X) = \beta, \mathrm{Var}(X) = \beta^2$
- m.g.f. $M_X(t) = \frac{1}{1-\beta t}, t < \beta^{-1}$

- notes: (1) β^{-1} is called the rate parameter. (2) Has the memoryless property. (3) Special case of gamma distribution. (4) $X^{1/\gamma}$ is Weibull.

F

- p.d.f. $f(x|\nu_1,\nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}(\frac{\nu_1}{\nu_2})^{\nu_1/2}\frac{x^{(\nu_1-\nu_2)/2}}{(1+\frac{\nu_1}{\nu_2}x)^{(\nu_1+\nu_2)/2}}\mathbb{1}_{[0,\infty)}(x)$, $\nu_1,\nu_2 = 1,\dots$
- $\mathrm{E}(X) = \frac{\nu_2}{\nu_2-2}, \nu_2 > 2; \mathrm{Var}(X) = 2\frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)}(\frac{\nu_2}{\nu_2-2})^2, \nu_2 > 4$
- m.g.f. does not exist.
- notes: (1) $F_{\nu_1,\nu_2} = (\chi^2_{\nu_1}/\nu_1)/(\chi^2_{\nu_2}/\nu_2)$ where the two χ^2s are independent. (2) $F_{1,\nu} = t^2_\nu$. (3) $\nu_1 F_{\nu_1,\nu_2} \to \chi^2_{\nu_1}$ if $\nu_2 \to \infty$. (4) If $X \sim F_{\nu_1,\nu_2}$, then $X^{-1} \sim F_{\nu_2,\nu_1}$.

Gamma(α,β)

- p.d.f. $f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}\mathbb{1}_{[0,\infty)}(x); \alpha,\beta > 0$ are shape and scale parameters.
- c.d.f. $F(x|\alpha,\beta) = \frac{\Gamma(\frac{x}{\beta};\alpha)}{\Gamma(\alpha)} = \frac{\int_0^{x/\beta} e^{-t}t^{\alpha-1}dt}{\Gamma(\alpha)}$.
- $\mathrm{E}(X) = \alpha\beta; \mathrm{Var}(X) = \alpha\beta^2$
- m.g.f. $M_X(t) = (\frac{1}{1-\beta t})^\alpha, t < \beta^{-1}$, $\mathrm{E}(X^n) = \beta^n\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}$; c.f. $\psi_X(t) = (\frac{1}{1-\beta it})^\alpha$
- notes: (1) Special cases are exponential ($\alpha = 1$) and chi squared $\chi^2(p)$ ($\alpha = p/2, \beta = 2$). (2) Gamma is a scale family in the second parameter, thus $\mathrm{Gamma}(\alpha,\beta)/\beta = \mathrm{Gamma}(\alpha,1)$. (3) The inverted gamma distribution $\mathrm{IG}(\alpha,\beta)$ is defined by $(\mathrm{Gamma}(\alpha,\beta))^{-1}$.

Inverted Gamma(α,β)

- p.d.f. $f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{-(\alpha+1)}e^{-1/(x\beta)}\mathbb{1}_{[0,\infty)}(x); \alpha,\beta > 0$ are shape and scale parameters.
- $\mathrm{E}(X) = \frac{1}{\beta(\alpha-1)}; \mathrm{Var}(X) = \frac{1}{\beta^2(\alpha-1)^2(\alpha-2)}$
- notes: The inverted Gamma X is obtained from $X = Y^{-1}$ where $Y \sim \mathrm{gamma}(\alpha,\beta)$ with shape and scale $\alpha,\beta > 0$.

Lognormal(μ,σ^2)

- p.d.f. $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{e^{-(\log x-\mu)^2/(2\sigma^2)}}{x}\mathbb{1}_{[0,\infty)}(x)$, $-\infty < \mu < \infty, \sigma > 0$.
- $\mathrm{E}(X) = e^{\mu+(\sigma^2/2)}; \mathrm{Var}(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$
- m.g.f. does not exist. $\mathrm{E}(X^n) = e^{n\mu+(n^2\sigma^2)/2}$
- notes: Exits another distribution with the same moments.

Normal(μ,σ^2)

- p.d.f. $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$, $-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$

- $\mathrm{E}(X) = \mu; \mathrm{Var}(X) = \sigma^2$
- m.g.f. $M_X(t) = e^{\mu t + (\sigma^2 t^2)/2}$; c.f. $\psi_X(t) = e^{i\mu t - (\sigma^2 t^2)/2}$
- notes: If $X \sim N(\mu, \sigma^2)$,then $aX + b \sim N(a\mu + b, a^2\sigma^2)$ for constants a, b.

Student's t

- p.d.f. $f(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+\frac{x^2}{\nu})^{(\nu+1)/2}}, -\infty < x < \infty, \nu = 1, \ldots$
- $\mathrm{E}(X) = 0, \nu > 1; \mathrm{Var}(X) = \frac{\nu}{\nu-2}, \nu > 2$
- m.g.f. does not exist. $\mathrm{E}(X^n) = \frac{\Gamma(\frac{n+1}{2})\Gamma(\frac{\nu-n}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})} \nu^{n/2}$ if $n < \nu$ and even; $\mathrm{E}(X^n) = 0$ if $n < \nu$ and odd.
- notes: (1) $F_{1,\nu} = t_\nu^2$; (2) $t(1)$ is the standard Cauchy distribution; (3) Z is standard normal random variable, X has chi squared distribution with ν degree of freedom. If X and Z are independent, then $\frac{Z}{\sqrt{X/\nu}}$ has t distribution with ν degree of freedom.

Uniform(a,b)

- p.d.f. $f(x|a, b) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$.
- $\mathrm{E}(X) = \frac{a+b}{2}; \mathrm{Var}(X) = \frac{(b-a)^2}{12}$
- m.g.f. $M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$; c.f. $\psi_X(t) = \frac{e^{ibt} - e^{iat}}{i(b-a)t}$
- notes: (1) When $a = 0, b = 1$, this is a special case of the beta ($\alpha = \beta = 1$). (2) If $X \sim \mathrm{Uniform}([0, 1])$, then $-\log(X) \sim \mathrm{gamma}(1, 1)$.

Weibull(γ, β)

- p.d.f. $f(x|\gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta} \mathbb{1}_{[0,\infty)}(x), \gamma > 0, \beta > 0$.
- $\mathrm{E}(X) = \beta^{1/\gamma}\Gamma(1 + \frac{1}{\gamma}); \mathrm{Var}(X) = \beta^{2/\gamma}(\Gamma(1 + \frac{2}{\gamma}) - \Gamma^2(1 + \frac{1}{\gamma}))$
- m.g.f. exists only for $\gamma \geq 1$; $\mathrm{E}(X^n) = \beta^{n/\gamma}\Gamma(1 + \frac{n}{\gamma})$
- notes: (1) When $\gamma = 1$, this becomes exponential distribution. (2) $\mathrm{Weibull}(\gamma, \beta) = (\mathrm{gamma}(1, \beta))^{1/\gamma}$. (3) The identity holds $\int_0^\infty x^{\gamma-1} \exp(-\frac{x^\gamma}{\beta}) dx = \frac{\beta}{\gamma}$.

Some Common Multivariate Distributions

Multivariate normal(μ, Σ)
Let X denote a random vector in $\mathbb{R}^k$.

- p.d.f. $f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{k/2}(\det \Sigma)^{1/2}} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)), \Sigma > 0$.
- $\mathrm{E}(X) = \mu; \mathrm{Var}(X) = \Sigma$.

- m.g.f. $M_X(t) = \exp(\mu^T t + \frac{1}{2}t^T \Sigma t)$; c.f. $\psi_X(t) = \exp(i\mu^T t - \frac{1}{2}t^T \Sigma t)$
- notes: (1)
$$\text{If } X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right],$$
then $X_1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}X_2 \perp\!\!\!\perp X_2$ and $X_1|X_2 \sim N(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$, where $\mu_{1\cdot 2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ and $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.
(2) Bivariate normal. Let Z_1, Z_2 are independent normal. For some given values $\mu_X, \mu_Y, \sigma_X > 0$, $\sigma_Y > 0$, $0 < \rho < 1$, define $X = \sigma_X Z_1 + \mu_X, Y = \sigma_Y(\rho Z_1 + (1-\rho^2)^{1/2}Z_2) + \mu_Y$, then (X, Y) has bivariate normal distribution whose density is given by
$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\Big\{-\frac{1}{2(1-\rho^2)}\Big[\Big(\frac{x-\mu_X}{\sigma_X}\Big)^2 + \Big(\frac{y-\mu_Y}{\sigma_Y}\Big)^2 - 2\rho\Big(\frac{x-\mu_X}{\sigma_X}\Big)\Big(\frac{y-\mu_Y}{\sigma_Y}\Big)\Big]\Big\}$$
Also, $f(x,y) = f_{X|Y}(x|y)f_Y(y), f_{X|Y} \sim N(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y), \sigma_X^2(1-\rho^2))$ and $f_Y \sim N(\mu_Y, \sigma_Y^2)$. Similarly, $f(x,y) = f_{Y|X}(y|x)f_X(x), f_{Y|X} \sim N(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2))$ and $f_X \sim N(\mu_X, \sigma_X^2)$.
(3) Assuming $X \sim N(\mu, \Sigma)$ where $\Sigma > 0$, then if $Y = AX + b$ for some matrix A and a vector of constant b, then $Y \sim N(A\mu + b, A\Sigma A^T)$. For $A\Sigma A^T > 0$, it must be that A is full row rank.
(4) If $X_1, \dots, X_k$ are independent normal random variables, then any linear combination of $X_1, \dots, X_k$ is also a normal random variable.
(5) $X = (X_1, \dots, X_k)^T$ is multivariate normal if and only if $a^T X$ is normal for all nonzero constant vectors a.
(6) If $X = (X_1, \dots, X_k)^T$ is multivariate normal, then each component X_i is also normal. The converse is not true in general unless $X_1, \dots, X_k$ are also independent.
(7) If a random vector X has multivariate normal distribution, then two or more of its components that are (pairwise) uncorrelated are also (mutually) independent. As a consequence, pairwise independence of the components will imply the (mutual) independence of the components.

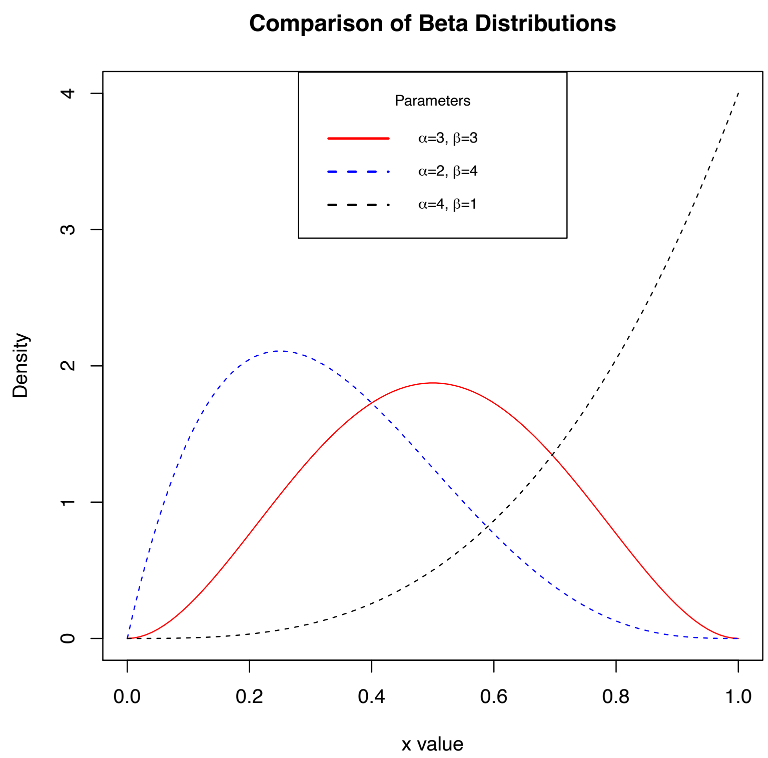
Multinomial$(n; p_1, \dots, p_k)$
Multinomial distribution models the outcomes of tossing a k-sided die for n times.

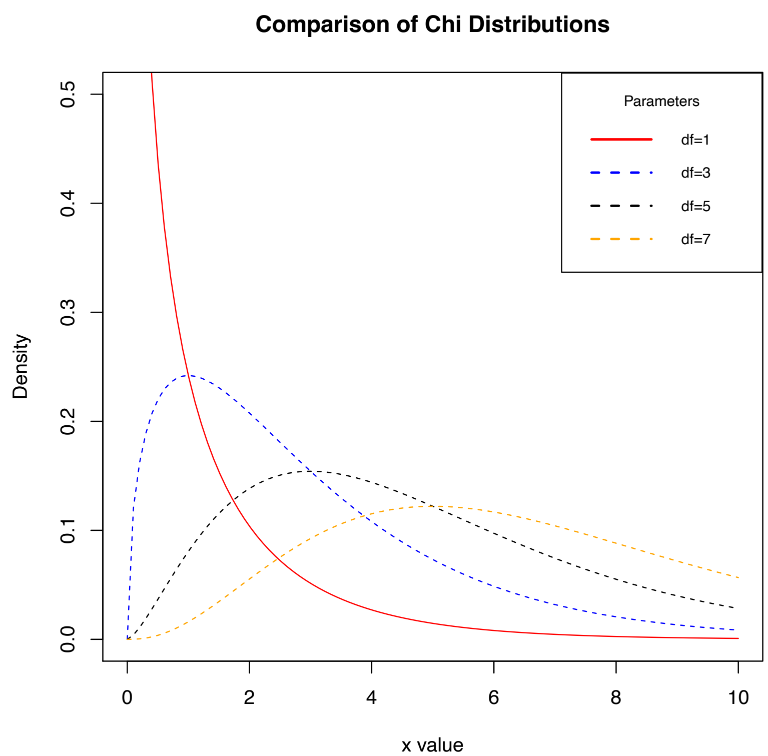
- pdf. $f(x_1, \dots, x_k|n; p_1, \dots, p_k) = \binom{n}{x_1, \dots, x_k}(\prod_{i=1}^k p_i^{x_i})\mathbb{1}(\sum_{i=1}^k x_i = n)$.
- $\mathrm{E}(X_i) = np_i; \mathrm{var}(X_i) = np_i(1-p_i)$.
- notes: (1) $\binom{n}{x_1, \dots, x_k}$ is the multinomial coefficient (the number of ways to divide a set of size $n = \sum_{i=1}^k x_i$ into subsets with sizes x_1 up to x_k).
(2) $(p_1 + \cdots + p_k)^n = \sum_{x_1+\cdots+x_k=n} \frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$.
(3) For $i \neq j$, $X_i|X_j \sim$Binomial$(n - X_j, \frac{p_i}{1-p_j})$, $\mathrm{cov}(X_i, X_j) = -np_ip_j$.
(4) Suppose $X_i \sim$ Poisson(λ_i), $i = 1, \dots, k$, are independent, then $P(X_1, \dots, X_k | \sum_{i=1}^k X_i =$

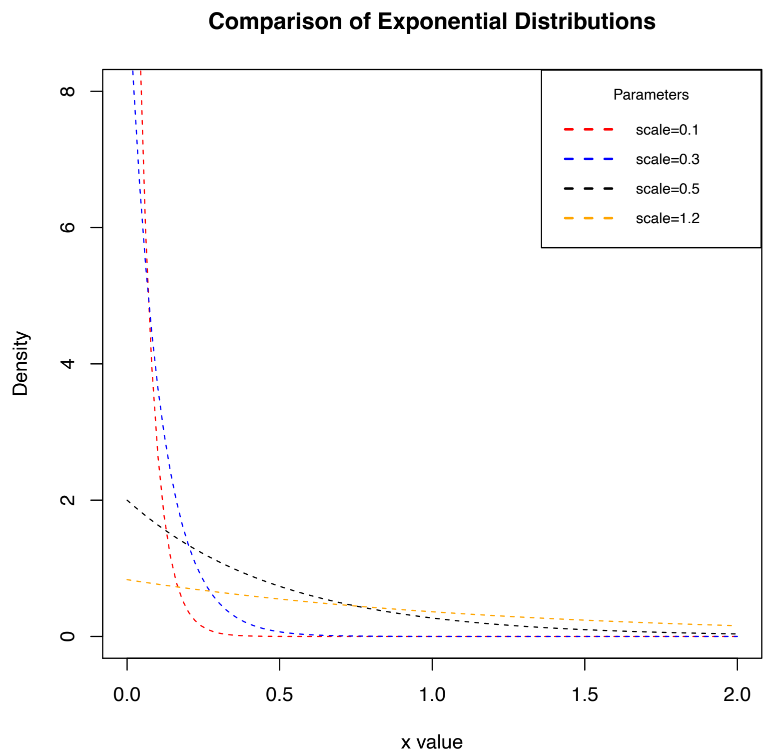
$n) = \text{Multinorm}(n; \frac{\lambda_1}{\sum_i \lambda_i}, \ldots, \frac{\lambda_k}{\sum_i \lambda_i})$
(5) When $n = 1$, $f(x_1, \ldots, x_k|1; p_1, \ldots, p_k) = \mathbb{1}(\sum_{i=1}^k x_i = 1) \prod_{i=1}^k p_i^{\mathbb{1}(x_i=1)}$. Since $(x_1, \ldots, x_k)$ can take k states, we think of x as being a scalar categorical random variables with k possible values. This gives categorical distribution (multinoulli distribution) using the notation $\text{Cat}(x|p_1, \ldots, p_k) = \text{Multinomial}(x_1, \ldots, x_k|1, p_1, \ldots, p_k)$. That is when $X \sim \text{Cat}(p_1, \ldots, p_k)$, $P(X = j|p_1, \ldots, p_k) = p_j$.

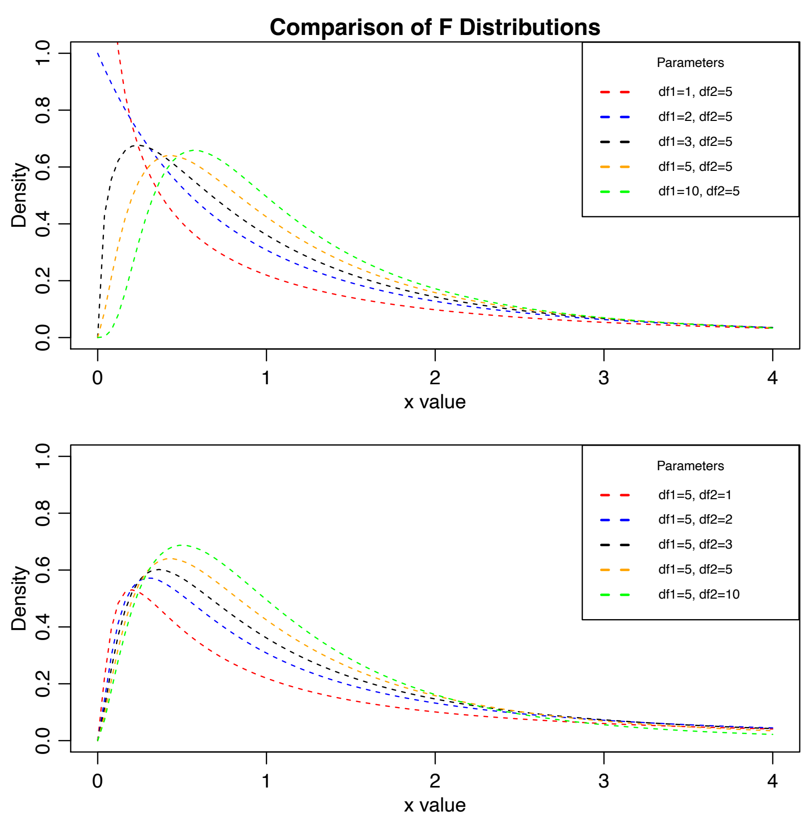
Dirichlet$(\alpha_1, \ldots, \alpha_k)$
Dirichlet distribution is a multivariate generalization of beta distribution. Let $\alpha_0 = \sum_{i=1}^k \alpha_i$.

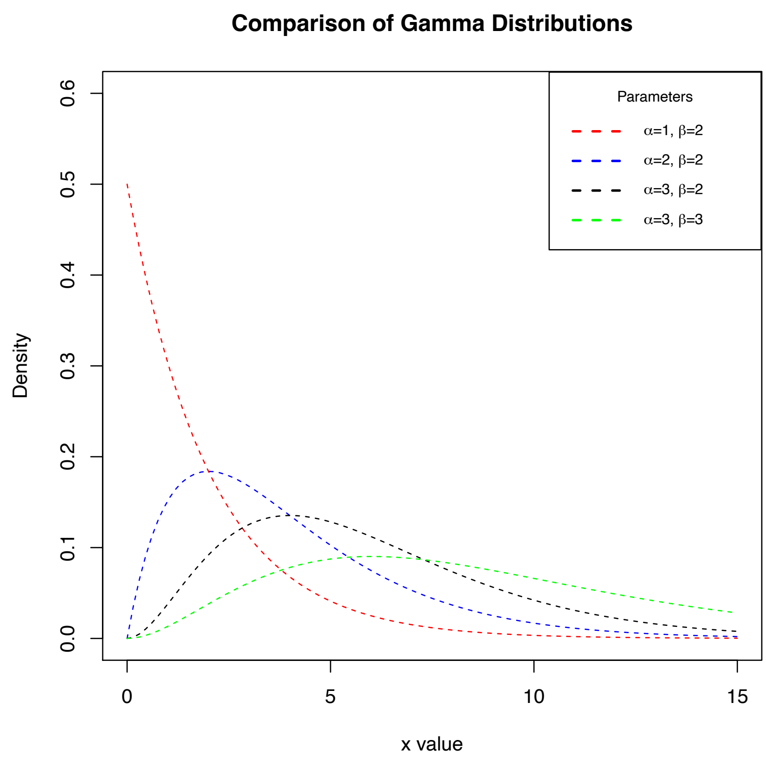
- pdf. $f(x_1, \ldots, x_k) = \frac{\Gamma(\alpha_0)}{\prod_i \Gamma(\alpha_i)} \prod_i x_i^{\alpha_i - 1} \mathbb{1}\{x_i > 0, \sum_i x_i = 1\}$.
- $\text{E}(X_i) = \frac{\alpha_i}{\alpha_0}; \text{var}(X_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)}; \text{mode}(X_i) = \frac{\alpha_i - 1}{\alpha_0 - k}$
- notes: (1) If $X \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$, then $X_1 \sim \text{beta}(\alpha_1, \sum_i \alpha_i - \alpha_1)$.
  (2) If $X \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$, then $X_i = Y_i / \sum_i Y_i$ where $Y_i, i = 1, \ldots, k$ are independent gamma$(\alpha_i, 1)$.
  (3) If $X \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$, then $(X_1, \ldots, X_i + X_j, \ldots, X_k) \sim \text{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_k)$.
  (4) If $X \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$, given two partitions A, B of index, then $(\sum_{i \in A} X_i, \sum_{i \in B} X_i) \sim \text{Dir}(\sum_{i \in A} \alpha_i, \sum_{i \in B} \alpha_i)$, i.e., $\sum_{i \in A} X_i \sim \text{beta}(\sum_{i \in A} \alpha_i, \sum_{i \in B} \alpha_i)$.
  (5) Conditional distribution of a subvector given remaining elements is also Dirichlet.
  (6) $\text{Cov}(X_i, X_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}, i \neq j$.
  (7) Large $\alpha_i > 1$ values push the x_i to some central value, where smaller $\alpha_i < 1$ values push x_i to the corners (x_i tending 0). If all α_i are equal, then the distribution is symmetric. If $\alpha_1 = \cdots = \alpha_k = 1$, the points are uniformly distributed. If $\alpha_1 = \cdots = \alpha_k \to \infty$, then $x_1 = \cdots = x_k = 1/k$ with probability 1.

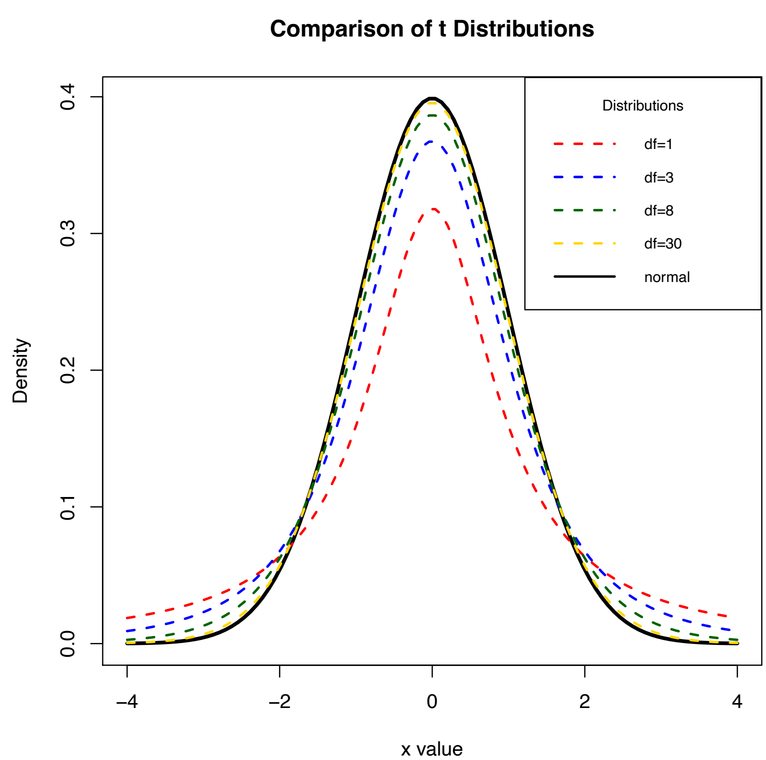

Comparison of Beta Distributions
Parameters
α=3, β=3
α=2, β=4
α=4, β=1
Density
x value
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

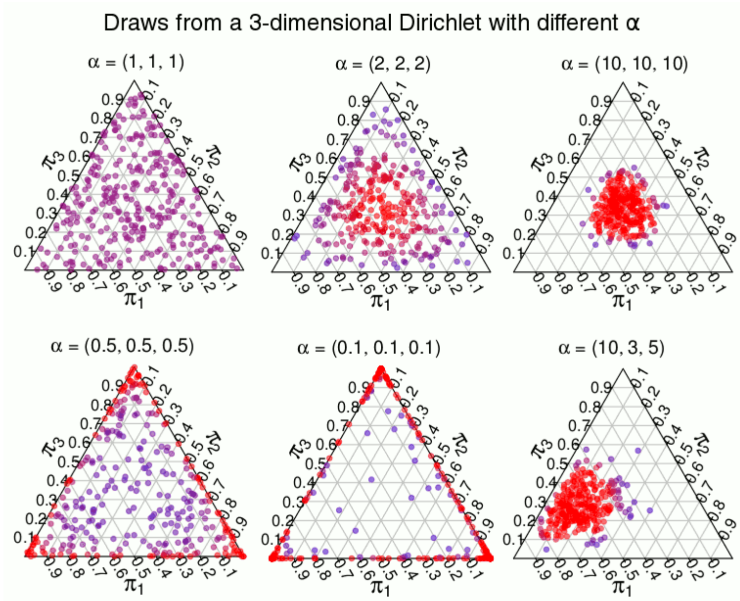

Comparison of Chi Distributions
Parameters
df=1
df=3
df=5
df=7
Density
x value
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10

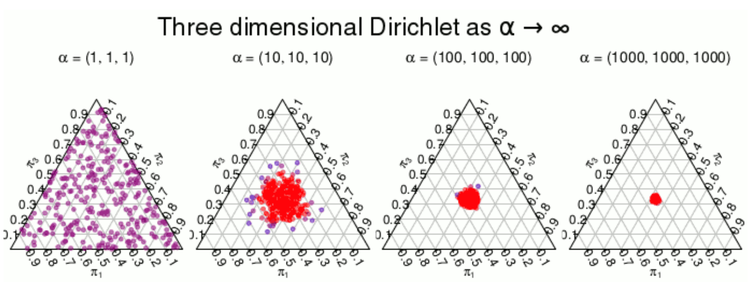

Comparison of Exponential Distributions
Parameters
scale=0.1
scale=0.3
scale=0.5
scale=1.2
Density
x value

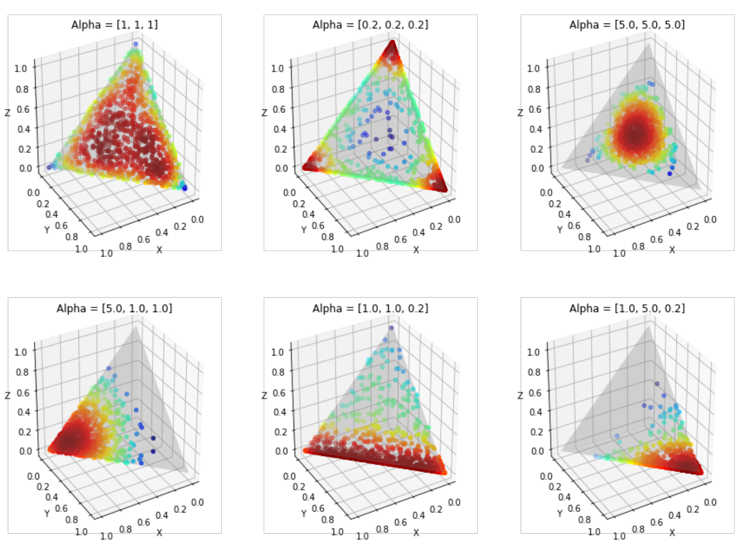

Comparison of F Distributions
Parameters
df1=1, df2=5
df1=2, df2=5
df1=3, df2=5
df1=5, df2=5
df1=10, df2=5
Parameters
df1=5, df2=1
df1=5, df2=2
df1=5, df2=3
df1=5, df2=5
df1=5, df2=10
Density
x value


Comparison of Gamma Distributions
Parameters
α=1, β=2
α=2, β=2
α=3, β=2
α=3, β=3
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
x value


Comparison of t Distributions
Distributions
df=1
df=3
df=8
df=30
normal
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x value


Draws from a 3-dimensional Dirichlet with different α
α = (1, 1, 1)
α = (2, 2, 2)
α = (10, 10, 10)
α = (0.5, 0.5, 0.5)
α = (0.1, 0.1, 0.1)
α = (10, 3, 5)


Three dimensional Dirichlet as α → ∞
α = (1, 1, 1)
α = (10, 10, 10)
α = (100, 100, 100)
α = (1000, 1000, 1000)


Alpha = [1, 1, 1]
Alpha = [0.2, 0.2, 0.2]
Alpha = [5.0, 5.0, 5.0]
Alpha = [5.0, 1.0, 1.0]
Alpha = [1.0, 1.0, 0.2]
Alpha = [1.0, 5.0, 0.2]