Notes on:

Probability and Statistics I

David Duan

Last Updated: September 27, 2023 (Draft)

Contents

Chapter 1

Probability and Distributions

Contents

1.1 Definitions

There are three main methods on defining probability.

Definition 1.1.1 In everyday life, **probability** is the measure of a person's belief in the occurrence of a future event.

This is an acceptable practical interpretation, but for statistics, we seek a better definition for a clearer understanding of how it can me measured and assists us in making inferences.

Definition 1.1.2 Probability is the **relative frequency** of an event happening, which is defined as the fraction of times an event occurs if it's repeated over and over infinitely.

The last method, which we will be focusing on for this course, is defined using **axioms of probability**.

1.2 Review of Set Theory

Definition 1.2.1

- We use capital letters to denote sets of objects: A, B, etc.
- S denotes the *universal set* which is the set of all possible objects.
- ϕ denotes the *empty set*.

Proposition 1.2.2 For any two sets A and B,

- A is a *subset* of B is denoted $A \subseteq B$.
- The *union* of A and B is denoted $A \cup B$, and the union of many sets can be written as:
$$A_1 \cup A_2 \cup \cdots = \bigcup_{i=1}^{\infty} A_i$$
- The *intersection* of sets A and B is denoted $A \cap B$.
- If $A \subseteq S$, then $A^c = \{x \in S : x \notin A\}$.
- Two sets A and B are said to be *disjoint* or *mutually exclusive* if $A \cap B = \phi$.

Theorem 1.2.3

- Distributive Laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cup C) = (A \cup B) \cap (A \cup C)$$

- De Morgan's Laws:

$$(A \cap B)^c = A^c \cup B^c$$

$$(A \cup B)^c = A^c \cap B^c.$$

- Distributive Law:

$$A \cup (B \cup C) = (A \cup B) \cup C.$$

$$A \cap (B \cap C) = (A \cap B) \cap C.$$

For non-decreasing or non-increasing sets, we have:

Theorem 1.2.4 A sequence of events $\{C_n\}$ is an increasing(resp. decreasing) sequence if

$$C_n \subset C_{n+1} \quad \text{resp. } (C_{n+1} \subset C_n)$$

for all n, in which we write

$$\lim_{n\to\infty} C_n = \bigcup_{n=1}^{\infty} C_n \quad \text{resp. } (\lim_{n\to\infty} C_n = \bigcap_{n=1}^{\infty} C_n)$$

A analogous definition for *nonincreasing* and *nondecreasing* arises where we replace the strict subset symbol with non-strict subset symbol $\subseteq$.

1.3 The Probability Set Function

A **random experiment** is a process of observation which leads to a random outcome, the set of all possible outcomes are called the *sample space*, and each point in this set is a *sample point*. For now we will simply focus on experiments whose sample space is *finite* or *countably infinite*.

Definition 1.3.1 A **discrete sample space** is one that contains countable number of distinct sample points.

Typically, *events* are just possible outcomes of an experiment, but for experiments with discrete sample spaces, we can utilize certain concepts from set theory.

Definition 1.3.2 An **event** in a discrete sample S is a collection of sample points, i.e. any subset of S.

- A **simple event** is an event that cannot be decomposed, i.e. the event corresponds to one and only one possible *sample point*.
- A **compound event** consists of two or more simple events.

Example 1.3.3 In the six-sided dice tossing experiment,

$$S = \{E_1, E_2, E_3, E_4, E_5, E_6\},$$

where each simple event E_i represents "observe number" i, notice each simple event corresponds to a sample point.The events "observe a 1", "observe a 3", "observe a 5", are all simple events while an event like "observe an odd outcome" is a compound event since it is composed of three simple events E_1, E_2, E_3.

On analyzing the relative frequency of events, we notice that three conditions must hold, these properties are so important we call them the **probability axioms**.

Definition 1.3.4 [Probability Axioms]
Suppose S is a sample space. To every event A in S, we assign a number $P(A)$, called the probability of A, so that the following axioms hold:

- $P(A) \geq 0$
- $P(S) = 1$
- If $A_1, A_2, A_3, ...$ are a sequence of pairwise mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + \cdots$$

1.4 Combinatorial tools

Theorem 1.4.1 [Multiplication rule/mn rule]
If we have two sets with m elements in one set and n elements in the other, then it is possible to form $m \times n$ pairs containing one item from each set.

Remark 1.4.2 The multiplication rule can be extended to any number of sets. That is, given n sets each with $k_1, ..., k_n$ elements, the number of unique combinations where we pick one element from each set is $k_1 \cdots k_n$.

In some instances, it is useful to know the number of distinct ways that a set of elements can be arranged in sequences, this brings us to another useful combinatorial result.

Theorem 1.4.3 [Permutation and Combinations]
Suppose we have n distinct objects. An *ordered* arrangement of these objects are called **permutations**. The number of permutations for r items is

$$P_r^n = \frac{n!}{(n-r)!}$$

The number of unordered subsets of r objects out of n is called the number of **combinations**, it can be found using

$$C_r^n = \binom{n}{r} = \frac{P_r^n}{r!} = \frac{n!}{r!(n-r)!}$$

The next result can be used to determine the number of subsets of various sizes that can be formed by partitioning a set into non-overlapping groups.

Theorem 1.4.4 [Combinatorics: Extension]
To find the number of ways of partitioning n distinct objects into k distinct groups containing $n_1, n_2, ..., n_k$ objects, respectively, where each object appears in exactly one group and $\sum_{i=1}^{k} n_i = n$, is

$$\binom{n}{n_1, ..., n_k} = \frac{n!}{n_1!...n_k!}$$

Theorem 1.4.5 [Continuity Theorem of Probability]
Let $\{C_n\}$ be a increasing sequence of events. Then

$$\lim_{n\to\infty} P(C_n) = P(\lim_{n\to\infty} C_n) = P\left(\bigcup_{n=1}^{\infty} C_n\right)$$

Let $\{C_n\}$ be a decreasing sequence of events. Then

$$\lim_{n\to\infty} P(C_n) = P(\lim_{n\to\infty} C_n) = P\left(\bigcap_{n=1}^{\infty} C_n\right)$$

1.5 Conditional Probability and Independence

The concept of *conditional probability* comes from the probability of an event happening based on our knowledge of other events which have occurred.

Example 1.5.1 Suppose we toss a fair dice, the unconditional probability of it landing on a 3 is 1/6, but if we know that an odd number has fallen,t hen the conditional probability of landing on a 3 becomes 1/3.

Definition 1.5.2 The *conditional probability* of an event A, given that an event B has occurred, is equal to

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

provided $P(B) > 0$.

Now suppose that the probability of the occurrence of an event A is unaffected by whether another event B has occurred, when this happens, we say that the two events A and B are independent.

Definition 1.5.3 Two events A and B are *independent* if any one of the following holds:

- $P(A|B) = P(A)$,
- $P(B|A) = P(B)$,
- $P(A \cap B) = P(A)P(B)$.

Otherwise, the events are *dependent*.

Example 1.5.4 Consider the following events for a single toss of a fair dice:

- A = Observe an odd number,
- B = Observe an even number,
- C = Observe a 1 or a 2.

Are A and B independent? No, because

$$P(A \cap B) = 0 \quad \text{while} \quad P(A)P(B) = \frac{1}{4}$$

Are A and C independent? Yes, as

$$P(A \cap C) = \frac{1}{6} \quad \text{and} \quad P(A)P(B) = \frac{1}{6}$$

1.6 Twos Laws of Probability

The first two laws gives the probabilities of unions and intersections of events.

Theorem 1.6.1 [Additive Law of Probability]
The probability of the union of two events A and B is

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

If A and B are mutually exclusive, then

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

Corollary 1.6.2 For an event A,

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

Proof. Since $A \subseteq S$, then $A^c \subseteq S$ and $A \cup A^c = S$. Then by second probability axiom we have

$$P(A \cup A^c) = 1$$

by properties of complement we know A and A^c are mutually exclusive, hence

$$P(A \cup A^c) = P(A) + P(A^c) \implies P(A) = 1 - P(A^c).$$

Notice that the additive law can be extended to k events by repeatedly applying the above theorem.

Theorem 1.6.3 [Inclusion-Exclusion Formula]
The probability of the union of k events, $A_1, ..., A_k$ is

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots + (-1)^{k-1} P(A_1 \cap \cdots \cap A_k)$$

Intuitively, we subtract the overlaps but have to add back things in the union which we've subtracted twice.

The multiplicative law of probability gives the probability of the intersection of two events.

Theorem 1.6.4 [The Multiplicative Law of Probability]
The probability of the intersection of two events A and B is

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

If A and B are independent, then

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

Similarly, the multiplicative law can be extended the find the probability of the intersection of any number of events. By twice applying the theorem,

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

And for k events, we apply it k times and see that

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1)P(A_2|A_1) \cdots P(A_k|A_1 \cap \cdots \cap A_{k-1}).$$

1.7 Bayes' Rule

Often times, it is useful to view a sample space as the union of mutually exclusive subsets, this brings us to the definition of a *partition*.

Definition 1.7.1 For some positive integer k, let the sets $B_1, B_2, ..., B_k$ be such that

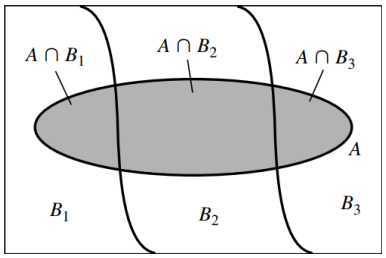
$$S = B_1 \cup \cdots \cup B_k \quad \text{and} \quad B_i \cap B_j = \emptyset, \text{ for } i \neq j.$$

Then the collection of sets $\{B_1, ..., B_k\}$ is a *partition* of S.

Now, given any subset A of S, if $\{B_1, ..., B_k\}$ is a partition of S, then A can be decomposed into

$$A = (A \cap B_1) \cup \cdots \cup (A \cap B_k)$$

Figure below illustrates this decomposition for $k = 3$.



This leads us to the **law of total probability.**

Theorem 1.7.2 Assume that $\{B_1, .., B_k\}$ is a partition of S such that $P(B_i) > 0$ for $i = 1, ..., k$. Then for any event A

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

Proof. Since A can be written as $A = (A \cap B_1) \cup \cdots \cup (A \cap B_k)$ and each bracket is disjoint with all other,

$$\begin{aligned} P(A) &= P(A \cap B_1) + \cdots + P(A \cap B_k) \\ &= P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k) \\ &= \sum_{i=1}^{k} P(A|B_i)P(B_i) \end{aligned}$$

Using this result, we can derive a famous result known as Bayes' Theorem

Theorem 1.7.3 [Bayes' Theorem]
Assume that $\{B_1, ..., B_k\}$ is a *partition* of S such that $P(B_i) > 0$ for $i = 1, ..., k$. Then

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

Intuitively, Bayes' Theorem is used when we know the probability of each B_i of a partition $\{B_1, ..., B_k\}$, and subsequently observe the conditional probability of some other event A happening given each of B_1,B_2,..., or B_k. From this, we can use Bayes' to calculate any one of the invert conditional probabilities $P(B_j|A)$ for some j.

1.8 Random Variables

Definition 1.8.1 For a random experiment with sample space S, a *random variable* is a real-valued function with domain S.

Remark 1.8.2 A random variable is simply the process of assigning a numerical value to each outcome of an experiment for ease of making calculations and inferences. They are called random variables become their input is the outcome of a random experiment.

There are two types of random variables which we will be focusing on: **discrete** and **continuous**.

1.9 Discrete Random Variables

Definition 1.9.1 A random variable is said to be *discrete* if it can assume only a finite or countably infinite number of distinct values.

Notationally, we will use capital letters, such as X, Y, Z to denote a *random variable*, and lowercase letters like x, y, z to denote a *particular value* that a random variable may assume. Furthermore, the expression $(X = x)$ can be understood as *the set of app points in S that is assigned the value x by the random variable X*. With this in mind, the notation $P(X = x)$ now makes sense.

Definition 1.9.2 The probability of X taking on the value x, denoted $P(X = x)$, is defined as the *sum of probabilities of all sample points in S* that are assigned to the value y. Someone we will write $p_X(x)$ instead of $P(X = x)$.

Because $p(x)$ is a function which assigns probabilities to each value x of the random variable X, it is usually called the *probability function* of X.

Definition 1.9.3 The *probability distribution* of a discrete random variable X can be represented by a formula, a table, or a graph that provides $p(x) = P(X = x)$ for all x.

Remark 1.9.4 Notice that $P(X = x) \geq 0$ for all x, but the probability distribution only assigns nonzero probabilities to only a countable number of distinct x values. And any value x that is not explicitly assigned a positive probability is assumed to have $P(X = x) = 0$.

For any discrete random variables, the *probability distribution* is called the **probability mass function (pmf)** and is denote $p_X(x)$, when there is no ambiguity, we may remove the subscript and just write $p(x)$.

Theorem 1.9.5 For any *discrete probability distribution*, the following must be true:

- $0 \leq p(x) \leq 1$ for all x.
- $\sum_x p(x) = 1$, where the summation is over all values of y with nonzero probability.

1.10 Cumulative Distribution Function

Definition 1.10.1 Let X denote any random variable. The *cumulative distribution function (cdf)* of X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P(X \leq x), \quad \text{for} \quad -\infty < x < \infty.$$

The subscript X can be removed if no ambiguity involved.

Proposition 1.10.2 If $F(x)$ is a cdf, then

- $F(-\infty) = \lim_{x \to -\infty} P(X \leq x) = \lim_{x \to -\infty} F(x) = 0.$
- $F(\infty) = \lim_{x \to \infty} P(X \leq x) = \lim_{x \to \infty} F(x) = 1.$
- $F(x)$ is a right-continuous non-decreasing function of x.

Remark 1.10.3 The definition for a cdf is the same for both *discrete* and *continuous random variables*. However, the cdf for a *discrete random variable* is a step function, while the cdf of a *continuous random variable* is a **continuous** non-decreasing line

1.11 Discrete Transformations

Suppose we have some discrete random variable X with known distribution, and we are interested in a random variable Y which is some transformation of X, say, $Y = g(X)$. How do we determine the distribution of Y? Assume X has suppose S_X, then Y has support $S_Y = \{g(x) : x \in S_X\}$. If g is bijective, then the pmf of Y can be obtained easily by

$$p_Y(y) = P(Y = y) = P[g(X) = y] = P[X = g^{-1}(y)] = p_X(g^{-1}(y)).$$

If g is not bijective, instead of developing an overall rule, we usually obtain the pmf of Y in a straightforward manner by finding patterns or brute-forcing.

Example 1.11.1 Suppose X have the pmf $p_X(x) = \frac{1}{3}$ for $x = -1, 0, 1$ and zero elsewhere, and we want to find the pmf of $Y = X^2$. Note $g(x) = x^2$ is not a bijective function on the domain as $g(-1) = g(1)$, we instead use the fact that $S_Y = \{0, 1\}$, so

$$p_Y(1) = P(X = 1) + P(X = -1) = \frac{2}{3} \quad \text{and} \quad p_Y(0) = P(X^2 = 0) = P(X = 0) = \frac{1}{3}$$

1.12 Continuous Random Variable

Continuous random variables can take on an uncountable infinite number of values, think of people's height.

Definition 1.12.1 A random variable X is *continuous* if and only if its cdf $F(x)$ is continuous for $-\infty < x < \infty$.

Remark 1.12.2 If X is a continuous random variable, then for any real number x,

$$P(X = x) = 0.$$

If this were not true, then $P(X = x_0) = p_0 > 0$ and so $F(x)$ would have a jump discontinuity of size p_0 at x_0 and so $F(x)$ would not be continuous. Intuitively, we should not be bothered with measure the probability of a continuous random variables at discrete point but rather on intervals.

The derivative of $F(x)$ is another important function.

Definition 1.12.3 Let $F(x)$ be the cdf of a continuous random variable X. Then $f(x)$, given by $f(x) = \frac{dF(x)}{dx} = F'(x)$ wherever the derivative exists, is called the *probability density function (pdf)* for the random variable X.

Corollary 1.12.4 From the above definition, we see that

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

Because of the properties that the cdf $F(x)$ holds, we can deduce two properties for the pdf $f(y)$ which is similar to theorem 1.9.5 in the discrete case.

Theorem 1.12.5 For any *continuous probability distribution*, the following must be true:

- $f(x) \geq 0$ for all x, $-\infty < y < \infty$,
- $\int_{-\infty}^{\infty} f(x)dx = 1$.

To calculate probabilities of a continuous random variable between a interval $P(a < X \leq b)$, notice that

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) = \int_{a}^{b} f(x)dx,$$

and since $P(X = a) = P(X = b) = 0$, we indeed have $P(a \leq X \leq B) = P(a < X < b)$.

Theorem 1.12.6 If a random variable X has pdf $f(x)$ and $a < b$, then the probability that X falls in the interval $[a, b]$ is

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

1.13 Quantiles and percentiles

Definition 1.13.1 Let X denote any random variable. If $0 < p < 1$, the pth *quantile* of X, denoted by ϕ_p, is the smallest value such that $P(X \leq \phi_p) = F(\phi_p) \geq p$. If X is continuous, then ϕ_p is the smallest value such that $F(\phi_p) = p$. Some prefer to call ϕ_p is the $100p$th *percentile* of X.

The most important case is $\phi_{0.5}$ which is the *median* or the *second quantile* of the random variable. We also have $\phi_{0.25}$ and $\phi_{0.75}$ which can also called the first, and third quantile. Lastly, the difference $iq = q_3 - q_1$ is called the *inter-quartile range* of X which measures the *spread* or *dispersion* of the distribution of X.

1.14 Continuous Transformations

If X is a continuous random variable with a known pdf f_X, if g is bijective, then we can find the pdf of a random variable $Y = g(X)$ by first obtaining its cdf.

Theorem 1.14.1 Let X be a continuous random variable with pdf $f_X(x)$ and support S_X. Let $Y = g(X)$, where $g(x)$ is a bijective differentiable function on S_X. Denote the inverse of g by $x = g^{-1}(y)$ and let $\frac{dx}{dy} = \frac{d[g^{-1}(y)]}{dy}$. Then the pdf of Y is given by

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}|, \quad \text{for} \quad y \in S_Y$$

where the support of Y is $S_Y = \{y = g(x) : x \in S_X\}$.

Remark 1.14.2 We refer to $\frac{dx}{dy}$ as the **Jacobian** of the transformation and the formula is derived from change of variables.

1.15 Expected values of Random Variable

Definition 1.15.1 Let X be a random variable. If X is discrete with pmf $p(x)$ and $\sum_x |x|p(x) < \infty$, then the expectation of X is

$$E(X) = \sum_x xp(x)$$

If X is continuous with pdf $f(x)$ and $\int_{-\infty}^{\infty} |x|f(x) < \infty$, then the expectation of X is

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Remark 1.15.2 The expectation $E(X)$ of a random variable is also called the *expected value* or *mean* of X, and is usually denoted by μ.

Theorem 1.15.3 [Expected Value of a Function of a Random Variable]
Let X be a random variable, and let $Y = g(X)$ for some real-valued function g.
If X is discrete with pmf $p_X(x)$ and support S_X, moreover $\sum_{x\in S_X} |g(x)|p_X(x) < \infty$, then the expectation of Y exists and is given by

$$E(Y) = \sum_{x\in S_X} g(x)p_X(x).$$

If X is continuous with pdf $f_X(x)$ and $\int_{-\infty}^{\infty} |g(x)|f_X(x) < \infty$, then the expectation of Y exists and is given by

$$E(Y) = \int_{-\infty}^{\infty} g(x)f_X(x).$$

Theorem 1.15.4 [Expectation of a Constant] Let X be any random variable and c be a constant. then $E(c) = c$.

Theorem 1.15.5 [Linearity of Expectations]
Let $g_1(X)$ and $g_2(X)$ be functions of a random variable X. If $E(g_1(X))$ and $E(g_2(X))$ exist. Then for any constants k, the expectation of $E(g_1(X) + kg_2(X))$ exists and is given by

$$E[g_1(X) + kg_2(X)] = E[g_1(X)] + kE[g_2(X)].$$

1.16 Special Expectations

Definition 1.16.1 For a random variable X with expected value μ, the *variance* of X is defined as the expectation of $(X-\mu)^2$. That is,

$$Var(X) = E[(X-\mu)^2] = E(X^2) - \mu^2.$$

Variance is also commonly noted as σ^2.

Remark 1.16.2 The variances measured how "spread out" the values are of a random variable around the mean.

Definition 1.16.3 [Standard Deviation]
The *standard deviation* of a random variable is the positive square root of the variance and is denoted σ, that is,

$$\sigma = \sqrt{Var(X)} = \sqrt{\sigma^2}.$$

Variance is not a linear function, however we can develop a nice formula using *linearity of expectations*.

Theorem 1.16.4 Let X be a random variable with finite mean μ and variance σ^2. Then for all constants a and b,

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

1.17 Moment Generating Functions

Before introducing what a moment generating function is, we need to know what is a moment. Parameters μ and σ are meaningful numerical measures associated with a random variable X. However, they do not provide a uniquely identify the distribution of X. In this section, we will discover a set of measures that does uniquely determine $p(y)$.

Definition 1.17.1 The *kth moment of a random variable X taken about the origin* is defined to be $E(Y^k)$ and is denoted by μ'_k.

In particular, notice that the first moment about about the origin, $E(Y)$ is the expected value, and the second moment $E(Y^2)$ is employed in finding σ^2.

Definition 1.17.2 the kth *moment of a random variable* X *taken about its mean*, or the kth *central moment of* X, is defined to be $E[(X-\mu)^k]$ and is denoted by μ_k.

Notice that σ^2 is the 2nd central moment.
The moments μ'_k, opposed to mean and variance, can be used to show that two random variables X and Y have identical probability distributions, under some fairly general conditions. So a major use of moments is to approximate the probability distribution of a random variable.
With this, we can introduce the *moment-generating function*, which essentially, just packages all the moments into a single

Definition 1.17.3 The *moment-generating function* $m(t)$ for a random variable X is defined to be $m(t) = E(e^{tX})$. The mgf of X exists if there exists a positive constant b such that $m(t)$ is finite for $|t| \leq b$.

The *moment-generating function* possesses two important applications. First, if the mgf of a random variable X exist, then we can find any of the moment of X.

Theorem 1.17.4 If $m(t)$ exists, then for any positive integer k,

$$\frac{d^k m(t)}{dt^k}\bigg|_{t=0} = m^{(k)}(0) = \mu'_k$$

In other words, if you find the kth derivative of $m(t)$ with respect to t and set it to 0 then the result will be $E(X^k)$.

Proof (Sketch). Expanding e^{tx} by its Taylor polynomial, we have

$$e^{tx} = 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \frac{(tx)^4}{4!} + \cdots .$$

and find then

$$m(t) = E(e^{tX}) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots .$$

it follows that

$$m^{(1)}(t) = t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots ,$$

$$m^{(2)}(t) = \mu'_2 + \frac{t^2}{2!}\mu'_3 + \frac{t^3}{3!}\mu'_4 + \cdots ,$$

and in general

$$m^{(k)}(t) = \mu'_k + \frac{t^2}{2!}\mu'_{k+1} + \frac{t^3}{3!}\mu'_{k+2} + \cdots .$$

Now setting $t = 0$ in each of the case we see that that $m^{(k)}(0) = \mu'_k$.

1.18 Important inequalities

Theorem 1.18.1 [Markov Inequality]
Let $u(X)$ be a nonnegative function of the random variable X. If $E[u(X)]$ exists, then for every positive constant c,

$$P(u(X) \geq c) \leq \frac{E[u(X)]}{c}.$$

Proof. We'll prove this for continuous random variables, but the proof for the discrete is essentially the same. Let $f(x)$ be the pdf of X and fix positive c, let $A = \{x : u(x) \geq c\}$, we have

$$\begin{aligned} E(u(X)) &= \int_A u(x)f(x)dx + \int_{A^c} u(x)f(x)dx \\ &\geq c\int_c^\infty f(x)dx + 0 \\ &= cP(u(x) \geq c) \end{aligned}$$

Divide by c on both side gives us the desired equation.

The preceding theorem is a generalization of an inequality which is called **Chebyshev's Inequality**.

Theorem 1.18.2 [Chebyshev's Inequality] Let X be a random variable with mean μ and finite variance σ^2. Then, for any constant $k > 0$,

$$P(|X - u| < k\sigma) \geq 1 - \frac{1}{k^2} \quad \text{or} \quad P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. Using Markov Inequality, take $u(X) = (X - \mu)^2$ and $c = k^2\sigma^2$. Then we have

$$P[(X - \mu)^2 \geq k^2\sigma^2] \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2}.$$

Since the numerator of the right-hand side is σ^2, it can be written as

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

which is the desired result. Naturally, we would take the positive number k to be greater than 1 to have an inequality of interest.

An alternative proof not using Markov Inequality is provided below.

Proof. The proof for a continuous random variable will be given below, but the proof for the discrete case will be very similar. Let $f(x)$ be the pdf of X. Then

$$\begin{aligned} V(X) = \sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\ &= \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)dx. \end{aligned}$$

The second integral is always greater than or equal to zero, moreover, $(x-\mu)^2 \geq k^2\sigma^2$ for all values of x between the limits of integration for the first and third integral, as when $x = \mu \pm k\sigma$,

$$[(\mu - k\sigma) - \mu]^2 = k^2\sigma^2 \geq k^2\sigma^2.$$

Hence,

$$V(X) = \sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x)dx = k^2\sigma^2\Big[\int_{-\infty}^{\mu-k\sigma} f(x)dx + \int_{\mu+k\sigma}^{\infty} f(x)dx\Big]$$

or

$$\sigma^2 \geq k^2\sigma^2[P(X \leq \mu - k\sigma) + P(X \geq \mu + k\sigma)] = k^2\sigma^2 P(|X-\mu| \geq k\sigma).$$

Dividing by $k^2\sigma^2$, we get

$$P(|X-u| < k\sigma) \geq 1 - \frac{1}{k^2} \quad \text{or} \quad P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

which are the desired equations.

Remark 1.18.3 This result applies for any probability distribution, whether it is normal or not. However, the results of the theorem are very conservative, that is, the bound could be quite far away from the actual probability.

Definition 1.18.4 [Convex]
A function ϕ defined on an interval (a,b), $-\infty \leq a < b \leq \infty$, is *convex* if and only if for all $x, y \in (a,b)$ and for all $0 < \gamma < 1$,

$$\phi[\gamma x + (1-\gamma)y] \leq \gamma\phi(x) + (1-\gamma)\phi(y).$$

If the inequality is strict, then ϕ is *strictly convex*.

Theorem 1.18.5 If ϕ is convex on an open interval I and X is a random variable such that $S_X \subseteq I$ with finite expectation, then

$$\phi[E(X)] \leq E[\phi(X)]$$

If ϕ is strictly convex, then the inequality is strict unless X is a constant random variable.

Chapter 2

Multivariate Distributions

Contents

2.1 Distributions of Two Random Variables

Definition 2.1.1 Give a random experiment with a sample space S, consider two random variables X_1 and X_2