Probability Notes

Compiled by Paul J. Hurtado

Last Compiled: January 25, 2016

About These Notes

These are course notes from a Probability course taught using *An Introduction to Mathematical Statistics and Its Applications* by Larsen and Marx (5th ed).

These notes are heavily based on notes provided to me by Professor Ania Panorska, who had previously taught that course, plus material I have added based on my own materials or material found in other Probability textbooks.

You will undoubtedly find these notes lack many important details. **I strongly urge you to seek out more detailed treatments of this material as needed** -- e.g., by reading them along side a textbook or similarly thorough resource -- especially if using these notes for more than a light refresher.

-- Paul J. Hurtado

Contents

Basic Definitions

Experiment: Any procedure that can be repeated under the same conditions (theoretically) infinite number of times, and such that its outcomes are well defined. By *well defined* we mean we can describe **all** possible outcomes.

Outcome (or **Sample Outcome**)**:** Any possible outcome of the experiment.

Sample Space: The set of all possible outcomes of an experiment. Usually denoted by S.

Event: A subset of the sample space S. Events are usually denoted by capital letters.

A **probability space** comprises three parts (S,$\mathcal{F}$,P):

1. S is the **sample space**, the set of all **outcomes**. (Some texts use Ω instead of S). Ex: For a coin toss experiment, $S = \{H, T\}$.
2. $\mathcal{F}$ is the σ-**algebra** associated with S. It is the collection of subsets of S (we call these subsets **events**), and includes S and the empty set $\emptyset$. This **set of events** is closed under countable unions, countable intersections and complementation. Furthermore, $\mathcal{F}$ satisfies:
   (a) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, and
   (b) if A_1, A_2, ... are in $\mathcal{F}$, then their union $\bigcup_i A_i$ is also in $\mathcal{F}$.

   NOTE: Together these conditions imply closure under countable intersections. If S is countable, $\mathcal{F}$ is the power set of S, i.e., it is all subsets of S. Mathematicians call these events the **measurable sets** in S.
3. P is our probability function $P : \mathcal{F} \to [0, 1]$. It associates each event (i.e., each subset of S included in $\mathcal{F}$) with a number between 0 and 1. Furthermore, we require that
   (a) P is non-negative ($P(A) \geq P(\emptyset) = 0,\ \forall A \in \mathcal{F}$),
   (b) P is countably additive, i.e., for for a countable, disjoint set of events A_1, A_2, ... then $P(\bigcup A_i) = \sum_i P(A_i)$), and
   (c) $P(S) = 1$.

Set Operations

Operations on events (sets): Union, Intersection, Complement.

Definition: Let A and B be two events from sample space S.

1. The **union** of A and B ($A \cup B$) is exactly all elements in A or B or both.
2. The **intersection** of A and B ($A \cap B$) is exactly all elements in both A and B.
3. The **complement** of A is the event (set) A^C which contains all elements in S not in A.

NOTE: We can extend the definition of a union (or intersection) of two events, to any finite number of events $A_1, A_2, \ldots, A_k$ defined over the sample space S.

Definition: Events A and B are **mutually exclusive** if their intersection is empty ($A \cap B = \emptyset$).

1. The union of $A_1, A_2, \ldots, A_k$ is the event (set) $\bigcup_{i=1}^{k} A_i = A_1 \cup A_2 \cup \ldots \cup A_k$ which elements belong to at least one of the sets $A_1, A_2, \ldots, A_k$.

2. The intersection of $A_1, A_2, \ldots, A_k$ is the event (set) $\bigcap_{i=1}^{k} A_i = A_1 \cap A_2 \cap \ldots \cap A_k$ which elements belong to all the sets $A_1, A_2, \ldots, A_k$.

Additional properties of unions and intersections:

1. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
2. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
3. $A \cup (B \cup C) = (A \cup B) \cup C$
4. $A \cap (B \cap C) = (A \cap B) \cap C$

DeMorgan's Laws: Treat complement of a union or intersection.

1. The complement of a union of $A_1, A_2, \ldots, A_k$ is the intersection of the complements $A_1^C, A_2^C, \ldots, A_k^C$, that is $(\bigcup_{i=1}^{k} A_i)^C = \bigcap_{i=1}^{k} A_i^C$.

2. The complement of an intersection of $A_1, A_2, \ldots, A_k$ is the union of the complements $A_1^C, A_2^C, \ldots, A_k^C$, that is $(\bigcap_{i=1}^{k} A_i)^C = \bigcup_{i=1}^{k} A_i^C$.

Probability Function, P()

The probability function $P(\cdot)$ is a function defined on the set of events (subsets of S) which assigns a value in [0,1] to each event A, that is, $P(A) \in [0, 1]$.

Kolmogorov's Axioms. A function P is a **probability function** if and only if it satisfies the following axioms:

1. Probability of any event A is nonegative: $P(A) \geq 0$.

2. Probability of the sample space is 1: $P(S) = 1$.

3. The probability of a union of two mutually exclusive events A and B is the sum of their probabilities: $P(A \cup B) = P(A) + P(B)$ for any mutually exclusive events A and B.

4. The probability of a union of infinitely many pairwise disjoint events, is the sum of their probabilities. That is, if $A_1, A_2, \ldots$ are events over S such that $A_i \cap A_j = \emptyset$ for $i \neq j$, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

NOTE: Axioms 1 - 3 are enough for finite sample spaces. Axiom 4 is necessary when the sample space is infinite (e.g. the real numbers, $\mathbb{R}$).

Properties of Probability Functions

Suppose P is probability function on the subsets of the sample space S, and A and B are events defined over S. Then, the following are true.

1. $P(A^C) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. If $A \subset B$, then $P(A) \leq P(B)$.
4. For any event A, $P(A) \leq 1$.
5. If events $A_1, A_2, \ldots, A_k$ are such that $A_i \cap A_j = \emptyset$ for $i \neq j$, then $P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i)$.
6. **Addition Rule:** For any two events A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Conditional Probability and Total Probability

Definition: The conditional probability of event A **given that event** B **occurred** is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0.$$

Theorem: Multiplication Rule: The probability of A and B, $P(A \cap B)$ can be found using conditional probability: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ for $P(A \neq 0)$ and $P(B \neq 0)$.

Theorem: Multiplication Rule for more than 2 events: Let $A_1, A_2, \ldots, A_n$ be events over S. then $P(A_1 \cap A_2 \cap, \ldots, A_n) = P(A_1)P(A_2|A_1) \cdots P(A_{n-1}|A_1 \cap, \ldots, A_{n-2})P(A_n|A_1 \cap, \ldots, A_{n-1})$.

Definition: Sets $B_1, B_2, \ldots, B_n$ form a **partition** of the sample space S if: (1) They "cover" S, i.e., $B_1 \cup B_2 \cup \ldots \cup B_n = S$; and (2) They are pairwise disjoint.

Theorem: Total Probability formula: Let the sets $B_1, B_2, \ldots, B_n$ form a partition of the sample space S. Let A be an event over S. Then

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

Bayes Rule

Theorem: Bayes Formula: (1) For any events A and B defined on sample space S and such that $P(B) \neq 0$ we have:

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

(2) More generally, if the sets $B_1, B_2, \ldots, B_n$ form a partition of the sample space S, we have

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

for every $j = 1, \ldots, n$.

Independence

Definition: Two events A and B are called **independent** if $P(A \cap B) = P(A)P(B)$.

NOTE: (1) If A and B are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

(2) If A and B are independent, then so are their complements A^C and B^C.

Definition: Events $A_1, A_2, \ldots, A_n$ are independent if for every subset of them we have

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

Independent vs Mutual Exclusive (aka Disjoint)

How is **independence** related to sets being **disjoint** (i.e., **mutually exclusive**)?

1. **Independence** deals with the relationship between the **probabilities** of events A and B, and the probability of their co-occurrence, $P(A \cap B)$. Independence says something about events that can co-occur, whereas disjoint events, by definition, never co-occur.

2. The notion of sets being **disjoint** relates to the elements in those events, and whether or any are shared (i.e, whether or not they have an empty intersection). **Mutual exclusivity** describes *which outcomes cannot co-occur*. Intuition should tell us that disjoint sets are NOT independent! Why? Suppose two events are disjoint. Then knowledge of one event occurring tells you quite a bit of information about whether or not the other has occurred (by definition, it has not!). For example, if you are 22 years old, I know that you are not 21 years old. In fact, *disjoint sets cannot be independent* except in the trivial case where one or both events has probability zero: since $P(A \cap B) = 0$ for disjoint events, they can only satisfy the definition of independence ($P(A \cap B) = P(B)\,P(A)$) if $P(A) = 0$ or $P(B) = 0$ (or both are true).

Example: Consider the experiment defined by one card out of a standard 52 card deck.
Let event A be that the card is red (i.e., A is the set of all 26 red cards) and B be the event that the card is a king (i.e., B is all four kings).
Are A and B independent? Check that they satisfy the definition, $P(A \cap B) = P(A)P(B)$:

$P(A \cap B) = P(\text{red king}) = 1/26$

$P(A)P(B) = 1/2 \cdot 4/52 = 1/26$

So they are independent events!
Are they disjoint? No. Their intersection $A \cap B = \{\text{red king}\}$ is not empty.

Combinatorics: Counting, Ordering, Arranging

Multiplication Rule: If operation A can be performed in n different ways and operation B can be performed in m different ways, then the sequence of these two operations (say, AB) can be performed in $n \cdot m$ ways.

Extension of Multiplication Rule to k operations: If operations A_i, $i = 1, \ldots, k$ can be performed in n_i different ways, then the ordered sequence (operation A_1, operation A_2, ..., operation A_k) can be performed in $n_1 n_2 \cdots n_k$ ways.

Permutations: An arrangement of k objects in a row is called *a permutation of length k.*

Number of permutations of k elements chosen from a set on n elements: The number of permutations of length k, that can be formed from a set of n distinct objects is

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

Number of permutations of n elements chosen from a set on n elements: The number of permutations of length n (ordered sequences of length n), that can be formed from a set of n distinct objects is

$$n(n-1)(n-2)\cdots(1) = n!.$$

Approximation for n! (Stirling's Formula): $n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}$.

Number of permutations of elements that are not all different: The number of permutations of length n, that can be formed from a set of n_1 objects of type 1, n_2 objects of type 2, ..., n_k objects of type k, where $\sum_{i=1}^{k} n_i = n$, is

$$\frac{n!}{n_1! n_2! \cdots n_k!}.$$

Combinations: A set of k unordered objects is called *a combination of size k.*

Number of combinations of size k of n distinct objects: The number of ways to form combinations of size k from a set of n distinct objects, no repetitions, is denoted by the Newton symbol (or *binomial coefficient*) $\binom{n}{k}$, and equal to

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

NOTE: The number of combinations of size k of n distinct objects is the number of different subsets of size k formed from a set of n elements.

Combinatorial probabilities - classical definition of probability: Suppose there are n simple outcomes in a sample space S. Let event A consist of m of those outcomes. Suppose also that all outcomes are qually likely. Then, the probability of event A is defined as P(A)=m/n.

Random Variables

Definition A **probability space** $(S, \mathcal{E}, P)$ is composed of a sample space S, the algebra $\mathcal{E}$ (the set of subsets of S), and a probability function $P : \mathcal{E} \to [0, 1]$ that satisfies Kolmogorov's axioms. In practice, we think of **random variables** (r.v.) in two ways.

1. We commonly think of a random variable as a "place holder" for the observed outcome of an experiment. Ex: *Let X be the number of heads in 10 coin tosses.*

2. Formally, if X is a random variable, it is a real-valued *measurable function* that maps one probability space into another (real-valued) probability space. That is, $X : (S, \mathcal{E}, P) \to (\Omega, \mathcal{F}, P_X)$ where $\Omega \subseteq \mathbb{R}$ and we define

$$P_X(A) = P(s \in \Omega : X(s) \in A) \text{ for all events } A \in \mathcal{F}$$

   Q: How are these consistent?

   A: We tend to only be explicit about the real-valued representation of the outcome, and focus on X and P_X instead of explicitly defining all of the other details.

Definition: We refer to P as the **distribution of the random variable** and this often is sufficient to imply the structure of the associated probability space and experiment.

Definition A real-valued function X that maps one probability space $(S, \mathcal{E})$ to another probability space $(\Omega, \mathcal{F})$ is called a **random variable** (r.v.) if

$$X^{-1}(E) \in \mathcal{E} \text{ for all } E \in \mathcal{F}$$

That is, each event in the "new" algebra corresponds to (measureable) events in the original space. This ensures that X induces a consistent probability measure on the new space.
Definition Suppose r.v. X maps $(S, \mathcal{E}, P) \to (\Omega, \mathcal{F}, P_X)$. The probability function (measure) P_X is called the **probability distribution of** X and is given by

$$P_X(A) = P(\{s \in S : X(s) \in A\}) \text{ for all } A \in \mathcal{F}.$$

NOTE: By X being real-valued, we mean that $\Omega \subseteq \mathbb{R}$ or $\Omega \subseteq \mathbb{R}^n$. In the latter case, we call X a **random vector**.
Example 1: Stating that "X is a Bernoulli r.v. with probability p of success" implies that $S = \{0, 1\}$ and $P(X = k) = p^k(1 - p)^{1-k}$. That is, $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

Definition: A **Bernoulli process** $X = (X_1, ..., X_n)$ is a series of n *independent and identically distributed* (*iid*) Bernoulli trials (X_i) each with probability p of success.

Example 2: Stating that "Y is a binomial r.v. with parameters n and p" implies that $Y = \sum_{i=0}^{n} X_i$ is the number of successes in a Bernoulli process of length n, and therefore that $S = \{0, 1, ..., n\}$ and $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ for $k \in S$ (zero otherwise).

Example 3: If X is a hypergeometric r.v., it represents the number of successes in n draws from a population of size N with K successes. Thus $S = \{0, ..., \min(K, n)\}$ and

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Theorem: The distribution of X is uniquely determined by the **cumulative distribution function** (cdf) of X, denoted ny F or F_X:

$$F(x) = P(X \leq x) = P((-\infty, x]).$$

Properties of cdf

1. F is nondecreasing: If $x_1 \leq x_2,$ then $F(x_1) \leq F(x_2)$;
2. F is right - continuous: for any x, $\lim_{y \to x^+} F(y) = F(x)$;
3. $\lim_{x \to \infty} F(x) = 1$;
4. $\lim_{x \to -\infty} F(x) = 0$.

NOTE: Here are two useful rules for computing probabilities:

1. For a sequence of *increasing* sets $A_1 \subset A_2 \subset \ldots$ the probability of their union is the limit of their probabilities, that is: $P(\bigcup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$.
2. For a sequence of *decreasing* sets $A_1 \supset A_2 \supset \ldots$ the probability of their intersection is the limit of their probabilities, that is: $P(\bigcap_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$.

Types of distributions: There are three main types of distributions / random variables:

1. **Discrete** r.v.: CDF is a step function, S has at most countable number of outcomes.
   Examples: Binomial, Poisson
2. **Continuous** r.v.: CDF is a smooth function, events are typically intervals.
   Examples: Normal, Exponential
3. Mixed r.v.: CDF is neither continuous nor step function.
   Example: Zero-inflated Normal

Discrete Random Variables

Definition. Suppose a sample space S has finite or countable number of simple outcomes. Let p be a real valued function on S such that

1. $0 \leq p(s) \leq 1$ for every element s of S;
2. $\sum_{s \in S} p(s) = 1$,

Then p is said to be a **discrete probability function**.

NOTE: For any event A defined on S: $P(A) = \sum_{s \in A} p(s)$.

Definition. A real valued function $X : S \to \mathbb{R}$ is called a **random variable**.

Definition. A random variable with finite or countably many values is called a **discrete random variable**.

Definition. Any discrete random variable X is described by its **probability density function** (or probability mass function), denoted $p_X(k)$, which provides probabilities of all values of X as follows:

$$p_X(k) = P(s \in S : X(s) = k).$$

NOTE: For any k not in the range (set of values) of X: $p_X(k) = 0$.

NOTE: For any $t \leq s$, $P(t \leq X \leq s) = \sum_{k=t}^{s} P(X = k)$.

NOTATION: For simplicity, we denote $p_X(k) = P(X = k)$ thus suppressing the dependence on the sample space.

Examples:
1. Binomial r.v. X with n trials and probability of success equal to p, i.e., $X \sim \text{binom}(n, p)$.

$$p_X(k) = P(\text{k successes in n trials}) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \text{ for } k = 0, 1, 2, \ldots, n.$$

2. Hypergeometric random variable X.

Definition. Let X be a discrete random variable. For any real number t, the cumulative distribution function F of X at t is given by

$$F_X(t) = P(X \leq t) = P(s \in S : X(s) \leq t).$$

Linear transformation: Let X be a discrete random variable (rv). let Y=aX+b, where a and b are real constants. Then $p_Y(y) = p_X(\frac{y-b}{a})$.

Continuous Random Variables

Suppose a sample space Ω is uncountable, e.g., $\Omega = [0,1]$ or $\Omega = \mathbb{R}$. We can define a random variable $X : (\Omega, \mathcal{E}) \to (S, \mathcal{B})$ where the new sample space S is a subset of $\mathbb{R}$ and the algebra $\mathcal{B}$ is the Borel sets (all unions, intersections and complements of the open and closed intervals in S). The probability structure on such a space can be described using a special function, f called *probability density function* (pdf).

Definition. If sample space $S \subseteq \mathbb{R}$ then we say P is a **continuous probability distribution** if there exists a function $f(t)$ such that for any closed interval $[a,b] \subset S$ we have that $P([a,b]) = \int_a^b f(t)dt$. It follows that $P(A) = \int_A f(t)dt$ for all events A.
For a function f to be a pdf, it is necessary and sufficient that the following properties hold:

1. $f(t) \geq 0$ for every t;
2. $\int_{-\infty}^{\infty} f(t)dt = 1$.

NOTE: If $P(A) = \int_A f(t)dt$ for all A, then P satisfies all the Kolmogorov probability axioms.

Definition: Any function Y that maps S (a subset of real numbers) into the real numbers is called **a continuous random variable**. The pdf of Y is a function f such that

$$P(a \leq Y \leq b) = \int_a^b f(t)dt.$$

For any event A defined on S: $P(A) = \int_A f(t)dt$.
Theorem: For any continuous random variable $P(X = a) = 0$ for any real number a.
Definition. The cdf of a continuous random variable Y (with pdf f) is $F_Y(t)$, given by

$$F_Y(y) = P(Y \leq y) = P(\{s \in S : Y(s) \leq y\}) = \int_{-\infty}^{y} f(t)dt \quad \text{for any real } \ y.$$

Theorem. If $F_Y(t)$ is a cdf and $f_Y(t)$ is a pdf of a continuous random variable Y, then

$$\frac{d}{dt}F_Y(t) = f_Y(t).$$

Linear transformation: Let X be a continuous random variable with pdf f. Let $Y = aX + b$, where a and b are real constants. Then the pdf of Y is: $g_Y(y) = \frac{1}{|a|} f_X(\frac{y-b}{a})$.

Expectation and Expected Values

We often quantify the *central tendency* of a random variable using its **expected value** (mean).

Definition Let X be a random variable.

1. If X is a discrete random variable with pdf $p_X(k)$, then the expected value of X is given by

$$E(X) = \mu = \mu_X = \sum_{\text{all k}} k \cdot p_X(k) = \sum_{\text{all k}} k \cdot P(X = k)$$

2. If X is a continuous random variable with pdf f, then

$$E(X) = \mu = \mu_X = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

3. If X is a mixed random variable with cdf F, then the expected value of X is given by

$$E(X) = \mu = \mu_X = \int_{-\infty}^{\infty} xF'(x)dx + \sum_{\text{all k}} k \cdot P(X = k),$$

where F' is the derivative of F where the derivative exists and k's in the summation are the "discrete" values of X.

NOTE: For the expectation of a random variable to exist, we assume that all integrals and sums in the definition of the expectation above converge **absolutely**.

Definition: The **median** of a random variable is the value at the midpoint distribution of X -- another way to characterize the central tendency of a random variable. Specifically, if X is a discrete random variable, then its **median** m is the point for which $P(X < m) = P(X > m)$. If there are two values m and m' such that $P(X \leq m) = 0.5$ and $P(X \geq m') = 0.5$, the median is the average of m and m', $(m + m')/2$.

If X is a continuous random variable with pdf f, the **median** is the solution of the equation:

$$\int_{-\infty}^{m} f(x)\, dx = \frac{1}{2}.$$

Expected Values of Functions of Random Variables

Theorem. Let X be a random variable. Let $g(\cdot)$ be a function of X.
If X is discrete with pdf $p_X(k)$, then the expected value of $g(X)$ is given by

$$E\left(g(X)\right) = \sum_{\text{all k}} g(k) \cdot p_X(k) = \sum_{\text{all k}} g(k) \cdot P(X = k),$$

provided that $\sum_{\text{all k}} |g(k)|\, p_X(k)$ is finite.

If X is a continuous random variable with pdf $f_X(x)$, and if g is a continuous function, then the expected value of $g(X)$ is given by

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

provided that $\int_{-\infty}^{\infty} |g(x)|\, f(x)\, dx$ is finite.

NOTE: Expected value is a linear operator, that is $E(aX + b) = aE(X) + b$, for any rv X.

Properties of Expectation, E()

1. Linearity: $E(aX + b) = aE(X) + b$, or in general, $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X)$
2. For an indicator function, $E(\mathbb{1}_A(X)) = P_X(A)$
3. For X a finite random variable, $S = \{1, ..., n\}$, then

$$E(X) = \sum_{j=1}^{n} P(X \geq j)$$

4. (*Markov Inequality*) For $X \geq 0$,

$$P(X \geq a) \leq \frac{E(X)}{a}$$

5. If X and Y are independent, $E(XY) = E(X)E(Y)$.

Variance

To get an idea about variability of a random variable, we look at the *measures of spread*. These include the **variance**, **standard deviation**, and **coefficient of variation**.

Definition. Variance of a random variable, denoted Var(X) or σ^2, is the average of its squared deviations from the mean μ. Let X be a random variable.

1. If X is a discrete random variable with pdf $p_X(k)$ and mean μ_X, then the variance of X is given by

$$Var(X) = \sigma^2 = E[(X - \mu_X)^2] = \sum_{\text{all k}} (k - \mu_X)^2 p_X(k) = \sum_{\text{all k}} (k - \mu_X)^2 P(X = k)$$

2. If X is a continuous random variable with pdf f and mean μ_X, then

$$Var(X) = \sigma^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx.$$

3. If X is a mixed random variable with cdf F and mean μ_X, then the variance of X is given by

$$Var(X) = \sigma^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 F'(x)dx + \sum_{\text{all k}} (k - \mu_X)^2 P(X = k),$$

where F' is the derivative of F where the derivative exists and k's in the summation are the "discrete" values of X.

NOTE: If $E(X^2)$ is not finite, then the variance does not exist.

Definition. The **standard deviation** ($sd(X)$ or σ) is $sd(X) = \sqrt{Var(X)}$.

NOTE: The units of variance are square units of the random variable. The units of standard deviation are the same as the units of the random variable.

Theorem: Let X be a random variable with variance σ^2. Then, we can compute σ^2 as follows:

$$Var(X) = \sigma^2 = E(X^2) - \mu_X^2 = E(X^2) - [E(X)]^2$$

Theorem: Let X be a r.v. with variance σ^2. Then variance of $aX + b$, for any real a and b, is given by:

$$Var(aX + b) = a^2 Var(X).$$

Definition: The **coefficient of variation (CV)** is the standard deviation divided by the mean:

$$CV(X) = E(X)/\sqrt{(Var(X)}$$

The *sd* gives an absolute measure of spread, while the CV quantifies spread *relative to the mean*.

Moments of a Random Variable

Expected value is called the *first moment* of a random variable. Variance is called the *second central moment* or *second moment about the mean* of a random variable. In general, we have the following definition of the central and ordinary moments of random variables.

Definition: Let X be an r.v. Then the

1. The r^{th} moment of X (about the origin) is $E(X^r)$, provided that the moment exists.
2. The r^{th} moment of X about the mean is $E[(X - \mu_X)^r]$, provided that the moment exists.

Multivariate Distributions

In statistics, we typically worth with data sets with sample sizes greater than one! This naturally leads us to consider all of these data not as replicates from a single univariate distribution, but as a single vector-valued observation from a multivariate distribution. Before we discuss how the above material generalizes to $N > 1$ dimensions, here is some motivation from statistics.

Motivating Examples: Multivariate vs Univariate

Before we discuss random vectors, here is some statistical motivation for caring about multivariate distributions. These examples emphasize two things: First, linear algebra is fundamental to applied statistics. Embrace it! Second, a common use of *density* and *probability mass functions* for parameter estimation are to define *likelihoods*, which are joint mass (or density) functions but where we flip-flop our notions about which quantities in these equations are constants vs variables.

ORDINARY LEAST SQUARES (OLS)

Suppose you have data y_i that are assumed to be observations of normally distributed random variables Y_i with standard deviation σ and a mean μ_i that depends on different factors X_i that can be manipulated (or that can otherwise vary) for each experiment. For example, heights of individuals (Y_i) as a function of age, gender, etc. might look like

$$Y_i = Normal(\mu_i = \beta_0 + \beta_i X_{i1} + \cdots + \beta_k X_{ik}, \sigma)$$

Since a normal r.v. with mean μ and standard deviation σ can be written as μ plus a normal r.v. with mean 0 (i.e., $\mu + N(0, \sigma)$) it follows that

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$$

where each ϵ_i are independent normals with mean 0 and standard deviation σ. Writing these n equations in matrix form yields

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 + X_{11} + \cdots + X_{1k} \\ 1 + X_{21} + \cdots + X_{2k} \\ \vdots \\ 1 + X_{n1} + \cdots + X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or written in more compact matrix and vector notation,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Note that $E(\mathbf{Y}) = \mathbf{X}\beta$. Assuming the observed outcomes (data) $y = (y_1, \cdots, y_n)^T$ and inputs $\mathbf{X}$ are known, and the goal is to estimate the (unknown) parameters β (call this estimate $\hat{\beta}$). Statistical theory says the best way to compute that estimate is to take the *sum of squared differences* (SSD) between the observed data and the expected model output for a given set of parameters β (i.e., $SSD = r^T r$ where $r = y - E(\mathbf{Y})$; a measure of "distance" between model and data) then use the β that minimizes that distance as our estimate $\hat{\beta}$. It can be shown with a little linear algebra that

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}.$$

Therefore we've used linear algebra and a little multivariate calculus to turn an optimization problem into a relatively simple matrix computation!

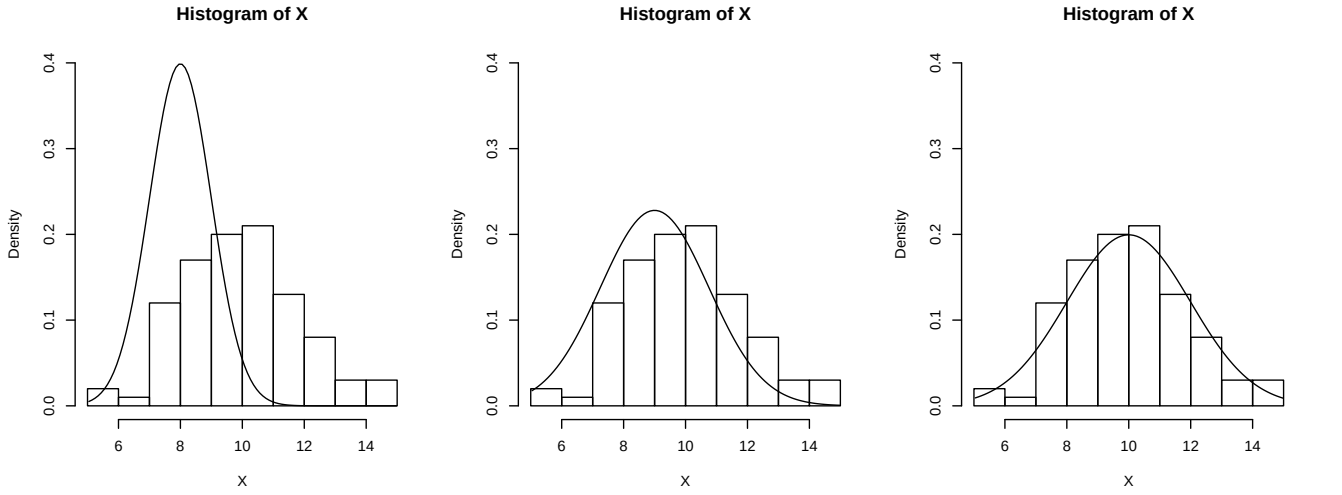
Concluding Remark: In practice, *statistics is a multivariate endeavor* and therefore you should be familiar with these basic probability concepts in a multivariate setting. Also, some basic tools from linear algebra are essential to thinking critically about both theoretical and applied statistics.

Density vs Likelihood

Definition: A **random sample** of size N is a set of N *independent and identically distributed* (*iid*) observations $X_1 = x_1$, $X_2 = x_2$, ..., $X_n = x_n$.

Here's a crude, graphical way of estimating the mean μ and variance σ^2 of a normal distribution from a random sample of data: Plot a histogram, choose an initial μ and σ and overlay the corresponding density curve. Iteratively adjust μ and σ until it looks like a good fit. In **R**...

```
set.seed(661); ## See ?set.seed or ask me :-)
X=rnorm(100,10,2); ## 100 replicates drawn from Normal(mean=10,sd=2)
par(mfrow=c(1,3));
xvals = seq(min(X),max(X),length=100); # for plotting...
hist(X,freq=FALSE,ylim=c(0,.4)); points(xvals,dnorm(xvals,8,1),type="l")
hist(X,freq=FALSE,ylim=c(0,.4)); points(xvals,dnorm(xvals,9,1.75),type="l")
hist(X,freq=FALSE,ylim=c(0,.4)); points(xvals,dnorm(xvals,10,2),type="l")
```



Formally, we'd like to compute some "goodness of fit" measure instead of just trusting our intuition with what "looks like a good fit". This might be the *SSD* (sometimes called the *sum of squared errors* [*SSE*]) from the OLS example above, but another options comes from some theoretical results in mathematical statistics: the *likelihood* of parameters μ and σ given the data X. Here, our estimates are the values of μ and σ that maximize the *likelihood*.

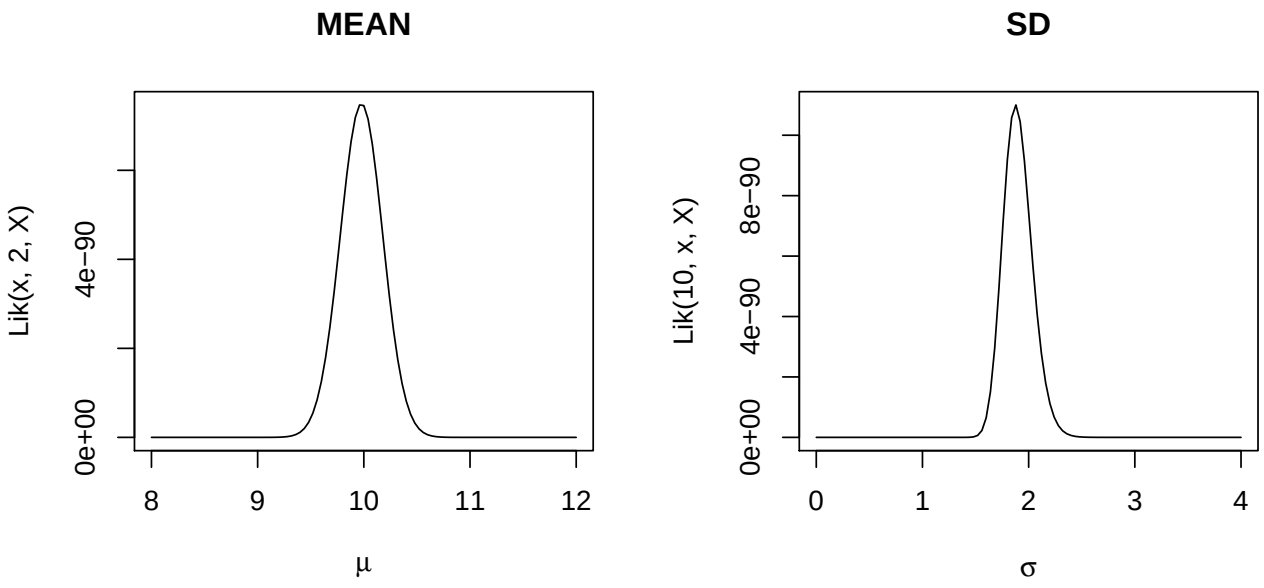
What is this *likelihood*? This is defined by the distribution for random vector X, but where we flip around our notion of what's fixed and what varies. That is, we treat the x values (our data) as fixed, and our candidate parameter estimates μ and σ are treated as variable quantities. Let us consider at a specific example to see how we define and use a likelihood function in practice.

Example: Assume all X_i are iid with normal density $f(x_i; \mu, \sigma)$. This implies the joint density $f_X(x_1, ..., x_n; \mu, \sigma) = \prod_{i=1}^{n} f(x_i, \mu, \sigma)$. Here we can write it as a simple product, thanks to the independence of the individual random variables. This density function defines the likelihood function for parmeters μ and σ

$$\mathcal{L}(\mu, \sigma; \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \mu, \sigma)$$

Note that we've gone from a function of n variables (number of data points) down to a function of 2 variables (number of parameters), and our domain is no longer the sample space but is instead the range of possible parameters ($\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$). Plotting likelihood values over a range of possible parameter values (here holding one parameter constant while varying the other) in **R** yields...

```
par(mfrow=c(1,2))
Lik=Vectorize(function(mu,sd,xs) prod(dnorm(xs,mean=mu,sd=sd)),"mu");
# fix sd=2, vary mu
curve(Lik(x,2,X),from=8,to=12, main="MEAN", xlab=expression(mu))
# fix mu, vary sd
Lik=Vectorize(function(mu,sd,xs) prod(dnorm(xs,mean=mu,sd=sd)),"sd");
curve(Lik(10,x,X),from=0,to=4, main="SD", xlab=expression(sigma))
# Optimization algorithms can then be used to refine estimates.
```



The *maximum likelihood estimates* of μ and σ are the pair of values that yield the maximum likelihood value. In this case, using the `optim()` function yields $\mu = 9.98$ and $\sigma = 1.88$.

Concluding Remark: In this example, we are inferring the parameters for a single distribution from our random sample of data. We do so by treating those data as a random vector -- a single observation from a multivariate distribution. We typically do statistics by *treating all of our data as a single outcome from a joint distribution.* Therefore, to have a deeper understanding of Statistics, we need to understand Probability from a multivariate perspective.

Random Vectors and Joint Densities

Joint densities describe probability distributions of *random vectors*. A random vector $\mathbf{X}$ is an n-dimensional vector where each component is itself a random variable, i.e., $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where all $X_i s$ are rvs.

Discrete random vectors are described by the *joint probability density function of* X_i (or joint pdf), $i = \{1, \cdots, n\}$ denoted by

$$P(X = x) = P(s \in S : X_i(s) = x_i \text{ for all } i) = p_X(x_1, .., x_n)$$

Another name for the joint pdf of a discrete random vector is *joint probability mass function (pmf).*

Computing probabilities for discrete random vectors. For any subset A of R^2, we have

$$P((X, Y) \in A) = \sum_{(x,y)\in A} P(X = x, Y = y) = \sum_{(x,y)\in A} p_{X,Y}(x, y).$$

Continuous random vectors are described by the *joint probability density function of X and Y (or joint pdf)* denoted by $f_{X,Y}(x, y)$. The pdf has the following properties:

1. $f_{X,Y}(x, y) \geq 0$ for every $(x, y) \in R^2$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$.
3. For any region A in the xy-plane $P((X, Y) \in A) = \int \int_A f_{X,Y}(x, y)dxdy$.

Marginal distributions. Let (X, Y) be a continuous/discrete random vector having a joint distribution with pdf/pmf $f(x, y)$. Then, the one-dimensional distributions of X and Y are called *marginal distributions*. We compute the marginal distributions as follows:
If (X, Y) is a discrete vector, then the distributions of X and Y are given by:

$$f_X(x) = \sum_{\text{all y}} P(X = x, Y = y) \text{ and } f_Y(y) = \sum_{\text{all x}} P(X = x, Y = y).$$

If (X, Y) is a continuous vector, then the distributions of X and Y are given by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

Joint cdf of a vector (X, Y). The joint cumulative distribution function of X and Y (or *joint cdf*) is defined by

$$F_{X,Y}(u, v) = P(X \leq u, Y \leq v).$$

Theorem. Let $F_{X,Y}(u, v)$ be a joint cdf of the vector (X, Y). Then the joint pdf of (X, Y), $f_{X,Y}$, is given by second partial deriveative of the cdf. That is $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$, provided that $F_{X,Y}(x, y)$ has continuous second partial derivatives.

Independence Revisited

Definition. Two random variables are called independent if and only if (*iff*) for any events A and B in S, it follows that $P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B)$.

Theorem. The random variables X and Y are independent iff

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

where $f(x, y)$ is the joint pdf of (X, Y), and $f_X(x)$ and $f_Y(y)$ are the marginal densities of X and Y, respectively.

NOTE: Random variables X and Y are independent iff $F_{X,Y}(x, y) = F_X(x)F_Y(y)$, where $F(x, y)$ is the joint cdf of (X, Y), and $F_X(x)$ and $F_Y(y)$ are the marginal cdf's of the X and Y, respectively.

Independence of more than 2 r.v.s A set of n random variables $X_1, X_2, \ldots, X_n$ are independent iff their joint pdf is a product of the marginal pdfs. That is

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)$$

where $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ is the joint pdf of the vector $(X_1, X_2, \ldots, X_n)$, and $f_{X_1}(x_1)$, $f_{X_2}(x_2), \cdots$, and $f_{X_n}(x_n)$ are the marginal pdf's of the variables $X_1, X_2, \ldots, X_n$.

Conditional Distributions Revisited

Let (X, Y) be a random vector with some joint pdf or pmf. Consider the problem of finding the probability that X=x **AFTER** a value of Y was observed. To do that we develop *conditional distribution* of X given Y=y.

Definition. If (X, Y) is a discrete random vector with pmf $p_{X,Y}(x, y)$, and if $P(Y = y) > 0$, then the *conditional distribution* of X given Y=y is given by the *conditional pmf*

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Similarily, if $P(X = x) > 0$, then the *conditional distribution* of Y given X=x is given by the *conditional pmf* $p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$.

Definition. If (X, Y) is a continuous random vector with pdf $f_{X,Y}(x, y)$, and if $f_Y(y) > 0$, then the *conditional distribution* of X given Y=y is given by the *conditional pdf*

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

Similarly, if $f_X(x) > 0$, then the *conditional distribution* of Y given X=x is given by the *conditional pdf* $f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.

Independence and conditional distributions. If random variables X and Y are independent, then their marginal pdf/pmf's are the same as their conditional pdf/pmf's. That is $f_{Y|X=x}(y) = f_Y(y)$ and $f_{X|Y=y}(x) = f_X(x)$, for all y and x where $f_Y(y) > 0$ and $f_X(x) > 0$, respectively.

Expected Values, Variance, and Covariance Revisited

Definition: Let (X, Y) be a random vector with pmf p (discrete) or pdf f (continuous). Let g be a real valued function of (X, Y). Then, the **expected value of random variable** $g(X, Y)$ is

$$E(g(X,Y)) = \sum_{allx}\sum_{ally} g(x,y)p(x,y), \text{ in the discrete case, or}$$

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy, \text{ in the continuous case,}$$

provided that the sums and the integrals converge absolutely.

Mean of a sum of random variables. Let X and Y be any random variables, and a and b real numbers. Then

$$E(aX + bY) = aE(X) + bE(Y),$$

provided both expectations are finite.

NOTE: Let $X_1, X_2, \ldots, X_n$ be any random variables with finite means, and let $a_1, a_2, \ldots, a_n$ be a set of real numbers. Then

$$E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n).$$

Mean of a product of independent random variables. If X and Y are independent random variables with finite expectations, then $E(XY) = E(X)\,E(Y)$.

Variance of a sum of independent random variables.
Let $X_1, X_2, \ldots, X_n$ be any independent random variables with finite second moments (i.e. $E(X_i^2) < \infty$). Then

$$Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n).$$

NOTE: Let $X_1, X_2, \ldots, X_n$ be any independent random variables with finite second moments, and let $a_1, a_2, \ldots, a_n$ be a set of real numbers. Then

$$Var(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2Var(X_1) + a_2^2Var(X_2) + \cdots + a_n^2Var(X_n).$$

Definition: The **covariance** of two jointly distributed *random variables* X and Y is defined as

$$cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

Sometimes the notation $\sigma(X, Y)$ is used instead of $cov(X, Y)$.

Definition: The **variance-covariance matrix** $cov(\mathbf{X}, \mathbf{Y})$ for *random vectors* $\mathbf{X}$ and $\mathbf{Y}$ is

$$cov(\mathbf{X},\mathbf{Y}) = E((\mathbf{X} - E(\mathbf{X}))\,(\mathbf{Y} - E(\mathbf{Y}))^T) = E(\mathbf{X}\mathbf{Y}^\mathbf{T}) - E(\mathbf{X})\,E(\mathbf{Y})^T$$

The ij^{th} entry in the covariance matrix is the covariance between X_i and Y_j. If $\mathbf{X} = \mathbf{Y}$, we often use the notation $cov(\mathbf{X}) = cov(\mathbf{X}, \mathbf{X})$. The diagonal entries of $\sigma(\mathbf{X})$ are $[Var(X_1),\ \ldots,\ Var(X_n)]$. The notation $\sigma(\mathbf{X}, \mathbf{Y})$ or sometimes $\Sigma(\mathbf{X}, \mathbf{Y})$ is often used for $cov(\mathbf{X}, \mathbf{Y})$.

Definition: The **correlation** *random variables* X and Y is a standardized covariance:

$$corr(X,Y) = \frac{E((X - E(X))(Y - E(Y)))}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{cov(X,Y)}{sd(X)\,sd(Y)}$$

Combining Random Variables: Sums, Products, Quotients

Let X and Y be **independent** random variables with pdf or pmf's f_X and f_Y or p_X and p_Y, respectively. Then...

If X and Y are discrete random variables, then the pmf of their **sum** W = X+ Y is

$$p_W(w) = \sum_{allx} p_X(x)p_Y(w-x).$$

If X and Y are continuous random variables, then the pdf of their **sum** W = X+ Y is the *convolution* of the individual densities:

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w-x)dx.$$

If X and Y are independent continuous random variables, then the pdf of their **quotient** $W = Y/X$ is given by:

$$f_W(w) = \int_{-\infty}^{\infty} \mid x \mid f_X(x)f_Y(wx)dx.$$

The above formula is valid, if X is equal to zero in at most a set of isolated points (no intervals).

If X and Y are independent continuous random variables, then the pdf of their **product** $W = XY$ is given by:

$$f_W(w) = \int_{-\infty}^{\infty} \frac{1}{\mid x \mid} f_X(w/x)f_Y(x)dx.$$

Special Distributions

Some useful distributions are "special" enough to be named. They include: Poisson, exponential, Normal (Gaussian), Gamma, geometric, negative binomial, Binomial and hypergeometric distributions. We already saw and used exponential, Binomial and hypergeometric distributions. We will now explore the definitions and properties of the other "special" distributions.

Things to remember when learning probability distributions:

Mathematical Details	Application Context
1. Is it **continuous**, or **discrete**? What's the sample space?	1. Corresponding experiment? Cartoon example?
2. Density(mass) function? Parameter ranges?	2. Any common applications? Why is it "special" or useful?
3. Expected value formula? Variance?	3. Relationships to other distributions?

Bernoulli. Sample space $\{0, 1\}$, mean p, variance $p(1-p)$, and mass function

$$p_x = p^x(1-p)^{1-x}$$

Binomial. The number of successes in n Bernoulli(p) trials. Discrete random variable with sample space $\{0, \ldots, n\}$, and mass function (with parameters n, p) given by

$$p_x = \binom{n}{k} p^x (1-p)^{n-x}$$

The mean is np and the variance is $np(1-p)$.

Multinomial (Generalized Binomial). Discrete random variable for the number of each of k types of outcomes in n trials. Sample space $\{0, ..., n\}^k$, and mass function (with parameters n,p_1,...,p_k where the $\sum p_i = 1$) given by

$$p_{x_1,\ldots,x_k} = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

The marginals are binomial, thus the means are $E(X_i) = np_i$ and the variances are $Var(X_i) = np_i(1-p_i)$.

Hypergeometric. Discrete r.v. with sample space $\{0, ..., w\}$, and mass function (with parameters N, w, n) given by

$$p_x = \frac{\binom{w}{x}\binom{N-w}{n-x}}{\binom{N}{n}}$$

The mean is $n\,w/N$ and the variance is $n\,w/N(1-w/N)(N-n)/(N-1)$.

Generalized Hypergeometric. Discrete random variable with parameters n, n_1, ..., n_k, $\sum n_i = N$, with mass function

$$p_{x_1,\ldots,x_k} = \frac{\binom{n_1}{x_1} \cdots \binom{n_k}{x_k}}{\binom{N}{n}}$$

The marginals are Hypergeometric.

Uniform (Continuous). Discrete Continuous random variable with sample space {}, and density function

$$f_x = (b-a)^{-1}$$

The mean is $(b+a)/2$ and the variance is $(b-a)^2/12$.

Poisson distribution. Discrete random variable with $\lambda > 0$, mass function

$$P(X=k) = \frac{e^{-\lambda}(\lambda^k)}{k!} \text{ for } k = 0, 1, 2, \ldots.$$

The mean and variance are the same, namely $E(X) = Var(X) = \lambda$.

Poisson Approximation to Binomial distribution. Let $X \sim Bin(n,p)$ be a binomial random variable with number of trials n and probability of success p. Then, for large n and small p, that is when $n \to \infty$ and $p \to 0$ in such a way that $np = \lambda$ is held constant, we have

$$\lim_{n\to\infty, p\to 0} P(X=k) = \frac{e^{-np}(np)^k}{k!} = \frac{e^{-\lambda}(\lambda)^k}{k!}.$$

For large values of n, small p, we can therefore approximate the Binomial distribution with a Poisson distribution: $P(X=k) \approx \frac{e^{-np}(np)^k}{k!}$.

Poisson Model. Suppose events can occur in space or time in such a way that:

1. The probability that two events occur in the same *small* area or time interval is zero.
2. The events in disjoint areas or time intervals occur independently.
3. The probability than an event occurs in a given area or time interval T depends only on the size of the area or length of the time interval, and not on their location.

Poisson Process. Suppose that events satisfying the Poisson model occur at the rate λ per unit time. Let X(t) denote the number of events occuring in time interval of length t. Then

$$P(X=k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}.$$

X(t) is called *Poisson process* with rate λ.

Exponential Continuing from above, the waiting time Y between consecutive events has an exponential distribution with parameter λ (that is with mean $1/\lambda$), that is $P(Y > t) = e^{-\lambda t}$, $t > 0$, or equivalently,

$$f(t) = \lambda e^{-\lambda t}, \text{ for } t > 0.$$

The mean is $1/\lambda$ and the variance is $1/\lambda^2$.

Geometric and Negative Binomial Distributions

Geometric experiment: Toss a fair coin until the first H appears. Let X=number of tosses required for the first H. Then X has geometric distribution with probability of success 0.5.

Definition: Geometric distribution. A random variable X has a geometric distribution with parameter p if its pmf is

$$P(X = k) = (1 - p)^{k-1}p, \text{ for } k = 1, 2, 3, \ldots.$$

It is denoted $X \sim Geo(p)$. The mean and variance of a geometric distribution are $EX = 1/p$ and $Var(X) = \frac{1-p}{p^2}$, respectively. The mgf of X is $M_X(t) = \frac{pe^t}{1-(1-p)e^t}$.

Memoryless property of geometric distribution. Let $X \sim Geo(p)$, then for any n and k, we have

$$P(X = n + k \mid X > n) = P(X = k).$$

Negative Binomial experiment. Think of geometric experiment performed until we get r successes. Let X = number of trials until we have r successes.

Definition: Geometric distribution. A random variable X has a negative binomial distribution with parameters r and p if its pmf is

$$p_X(k) = P(X = k) = \begin{pmatrix} k-1 \\ r-1 \end{pmatrix} (1 - p)^{k-r}p^r, \text{ for } k = r, r + 1, r + 2, \ldots.$$

A common notation for X is $X \sim NegBin(r, p)$. The mean and variance of X are $EX = r/p$ and $VarX = \frac{r(1-p)}{p^2}$. The mgf of X is $M_X(t) = \left[\frac{pe^t}{1-(1-p)e^t}\right]^r$.

Connection between negative binomial and geometric distributions. If $X_1, X_2, \ldots, X_r$ are iid $Geo(p)$, then $\sum_{i=1}^{r} X_i \sim NegBin(r, p)$.

Exponential and Gamma Distributions

Definition. The Gamma function. For any positive real number $r > 0$, the *gamma function* of r is denoted $\Gamma(r)$ and equal to

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

Theorem. Properties of Gamma function. The Gamma(r) function satifies the following properties:

1. $\Gamma(1) = 1$.
2. $\Gamma(r) = (r-1)\Gamma(r-1)$.
3. For r integer, we have $\Gamma(r) = (r-1)!$.

Definition of the Gamma random variable. For any real positive numbers $r > 0$ and $\lambda > 0$, a random variable with pdf

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \;\; x > 0,$$

is said to have a Gamma distribution with parameters r and λ, denoted $X \sim \Gamma(r, \lambda)$.
Theorem: moments and mgf of a gamma distribution. If $X \sim \Gamma(r, \lambda)$ then

1. EX= r/λ.
2. Var(X)= r/λ^2.
3. Mgf of X is $M_X(t) = (1 - t/\lambda)^r$.

Theorem. Let $X_1, X_2, \ldots, X_n$ be iid exponential random variables with parameter λ, that is with mean $1/\lambda$. The the sum of X_i's has a gamma distribution with parameters n and λ. More precisely, $\sum_{i=1}^n X_i \sim \Gamma(r, \lambda)$.

Theorem. A sum of independent gamma random variables $X \sim \Gamma(r, \lambda)$ and $Y \sim \Gamma(s, \lambda)$ with the same λ has a gamma distribution with $r' = r + s$ and the same λ. That is $X + Y \sim \Gamma(r + s, \lambda)$.

Note: In a sequence of Poisson events occurring with rate λ per unit time/area, the waiting time for the r'th event has a $\Gamma(r, \lambda)$ distribution.

Normal (Gaussian) Distribution

Normal (Gaussian) distribution. Continuous random variable X has a normal distribution with mean μ and variance σ^2 if its pdf is of the form:

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

where μ and σ^2 are real valued constants. If X has pdf as above, we denote it: $X \sim N(\mu, \sigma^2)$. The mgf of X is $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$, for any real t.
The normal pdf is bell shaped and centered around the mean μ. There is a special Normal distribution with mean 0 and variance 1, called standard normal distribution, and denoted by $Z \sim N(0, 1)$. The standard normal pdf is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

The values of the standard normal cdf are tabulated. To find probabilities related to general normal random variables, use the following fact:

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

Theorem: Linear combinations of independent normal r.v.s are themselves normal.

1. Let $X_1 \sim N(\mu_1, \sigma_1^2)$, and $X_2 \sim N(\mu_2, \sigma_2^2)$, with X_1 and X_2 independent. Then $X_1 \pm X_2 \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$, and more generally:
2. Let $X_i \sim N(\mu_i, \sigma_i^2)$, for $i = 1, \ldots, n$, and X_i's ind. Then $Y = \sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$, and
3. For any real numbers $a_1, a_2, \ldots a_n$, $Y = \sum_{i=1}^n a_i X_i \sim N(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2)$.
4. Let $X_i \sim N(\mu, \sigma^2)$ iid for $i = 1, \ldots, n$. Then $\bar{X} \sim N(\mu, \sigma^2/n)$.

Normal Approximation to Binomial. Let $X \sim Bin(n, p)$ and $Y \sim N(np, np(1-p))$. Then for large n

$$P(a \leq X \leq b) \approx P(a \leq Y \leq b).$$

Continuity correction for the normal approximation to binomial. To "correct" for the fact that binomial is discrete and normal is a continuous distribution, we do the following *correction for continuity*: $P(X = x) \approx P(x - 0.5 < Y < x + 0.5)$.

Convergence Concepts & Laws of Large Numbers

Before discussing the Central Limit Theorem (CLT), *Weak Law of Large Numbers* (WLLN) and *Strong Law of Large Numbers* (SLLN) it helps to know some different convergence concepts that exist in probability (and measure theory).

We begin with two results that help us bound probabilities when only the mean is known:

Markov Inequality: For any non-negative valued r.v. Y with $E(Y) = \mu$, then for $a > 0$

$$P(Y \geq a) \leq \frac{E(Y)}{a}.$$

Proof (finite-variance, continuous case):

$$\frac{E(Y)}{a} = \frac{1}{a}\int_0^\infty y\, f(y)\, dy \geq \frac{1}{a}\int_a^\infty y\, f(y)\, dy \geq \frac{1}{a}\int_a^\infty a\, f(y)\, dy = P(Y \geq a) \qquad \blacksquare$$

Chebychev Inequality: For r.v. X with $E(X) = \mu$ and $Var(X) = \sigma^2 < \infty$, then for any $k > 0$ the probability that X deviates more than k from the mean is bounded by

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Sketch of Proof: Apply the Markov Inequality using $Y = (X - \mu)^2$ and $a = k^2$.

Convergence Concepts in Probability

Definition: The r.v.s X_n **converge in distribution** to r.v. X ($X_n \xrightarrow{D} X$) if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all x where $F_X(x)$ is continuous. This is point-wise convergence of cdfs.

Definition: The r.v.s X_n **converge in probability** to r.v. X ($X_n \xrightarrow{P} X$) if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \quad \Leftrightarrow \quad \lim_{n\to\infty} P(|X_n - X| \leq \epsilon) = 1.$$

This convergence of probability values is in measure theory called *convergence in measure*.

Definition: The r.v.s X_n **converges almost surely** to r.v. X ($X_n \xrightarrow{a.s.} X$) if for all $\epsilon > 0$

$$P\left(\lim_{n\to\infty} |X_n - X| \leq \epsilon\right) \equiv P\left(\{\text{all } \omega \in S \text{ such that } \lim_{n\to\infty} |X_n(\omega) - X(\omega)| \leq \epsilon\}\right) = 1$$

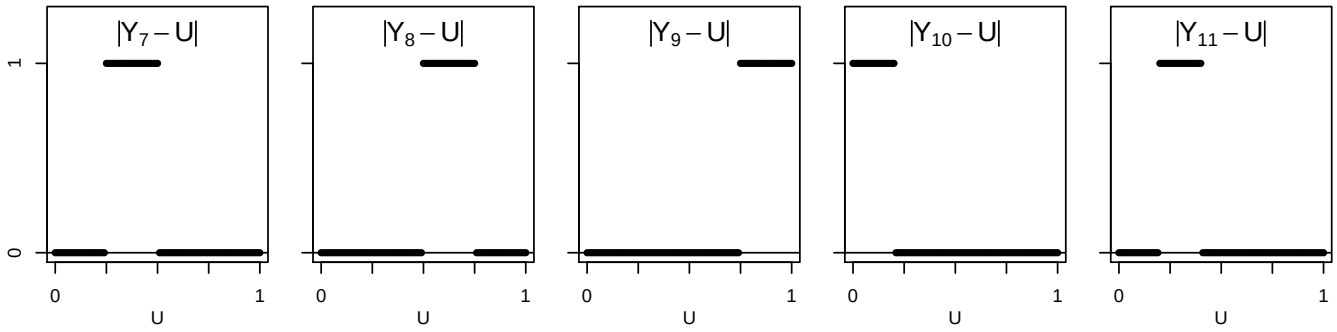
In measure theory, *almost everywhere* means a statement holds true for all but a set of measure zero. Thinking of random variables as functions on our sample space, this is just *pointwise convergence of the random variables* except perhaps on some set of measure zero.

Theorem: If X_n *converges almost surely* to X, then it also *converges in probability*. If X_n *converges in probability* to X, the it also *converges in distribution*.

Example 1 (Convergence in probability, but not almost surely.)
Let U be uniform on [0,1], and define the sequence of random variables Y_n to all depend directly on U according to $Y_n = U + \mathbb{1}_{A_n}(U)$ where intervals A_n are defined as the n^{th} interval in the sequence [0,1/2], [1/2,1], [0,1/3], [1/3,2/3], [2/3,1], [0,1/4],... That is, for observation $U = u$, $Y_n = u + 1$ if $u \in A_n$, otherwise $Y_n = u$. Note these r.v.s Y_n are not independent, since each depends directly on U! Observing that, as $n \to \infty$, the width of interval $A_n \to 0$, it follows that Y_n converges in probability to U since

$$\lim_{n\to\infty} P(|Y_n - U| \geq \epsilon) = \lim_{n\to\infty} P(U \in A_n) = 0.$$

But for a given outcome $U = u$, $Y_n(u)$ never converges since for any $N > 0$ there is always some $k > N$ where $Y_k(u) = 1 + u$.



Since $|Y_n - U|$ converges nowhere on [0,1],

$$P\left(\lim_{n\to\infty} |Y_n - U| \leq \epsilon\right) = P(\emptyset) = 0 \neq 1$$

That is, there is no almost sure convergence.

Example 2 (Convergence in distribution, but not in probability.)
Let X be a standard Normal r.v. ($E(X) = 0$, $Var(X) = 1$). Let $X_n = -X$ for all n. Then all X_n and X have the same distribution (i.e., $F_{X_n}(x) = F_X(x)$ for all x and n), so trivially X_n *converges in distribution* to X. However, for $\epsilon > 0$, symmetry gives that

$$P(|X_n - X| \geq \epsilon) = P(|2X| \geq \epsilon) = P(|X| \geq \epsilon/2) = P\left(X \notin \left[\frac{-\epsilon}{2}, \frac{\epsilon}{2}\right]\right).$$

Since $P\left(X \notin \left[\frac{-\epsilon}{2}, \frac{\epsilon}{2}\right]\right) > 0$ for all $\epsilon > 0$, it follows that

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = P\left(X \notin \left[\frac{-\epsilon}{2}, \frac{\epsilon}{2}\right]\right) > 0.$$

Therefore X_n does not converge in probability to X.

Laws of Large Numbers

Weak Law of Large Numbers (WLLN): Let X_i be iid with mean μ. Then $\overline{X_n} = \sum_{i=1}^{n} X_i$ *converges in probability* to μ, i.e., $\overline{X_n} \xrightarrow{P} \mu$. That is, for all positive ϵ near zero,

$$\lim_{n\to\infty} P(|\overline{X_n} - \mu| \geq \epsilon) = 0 \quad \Leftrightarrow \quad \lim_{n\to\infty} P(|\overline{X_n} - \mu| \leq \epsilon) = 1.$$

Proof (when $Var(X) = \sigma^2 < \infty$): Apply the Chebychev Inequality. This was first proven in the 1700s by Bernoulli, and incrementally generalized by Markov then Chebychev.

Strong Law of Large Numbers (SLLN): Let X_i be iid with mean μ, and let $\overline{X_n} = \sum_{i=1}^{n} X_i$. Then $\overline{X_n}$ *converges almost surely* to μ. That is, for all positive ϵ near zero,

$$P\left(\lim_{n\to\infty} |\overline{X_n} - \mu| \geq \epsilon\right) = 0 \quad \Leftrightarrow \quad P\left(\lim_{n\to\infty} |\overline{X_n} - \mu| \leq \epsilon\right) = 1.$$

NOTE: Borel gave the first proof of the SLLN, 200 years later, in 1909. It was incrementally improved by Cantelli, Khintchine (who named it the SLLN) and Kolmogorov (in the 1930s).

Weak vs Strong: Accordingly, *almost sure convergence* is called a *stronger* form of convergence than *convergence in probability*, and *convergence in distribution* is even more *weak*.

NOTE: The WLLN and SLLN basically both state that the average of n *iid* random variables (with mean $\mu < \infty$) converges to μ as $n \to \infty$. The Weak LLN states this in the *weaker* form ($X_n \xrightarrow{P} \mu$), while the Strong LLN states this in the (stronger) form ($X_n \xrightarrow{a.s.} \mu$).

Central Limit Theorems (CLTs)

Classic CLT (Lindberg-Levy): Suppose random variables X_1, ..., X_n are (1) *independent* and (2) *identically distributed* (*iid*) with (3) finite mean $E(X_i) = \mu$ and (4) finite variance $Var(X_i) = \sigma^2$. Then the quantity

$$S_n = \frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) - \mu\right)$$

converges *in distribution* to a standard Normal r.v., i.e., $S_n \xrightarrow{D} \mathcal{N}(0,1)$.

NOTE: Other CLTs relax the *iid* assumptions, but require additional conditions that must be met. The Lyapunov CLT, for example, relaxes the assumption of identical distributions:

CLT (Lyapunov): Suppose r.v.s X_1, ..., X_n are (1) *independent*, (2) each have finite mean $E(X_i) = \mu_i$ and (3) variance $Var(X_i) = \sigma_i^2$. Define $s_n = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$. Then the quantity

$$S_n = \frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i)$$

converges *in distribution* to a standard Normal *if* the following condition holds for some $\delta > 0$ (usually checking $\delta = 1$ is all it takes):

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E\left(|X_i - \mu_i|^{2+\delta}\right) = 0.$$