STAT 160A

Elements of Probability and Statistical Theory

1 Probability and Distribution

Introduction

- A phenomenon is *random* if individual outcomes of the experiment are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

- Three key words for describing random phenomenon:
  experiment– any procedure that (1) can be repeated, theoretically, an infinite number of times; and (2) has a well-defined set of possible outcomes.
  sample space– set of all possible outcomes of an experiment. Denoted by $\mathcal{C}$.
  event– any subset of the sample space $\mathcal{C}$, denoted by C. An event is said to occur if the outcome of the experiment is in C.

- Probability theory is a mathematical model for random phenomenon

Examples

1. Flip a coin one time.
   - It is a experiment: (1) coin may be repeatedly tossed under the same conditions and (2) only two possible outcomes
   - Sample space $\mathcal{C}$={T,H}
   - Event: $C_1 = \{T\}$,$C_2 = \{H\}$
2. Flip a coin three times.
   - Sample space $\mathcal{C}$={TTT,HTT,THT,TTH,HHT,HTH,THH,HHH}
   - Event: majority of coin show heads C={HHT,HTH,THH,HHH}
3. Flip a coin until the first tail appears.
   - Sample space $\mathcal{C}$={HT,HHT,HHHT,HHHHT,....}

Set Theory

A ***set*** is a collection of objects under consideration, sometime it is also called the ***space***. If an object x belongs to a set $\mathcal{C}$, it is said to be an ***element*** of the set, denoted by $x \in \mathcal{C}$.

- Definition 1.2.1 (subset)

  If for every $x \in C_1$ it is also true that $x \in C_2$, then the set C_1 is called a **subset** of the set C_2, denoted by $C_1 \subset C_2$. If $C_1 \subset C_2$ and $C_2 \subset C_1$, then $C_1 = C_2$.

- Definition 1.2.2 (null set)

  If a set C has no elements, C is called the **null set** and denoted by $C{=}\phi$.

- Definition 1.2.3 (union)

  The **union** of C_1 and C_2, written $C_1{\cup}C_2$, is the set of all elements belong to either C_1 or C_2 or both.

- Definition 1.2.4 (intersection)

  The **intersection** of C_1 and C_2, written $C_1 \cap C_2$, is the set of all elements belong to both C_1 and C_2.

  Note: The union and intersection of multiple sets are defined in a similar manner

- Definition 1.2.6 (Complement)

  If $C \subset \mathcal{C}$, then the **complement** of C consists of all elements of $\mathcal{C}$ that are not elements of C, denoted by C^c. In particular, $\mathcal{C}^{\mathsf{c}} = \phi$.

- Mutually exclusive sets

  Two sets A and B are said to be mutually exclusive if they have no elements in common - that is $A \cap B = \phi$.

- DeMorgan's Laws

  $(C_1 \cap C_2)^c = {C_1}^c \cup {C_2}^c$

  $(C_1 \cup C_2)^c = C_1^c \cap C_2^c$

- Manipulating sets

  Commutative law: $A \cup B = B \cup A$, $A \cap B = B \cap A$

  Associative law:

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

- Some examples of *set function*.

- Integral and Sum

  integral over a one(two)-dimensional set C:

  $\int_C f(x)dx$ $(\int \int_C g(x, y)dxdy)$

  sum extended over all $x \in C$ $((x, y) \in C)$:

  $\sum_C f(x)(\sum \sum_C g(x, y))$

- Exercise

  1.2.1,1.2.2,1.2.4,1.2.5,1.2.8,1.2.13,1.2.14

The Probability Set Function

- Frequency interpretation of probability

The *relative frequency* f_C of an event C is a proportion measuring how often, or how frequently, the event occurs in an experiment repeated N times. That is, $f_C = \#\{C\}/N$.

Example (page 2): In the cast of one red die and one while die, let $\mathcal{C}$ denote the sample space consisting of the ordered pairs. Let C be event that the sum of the pair is equal to seven. Suppose that the dice are cast N=400 times and f_C=60. Then the relative frequency is 0.15.

Note: Three properties of relative frequency:

(1)$f_C \geq 0$.

(2)$f_C \leq 1$.

(3)Suppose that A and B are mutually exclusive events, then $f_{\mathtt{A}\cup\mathtt{B}} = f_{\mathtt{A}} + f_{\mathtt{B}}$

The frequency interpretation of **probability** is that the probability of an event C is the expected relative frequency of C in a large number of trials. In symbols, the proportion of times occurs in trials, call it $P_n(C)$, is expected to be roughly equal to the theoretical probability $P(C)$ if n is large:$P_n(C) \approx P(C)$ for large n.

Example: Observation of the sex of a child. The sample space is C={boy,girl}. The following table shows the proportion of boys among live births to residents of the U.S.A. over the past 13 years. The relative frequency of boys among newborn children in the U.S.A. appears to be stable at around 0.512. This suggests that a reasonable model for the outcome of a single birth is P(boy)=0.512 and P(girl)=0.488.

Year	Number of Births	Proportion of Boys
1990	4,158,212	0.5121179
1991	4,110,907	0.5112054
1992	4,065,014	0.5121992
1993	4,000,240	0.5121845
1994	3,952,767	0.5116894
1995	3,926,589	0.5084196
1996	3,891,494	0.5114951
1997	3,880,894	0.5116337
1998	3,941,553	0.5115255
1999	3,959,417	0.5119072
2000	4,058,814	0.5117182
2001	4,025,933	0.5111665
2002	4,021,726	0.5117154

- σ-Field:

Let $\mathcal{B}$ be a collection of subsets of sample space $\mathcal{C}$. $\mathcal{B}$ is a σ-Field if

(1) $\phi \in \mathcal{B}$

(2) If $C \in \mathcal{B}$ then $C^c \in \mathcal{B}$

(3) If the sequence of sets $\{C_1, C_2, ...\}$ is in $\mathcal{B}$ then $\bigcup_{i=1}^{\infty} C_i \in \mathcal{B}$.

Note:

(1)σ-Field always contains ϕ and sample space $\mathcal{C}$

(2)σ-Field is also closed under countable intersection

Some examples of σ-Field: Let $\mathcal{C}$ be any set

1. Let $C \subset \mathcal{C}$, $\mathcal{B} = \{C, C^c, \phi, \mathcal{C}\}$ is a σ-Field.
2. $\mathcal{B}$=\{the collection of all subsets of $\mathcal{C}$\} is a σ-Field.
3. Suppose $\mathcal{D}$ is a nonempty collection of subsets of $\mathcal{C}$,
$\mathcal{B} = \cap\{\varepsilon : \mathcal{D} \subset \varepsilon$ and ε is a σ-Field\} is a σ-Field.
4. Let $\mathcal{C}$=R, where R is the set of all real numbers. Let $\mathcal{I}$ be the

set of all open intervals in R. $\mathcal{B}_0 = \cap\{\varepsilon : \mathcal{I} \subset \varepsilon$ and ε is a σ-Field$\}$ is a σ-Field. $\mathcal{B}_0$ is often referred to as the **Borel** σ-Field on the real line.

- Axiomatic definition of probability

  Although the frequency interpretation of probability is the way what probability represents but it is hard to make it into a rigorous mathematical definition of probability. Kolmogorov (1933) developed an axiomatic definition of probability which he then showed can be interpreted, in a certain sense, as the limit of the relative frequency in a large number of experiments.

  - Definition 1.3.2 (Probability)

    Let $\mathcal{C}$ be a sample space and let $\mathcal{B}$ be a σ-Field on $\mathcal{C}$. A *probability function*(measure) on the event $C \in \mathcal{B}$ is a real valued function defined on $\mathcal{B}$ which satisfies the following three axioms:

    1. $0 \leq P(C)$ for all $C \in \mathcal{B}$

2. P($\mathcal{C}$)=1 where $\mathcal{C}$ is the sample space .

3. For any sequence of mutually exclusive events $\{C_n\}$,
$P(\bigcup_{n=1}^{\infty} C_n) = \sum_{n=1}^{\infty} P(C_n)$

We refer to P(C) as the probability of an event C.

Using these axioms and strong law of large numbers, we will prove it in Chapter 4 that if an experiment is repeated over and over again, then with probability 1, the proportion of times that a specific event C occurs converges to P(C), which is essentially the frequency interpretation of probability.

- Theorem 1.3.1 For each event $C \in \mathcal{B}$, $P(C) = 1 - P(C^c)$
- Theorem 1.3.2 $P(\phi) = 0$
- Theorem 1.3.3 If $C_1 \subset C_2$, $P(C_1) \leq P(C_2)$
- Theorem 1.3.4 $0 \leq P(C) \leq 1$ for each $C \in \mathcal{B}$
- Theorem 1.3.5 $P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2)$

Note: Theorem 1.3.5 can be extended to provide an expression for $P(C_1 \cup C_2 \cdots \cup C_k)$; see remark 1.3.2 (the inclusion-exclusion identity).

- Permutation and Combinations

  The basic principle of counting: Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if for each outcome of experiment 1 there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments.

  Generalized basic principle of counting: If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes and if for each of these possible outcomes, there are n_2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments, there are n_3 possible outcomes of the third experiment, and if ... then there is a total of $n_1 * n_2 * \cdots n_r$ possible outcomes of the r experiments.

– A permutation of n distinct object is an arrangement of these objects on a line and the number of permutations of n distinct objects is equal to n! ($=n(n-1)\cdots(3)(2)(1)$)

– An k permutation of n distinct object ($k \leq n$) is an arrangement of k objects chosen from n distinct objects and the number of k permutations from n distinct object, denoted P_k^n and equal to n(n-1)(n-(k-1))=n!/(n-k)!

– An k combination from n distinct object ($k \leq n$) is a subset containing k objects taken from the set containing these n distinct objects. Note that the order to choose the objects out from the given set is not on account for a combination. The number of k combinations from n distinct object, denoted $\left(\begin{array}{c} n \\ k \end{array}\right)$ or C_k^n (also refered to as a **binomial coefficient**) and equal to n!/k!(n-k)!

- **Example 1.3.4** Let a card be drawn at random from an ordinary deck of 52 playing cards which has been well shuffled.
(1) The probability of drawing a card that is a spade is 0.25
(2) The probability of drawing a card that is a king is 1/13.
(3) Suppose 5 cards are taken at random without replacement and order is not important. Then the probability of getting a flush, all 5 cards of the same suit, is 0.00198.
(4) The probability of getting exactly 3 of a kind and the other two cards are distinct and are of different kinds is 0.0211
(5) The probability of getting exactly three cards that are kings and exactly two cards that are queens is 0.0000093.
Note: The case discussed above is assuming that all the outcomes in the sample space are equally likely.

- **A loaded dice example** A die is loaded in such a way that the probability of any particular face's showing is directly proportional to the number on that face. What is the probability of observing 1,2 or 3?

Solution: The experiment generates a sample space containing six outcomes that are not equally likely. By assumption, P(“i“face appears)=P(i)=ki, i=1,$\cdots$,6, where k is a constant. Since $\sum_{i=1}^{6} P(i)$=1, we have k=1/21. Therefore,P(1)+P(2)+P(3)=2/7.

- Theorem 1.3.6 (Continuity theorem of probability)
  Let C_n be an increasing (decreasing) sequence of events. Then

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

- Theorem 1.3.7 (Boole’s Inequality) Let C_n be an arbitrary sequence of events. Then

$$P\left(\bigcup_{n=1}^{\infty} C_n\right) \leq \sum_{n=1}^{\infty} P(C_n) \quad (2)$$

- Exercise 1.3.2, 1.3.4,1.3.6,1.3.16, 1.3.24

Conditional Probability and Independence

- Intuitive Definition: the conditional probability of an event is the revised probability of the event when there is additional information about the outcome of the random experiment.

- A motivated example (1.4.2)
  A bowl contains 8 chips: 3 red, 5 blue. Draw two chips successively at random and without replacement. C_1={1st draw is red chip}, C_2={2nd draw is blue chip}. How to calculate P(2nd draw is blue chip(C_2) given the 1st draw is red chip(C_1))=?
  Now, sample space=C_1={(R,R),(R,B)}, event={(R,B)} and P(2nd draw is blue given the 1st draw is red)=$\frac{P_1^3 P_1^5}{P_1^3 P_1^7} = \frac{P_1^3 P_1^5 / P_2^8}{P_1^3 P_1^7 / P_2^8}$
  As we can see $P(C_2|C_1) = P(C_1 \cap C_2|C_1) = \frac{P(C_1 \cap C_2|\mathcal{C})}{P(C_1|\mathcal{C})}$

- Definition of conditional probability

  The conditional probability of an event C_2, given an event C_1, denoted by $P(C_2|C_1)$, is defined as $P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)}$ provided $P(C_1) > 0$, $C_1, C_2 \subset \mathcal{C}$.

  Q: Is the conditional probability function a probability set function?

  A: It is a probability set function defined on σ-Field on sample space C_1.

  (i)$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)} \geq 0$

  (ii)$P(C_1|C_1) = 1$

  (iii)Let $\{C_i\}$, i=2,$\cdots, \infty$ be a pairwise mutually exclusive sequence of events, Then

$$P(\bigcup_{i=2}^{\infty} C_i | C_1) = \frac{P((\bigcup_{i=2}^{\infty} C_i) \cap C_1)}{P(C_1)} = \frac{P(\bigcup_{i=2}^{\infty} (C_i \cap C_1))}{P(C_1)}$$

$$= \frac{\sum_{i=2}^{\infty} P(C_i \cap C_1)}{P(C_1)} = \sum_{i=2}^{\infty} \frac{P(C_i \cap C_1)}{P(C_1)}$$

$$= \sum_{i=2}^{\infty} P(C_i | C_1)$$

- Multiplication rule for probability

  For any events C_1, C_2 with $P(C_1), P(C_2) > 0$, then
  $P(C_1 \cap C_2) = P(C_2|C_1)P(C_1) = P(C_1|C_2)P(C_2)$
  Note: The multiplication rule can be extended to three or more events.

- Law of Total probability

  If $C_1, C_2, ...C_k$ is a collection of pairwise mutually exclusive and exhaustive events, that is $C_i \cap C_j = \phi$ for $i \neq j$ and $\mathcal{C} = \bigcup_{i=1}^{k} C_i$, and $P(C_i) > 0$, i=1,...,k. Then for any event C,

$$P(C) = \sum_{i=1}^{k} P(C|C_i)P(C_i)$$

  The law of total probability enable us to evaluate the probability of certain events by breaking them into subevents for which we know (or can determine) the probabilities

- Bayes' theorem

  If $C_1, C_2, ...C_k$ is a collection of pairwise mutually exclusive and exhaustive events with $P(C_i) > 0$, i=1,...,k. Then for any event C with $P(C) > 0$, and for a given i, $0 \leq i \leq k$,

$$P(C_i|C) = \frac{P(C|C_i)P(C_i)}{\sum_{i=1}^{k} P(C|C_i)P(C_i)}$$

Here, $P(C_i)$ is called prior probability and $P(C_i|C)$ is called posterior probability.

- Independent Events

  Two events C_1, C_2 are independent if any one of the following hold:

  (1) $P(C_1 \cap C_2) = P(C_1)P(C_2)$.

  (2) $P(C_1|C_2) = P(C_1)$

  (3) $P(C_2|C_1) = P(C_2)$

- Independency of n events

  Let $C_1, C_2 \cdots C_n$ be n given events. These events are mutually independent events if and only if for any $2 \leq k \leq n$,

  $P(C_{d_1} \cap C_{d_2} \cap \cdots \cap C_{d_k}) = P(C_{d_1})P(C_{d_2}) \cdots P(C_{d_k})$

Random Variables

It is more convenient to describe the elements of a sample space $\mathcal{C}$ with numbers.

- A simple example

  A sample space $\mathcal{C}$={c: where c is Tail or c is Head }. If we define a function X such that X(c)=0 if c is T and X(c) =1 if c is H, this sample space can be described by a sample space on the real numbers $\mathcal{D}$={x: 0,1}.

- Definition 1.5.1 (random variable)

  A random variable X is a real-valued function defined on the sample space $\mathcal{C}$, which assigns to each element $c \in \mathcal{C}$ one and only one number contained in the set of real numbers

  $\mathcal{D} = \{x : x = X(c), c \in \mathcal{C}\}$

- Probability model of X

  If $B \subset \mathcal{D}$ and $C = \{c : c \in \mathcal{C} \text{ and } X(c) \in B\}$, then the probability of event B, denoted by $P_X(B)$, is equal to P(C).

  $P_X(B)$ is also a probability set function

  (1) $P_X(B) = P(C) \geq 0$

  (2) $P_X(\mathcal{D}) = P(\mathcal{C}) = 1$

  (3) For a sequence of mutually exclusive events $\{B_n\}$, let

  $C_n = \{c : c \in \mathcal{C} \text{ and } X(c) \in B_n\}$. $\{C_n\}$ are mutually exclusive.

$$P_X(\bigcup_{n=1}^{\infty} B_n) = P(\bigcup_{n=1}^{\infty} C_n) = \sum_{n=1}^{\infty} P(C_n) = \sum_{n=1}^{\infty} P(B_n)$$

- Definition 1.5.2 (Cumulative Distribution Function) Let X be a random variable, the cumulative distribution function (cdf) of X is defined by,

  $F_X(x) = P_X((-\infty, x]) = P(\{c \in \mathcal{C} : X(c) \leq x\}) = P(X \leq x)$

- Theorem 1.5.1 Let X be a r.v. with cdf F(x). Then
(a) For all a and b, if $a < b$ then $F(a) \leq F(b)$. (F is a nondecreasing function).
hint: $\{X \leq a\} \subset \{X \leq b\}$
(b) $\lim_{x \to -\infty} F(x) = 0$
(c) $\lim_{x \to \infty} F(x) = 1$
hint:$\{X \leq -\infty\} = \phi, \{X \leq \infty\} = samplespace$
(d) $\lim_{x \downarrow x_0} F(x) = F(x_0)$, (F is right continuous).
hint: let $\{x_n\}$ be any sequence of real numbers such that $x_n \downarrow x_0$. Let $C_n = \{X \leq x_n\}$. Then $\bigcap_{n=1}^{\infty} C_n = \{X \leq x_0\}$.

- Theorem 1.5.2 Let X be a r.v. with cdf F(x). Then for $a < b$, $P[a < X \leq b] = F_X(b) - F_X(a)$.

- Theorem 1.5.3 For any random variable, $P[X = x] = F_X(x) - F_X(x-)$, for all $x \in R$, where $F_X(x-) = lim_{z \uparrow x} F_X(z)$.

Discrete and Continuous Random Variables

A r.v. is a discrete r.v. if its space is either finite or countable. A r.v. is a continuous r.v. if its cdf F(x) is a continuous function for all $x \in R$

- Probability Mass Function Let X be a discrete r.v. with space $\mathcal{D}$. The probability mass function (pmf) of X is given by $p_X(x) = P[X = x]$, for $x \in \mathcal{D}$.
  Note: $0 \leq p_X(x) \leq 1, x \in \mathcal{D}$ and $\sum_{x \in \mathcal{D}} p_X(x) = 1$
- Probability Density Function Let X be a continous r.v. with cdf F(x). If there exists a function f(x) such that the F(x) of X can be written as $F(x) = \int_{-\infty}^{x} f(t)dt$ for all $x \in \mathcal{D}$. f(x) is called the pdf of X.
  (1) If f(x) is also continuous then $\frac{d}{dx}F(x) = f(x)$
  (2)$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b) = \int_a^b f_X(t)dt$. (P(X=a)=0 and P(X=b)=0)

(3)$f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(t)dt = 1$

- Transformations Occasionally we may know the distribution of a random variable X but require the distribution of a function Y=g(X).
  (1) If g is one-to-one and X is discrete r.v. then, $p_Y(y) = P[Y = y] = P[g(X) = y] = P(X = g^-1(y)] = p_X(g^-1(y))$
  (2) If g is not one-to-one, usually we will have single-valued inverse function $g^{-1}(y)$ and the pmf of Y can be obtained easily. See example 1.6.4.
  (3) If g is one-to-one and X is continuous r.v. then, $f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}|$, where $x = g^{-1}(y)$ and $dx/dy = d[g^{-1}(y)]/dy$.

Expectation of a Random Variable

- A example: Suppose N observations of a r.v. X consist of n_0 zeros, n_1 ones, n_2 twos,... Then the sample mean, or average, can be written as
$$\bar{X} = \frac{0n_0 + 1n_1 + 2n_2 + \cdots}{N} = 0(\frac{n_0}{N}) + 1(\frac{n_1}{N}) + 2(\frac{n_2}{N}) + \cdots$$
$$= 0p_0 + 1p_1 + 2p_2 + \cdots = \sum_x xp_x$$
where $p_x = n_x/N$, the observed relative frequency of x's. Now let $N \to \infty$, then $p_x \to P(x)$ for all x, so that $\bar{X} \to \sum_x xP(x)$. This limit is the population mean of X and is denoted by E(X), the expectation of X.

- Definition 1.8.1 (Expectation).
If X is a continuous r.v. and $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$, then the **expectation** of X is $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.

If X is a discrete r.v. and $\sum_x |x|p(x) < \infty$, then the **expectation** of X is $E(X) = \sum_x xp(x)$.

- Theorem 1.8.1 (Expectation of a function of a r.v.)
  Let Y=g(X) ,
  If X is continuous r.v. with pdf $f_X(x)$ and
  $\int_{-\infty}^{\infty} |g(x)|f_X(x)dx < \infty$, then $E(Y) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$
  If X is a discrete r.v with pmf $p_X(x)$ and $\sum_x |g(x)|p_X(x) < \infty$,
  then $E(Y) = \sum_x g(x)p_X(x)$.

- Theorem 1.8.2 $E[k_1g_1(X) + k_2g_2(X)] = k_1E[g_1(X)] + k_2E[g_2(X)]$

Some Special Expectation

- Definition 1.9.1 (Mean)
  $\mu = E(X)$, where μ is called the mean of X.
- Definition 1.9.2 (Variance)
  $\sigma^2 = Var(X) = E[(X-\mu)^2]$, where σ^2 or $Var(X)$ is called the variance of X.
- $Var(X) = E(X^2) - \mu^2$
- $E(X^2) \geq [E(x)]^2$
- If X is a r.v. with mean μ and variance σ^2, then for any real constants a and b, $Var(aX+b) = a^2 Var(X) = a^2\sigma^2$
- Moment
  The kth moment about the origin of a r.v. X (if the expectation exists): $\mu_k^{'} = E(X^k)$

The kth moment about the mean of a r.v. X(if the expectation exists): $\mu_k = E[(x-\mu)^k]$
$\mu_1^{'}$=mean, μ_2=variance

- Definition 1.9.3 (Moment Generating Function)
$M(t) = E(e^tX)$. The domain of M(t) is the set of all t such that $E[e^tX]$ exists. This domain is an interval containing 0. M(t) is called moment generating function (mgf). Other way to say, there is an $h > 0$ such that $E[e^tX]$ exists for all $t \in (-h, h)$. If such h doesnt exist, then X doesnt have a mgf.

- Theorem 1.9.1 (Uniqueness of mgf.) Let X and Y be random variables with mgf M_X and M_Y, respectively, existing in open intervals about 0. Then $F_X(z) = F_Y(z)$ for all
$z \in R \Leftrightarrow M_X(t) = M_Y(t)$ for all $t \in (-h, h)$ for some $h > 0$
Note: mgf uniquely and completely defines the distribution of a r.v.

- If the mgf of a r.v. X exists, then $E(X^r) = M^{(r)}(0)$ and

$$M(t) = 1 + \sum_{r=1}^{\infty} M^r(0) \frac{t^r}{r!}$$

for $|t| < h$ for some $h > 0$.

- If X is a r.v. with mgf $M_X(t)$, then Y=aX+b, a and b are constants, will have the mgf $M_Y(t) = e^{bt} M_X(at)$.

Important Inequalities

- Theorem 1.10.1 Let X be a random variable and let m, k are positive integers, where $k \leq m$. If $E[X^m]$ exists, then $E[X^k]$ exists.
- Theorem 1.10.2 (Markov's Inequality) Let u(X) be a nonnegative function of the random variable X. If E[u(X)] exists, then for every positive constant c, $P[u(X) \geq c] \leq \frac{E[u(X)]}{c}$
- Theorem 1.10.3 (Chebyshev's Inequality). Let r.v. X have a distribution of probability about which we assume only that there is a finite variance σ^2 (this implies the mean $\mu = E(X)$ exists,why?). Then for every $k > 0$,

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

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

- Definition 1.10.1 (Convex function) A function ϕ defined on an interval (a,b), $-\infty \leq a < b \leq \infty$, is a convex function if for all x,y in (a,b) and for all $0 < \gamma < 1$,

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

  We say ϕ is strictly convex if the above inequality is strict.

- Theorem 1.10.4. If ϕ is differentiable on (a,b) then
  (a) ϕ is convex $\Longleftrightarrow \phi^{'}(x) \leq \phi^{'}(y)$, for all $a < x < y < b$.
  (b) ϕ is strictly convex $\Longleftrightarrow \phi^{'}(x) < \phi^{'}(y)$, for all $a < x < y < b$.
  If ϕ is twice differentiable on (a,b) then
  (a) ϕ is convex $\Longleftrightarrow \phi^{''}(x) \geq 0$, for all $a < x < b$.
  (b) ϕ is strictly convex $\Longleftrightarrow \phi^{''}(x) > 0$, for all $a < x < b$.

- Theorem 1.10.5 (Jensen's Inequality). If ϕ is convex on an open interval I and X is a r.v. whose support is contained in I and has 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.

2 Multivariate Distribution

Distribution of Two Random Variables

- A example:Toss a coin three times, then the sample space
  $\mathcal{C} = \{c : TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$
  X_1: number of H's on first two tosses.
  $X_1(c) = 0$ if c=TTT,TTH
  $X_1(c) = 1$ if c=THT,HTT,THH,HTH
  $X_1(c) = 2$ if c=HHT,HHH
  X_2: number of H's on three tosses.
  $X_2(c) = 0$ if c=TTT
  $X_2(c) = 1$ if c=TTH,THT,HTT
  $X_2(c) = 2$ if c=THH,HTH,HHT
  $X_2(c) = 3$ if c=HHH
  Thus, $X = (X_1, X_2) : \mathcal{C} \to \mathcal{A}$,where
  $\mathcal{A}$={(0,0),(0,1),(1,1),(1,2),(2,2),(2,3)}

- Definition 2.1.1 (Bivariate r.v.)A bivariate random variable $X = (X_1, X_2)$ is a real-valued function which assigns to each element c of sample space $\mathcal{C}$ one and only one ordered pair of numbers $X_1(c) = x_1, X_2(c) = x_2$. The space of $X = (X_1, X_2)$ is $\mathcal{A} = \{(x_1, x_2) : X_1(c) = x_1, X_2(c) = x_2, c \in \mathcal{C}\}$
- Definition If event $A \subset \mathcal{A}, C = \{c : c \in C$ and $(X_1(c), X_2(c)) \in A\}$, then $P((X_1, X_2) \in A) = P(C)$.
- A bivariate random variable is of the discrete type or of the continuous type

$$\begin{aligned}
p_{X_1,X_2}(x_1, x_2) &= P[X_1 = x_1, X_2 = x_2] \quad \text{joint pmf for discrete case} \\
F_{X_1,X_2}(x_1, x_2) &= P(X_1 \le x_1, X_2 \le x_2) \\
&= \sum_{u \le x_1} \sum_{v \le x_2} p_{X_1,X_2}(x_1, x_2) \quad \text{joint cdf for discrete case}
\end{aligned}$$

$Note:$ (i) $0 \leq p_{X_1,X_2}(x_1, x_2) \leq 1$

(ii) $\sum \sum_{\mathcal{A}} p_{X_1,X_2}(x_1, x_2) = 1$

(iii) for an event $A \in \mathcal{A}$, $P[(X_1, X_2) \in A] = \sum \sum_{A} p_{X_1,X_2}(x_1, x_2)$

Example cont. Consider the previous example, what is the joint pmf of X_1 and X_2? What's the probability of event A={(1,1),(1,2)}?

Another example: A bin contains 1000 flower seeds consisting of 400 red, 400 white and 200 pink when flowering. If 10 seeds are selected at random without replacement, and if X: number of red, Y: number of white. Find the joint pmf of X and Y.

$$f_{X_1,X_2}(x_1,x_2) = \frac{\partial^2 F_{X_1,X_2}(x_1,x_2)}{\partial x_1 \partial x_2} \quad \text{joint pdf for continuous case}$$

$$F_{X_1,X_2}(x_1,x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1,X_2}(w_1,w_2)dw_1dw_2$$

joint cdf for continuous case

Note : $(i) 0 \le f_{X_1,X_2}(x_1,x_2) \le 1$

$$(ii) \int\int_{\mathcal{A}} f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = 1$$

(iii) for an event $A \in \mathcal{A}$, $P[(X_1,X_2) \in A] = \int\int_{A} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$

Example 2.1.2

- Theorem

$$P(a < X_1 \le b, c < X_2 \le d) = F(b,d) - F(b,c) - F(a,d) + F(a,c)$$

- Marginal Distribution

  The marginal pmf for a single discrete r.v. can be obtained from the joint discrete pmf by summing

$$f_{X_1}(x) = \sum_{\text{all} x_2} p_{X_1,X_2}(x_1, x_2), f_{X_2}(x) = \sum_{\text{all} x_1} p_{X_1,X_2}(x_1, x_2).$$

  The marginal pdf for a single continuous r.v. can be obtained from the joint continuous pdf by integrating

$$f_{X_1}(x) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) dx_2, f_{X_2}(x) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) dx_1.$$

  Example 2.1.3, 2.1.4

- Expectation

  Let (X_1, X_2) be a bivariate r.v. and let $Y = g(X_1, X_2)$ for some real valued funtion

Suppose (X_1, X_2) is discrete, then E(Y) exists if

$$\sum_{x_1}\sum_{x_2}|g(x_1, x_2)|p_{X_1,X_2}(x_1, x_2) < \infty$$

Then

$$E(Y) = \sum_{x_1}\sum_{x_2} g(x_1, x_2)p_{X_1,X_2}(x_1, x_2)$$

Suppose (X_1, X_2) is continuous, then E(Y) exists if

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|g(x_1, x_2)|f_{X_1,X_2}(x_1, x_2)dx_1dx_2 < \infty$$

Then

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1, x_2)f_{X_1,X_2}(x_1, x_2)dx_1dx_2$$

- Theorem 2.1.1. Let (X_1, X_2) be a bivariate r.v. Let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ be random variables whose expectations exist. Then for any real numbers k_1, k_2,

$$E(k_1 Y_1 + k_2 Y_2) = k_1 E(Y_1) + k_2 E(Y_2)$$

Example 2.1.5, 2.1.6

- Definition 2.1.2 (Moment Generating Function of a Bivariate r.v.) Let $\mathbf{X} = (X_1, X_2)'$ be a bivariate r.v. If $E(e^{t_1 X_1 + t_2 X_2})$ exists for $|t_1| < h_1$ and $|t_2| < h_2$, where h_1, h_2 are positive, then $M_{X_1,X_2}(t_1, t_2) = E(e^{t_1 X_1 + t_2 X_2})$ is called the moment-generating function (mgf) of $\mathbf{X}$.
  Note: $M_{X_1}(t_1) = M_{X_1,X_2}(t_1, 0), M_{X_2}(t_2) = M_{X_1,X_2}(0, t_2)$
  Example 2.1.7

- Definition 2.1.3 (Expected value of a Bivariate r.v.) Let $\mathbf{X} = (X_1, X_2)'$ be a bivariate r.v. Then the expected value of $\mathbf{X}$ exists if the expectations of X_1 and X_2 exist.
  $$E\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix}$$

Transformations: Bivariate Random Variables

- Discrete Case

  Let $p_{X_1,X_2}(x_1,x_2)$ be the joint pmf of two discrete-type r.v. X_1 and X_2. Let $y_1 = \mu_1(x_1,x_2)$ and $y_2 = \mu_2(x_1,x_2)$ define a one-to-one transformation. What are the joint pmf of the two new random variables $Y_1 = \mu_1(X_1,X_2)$ and $Y_2 = \mu_2(X_1,X_2)$?

  $$(1) \quad \begin{array}{l} y_1 = \mu_1(x_1,x_2) \\ y_2 = \mu_2(x_1,x_2) \end{array} \quad \Rightarrow \quad \begin{array}{l} x_1 = \omega_1(y_1,y_2) \\ x_2 = \omega_2(y_1,y_2) \end{array}$$

  (2) $p_{Y_1,Y_2}(y_1,y_2) = p_{X_1,X_2}[\omega_1(y_1,y_2),\omega_2(y_1,y_2)]$

  example 2.2.1

- Continuous Case

  Let $f_{X_1,X_2}(x_1,x_2)$ be the joint pdf of two continuous-type r.v. X_1 and X_2. Let $y_1 = \mu_1(x_1,x_2)$ and $y_2 = \mu_2(x_1,x_2)$ define a one-to-one transformation. What are the joint pdf of the two new random variables $Y_1 = \mu_1(X_1,X_2)$ and $Y_2 = \mu_2(X_1,X_2)$?

$$(1) \quad \begin{matrix} y_1 = \mu_1(x_1, x_2) \\ y_2 = \mu_2(x_1, x_2) \end{matrix} \Rightarrow \begin{matrix} x_1 = \omega_1(y_1, y_2) \\ x_2 = \omega_2(y_1, y_2) \end{matrix}$$

$$(2)\ J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \qquad \text{Jacobian of the transformation}$$

$$(3) f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}[\omega_1(y_1, y_2), \omega_2(y_1, y_2)]|J|$$

Example 2.2.3, 2.2.4, 2.2.5

- In addition to the change-of-variable techniques for finding distributions of functions of random variables, there are two other techniques: cdf techniques and mgf techniques.
  Example 2.2.2, 2.2.6,2.2.7

Conditional Distributions and Expectations

Conditional distribution considers the distribution of one of the random variables when the other random variable has assumed a specific value.

For example, the discrete random variable X and Y have joint probability mass function (pmf) defined by the following table. what is $P(1 \leq X \leq 3|Y = 2)$? what is $E(Y|X = 3)$?

		y		
		1	2	3
	1	0	0.1	0.05
x	2	0.3	0	0.1
	3	0.05	0.05	0
	4	0.2	0.05	0.1

- Conditional pmf of the discrete r.v.

  X_1, X_2: discrete r.v.

  $p_{X_1,X_2}(x_1, x_2)$: joint pmf

  p_{X_1}, p_{X_2}: marginal pmf

  Then for any $p_{X_1}(x_1) > 0$,

$$
\begin{aligned}
P(X_2 = x_2 | X_1 = x_1) &= \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_1 = x_1)} \\
&= \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_1}(x_1)} \\
&= p_{X_2|X_1}(x_2|x_1)
\end{aligned}
$$

  $p_{X_2|X_1}(x_2|x_1)$ is called the conditional pmf of X_2 given $X_1 = x_1$. Similarly, $p_{X_1|X_2}(x_1|x_2)$ is called the conditional pmf of X_1 given $X_2 = x_2$.

  Question: Is the conditional pmf a probability mass function?

  (i)$p_{X_2|X_1}(x_2|x_1) = \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)} > 0$

(ii)

$$\sum_{x_2} p_{X_2|X_1}(x_2|x_1) = \sum_{x_2} \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)}$$
$$= \frac{1}{p_{X_1}(x_1)} \sum_{x_2} p_{X_1,X_2}(x_1,x_2)$$
$$= \frac{p_{X_1}(x_1)}{p_{X_1}(x_1)} = 1$$

- Conditional pdf of the continuous r.v.

  X_1, X_2: continuous r.v.
  $f_{X_1,X_2}(x_1,x_2)$: joint pdf
  f_{X_1}, f_{X_2}: marginal pdf
  Then for any $f_{X_1}(x_1) > 0$, $f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}$ is called the conditional pdf of X_2 given $X_1 = x_1$. Similarly, $f_{X_1|X_2}(x_1|x_2)$ is called the conditional pdf of X_1 given $X_2 = x_2$.

Question: Is the conditional pdf a probability density function?

(i) $f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} > 0$

(ii)

$$\begin{aligned}\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2|x_1)dx_2 &= \int_{-\infty}^{\infty} \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}dx_2 \\ &= \frac{1}{f_{X_1}(x_1)} \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_2 \\ &= \frac{f_{X_1}(x_1)}{f_{X_1}(x_1)} = 1\end{aligned}$$

- Conditional probability

  discrete case:

$$P(a < X_2 < b|X_1 = x_1) = \sum_{a<x_2<b} p_{X2|X_1}(x_2|x_1)$$

$$P(c < X_1 < d|X_2 = x_2) = \sum_{c<x_1<d} p_{X1|X_2}(x_1|x_2)$$

continuous case:

$$P(a < X_2 < b | X_1 = x_1) = \int_a^b f_{X2|X_1}(x_2|x_1)dx_2$$
$$P(c < X_1 < d | X_2 = x_2) = \int_c^d f_{X1|X_2}(x_1|x_2)dx_1$$

- Conditional Expectation
  If $u(X_2)$ is a function of X_2, then
  discrete case: $E(u(X_2)|X_1) = \sum_{x_2} u(x_2)p_{X_2|X_1}(x_2|x_1)$
  continuous case: $E(u(X_2)|X_1) = \int_{-\infty}^{\infty} u(x_2)f_{X_2|X_1}(x_2|x_1)dx_2$

- Conditional Variance
  If $u(X_2)$ is a function of X_2, then

$$\begin{aligned} var(u(X_2)|x_1) &= E\{[u(X_2) - E(u(X_2)|x_1)]^2|x_1\} \\ &= E(X_2^2|x_1) - [E(X_2|x_1)]^2 \end{aligned}$$

- Theorem 2.3.1 Let (X_1, X_2) be a random vector such that the variance of X_2 is finite. Then,
  (a) $E[E(X_2|X_1)] = E(X_2)$
  (b) $var[E(X_2|X_1)] \leq Var(X_2)$

  Example 2.3.1, 2.3.2, Exercise 2.3.1,2.3.4,2.3.5,2.3.6,2.3.8

The Correlation Coefficient

- Covariance and Correlation coefficient Let X and Y be two given r.v.'s. The covariance of X and Y, denoted cov(X,Y), is defined as

$$cov(X,Y) = E[(X - \mu_1)(Y - \mu_2)] = E(XY) - \mu_1\mu_2$$

where $\mu_1 = E(X)$,$\mu_2 = E(Y)$.

The correlation coefficient of X and Y, denoted ρ, is defined as

$$\rho = cov(X,Y)/\sigma_1\sigma_2$$

where $\sigma_1 = \sqrt{var(X)}$,$\sigma_2 = \sqrt{var(Y)}$.

Example 2.4.1, Exercise 2.4.2

- Theorem 1 If X and Y are two given r.v.'s and a and b are two given constant, then

(a) cov(aX,bY)=ab*cov(X,Y)

(b) cov(X+a,Y+b)=cov(X,Y)

(c) cov(X, aX+b)=a*var(X)
(d) var(aX+bY)=a^2var(X)+b^2var(Y)+2ab*cov(X,Y)
exercise 2.4.11

- Theorem 2 If the correlation coefficient ρ of two r.v.'s X and Y exists, then $-1 \leq \rho \leq 1$, and $\rho = \pm 1$ if and only if Y=aX+b ($a \neq 0$). (That is, Y is a linear function of X).

- Theorem 2.4.1 Suppose (X,Y) have a joint distribution with the variance of X and Y finite and positive. Denote the means and variances of X and Y by μ_1, μ_2, and σ_1^2, σ_2^2, respectively, and let ρ be the correlation coefficient between X and Y. If $E(Y|X)$ is linear in X then

$$E(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1)$$

and

$$E(Var(Y|X)) = \sigma_2^2(1 - \rho^2)$$

Example 2.4.2, 2.4.3, Exercise 2.4.3,2.4.4,

- Theorem 4 Two dimensional moment-generating function(mgf) of the joint distribution of X and Y is defined as
  $M(t_1, t_2) = E(e^{t_1 X + t_2 Y})$, then
  $M_1(t_1) = M(t_1, 0)$
  $M_2(t_2) = M(0, t_2)$
  $\mu_1 = E(X) = \frac{\partial M(0,0)}{\partial t_1}$
  $\mu_2 = E(Y) = \frac{\partial M(0,0)}{\partial t_2}$
  $\sigma_1 = E(X^2) - \mu_1^2 = \frac{\partial^2 M(0,0)}{\partial t_1^2} - \mu_1^2$
  $\sigma_2 = E(Y^2) - \mu_2^2 = \frac{\partial^2 M(0,0)}{\partial t_2^2} - \mu_2^2$
  $E[(X - \mu_1)(Y - \mu_2)] = \frac{\partial^2 M(0,0)}{\partial t_1 \partial t_2} - \mu_1 \mu_2$
  Example 2.4.4

Independent Random Variables

Motivated example: $f(x_1, x_2) = f_{2|1}(x_2|x_1)f_1(x_1)$, what happens if $f_{2|1}(x_2|x_1)$ does not depend upon x_1?
Answer: $f_{2|1}(x_2|x_1) = f_2(x_2)$ and $f(x_1, x_2) = f_2(x_2)f_1(x_1)$

- Definition 2.5.1 (Independence) Let X_1 and X_2 have the joint pdf $f(x_1, x_2)$(joint pmf $p(x_1, x_2)$) and the marginal pdfs (pmfs) $f_1(x_1)(p_1(x_1))$ and $f_2(x_2)(p_2(x_2))$, respectively.
  X_1 and X_2 are independent $\Leftrightarrow f(x_1, x_2) = f_1(x_1)f_2(x_2)$ for continuous case
  X_1 and X_2 are independent $\Leftrightarrow p(x_1, x_2) = p_1(x_1)p_2(x_2)$ for discrete case
  Remark: if $f_1(x_1)$ and $f_2(x_2)$ are positive on and only on, the respective spaces $\mathcal{A}_1$ and $\mathcal{A}_2$, then $f_1(x_1)f_2(x_2)$ is positive on, and only on, the product space
  $\mathcal{A} = \{(x_1, x_2) : x_1 \in \mathcal{A}_1, x_2 \in \mathcal{A}_2\}$. To check whether two r.v. X_1

and X_2 are independent, check the joint range first. If $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, then go to check if $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. If not, we stop and conclude that they are not independent.

Example: Check whether the two r.v. X_1 and X_2 are independent, where the joint p.d.f of X_1 and X_2 is given by $f(x_1, x_2) = 2$ if $0 < x_1 < x_2 < 1$.

Solution: we can prove that $f_1(x_1) = 2(1 - x_1)$ if $0 < x_1 < 1$ and $f_2(x_2) = 2x_2$ if $0 < x_2 < 1$. So,

the joint range is $\mathcal{A} = \{(x_1, x_2) : 0 < x_1 < x_2 < 1\}$,

the range of X_1 is $\mathcal{A}_1 = \{x_1 : 0 < x_1 < 1\}$,

the range of X_2 is $\mathcal{A}_2 = \{x_2 : 0 < x_2 < 1\}$.

Obviously $\mathcal{A} \neq \mathcal{A}_1 \times \mathcal{A}_2$. X_1 and X_2 are dependent.

Example 2.5.1, Exercise 2.5.2,2.5.3

- Theorem 2.5.1 Let the random variables X_1 and X_2 have support $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively, and have the joint pdf (joint pmf) $f(x_1, x_2)$ $(p(x_1, x_2))$.

  X_1,X_2 are independent $\Leftrightarrow f(x_1, x_2) = g(x_1)h(x_2)$ for continuous case

X_1,X_2 are independent $\Leftrightarrow p(x_1, x_2) = g(x_1)h(x_2)$ for discrete case where $g(x_1) > 0$, $x_1 \in \mathcal{S}_1$, and $h(x_2) > 0$, $x_2 \in \mathcal{S}_2$.

Example 2.5.1, Example 2.5.2, Exercise 2.5.1

- Theorem 2.5.2 Let the r.v X_1 and X_2 have the joint cdf $F(x_1, x_2)$ and the marginal cdfs $F_1(x_1)$ and $F_2(x_2)$, respectively.

  X_1,X_2 are independent $\Leftrightarrow F(x_1, x_2) = F_1(x_1)F_2(x_2)$ for all $(x_1, x_2) \in R^2$

- Theorem 2.5.3 X_1,X_2 are independent

  $\Leftrightarrow P(a < X_1 \leq b, c < X_2 \leq d) = P(a < X_1 \leq b)P(c < X_2 \leq d)$ for every $a < b$ and $c < d$, where a,b,c,d are constant.

  Example 2.5.3, Exercise 2.5.5

- Theorem 2.5.4 If X_1 and X_2 are independent r.v. and that $E[u(X_1)]$ and $E[v(X_2)]$ exist. Then

  $E[u(X_1)v(X_2))] = E[u(X_1)]E[v(X_2)]$

  Example 2.5.4 Note that the converse is not true. That is if

$cov(X_1, X_2) = 0$, then X_1 and X_2 could be dependent.

- Theorem 2.5.5 Suppose the joint mgf, $M(t_1, t_2)$ exists for the random variables X_1 and X_2, then
  X_1,X_2 are independent $\Leftrightarrow M(t_1, t_2) = M(t_1, 0)M(0, t_2)$
  Example 2.5.5, 2.5.6, Exercise 2.5.6
  More Examples: Exercise 2.5.9, 2.5.12

Extension to Several Random Variables

The notion about two random variables can be extended immediately to n random variables.

- Definition 2.6.1 A n variate random variable $X = (X_1, ..., X_n)$ is a real-valued function which assigns to each element c of sample space $\mathcal{C}$ one and only one *ordered n-tuples* of numbers $X_1(c) = x_1, ..., X_n(c) = x_n$. The space of $X = (X_1, ..., X_n)$ is $\mathcal{A} = \{(x_1, ..., x_n) : X_1(c) = x_1, ..., X_n(c) = x_n\}$. Furthermore, if event $A \subset \mathcal{A}, C = \{c : c \in \mathcal{C}$ and $(X_1(c), ...X_n(c)) \in A\}$, $P((X_1, ...X_n) \in A) = P(C)$

- Joint pdf(pmf) and CDF

  Discrete case:

$$p(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n)$$

$$F(x_1, ..., x_n) = P(X_1 \le x_1, ..., X_n \le x_n) = \sum_{u_1 \le x_1} \cdots \sum_{u_n \le x_n} p(x_1, ..., x_n)$$

Continuous case:

$$f(x_1, ..., x_n) = \frac{\partial^n F(x_1, ..., x_n)}{\partial x_1 ... \partial x_n}$$

$$F(x_1, ..., x_n) = P(X_1 \le x_1, ..., X_n \le x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(x_1, ..., x_n) dx_1 ... dx_n$$

Example 2.6.1

- Expectation

$$E[u(X_1, ...X_n)] = \sum_{x_n} \cdots \sum_{x_1} u(x_1, ...x_n) p(x_1, ..., x_n)$$

discrete case

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, ...x_n) f(x_1, ..., x_n) dx_1 \cdots dx_n$$

continuous case

$$E[\sum_{j=1}^{m} k_j u_j(X_1, ..., X_n)] = \sum_{j=1}^{m} k_j E[u_j(X_1, ..., X_n)]$$

- Marginal pdf(pmf) of one random variable

$$p(x_1) = \sum_{x_n} \cdots \sum_{x_2} p(x_1, ..., x_n) \quad \text{discrete case}$$

$$f(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n) dx_2 \cdots dx_n \quad \text{continuous case}$$

- Marginal pdf(pmf) of k ($k < n$) random variable

$$p(x_1, ..., x_k) = \sum_{x_n} \cdots \sum_{x_{k+1}} p(x_1, ..., x_n) \quad \text{discrete case}$$

$$f(x_1, ...x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n) dx_{k+1} \cdots dx_n \quad \text{continuous case}$$

- Joint conditional pdf (pmf) of $(X_2 \cdots X_n)$ given $X_1 = x_1$

$$f(x_2, ..., x_n | x_1) = \frac{f(x_1, x_2, ..., x_n)}{f(x_1)} \quad \text{discrete case}$$

$$p(x_2, ..., x_n | x_1) = \frac{p(x_1, x_2, ..., x_n)}{p(x_1)} \quad \text{continuous case}$$

What is joint conditional p.d.f of any n-1 random variables, say $(X_1 \cdots X_{i-1}, X_{i+1} \cdots X_n)$, given $X_i = x_i$?
What is joint conditional p.d.f of any n-k random variables, for given values of the remaining k variables?

- Conditional Expectation of $u(X_2 \cdots X_n)$ given $X_1 = x_1$

$$E[u(X_2 \cdots X_n)|x_1] = \sum_{x_n} \cdots \sum_{x_2} u(x_2, \cdots, x_n) p(x_2, \cdots, x_n|x_1)$$

discrete case

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_2, \cdots, x_n) f(x_2, \cdots, x_n|x_1) dx_2 \cdots dx_n$$

continuous case

- Independence

(1) The r.v.'s $X_1, ..., X_n$ are mutually independent if and only if
$f(x_1, ..., x_n) \equiv f(x_1) \cdots f(x_n)$ or $p(x_1, ..., x_n) \equiv p(x_1) \cdots p(x_n)$
(2)if $X_1, ..., X_n$ are mutually independent then
$P(a_1 < X_1 < b_1, a_2 < X_2 < b_2, ..., a_n < X_n < b_n)$
$= P(a_1 < X_1 < b_1) P(a_2 < X_2 < b_2) \cdots P(a_n < X_n < b_n)$

(3)if $X_1, ..., X_n$ are mutually independent then

$$E[\prod_{i=1}^{n} u_i(X_i)] = \prod_{i=1}^{n} E[u_i(X_i)]$$

- Moment Generating Function (mgf)

  (1) $M(t_1, t_2, ..., t_n) = E(e^{t_1 X_1 + t_2 X_2 \cdots + t_n X_n})$

  (2) The mgf of the marginal distribution of X_i is

  $M_i(t_i) = M(0, ..., 0, t_i, 0, ..., 0)$

  (3) The mgf of the marginal distribution of X_i and X_j is

  $M(t_i, t_j) = M(0, ..., 0, t_i, 0, ..., 0, t_j, 0, ..., 0)$

  (4) X_1,...,X_n are independent if and only if

  $M(t_1, t_2, ..., t_n) = \prod_{i=1}^{n} M(0, ..., 0, t_i, 0, ...0)$

  Example 2.6.2, Remark 2.6.1. Exercise 2.6.1, 2.6.2,2.6.3

Transformations: Random Vectors

- Discrete Case

  Let $p_{X_1,\cdots,X_n}(x_1,...,x_n)$ be the joint pmf of n discrete-type r.v. X_1,...,X_n. Let $y_1 = \mu_1(x_1,...,x_n), \cdots, y_n = \mu_n(x_1,...,x_n)$ define a one-to-one transformation. What are the joint pmf of the n new random variables $Y_1 = \mu_1(X_1,...,X_n), \cdots, Y_n = \mu_n(X_1,...,X_n)$?

(1)
$$
\begin{aligned}
y_1 &= \mu_1(x_1,...,x_n) & & x_1 = \omega_1(y_1,...,y_n) \\
&\vdots & \Rightarrow & \qquad \vdots \\
y_n &= \mu_n(x_1,...,x_n) & & x_n = \omega_n(y_1,...,y_n)
\end{aligned}
$$

(2)

$$p_{Y_1,\cdots,Y_n}(y_1,\cdots,y_2) = p_{X_1,\cdots,X_n}(\omega_1(y_1,\cdots,y_n),\cdots,\omega_n(y_1,\cdots,y_n))$$

- Continuous Case

  Let $f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)$ be the joint pdf of n continuous-type r.v. $X_1,\cdots,X_n$. $y_1=\mu_1(x_1,...,x_n),\cdots,y_n=\mu_n(x_1,...,x_n)$ define a one-to-one transformation. What are the joint pdf of the n new random variables $Y_1=\mu_1(X_1,...,X_n),\cdots,$ $Y_n=\mu_n(X_1,...,X_n)$?

$$(1)\quad \begin{array}{ccc} y_1=\mu_1(x_1,...,x_n) & & x_1=\omega_1(y_1,...,y_n) \\ \vdots & \Rightarrow & \vdots \\ y_n=\mu_n(x_1,...,x_n) & & x_n=\omega_n(y_1,...,y_n) \end{array}$$

$$(2)\quad J=\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2}\cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2}\cdots & \frac{\partial x_2}{\partial y_n} \\ & \vdots & \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2}\cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \qquad \text{Jacobian of the transformation}$$

(3)$f_{Y_1,\cdots,Y_n}(y_1,\cdots,y_n) =$
$f_{X_1,\cdots,X_n}[\omega_1(y_1,\cdots,y_n),\cdots,\omega_n(y_1,\cdots,y_n)]|J|$
Example 2.7.1-2.7.5, Exercise 2.7.1,2.7.4

3 Some Special Distributions

The Binomial and Related Distribution

- Bernoulli Distribution

  A **Bernoulli experiment** is a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways. For example,
  rain or not rain tomorrow? ($X = 0 \rightarrow$ no rain, $X = 1 \rightarrow$ rain)
  Head turning up or tail turning up after flipping a coin once? ($X = 0 \rightarrow$ tail, $X = 1 \rightarrow$ head)
  Bernoulli Distribution: The r.v. X has a Bernoulli distribution with parameter p, $0 \leq p \leq 1$, if its pmf is given by P(X=1)=p, P(X=0)=1-p. This pmf can be written more succinctly as $p_X(x) = p^x(1-p)^{1-x}, x = 0, 1$
  Mean: E(X)=p
  Variance: Var(X)=p(1-p)
  mgf: $M(t) = pe^t + q, \forall t$

- Binomial Distribution

  Repeat the Bernoulli experiments in previous Example many times. Say, n times. Each time there is probability=p to observe 1 (rain or head turning up). If X is the number of 1 observed, then

  $$p(x) = P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, x = 0, 1, 2, \cdots, n$$

  Binomial Distribution: The r.v. X has a Binomial distribution b(n,p) with parameters n,p, where n is the number of trials, p is the probability of observing 1 in each independent trial, $0 \leq p \leq 1$, if the pmf of X is given by

  $$p_X(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, x = 0, 1, 2, \cdots, n.$$

  Mean: E(X)=np

  Variance: Var(X)=np(1-p)

  mgf: $M(t) = (pe^t + q)^n, \forall t$

Q1: Is b(n,p) a pmf?

Q2: How to use mgf to compute E(X) and Var(X)?

Q3: If n=1, Binomial distribution is also another special distribution. What is this distribution?

Example 3.1.1-3.1.5

- Theorem 3.1.1 Let $X_1, X_2, ..., X_m$ be independent random variables such that X_i has binomial $b(n_i, p)$ distribution, for $i = 1, 2, \cdots, m$. Let $Y = \sum_{i=1}^{m} X_i$. Then Y has a binomial $b(\sum_{i=1}^{m} n_i, p)$ distribution.

- Multinomial Distribution

  The binomial distribution can be generalized to the multinomila distribution. Let a random experiment be repeated n independent times. On each repetition, the experiment results in but one of k mutually exclusive and exhaustive ways, say $C_1, C_2, \cdots, C_k$. Let p_i be the probability that the outcome is an element of C_i and let p_i remain constant throughout the n independent repetitions, $i = 1, 2, \cdots, k$. If X_i are the number of outcomes that are elements of $C_i, i = 1, 2, \cdots, k-1$, then

$p(x_1, x_2, \cdots, x_{k-1}) = P(X_1 = x_1, X_2 = x_2, \cdots, X_{k-1} = x_{k-1})$
$= \frac{n!}{x_1! \cdots x_{k-1}!(n-(x_1+\cdots+x_{k-1}))!} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} p_k^{n-(x_1+\cdots+x_{k-1})}$

Example: trinomial distribution page 138-139

- Negative Binomial Distribution

Repeat the Bernoulli experiments in the first example until observing 1 (rain or head turning up) for r times. Each time there is probability=p to observe 1. If Y is the number of 0 observed (no rain or tail turning up), then

$$p_Y(y) = P(Y = y) == \begin{pmatrix} y + r - 1 \\ r - 1 \end{pmatrix} p^r (1-p)^y, y = 0, 1, 2, \cdots.$$

Negative Binomial Distribution: The r.v. Y has a Negative Binomial distribution with parameters r,p, where r is the number of trials observing 1, p is the probability of observing 1 in each independent trial, $0 \le p \le 1$, if the pmf of Y is given by

$$p_Y(y) = P(Y = y) = \begin{pmatrix} y + r - 1 \\ r - 1 \end{pmatrix} p^r (1-p)^y, y = 0, 1, 2, \cdots.$$

Mean: $E(X) = \frac{r(1-p)}{p}$

Variance: $\frac{r(1-p)}{p^2}$

mgf: $M(t) = p^r[1-(1-p)e^t]^{-r}$ for $t < -ln(1-p)$

- Geometric Distribution

  Repeat the Bernoulli experiments in the first example until observing 1 (rain or head turning up) for the first times. Each time there is probability=p to observe 1. If Y is the number of 0 observed (no rain or tail turning up), then

  $p_Y(y) = P(Y = y) = p(1-p)^y, y = 0, 1, 2, \cdots.$

  Geometric Distribution: The r.v. Y has a Geometric distribution with p, where p is the probability of observing 1 in each independent trial, $0 \le p \le 1$, if the pmf of Y is given by

  $p_Y(y) = P(Y = y) = p(1-p)^y, y = 0, 1, 2, \cdots.$

Mean: $E(X) = \frac{1-p}{p}$
Variance: $\frac{1-p}{p^2}$
mgf: $M(t) = p[1-(1-p)e^t]^{-1}$ for $t < -ln(1-p)$
Geometric distribution is a special case of Negative Binomial Distribution when r=1.

- Hypergeometric Distribution

An urn containing N objects in which M objects are defective $(M \leq N)$. A sample of n objects are chosen at random without replacement, and let X be the number of defective objects in the n objects chosen out. Then

$$P(X=x) = \frac{\begin{pmatrix} M \\ x \end{pmatrix}\begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, x = L, L+1, \cdots, U$$

$L = max\{0, n-N+M\}, U = min\{M, n\}$

Hypergeometric Distribution: The r.v. X has a

Hypergeometric distribution with parameters M,n,N, where $M, n \leq N$, if the pmf of X is given by

$$p_X(x) = \frac{\begin{pmatrix} M \\ x \end{pmatrix} \begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, x = L, L+1, \cdots, U$$

L=max{0,n-N+M}, U=min{M,n}

Mean:$E(X) = \frac{Mn}{N}$

Variance:$Var(X) = \frac{n\frac{M}{N}(1-\frac{M}{N})(N-n)}{N-1}$

- Asymptotic distribution of Hypergeometric Distribution

If $M, N \to \infty$, and $\lim_{M,N\to\infty} \frac{M}{N} = p, 0 \leq p \leq 1$,Then

$$\lim_{M,N\to\infty} \frac{\begin{pmatrix} M \\ x \end{pmatrix}\begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}$$

$\lim_{M,N\to\infty} \frac{M}{N} = p$

- Poisson Distribution

  Some events are rather rare, they don't happen that often. For instance, car accidents are the exception rather than the rule. Still, over a period of time, we can say something about the nature of rare events.

  eg1. If wearing seat belts reduce the number of death in car accidents. Here, the Poisson distribution can be a useful tool to answer question about benefits of seat belt use.

  eg2. Death of infants

  eg3. The number of misprints in a book

  The Poisson distribution is a mathematical rule that assigns probabilities to the number occurrences in a fixed interval(X). The only thing we have to know to specify the Poisson distribution is the mean number of occurrences for which the symbol λ is often used.

$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

Poisson Distribution The r.v. X has a Possion distribution with parameters λ, if the pmf of X is given by

$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0, 1, 2, \cdots, \infty$

Mean: E(X)=λ

Variance: Var(X)=λ

mgf: $M(t) = e^{\lambda(e^{t}-1)} \forall t$

The Poisson distribution resembles the binomial distribution in that it models counts of events. For example, a Poisson distribution could be used to model the number of accidents at an intersection in a week. However, if we want to use the binomial distribution we have to know both the number of people who make enter the intersection, and the number of people who have an accident at the intersection, whereas the number of accidents is sufficient for applying the Poisson distribution. Thus, the Poisson distribution is cheaper to use because the number of accidents is usually recorded by the police department, whereas the total number of drivers is not. This is supported by the following theorem.

- Asymptotic distribution of Binomial distribution is Poisson distribution

  If r.v. X has a binomial distribution with parameter n and p, $n \to \infty$, $p \to 0$, $\lim_{n\to\infty} np = \lambda$, then

$$\lim_{n\to\infty} p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

- Conditions under which a Poisson distribution holds
  - counts of rare events
  - all events are independent
  - average rate does not change over the period of interest

- Theorem 3.2.1 Let $X_1, X_2, ..., X_n$ be independent random variables such that X_i has Poisson distribution with parameter m_i, for $i = 1, 2, \cdots, n$. Then $Y = \sum_{i=1}^{n} X_i$ has a Poisson distribution with parameter $\sum_{i=1}^{n} m_i$.

  Example 3.2.1-3.2.4

Special Continuous Distribution

- Gamma Distribution: $\Gamma(\alpha, \beta)$

  Motivation:

  gamma function of α: $\Gamma(\alpha) = \int_0^{+\infty} y^{\alpha-1} e^{-y} dy \quad \alpha > 0$

  properties:

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

  $(2)\Gamma(\alpha) = \int_0^{+\infty} y^{\alpha-1} e^{-y} dy =$

  $-y^{\alpha-1} e^{-y}|_0^{\infty} + (\alpha - 1) \int_0^{+\infty} y^{\alpha-2} e^{-y} dy = (\alpha - 1)\Gamma(\alpha - 1)$

  (3)If α is positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$

  Now if we introduce a new variable $y = x/\beta$, where $\beta > 0$, then

$$\Gamma(\alpha) = \int_0^{+\infty} (\frac{x}{\beta})^{\alpha-1} e^{-\frac{x}{\beta}} \frac{1}{\beta} dx,$$

$$\Longrightarrow \int_0^{+\infty} \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-\frac{x}{\beta}} dx = 1$$

Gamma Distribution: A continuous r.v. X has a Gamma distribution with parameters $\alpha > 0, \beta > 0$, if and only if

$$f(x) = \left\{ \begin{array}{cc} \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-\frac{x}{\beta}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{array} \right.$$

Here α is called shape parameter, β is called scale parameter
Mean: E(X)=$\alpha\beta$
Variance: Var(X)=$\alpha\beta^2$
mgf: M(t)=$(1-\beta t)^{-\alpha}$ for $t < \frac{1}{\beta}$
1. How to find mgf, mean and variance (page 151)?
2. Example 3.3.1,3.3.2
The gamma distribution is frequently used to model waiting times; for instance, in life testing, the waiting time until "death" is the random variable which is frequently modeled with a gamma distribution.

Let

w- time interval

W - a r.v. , time needed to obtain exactly k deaths (e.g. k=1)

k - a fixed positive integer

X - a r.v., the count of deaths within the time interval w, following poisson distribution with average count λw at time interval w,

The cdf for W is $P(W \leq w) = 1 - P(W > w)$.

Since the event $\{W > w\}$ means obtaining at most k-1 deaths within time interval w, we have

$$P(W > w) = \sum_{x=0}^{k-1} P(X = x) = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!} = \int_{\lambda w}^{\infty} \frac{z^{k-1} e^{-z}}{(k-1)!} dz$$

The last equation is obtained through integrating by part k-1 times. Therefore,

$$G(w) = P(W \le w) = \begin{cases} 1 - \int_{\lambda w}^{\infty} \frac{z^{k-1}e^{-z}}{(k-1)!}dz = \int_0^{\lambda w} \frac{z^{k-1}e^{-z}}{\Gamma(k)}dz & \text{if } w > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then, pdf is

$$g(w) = \begin{cases} \frac{\lambda^k w^{k-1} e^{-\lambda w}}{\Gamma(k)} & \text{if } w > 0 \\ 0 & \text{otherwise} \end{cases}$$

As a result, W is having a gamma distribution with $\alpha = k, \beta = \frac{1}{\lambda}$.
If W is the waiting time until the first death, that is, if k=1, then pdf of W is

$$g(w) = \begin{cases} \lambda e^{-\lambda w} & \text{if } w > 0 \\ 0 & \text{otherwise} \end{cases}$$

W is said to have an exponential distribution.

- Exponential distribution: $exp(\lambda)$

  A continuous r.v. X has a exponential distribution with parameter $\lambda > 0$, if and only if

$$f(x) = \{ \begin{array}{cc} \lambda e^{-\lambda w} & \text{if } w > 0 \\ 0 & \text{otherwise} \end{array}$$

  Mean: $E(X) = \frac{1}{\lambda}$
  Variance: $Var(X) = (\frac{1}{\lambda})^2$
  mgf: $M(t) = \frac{1}{1-\frac{t}{\lambda}}$ for $t < \lambda$
  Remark:
  1. exponential distribution is a special case of gamma distribution for $\alpha = 1, \beta = \frac{1}{\lambda}$
  2. exponential distribution is often used in survival analysis. Denote survival function as S(x) and X as the survival time,

$$S(x) = P(X > x) = 1 - P(X \le x) = \begin{cases} e^{-\lambda x} & \text{if } x \ge 0 \\ 1 & \text{otherwise} \end{cases}$$

3. Memoryless properties of exponential distribution

X has a $exp(\lambda)$ distribution $\Leftrightarrow P(X > a + t | X > a) = P(X > t)$ for any $a > 0, t > 0$

From this property, we can say that the life length of a cancer patient doesn't depend on how long he/she has survived.

- Chi-square distribution: $\chi^2(r)$

  A continuous r.v. X has a χ^2 distribution with parameter $r > 0$, where r is a positive integer, if and only if

$$f(x) = \begin{cases} \frac{1}{2^{r/2}\Gamma(r/2)} x^{r/2-1} e^{-x/2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean: E(X)=r

Variance: Var(X)=2r

mgf: $M(t) = (1-2t)^{-r/2}$ for $t < 0.5$

Remark:

$\chi^2(r)$ distribution is a special case of gamma distribution $\Gamma(r/2, 2)$

Example 3.3.3, 3.3.4

- Theorem 3.3.1 Let X have a $\chi^2(r)$ distribution. If $k > -r/2$ then $E(X^k)$ exists and it is given by

$$E(X^k) = \frac{2^k \Gamma(\frac{r}{2}+k)}{\Gamma(\frac{r}{2})} \quad \text{if} \quad k > -r/2$$

  Example 3.3.5,3.3.6

- Theorem 3.3.2. Let $X_1, ..., X_n$ be independent random variables. Suppose, for i=1,...,n, that X_i has a $\Gamma(\alpha_i, \beta)$ distribution. Let $Y = \sum_{i=1}^n X_i$. Then Y has $\Gamma(\sum_{i=1}^n \alpha_i, \beta)$ distribution.
  - Corollary 3.3.1 Let $X_1, ..., X_n$ be independent random variables. Suppose, for i=1,...,n, that X_i has a $\chi^2(r_i)$

distribution. Let $Y = \sum_{i=1}^{n} X_i$. Then Y has $\chi^2(\sum_{i=1}^{n} r_i)$ distribution.

- Beta distribution :$\beta(\alpha, \beta)$

  Motivation: Let X_1 and X_2 be two independent random variables that have Γ distribution $(\Gamma(\alpha, 1), \Gamma(\beta, 1))$ and the joint pdf

  $$h(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} e^{-x_1-x_2}$$

  for $0 < x_1 < \infty, 0 < x_2 < \infty$, and zero elsewhere, where $\alpha > 0, \beta > 0$.

  Let $Y_1 = X_1 + X_2, Y_2 = X_1/(X_1 + X_2)$, the joint pdf of Y_1, Y_2 is then

  $$g(y_1, y_2) = \begin{cases} \frac{y_2^{\alpha-1}(1-y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha+\beta-1} e^{-y_1} & \text{for } 0 < y_1 < \infty, 0 < y_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

  where, $\alpha > 0, \beta > 0$.

  Obviously, Y_1 and Y_2 are independent. We can also prove that

The marginal pdf of Y_2 is

$$g_2(y_2) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha-1}(1-y_2)^{\beta-1} & \text{if } 0 < y_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

this pdf is called beta distribution $\beta(\alpha, \beta)$

The marginal pdf of Y_1 is

$$g_1(y_1) = \begin{cases} \frac{1}{\Gamma(\alpha+\beta)} y_1^{\alpha+\beta-1} e^{-y_1} & \text{if } 0 < y_1 < \infty \\ 0 & \text{otherwise} \end{cases}$$

this pdf is $\Gamma(\alpha + \beta, 1)$

Beta Distribution: A continuous random variable X has a β distribution with parameters $\alpha > 0, \beta > 0$, if and only if

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where, $\alpha > 0, \beta > 0$.

Mean: $E(X) = \frac{\alpha}{\alpha+\beta}$
Variance: $Var(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$
Remark: uniform distribution unif(0,1) is a special case of beta distribution, $\beta(1,1)$

- Dirichlet Distribution Let $X_1, X_2, ..., X_{k+1}$ be independent random variables, each having a gamma distribution with $\beta = 1$. Let

$$\begin{aligned} Y_i =& \frac{X_i}{X_1 + X_2 + \cdots + X_{k+1}}, i = 1, 2, ..., k, \\ Y_{i+1} =& X_1 + X_2 + \cdots + X_{k+1} \end{aligned}$$

Then the joint pdf of $Y_1, ..., Y_k, Y_{k+1}$ is given by

$$\frac{y_{k+1}^{\alpha_1+\cdots+\alpha_{k+1}-1} y_1^{\alpha_1-1} \cdots y_k^{\alpha_k-1} (1 - y_1 - \cdots - y_k)^{\alpha_{k+1}-1} e^{-y_{k+1}}}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)\Gamma(\alpha_{k+1})}$$

if $(y_1, ..., y_k, y_{k+1}) \in \{0 < y_i, i = 1, ..., k, y_1 + \cdots + y_k < 1, 0 < y_{k+1} < \infty\}$ and is equal to zero elsewhere.

Random variables $Y_1, .., Y_k$ then have a joint pdf which is called Dirichlet pdf and is given by

$$g(y_1, ..., y_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_{k+1})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})} y_1^{\alpha_1 - 1} \cdots y_k^{\alpha_k - 1}(1 - y_1 - \cdots - y_k)^{\alpha_{k+1}-1}$$

when $0 < y_i, i = 1, ..., k, y_1 + \cdots + y_k < 1$, and is equal to zero elsewhere. Moreover, Y_{k+1} has a gamma distribution $\Gamma(\sum_{i=1}^{k+1} \alpha_i, 1)$ and Y_{k+1} is independent of $Y_1, .., Y_k$.

- F-distribution:$F(r_1, r_2)$
  Motivation: Let U and V be two independent random variables that have $\chi^2(r_1)$ and $\chi^2(r_2)$ distribution respectively. Let $W = \frac{U/r_1}{V/r_2}, Z = V$, the joint pdf of W, Z is then

$$g(w, z) = \begin{cases} \frac{1}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})2^{\frac{r_1+r_2}{2}}} \left(\frac{r_1 z w}{r_2}\right)^{\frac{r_1-2}{2}} z^{\frac{r_2-2}{2}} exp[-\frac{z}{2}(\frac{r_1 w}{r_2} + 1)] \frac{r_1 z}{r_2} \\ \text{for } 0 < w < \infty, 0 < z < \infty \\ 0 \quad \text{otherwise} \end{cases}$$

The marginal pdf of W is then

$$g_1(w) = \begin{cases} \frac{\Gamma(\frac{r_1+r_2}{2})(\frac{r_1}{r_2})^{r_1/2}}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})} \frac{w^{r_1/2-1}}{(1+r_1 w/r_2)^{(r_1+r_2)/2}} & \text{for } 0 < w < \infty \\ 0 & \text{otherwise} \end{cases}$$

This pdf is called F distribution $F(r_1, r_2)$.

Example 3.6.2.

- Normal Distribution :N(μ, σ^2)

Motivation: consider the integral: $I = \int_{-\infty}^{+\infty} e^{-y^2/2} dy$. We have

$$I^2 = \int_{-\infty}^{+\infty} e^{-y^2/2} dy \int_{-\infty}^{+\infty} e^{-x^2/2} dx = 2\pi$$

$$\therefore \int_{-\infty}^{+\infty} e^{-y^2/2} dy = \sqrt{2\pi} \Leftrightarrow \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = 1.$$

$f(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$ for $-\infty < y < \infty$ is called a pdf of a standard normal distribution N(0,1)for a random variable Y. Now consider a new random variable $X = y\sigma + \mu, \sigma > 0$, then

$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $-\infty < y < \infty$ is called a pdf of a normal distribution $N(\mu, \sigma^2)$ for random variable X.

Normal Distribution A continuous r.v. X has a normal distribution $N(\mu, \sigma^2)$ with parameter μ and $\sigma > 0$, if its pdf is

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

Mean: $E(X) = \mu$

Variance: $Var(X) = \sigma^2$

mgf: $M(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$

Example 3.4.1

Remark:

(1) when $\mu = 0$ and $\sigma = 1$, it is standard normal distribution

(2) $X \sim N(\mu, \sigma^2) \Leftrightarrow (X - \mu)/\sigma \sim N(0, 1)$

This property simplifies the calculation of probability concerning normally distributed variables

Example 3.4.3, 3.4.4, 3.4.5

(3)μ is called a location parameter since change its value simply changes the location of the middle of the normal pdf; σ is called a scale parameter because changing its value changes the spread of the distribution.

- Theorem 3.4.1 If r.v. X is $N(\mu, \sigma^2), \sigma^2 > 0$, then the r.v. $V = (X - \mu)^2/\sigma^2$ is $\chi^2(1)$.
- Theorem 3.4.2 Let $X_1, ..., X_n$ be independent random variables such that, for i=1,...,n, X_i has a $N(\mu_i, \sigma_i^2)$ distribution. Let $Y = \sum_{i=1}^n a_i X_i$, where $a_1, ..., a_n$ are constants. Then the distribution of Y is $N(\sum_{i=1}^n a_i u_i, \sum_{i=1}^n a_i^2 \sigma_i^2)$.
- Corollary 3.4.1. Let $X_1, ..., X_n$ be iid random variables with a common $N(\mu, \sigma^2)$ distribution. Let $\bar{X} = n^{-1} \sum_{i=1}^n X_i$. Then $\bar{X}$ has a $N(\mu, \sigma^2/n)$ distribution.
- Contaminated Normals Suppose we are observing a random variable which most of the time follow a standard normal distribution but occasionally follows a normal distribution with a

larger variance. In application, we might say that most of the data are "good" but that there are occasional outliers. This can be described as follows:

$Z \longrightarrow$ "good data" $\sim N(0,1)$

$I_{1-\epsilon} \longrightarrow$ indicator r.v. with bernoulli distribution

$P(I_{1-\epsilon}=1)=1-\epsilon, \quad P(I_{1-\epsilon}=0)=\epsilon$

Assume Z and $I_{1-\epsilon}$ are independent, and let

$W=ZI_{1-\epsilon}+\sigma_c Z(1-I_{1-\epsilon})$. The distribution of W is of interest and we can prove that it is a mixture of normals.

$$f(w)=\phi(w)(1-\epsilon)+\phi(w/\sigma_c)\frac{\epsilon}{\sigma_c}$$

where ϕ is the pdf of a standard normal.

$E(W)=0, Var(W)=1+\epsilon(\sigma_c^2-1)$

This can be easily extended to the general situation where X=a+bW.

- t-Distribution: t(r)

  Motivation: If $X_1 \sim N(0,1)$, $X_2 \sim \chi^2(r)$, and X_1 and X_2 are

independent. What is the distribution of r.v. $T = \frac{X_1}{\sqrt{(X_2/r)}}$?
Using change-of-variable technique, we can prove that

$$g(t) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{(\pi r)}\Gamma(\frac{r}{2})} \frac{1}{(1+t^2/r)^{(r+1)/2}}$$

for $-\infty < t < \infty$.
the pdf g(t) of T is called t distribution with degree of freedom r.
Example 3.6.1.

- Theorem 3.6.1 Student's Theorem
  Let $X_1, ..., X_n$ be iid random variables each having a normal distribution with mean μ and variance σ^2. Define the random variable, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$
  (a)$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
  (b) $\bar{(}X)$ and S^2 are independent
  (c) $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$
  (d) $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

- Mixture Distribution

  Example 3.7.1-3.7.4