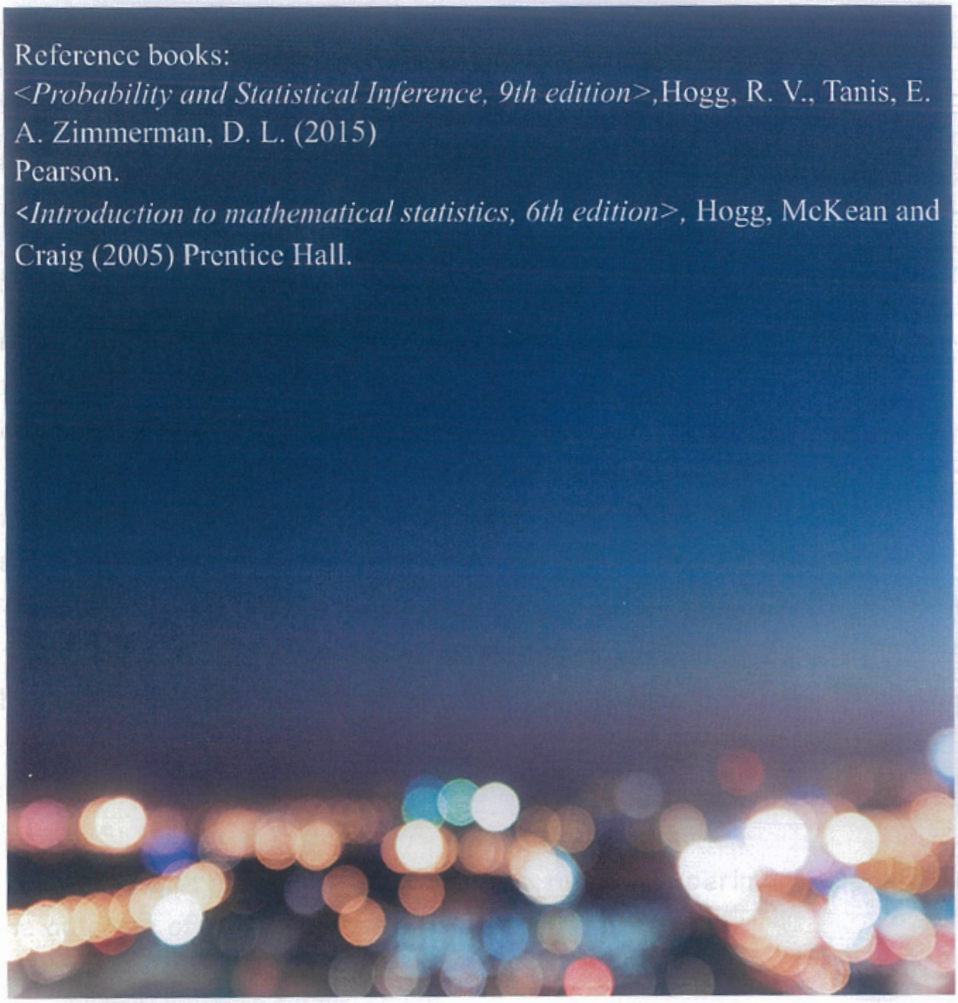
Reference books:

<*Probability and Statistical Inference, 9th edition*>,Hogg, R. V., Tanis, E. A. Zimmerman, D. L. (2015)
Pearson.

<*Introduction to mathematical statistics, 6th edition*>, Hogg, McKean and Craig (2005) Prentice Hall.



STA 2001 PROBABILITY AND STATISTICS I

STA2001 Section ONE Probability

O Probability theory and Statistics

1.Probability theory: a branch of mathematics and the theory for the probability and analysis of random phenomena

2.Statistics: the theory for the analysis of the data. How to extract information from data is the core of statistics. Information can be used for making decisions and predictions

3.Data: observation/measurements of random phenomena

4.Information: data becomes information once it has been analyzed in some fashion

5.Relationship: Probability theory is the mathematical foundation of statistics, while statistics is the application of probability theory.

O Fundamental Concepts

1.**Random experiment/phenomenon:** Any procedure that can be repeated infinitely many times and has a well-defined set of possible outcomes. At the same time, the outcome cannot be predicted with certainty.

2.Outcome/Sample space: the collection of all possible outcomes of a given random experiment, denoted by " S ".

3.Event/Set: Given an outcome space S, let A be a part of the collection of outcomes in S; that is, $A \subset S$. Then A is called an event.

4.An event is occurred. The time when the random experiment is performed and the outcome of the experiment is in A.

O Set Theory

1.A': the complement of A in S is the set of all elements in S that are not in A.

2.Commutative laws: $A \cup B = B \cup A$, $A \cap B = B \cap A$

3.Associative laws: $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$

4.Distributive laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,

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

5.De Mogan's laws: $(\bigcup_{n\geq 1} A_n)' = \bigcap_{n\geq 1} A_n'$; $(\bigcap_{n\geq 1} A_n)' = \bigcup_{n\geq 1} A_n'$

6.Special terminology associated with events that is often used by statisticians includes the following:

①$A_1, A_2, \ldots, A_k$ are mutually exclusive events means that $A_i \cap A_j = \varnothing$, i doesn't equal j; that is, $A_1, A_2, \ldots, A_k$ are disjoint sets;

②$A_1, A_2, \ldots, A_k$ are exhaustive events means that $A_1 \cup A_2 \cup \cdots \cup A_k = S$.

So if $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive events, we know that $A_i \cap A_j = \varnothing$, i doesn't equal j, and $A_1 \cup A_2 \cup \cdots \cup A_k = S$.

O Definition of Probability

1. An intuitive definition of probability: the probability of event A, denoted by P(A), often called the chance of A occurring. (1)Consider repeating the experiment a number of times—say, "n" times. (2)Count the number of times that event A actually occurred throughout these n performances, called the frequency of event A and is denoted by N (A). (3)The ratio N (A) / n is called the relative frequency of event A in these n repetitions of the experiment.

So it can be defined as " $P(A) = \lim_{n\to\infty} N(A)/n$" **(does not exist because the sequence does not converge. It is held by law of large numbers)**

2. Definition of Probability

★ the sequence $S(n) = \frac{N(A)}{n}$ does not converge.

So $\lim_{n\to\infty} \frac{N(A)}{n}$ does not exist.

Definition[Probability]

A real-valued, set function P that assigns to each event A in the sample space S, a number $P(A)$, called the probability of the event A such that the following properties are satisfied:

1. $P(A) \geq 0$.
2. $P(S) = 1$.
3. if $A_1, A_2, A_3, \cdots$ are countable and mutually exclusive events (we can give one-to-one correspondence with natural set $\mathbb{N}$)

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

or equivalently,

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

• The intuitive definition is based on Bernoulli Law of Large numbers (大数定律)

$\forall \varepsilon > 0$

$P(|S(n) - p_0| > \varepsilon) < \frac{\varepsilon}{n^2}$

& $P(|S(n) - p_0| < \varepsilon) \geq 1 - \frac{\varepsilon}{n^2}$

O Properties of Probability

Property 1: For each event A, $P(A) = 1 - P(A')$

Property 2: $P(\varnothing) = 0$.

Property 3: If events A and B are such that $A \subseteq B$, then $P(A) \leq P(B)$ —(proof)

$B = B \cap S = B \cap (A \cup A') = (B \cap A) \cup (B \cap A') = A \cup (B \cap A')$ mutually exclusive

So $P(B) = P(A) + P(B \cap A') \geq P(A)$

Property 4: For each event A, $P(A) \leq 1$.

Property 5: For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(proof) $A \cup B = (A \cup B) \cap S = (A \cup B) \cap (A \cup A') = A \cup (B \cap A')$ mutually exclusive

So $P(A \cup B) = P(A) + P(B \cap A') \xrightarrow[\text{process of proof}]{\text{Property 3}} P(A) + P(B) - P(A \cap B)$

Probability Space*

A probability space is a triple (S, F, P)

1. S: the sample space
2. F is σ-algebra on S and a set of subsets of S and called the event space
   - $S \in F$
   - F is closed under complement
   - F is closed under countable unions
3. $P : F \to [0,1]$ is the probabilty measure such that

$$P(A) \geq 0, \forall A \in F, \quad P(S) = 1, \quad P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

for countable and mutually exclusive $A_1, A_2, \cdots$

Note*:
1."F is σ-algebra on S and a set of subsets of S"—One possible F contains all the subsets of S.
2."closed under complement"—if A is in F, then A' is in F.
3."closed under countable unions"—if $A_1, A_2 \ldots\ldots A_k$ are in F, $\cup_{i=1}^{k} A_i$ is in F.

○ The probability function of equally likely events

Let $S = \{e_1, e_2, \ldots, e_m\}$, where each e_i is a possible outcome of the experiment. If each of these outcomes has the same probability of occurring, we say that the m outcomes are **equally likely**. That is, $P(\{e_i\}) = 1/\mathrm{m}$, $i = 1, 2, \ldots, m$.

In this case, $P(A)$ is equal to the number of ways favorable to the event A divided by the total number of ways in which the experiment can terminate. That is, under this assumption of equally likely outcomes, we have $P(A) = \mathrm{N(A)}/\, N(S)$

○ Multiplication Principle

Suppose that an experiment (or procedure) E_1 has n_1 outcomes and, for each of these possible outcomes, an experiment (procedure) E_2 has n_2 possible outcomes. Then the composite experiment (procedure) E_1E_2 that consists of performing first E_1 and then E_2 has n_1n_2 possible outcomes.

○ Method of Enumeration (Permutation and Combination)

Permutation of n distinct objects:

Definition 1.2-2
Each of the ${}_nP_r$ arrangements is called a **permutation of n objects taken r at a time.**

$${}_nP_r = n! \,/\, (n-r)!$$

• ☆ Fractorial $x!$ is defined by Gamma function

$$\Gamma(x+1) = \int_0^{\infty} t^x e^{-t} dt, \quad x > -1$$

$\Gamma(x+1) = x\,\Gamma(x)$ by integration by parts

Combination of n objects taken r at a time → using ${}_nP_r$ & multiplication principle

Definition 1.2-6
Each of the ${}_nC_r$ unordered subsets is called a **combination of n objects taken r at a time**, where

$${}_nC_r = \binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

Definition 1.2-7
Each of the ${}_nC_r$ permutations of n objects, r of one type and $n-r$ of another type, is called a **distinguishable permutation**.

The argument used in determining the binomial coefficients in the expansion of $(a+b)^n$ can be extended to find the expansion of $(a_1+a_2+\cdots+a_s)^n$. The coefficient of ${a_1}^{n_1}{a_2}^{n_2}\cdots{a_s}^{n_s}$, where $n_1+n_2+\cdots+n_s=n$, is

$$\binom{n}{n_1,n_2,\ldots,n_s}=\frac{n!}{n_1!n_2!\cdots n_s!}.$$

This is sometimes called a **multinomial coefficient**.

O Conditional Probability

Definition 1.3-1
The **conditional probability** of an event A, given that event B has occurred, is defined by

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

sample space → shrinks to B

probability with two different experiments

provided that $P(B)>0$.

mutually exclusive

Conditional probability satisfies the axioms for a probability function.

Prof. $P(A_1\cup A_2\cup\cdots\cup A_n\mid B)=\frac{P((A_1\cup A_2\cup\cdots\cup A_n)\cap B)}{P(B)}=\frac{P((A_1\cap B)\cup(A_2\cap B)\cup\cdots\cup(A_n\cap B))}{P(B)}=\sum_{i=1}^{n}\frac{P(A_i\cap B)}{P(B)}$

Definition 1.3-2
The probability that two events, A and B, both occur is given by the **multiplication rule**,

$$P(A\cap B)=P(A)P(B\mid A),$$

$=\sum_{i=1}^{n}P(A_i\mid B)$ (By probability functions)

provided $P(A)>0$ or by

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

provided $P(B)>0$.

Definition

The probability that three events, A, B and C all occur is given by the multiplication rule

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

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

$$\Rightarrow P(A\cap B\cap C)=P(A)P(B|A)P(C|A\cap B)$$

O Independent events

Events that are independent are sometimes called **statistically independent, stochastically independent, or independent in a probabilistic sense.**

Definition 1.4-1
Events A and B are **independent** if and only if $P(A\cap B) = P(A)P(B)$. Otherwise, A and B are called **dependent** events.

$\Rightarrow P(A|B) = P(A)$ & $P(B|A) = P(B)$ ($P(B)>0$) ($P(A)>0$)

★ If A and B are independent, A' and B, A and B', A' and B' are all independent.

Prof. $P(A'\cap B) = P(B)\cdot P(A'|B) = P(B)\cdot(P(A\cup A'|B) - P(A|B)) = P(B)(1-P(A)) = P(B)\cdot P(A')$

Definition 1.4-2
Events A, B, and C are **mutually independent** if and only if the following two conditions hold:

(a) A, B, and C are pairwise independent; that is,

$$P(A\cap B) = P(A)P(B), \quad P(A\cap C) = P(A)P(C),$$

and

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

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

independent

the same way to prove others.

A, B are independent $\Leftrightarrow$ A', B are independent $\Leftrightarrow$ A, B' ..... $\Leftrightarrow$ A', B' .....

if and only if

There are some events which are pairwise indepent and not mutually independent.

meaning that (a) $\not\Rightarrow$ (b) & (b) $\not\Rightarrow$ (a)

For four / more events $\Rightarrow$ the same! (each pair, triple, quartet .....) and $P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$

O Bayes' Theorem

Let $B_1, B_2, \ldots, B_m$ constitute a partition of the sample space S.

That is, $S = B_1 \cup B_2 \cup \cdots \cup B_m$ and $B_i \cap B_j = \emptyset$, i is not equal to j.

The events $B_1, B_2, \ldots, B_m$ are for sure mutually exclusive and exhaustive.

Suppose the **prior probability** of the event B_i is positive; that is, $P(B_i) > 0$, i = 1, ..., m. If A is an event, then A is the union of m mutually exclusive events, namely $A = (B_1 \cap A)\cup(B_2 \cap A)\cup\cdots\cup(B_m \cap A)$

Thus, we get which is sometimes called the **law of total probability.** (On the right hand side)

$$P(A) = \sum_{i=1}^{m} P(B_i \cap A) = \sum_{i=1}^{m} P(B_i)P(A \mid B_i),$$

If P(A) is larger than 0, we get the **Bayes' Theorem:**

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{\sum_{i=1}^{m} P(B_i)P(A \mid B_i)}, \qquad k = 1, 2, \ldots, m.$$

$P(B_k|A)$ — likelihood of B_k

$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)}$ (Baye's Theorem)

The conditional probability $P(B_k \mid A)$ is often called the **posterior probability** of B_k.

- If A, B, C are mutually independent,

(1) A, (B∩C) are independent

prof. $P(A\cap(B\cap C)) = P(A)\cdot P(B)\cdot P(C) = P(A)\cdot P(B\cap C)$

(2) A', (B∩C') are independent

prof. $P(A'\cap(B\cap C')) = P((A'\cap C')\cap B) = P((A\cup C)'\cap B) = P(B) - P(B\cap(A\cup C))$

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

$= P(B)(1-P(A)-P(C)+P(A)P(C))$

$= P(B)(1-P(A))(1-P(C))$

$= P(A')\cdot P(B\cap C')$

⇔ A, (B∩C') are independent

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

(3) A, B∪C are independent

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

$= P(A)P(B\cup C)$ (by $P(B)+P(C)-P(B)P(C)$)

(4) $P(A'\cap B'\cap C') = P(A'\cap(B'\cap C'))$

$= P(A')\cdot P(B'\cap C')$

$= P(A')P(B')P(C')$

(A, B∪C independent ⇕ A', (B∪C)' ≠ (B'∩C') independent)

Principle of inclusion-exclusion ⇕ "Property S"

*Note: Morgan Law & repulsion principle

Two questions deserving discussion

(i) Monte Hall Problem

(Three curtains, just one prize behind, a contestant selects one at random, then Hall opens one to reveal a worthless one, should the contestant change her choice?)

* Error: On considering, we'll think it as $\frac{1}{2}$ / $\frac{1}{2}$ probability. But it is not true. It is not choosing one from two at random.

* Analyze: First, she has $\frac{1}{3}$ probability choose a right one $\Rightarrow$ so P(not change) $= \frac{1}{3}$
she has $\frac{2}{3}$ probability choose a wrong one $\not\Rightarrow$ so P(change) $= \frac{2}{3}$
just change can will

* Ways to solve this problem:
Assume the contestant chooses door "1" (no effect)
P(prize in door "i") $= \frac{1}{3}$, $i = 1, 2, 3$

$$P(\text{hold \& win}) = P(\text{prize in door "i"} \mid \text{host opening door "3"}) = \frac{P(\text{prize in door "1"} \cap \text{host opening door "3"})}{P(\text{host opening door "3"})}$$

no matter because prize can be anywhere.

$$= \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(\text{hold \& loose}) = P(\text{prize in door "2"} \mid \text{host open door "3"}) = \cdots = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

(ii) Roll a pair of 4-sided dice, observing the sum & getting P(C)

A = {a sum of 3}; B = {a sum of 3 or 5}; C = {a sum of 3 before the sum of 5}

* Understand: ① C means that keep rolling until 3 or 5 happens.
② Deeply means that we just care about 3 & 5.

* Solve this problem:

1) 3 can emerge under i^{th} rolling without 5:

$$P = \frac{2}{16} + \frac{10}{16} \times \frac{2}{16} + \left(\frac{10}{16}\right)^2 \times \frac{2}{16} + \cdots = \frac{2}{16} \times \frac{1}{1 - \frac{10}{16}} = \frac{1}{3}$$

2) Conditional probability:

comphrehension

$$P(C) = P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{N(A \cap B)}{N(B)} = \frac{2}{6} = \frac{1}{3}$$

B — whole: 3 or 5 & A: just 3

STA2001 Section TWO Discrete Distributions

Random Variable

Definition 2.1-1
Given a random experiment with an outcome space S, a function X that assigns one and only one real number $X(s) = x$ to each element s in S is called a **random variable**. The **space** of X is the set of real numbers $\{x : X(s) = x, s \in S\}$, where $s \in S$ means that the element s belongs to the set S.

Definition[Random Variable]

Given a random experiment with sample space S, a function $X : S \to \overline{S} \subseteq R$ that assign one real number $X(s) = x$ to each $s \in S$ is called Random Variable (RV).

Remarks:

- Repeat a random experiment $\Leftrightarrow$ generate a number from $\overline{S}$
- X can be not one to one, old experiment with S new random experiment with $\overline{S}$

Universal expression

- uppercase letters, e.g. $X, Y, Z \to$ RVs
- lowercase letters, e.g. $x, y, z \to$ the numeric values of RV X, Y, Z, respectively

For a given random experiement, two probability functions are involved through $X : S \to \overline{S}$,

- $P_r(\cdot)$ is the probability function associated with S
- $P(\cdot)$ is the probability function associated with $\overline{S}$

$$P(X = x) \triangleq P(\{X = x\}) = P_r(\{s | X(s) = x, s \in S\})$$
$$P(X \in A) \triangleq P(\{X \in A\}) = P_r(\{s | X(s) \in A, s \in S\})$$

think of the example of flip a coin to understandard the notations.

Discrete Random Variable

Suppose that the space S contains a countable number of points; that is, either S contains **a finite number of points**, or t**he points of S can be put into a one-to-one correspondence with the positive integers.** Such a set S is called a set of ***discrete points*** or simply ***a discrete outcome space.*** Furthermore, any random variable defined on such

an S can assume at most a countable number of values, and is therefore called ***a random variable of the discrete type.***

○ Probability Mass Function (pmf)

Definition 2.1-2

The pmf $f(x)$ of a discrete random variable X is a function that satisfies the following properties:

(a) $f(x) > 0, \qquad x \in S;$

(b) $\sum_{x \in S} f(x) = 1;$

(c) $P(X \in A) = \sum_{x \in A} f(x), \qquad$ where $A \subset S.$

Of course, we usually let f(x) = 0 when x $\in$ S'; thus, the domain of f(x) is the set of real numbers.

○ Cumulative Distribution Function (CDF)

Definition[cdf]

The function $F(x) : R \to [0\ 1]$:

$$F(x) = P(X \leq x)$$

is called the cumulative distribution function (cdf).

1. $F(x)$ is nondecreasing and moreover,

$$P(X \leq x) = \sum_{x' \leq x, x' \in S} f(x').$$

2. relation between the probability function and the cdf

$$P(a \leq X \leq b) = F(b) - F(a)$$

○ Uniform Distribution

When a pmf is constant on the space or support, we say that the distribution is **uniform** over that space.

○ Mathematical Expectation

Definition 2.2-1

If $f(x)$ is the pmf of the random variable X of the discrete type with space S, and if the summation

$$\sum_{x \in S} u(x)f(x), \qquad \text{which is sometimes written} \qquad \sum_{S} u(x)f(x),$$

exists, then the sum is called the **mathematical expectation** or the **expected value** of $u(X)$, and it is denoted by $E[u(X)]$. That is,

$$E[u(X)] = \sum_{x \in S} u(x)f(x).$$

Theorem 2.2-1 When it exists, the mathematical expectation E satisfies the following properties:

(a) If c is a constant, then $E(c) = c$.

(b) If c is a constant and u is a function, then

$$E[c\,u(X)] = cE[u(X)].$$

(c) If c_1 and c_2 are constants and u_1 and u_2 are functions, then

$$E[c_1u_1(X) + c_2u_2(X)] = c_1E[u_1(X)] + c_2E[u_2(X)].$$

Let g(X) = (X – b)^2 where b is a constant to be chosen and suppose E [(X – b)^2] exists. The value of b = E(X) makes E [(X – b)^2] minimized. Mean is a minimum min square error (MMSE) estimator.

O **Special Mathematical Expectation**

(i) Let g(x) = x, we have the **mean** of random variable X:

E(X)= $$\mu = \sum_{x\in S} xf(x) = u_1f(u_1) + u_2f(u_2) + \cdots + u_kf(u_k).$$

(ii) Let g(x) = (x - E(X))^2, we have the **variance** of random variable X:

$$\sigma^2 = \mathrm{Var}(X) = \sum_{x\in S} (x-\mu)^2 f(x) = (u_1-\mu)^2 f(u_1) + (u_2-\mu)^2 f(u_2) + \cdots + (u_k-\mu)^2 f(u_k).$$

And also

Var(X) = E(X^2) - E(X)^2

The positive square root of the variance is called the **standard deviation** of X and is denoted by the Greek letter σ (sigma)

$$Var(c) = 0, \quad Var(cX) = c^2\,Var(X)$$

Properties of Variance: Let c be a constant

(iii) Let g(x) = x^r, we have the **rth moment** of random variable X about the origin/b (r is a positive integer):

$$E(X^r) = \sum_{x\in S} x^r f(x) \qquad E[(X-b)^r] = \sum_{x\in S} (x-b)^r f(x)$$

Definition 2.3-1

Let X be a random variable of the discrete type with pmf $f(x)$ and space S. If there is a positive number h such that

$$E(e^{tX}) = \sum_{x\in S} e^{tx} f(x)$$

exists and is finite for $-h < t < h$, then the function defined by

$$M(t) = E(e^{tX})$$

is called the **moment-generating function of** X (or of the distribution of X). This function is often abbreviated as mgf.

- **Moment Generating Function (mgf)**
- **Properties of mgf**

(i) M(0)=1

(ii) If two RVs have the same mgf, they have the same probability distribution. (i.e. pmf or pdf) —— a crucial theorem to be proved.

(iii) Derivatives:

$$M'(t) = \sum_{x \in \overline{S}} x e^{tx} f(x)$$

$$M''(t) = \sum_{x \in \overline{S}} x^2 e^{tx} f(x)$$

$$M^{(r)}(t) = \sum_{x \in \overline{S}} x^r e^{tx} f(x)$$

Setting t = 0, we get

$$M'(0) = \sum_{x \in S} x f(x) = E(X),$$

$$M''(0) = \sum_{x \in S} x^2 f(x) = E(X^2),$$

$$M^{(r)}(0) = \sum_{x \in S} x^r f(x) = E(X^r).$$

Bernoulli Distribution

- Description: A **Bernoulli experiment** is a random experiment, the outcome of which can be classified in one of two mutually exclusive and exhaustive ways—say, success or failure (e.g., female or male, life or death)
- Pmf: Let X be a random variable associated with a Bernoulli trial by defining it as follows: **X (success) = 1 and X (failure) = 0.**

  That is, the two outcomes, success and failure, are denoted by one and zero, respectively. The pmf of X can be written as

  $$f(x) = p^x(1-p)^{1-x} \; ; \; x = 0, 1$$

we say that X has a Bernoulli distribution.

- Expectation, Variance & mgf:

$$\mu = E(X) = \sum^{1} x p^x (1-p)^{1-x} = (0)(1-p) + (1)(p) = p,$$

$$\sigma^2 = \mathrm{Var}(X) = \sum_{x=0}^{1} (x-p)^2 p^x (1-p)^{1-x}$$

$$= (0-p)^2(1-p) + (1-p)^2 p = p(1-p) = pq.$$

. Mgf: $M(t) = E[e^{tX}] = e^t \cdot p + (1-p), \quad t \in (-\infty, \infty)$

Binomial Distribution

- Bernoulli Trials: A Bernoulli experiment performed n times **independently** (all trials are independent) and the probability of success, say p, **remains the same** from trial to trial.
- Random sample of size n from a Bernoulli distribution: In a sequence of n Bernoulli trials, let X_i denote the Bernoulli RV associated with the i^{th} trial.

  An observed sequence of n Bernoulli trials will be **n-tuple of zeros and ones,** which is called a random sample of size n from a Bernoulli distribution.

  $$P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$$

- Definition (pmf)

Definition[Bionmial distribution]

A RV X is said to have a binomial distribution, if the range space $S = \{0, 1, \cdots, n\}$ and the pmf

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \cdots, n$$

and denoted by $X \sim b(n, p)$, where the constants n, p are parameters of the distribution.

The constants n and p are called the **parameters** of the binomial distribution.

CDF:

$$F(x) = P(X \le x) = \sum_{y \in \{X \le x\}} f(y) = \sum_{y=0}^{[x]} \binom{n}{y} p^y (1-p)^{n-y},$$

where $x \in (-\infty, \infty)$ and $[x]$ is the largest integer $\le x$.

- mgf, Expectation and Variance

$$M(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x}$$

$$= \left[(1-p) + pe^t\right]^n, \qquad -\infty < t < \infty,$$

$$\mu = E(X) = M'(0) = np$$

$$\sigma^2 = E(X^2) - [E(X)]^2 = M''(0) - [M'(0)]^2$$

$$= n(n-1)p^2 + np - (np)^2 = np(1-p).$$

Negative Binomial Distribution

- Description: We are interested in the situation that we observe a **sequence of Bernoulli trials** until **exactly r successes** occur, where r is **a fixed positive integer. (number of Bernoulli trials** given **a fixed success number r)**

- Getting pmf: Define a **Random Variable X** to denote the trial number at which the r^{th} success is observed. Then X has the rangeS={r, r+1, ⋯}.

Let f (x) denote the pmf of X.

Then f (x) = P({At the x^{th} trial, r^{th} success is observed})

= P({For the first x-1 trials, r-1 success have been observed}) ∩ {At the x^{th} trial, the outcome is success})

= P(A∩B) = P(A)*P(B)(Because A and B are independent.)

Therefore, { P(A)=p, P(B) can be calculated by X~b(r-1, x-1) }

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \cdots$$

- Formal Definition:

Definition[Negative Binomial Distribution]

If a RV X has its pmf in the form of

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \cdots$$

then X is said to have a negative binomial distribution with the probability of success p and the number of successes r we are interested in.

This distribution get its name due to the negative binomial series

$$(1-w)^{-r} = \sum_{x=r}^{\infty} \binom{x-1}{r-1} w^{x-r}$$

negative binomial expansion

Method I: (Show the expansion)

$(1-w)^{-r} = \left(\frac{1}{1-w}\right)^r = \left(\sum_{i=0}^{\infty} w^i\right)^r$ (Talor expansion)

Consider w^k: $\sum_{k=0}^{\infty} \binom{k+r-1}{r-1} w^k$ (k balls, r-groups, r can be 0) rewrite $\sum_{k=r}^{\infty} \binom{k-1}{r-1} \cdot w^{k-r}$

(One way to expand)

Method II: Binomial Theorem

$(1-w)^{-r} = \sum_{k=0}^{\infty} \binom{-r}{k} (-w)^k = \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} w^k = \sum_{k=0}^{\infty} \binom{r+k-1}{k} w^k$

- Geometric Distribution: (the special case when r=1)

Definition[Geometric Distribution]

If a RV X has its pmf in the form of

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \cdots$$

then X is said to have a geometric distribution with the probability of success p.

For the case $r = 1$,

$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \cdots \rightarrow$ geometric distribution

For a positive integer k,

$$P(X > k) = \sum_{x=k+1}^{\infty} p(1-p)^{x-1} = \frac{(1-p)^k p}{1-(1-p)} = (1-p)^k$$

$$P(X \le k) = \sum_{x=1}^{k} p(1-p)^{x-1} = 1 - P(X > k) = 1 - (1-p)^k$$

1) $\sum_{x=r}^{\infty} f(x) = p^r \cdot \sum_{x=r}^{\infty} \binom{x-1}{r-1} (1-p)^{x-r}$

$= p^r \cdot (1-(1-p))^{-r} = p^0 = 1$

2) mgf: $M(t) = \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r}$

$= \frac{(pe^t)^r}{(1-(1-p)e^t)^r} = \left(\frac{pe^t}{1-(1-p)e^t}\right)^r$

3) Expectation & Variance:

$E(X) = M'(0) = \frac{r}{p}$, $E(X^2) = \frac{r^2}{p^2} + \frac{r(1-p)}{p^2}$

$Var(X) = \frac{r(1-p)}{p^2}$ Mean and Variance of Negative Binomial Distribution: (do it by hand)

Poisson Distribution

- Description: There are experiments that result in **counting the number of times that particular events occur within a given period or for a given physical object.**

Counting such events can be seen as observations of a RV associated with an **approximate Poisson process (APP).**

- Approximate Poisson Process (APP):

Definition 2.6-1
Let the number of occurrences of some event in a given continuous interval be counted. Then we have an **approximate Poisson process** with parameter $\lambda > 0$ if the following conditions are satisfied:

(a) The numbers of occurrences in nonoverlapping subintervals are independent.

(b) The probability of exactly one occurrence in a sufficiently short subinterval of length h is approximately λh.

(c) The probability of two or more occurrences in a sufficiently short subinterval is essentially zero.

- Consider a random experiment described by APP. Let X denote the number of occurrences in **an interval with length 1/a unit interval.** We aim to find an approximation for f(x) = P(X = x) with x = 0, 1 ,2, ⋯

To this goal,

1. **Partition** the unit interval into n **equally spaced subintervals.** In each unit we have a Bernoulli Distribution (according to definition 2.6-1 (c))
2. If n is sufficiently large (n >> x), P(X = x) can be approximated by the probability that **exactly x of these n subintervals each has one occurrence.** The probability of one occurrence in any subinterval (with length 1/n) is approximately λ*(1/n) (according to definition 2.6-1 (b))
3. The n Bernoulli experiments are independent. Therefore occurrence and nonoccurrence in the n subintervals are n Bernoulli trials with probability of success λ/n. (according to definition 2.6-1 (a))

n^{th} Bernoulli trials ⇒ Binomial Distribution

- Definition (pmf):

Definition

It can be verified

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, , \cdots$$

is a well-defined pmf. If a RV X has $f(x)$ as its pmf, then X is said to have a Poisson distribution with the parameter λ and denoted by $X \sim$ Poission(λ).

(Show how to get the pmf by hand)

$$f(n,x) = \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$f(x) = \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty}\frac{n!}{x!(n-x)!}\frac{\lambda^x}{n^x}\cdot\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n!}{(n-x)!\,n^x} \cdot \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} \cdot \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}$$

$\lim_{n\to\infty}\left(\left(1+\frac{1}{\frac{n}{-\lambda}}\right)^{\frac{n}{-\lambda}}\right)^{-\lambda} = \lim_{\frac{n}{-\lambda}\to-\infty}\left(\left(1+\frac{1}{(-\frac{n}{\lambda})}\right)^{-\frac{n}{\lambda}}\right)^{-\lambda} = e^{-\lambda}$

"$\to 1$" by comparison $\in$ "1" "$e^{-\lambda}$" by Talor series.

So $f(x) = \frac{\lambda^x}{x!}e^{-\lambda}$, $x = 0, 1, \cdots$

Poisson Distribution is a good approximation of Binomial Distribution

- mgf, Mean and Variance of Poisson Distribution:

Conclusion: λ is the mean and variance of $X \sim$ Poisson(λ): the average number of occurrences in **the unit interval.**

(Show the process by hand)

mgf: $M(t) = E(e^{tX}) = e^{\lambda e^t - \lambda}$

- ***Note: choose appropriate λ & the unit interval**

* Origin of negative binomial:

- $\binom{r+y-1}{y} = (-1)^y \cdot \binom{-r}{y}$

proof. expansion: $\binom{r+y-1}{y} = \frac{(r+y-1)(r+y-2)\cdots(r+1)r}{y(y-1)(y-2)\cdots 2\cdot 1} = \frac{(-r)(-r-1)(-r-2)\cdots(-r-y+1)}{y!}\cdot(-1)^y = (-1)^y\binom{-r}{y}$

Similar to binomial coefficiency.

* mgf, Mean & Variance of Poisson Distribution

$$M(t) = \sum_{x=0}^{\infty} e^{tx}\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(e^t\cdot\lambda)^x}{x!} = e^{-\lambda}\cdot e^{\lambda e^t} = e^{\lambda e^t-\lambda}$$

Talor Expansion

$\mu = E(X) = M'(0) = \left(e^{\lambda e^t-\lambda}\right)'\Big|_{t=0} = \lambda$

$E(X^2) = \lambda^2+\lambda$, So

$Var(X) = E(X^2) - E^2(X) = \lambda$ ------ We get the meaning of "λ".

* Geometric RV Y (with p) (negative binomial RV: X)

$X = \sum_{i=1}^{r} Y_i$: 1th, 2th, ..., rth; Y_1, Y_2, ..., Y_r

$\{Y_i\}_{i=1}^{r}$ mutually independent

$E(X) = \frac{r}{p}$ (every Y_i: $E(Y_i) = \frac{1}{p}$)

$Var(X) = \sum_{i=1}^{r} Var(Y_i) = \frac{r(1-p)}{p^2}$

$M_X(t) = E(e^{tX}) = E(e^{t\sum_{i=1}^{r} Y_i}) = E(\prod_{i=1}^{r} e^{tY_i}) = \prod_{i=1}^{r} E(e^{tY_i}) = \prod_{i=1}^{r} M_{Y_i}(t)$

(mutually independent)

Riemann-Stieltjes integral

Abstract Riemann-Stieltjes integration is an optional topic for a first course in real analysis. In this paper, we examine some of the pedagogical reasons in favor of its inclusion and some of the technical anachronisms associated with it.

- **Introduction:** As the name suggests, Riemann-Stieltjes (RS) integration is a notion of integration properly generalizing Riemann integration — the type of integration covered in basic calculus. It will be in an upper-level course in real analysis. When studying Riemann integration, one may make use of either Riemann sums or upper and lower sums. This rather subtle point can be brought into sharp focus when one discovers that the same does not hold for RS integration, at least when the distribution function is discontinuous. Far from being pathological, the discontinuous distribution function is the vehicle by which RS integration unites the study of Riemann integration with that of probability theory or numerical integration.

The curriculum of a mathematics major usually includes a course in calculus-based probability theory. A student who has mastered this material is in a good position to appreciate the power of the RS integral. RS integration not only unites the apparently disconnected topics of discrete and continuous probability distributions, but facilitates the study of mixed-type distributions.

- **Two Competing Definitions:** The symbol F will always be used to represent a distribution function on a non-empty closed interval I = [a, b] of the real line R. This is a real-valued non-decreasing function, necessarily bounded on the interval I.

A partition of the interval I is a set P = $\{x_0, x_1, \ldots, x_n\}$ with $a=x_0 < x_1 < \cdots < x_n=b$. Let $I_k = [x_k, x_{k+1}]$ for $k=0,1,2,\ldots,n-1$. Define the mesh of P, $\|P\|$, to be the minimum value of $\Delta x_k = x_{k+1} - x_k$ for $k = 0,1,\ldots, n-1$. Given a distribution function F, let $\Delta F_k = F(x_{k+1}) - F(x_k)$ for $k=0,1,\ldots,n-1$.

Suppose that g is a bounded real-valued function defined on I. Let $m_k = \inf\{g(x) | x \in I_k\}$ and $M_k = \sup\{g(x) | x \in I_k\}$.

We define the lower and upper sums of g corresponding to P with respect to F by

$$S_*(g, F, \mathcal{P}) = \sum_{k=0}^{n-1} m_k \Delta F_k \quad \text{and} \quad S^*(g, F, \mathcal{P}) = \sum_{k=0}^{n-1} M_k \Delta F_k.$$

We say the partition P_2 refines P_1 (or that P_2 is finer than P_1) if $P_1 \subseteq P_2$. By induction on the number of points in $P_2 - P_1$, it is easy to show that if P_2 refines P_1, then for any g and F. $S_*(g, F, \mathcal{P}_1) \le S_*(g, F, \mathcal{P}_2) \le S^*(g, F, \mathcal{P}_2) \le S^*(g, F, \mathcal{P}_1)$.

By considering a common refinement P = P1 ∪ P2, we see that for any partitions P1 and P2 we have $S_*(g,F,\mathcal{P}_1) \le S^*(g,F,\mathcal{P}_2)$.

Thus, the quantities

$$L(g,F,I) = \sup\{S_*(g,F,\mathcal{P})\} \quad \text{and} \quad U(g,F,I) = \inf\{S^*(g,F,\mathcal{P})\}$$

are well-defined, the supremum and infimum being taken over all partitions P of I. Furthermore, L(g, F, I) ≤ U(g, F, I)

Definition 1. We say that g is *Darboux-Stieltjes* integrable on I with respect to F, denoted $g \in \mathcal{R}_1(F,I)$, if $L(g,F,I) = U(g,F,I)$.

Given a partition $\mathcal{P}$ of I, choose $c_k \in I_k$ for each $k = 0, 1, \dots, n-1$. Let $\mathcal{C} = \{c_0, c_1, \dots, c_{n-1}\}$. The *Riemann sum* of g corresponding to $\mathcal{P}$ and $\mathcal{C}$ with respect to F is

$$S(g,F,\mathcal{P},\mathcal{C}) = \sum_{k=0}^{n-1} g(c_k)\Delta F_k.$$

Definition 2. Suppose there is a real number A with the property that for any $\varepsilon > 0$ there exists a $\delta > 0$ such that if $\|\mathcal{P}\| < \delta$ then $|A - R(g,F,\mathcal{P},\mathcal{C})| < \varepsilon$ for any choice of $\mathcal{C}$. We say that g is *Riemann-Stieltjes integrable* on I with respect to F and denote this $g \in \mathcal{R}_2(F,I)$.

• Two Classes of Functions:

Theorem 1. $\mathcal{R}_2(F,I) \subseteq \mathcal{R}_1(F,I)$.

Proof. Suppose that $g \in \mathcal{R}_2(F,I)$ and let $\varepsilon > 0$ be given. Let A be the number given in Definition 2 and choose $\delta > 0$ corresponding to $\varepsilon/2$. Clearly, $L(g,F,I) \le A \le U(g,F,I)$. Let $\mathcal{P}$ be a partition satisfying $\|\mathcal{P}\| < \delta$. If $F(a) = F(b)$, then F is constant on I and so all sums evaluate to zero. Otherwise, choose a point c_k in each I_k so that

$$g(c_k) - m_k \le \frac{\varepsilon}{2(F(b) - F(a))}.$$

It then follows that

$$S(g,F,\mathcal{P},\mathcal{C}) - S_*(g,F,\mathcal{P}) < \frac{\varepsilon}{2}.$$

Then

$$\begin{aligned} |A - S_*(g,F,\mathcal{P})| &\le |A - S(g,F,\mathcal{P},\mathcal{C})| + |S(g,F,\mathcal{P},\mathcal{C}) - S_*(g,F,\mathcal{P})| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2}. \end{aligned}$$

Similarly, we can find a partition $\mathcal{P}'$ so that $|A - S^*(g,F,\mathcal{P}')| < \varepsilon$. As ε is arbitrary, we have $L(g,F,I) = U(g,F,I) = A$.

If $F(x) = x$, then the two definitions coincide and we have the usual notion of Riemann integrability. We can say even more.

Theorem 2. If F is continuous on I then for any bounded function g on I, $g \in \mathcal{R}_2(F, I)$ if and only if $g \in \mathcal{R}_1(F, I)$.

Proof. The implication follows from Theorem 1. For the converse, suppose that $g \in \mathcal{R}_1(F, I)$ and that $\varepsilon > 0$ is given. Suppose that $\mathcal{P}$ is an $n + 1$ point partition of I such that $S^*(g, F, \mathcal{P}) - S_*(g, F, \mathcal{P}) < \varepsilon/2$. Let M be an upper bound for $|g|$ on I. Because F is continuous on a closed interval, it is uniformly continuous. Therefore, there is a $\delta > 0$ so that

$$|F(x) - F(y)| < \frac{\varepsilon}{2Mn} \quad \text{whenever} \quad |x - y| < \delta.$$

Let Q be any partition of I such that $\|Q\| < \delta$. Let $J_0, J_1, \ldots, J_{m-1}$ be the subintervals of I determined by Q and let $d_l \in J_l$ for $l = 0, 1, \ldots, m - 1$. Let $\mathcal{D} = \{d_0, d_1, \ldots, d_{m-1}\}$. There are, at most, $n - 1$ intervals J_l which contain points from $\mathcal{P}$ in their interiors. Call these the bad intervals of Q and the remainder the good intervals. The total contribution to $S(g, F, Q, \mathcal{D})$ from the bad intervals is bounded in absolute value by $\varepsilon/2$. Let $\mathcal{U} = Q \cup \mathcal{P}$ and choose a sequence of points $\mathcal{E}$ from the intervals so determined. If we do this in such a way that the points in $\mathcal{E}$ coincide with those in $\mathcal{D}$ in all of the good intervals of Q, then

$$|S(g, F, Q, \mathcal{D}) - S(g, F, \mathcal{U}, \mathcal{E})| < \frac{\varepsilon}{2}.$$

Now every interval of $\mathcal{U}$ is a subset of some I_k, so for the corresponding element $e \in \mathcal{E}$, we have $m_k \leq g(e) \leq M_k$. From this it easily follows that

$$S_*(g, F, \mathcal{P}) \leq S(g, F, \mathcal{U}, \mathcal{E}) \leq S^*(g, F, \mathcal{P}).$$

Therefore, if we let A be the common value of $L(g, F, I)$ and $U(g, F, I)$, we have shown that

$$|A - S(g, F, Q, \mathcal{D})| < \varepsilon \quad \text{whenever} \quad \|Q\| < \delta.$$

Theorem 3. Suppose $I = [a, b]$ and $c \in (a, b)$. If $g \in \mathcal{R}_1(F, [a, c])$ and $g \in \mathcal{R}_1(F, [c, b])$ then $g \in \mathcal{R}_1(F, I)$. The corresponding statement concerning $\mathcal{R}_2$ is false.

Proof. Let $L' = \sup\{S_*(g, F, Q)\}$ and $U' = \inf\{S^*(g, F, Q)\}$, with the supremum and infimum taken over all partitions Q of I containing c. Since g is contained in both $\mathcal{R}_1(F, [a, c])$ and $\mathcal{R}_1(F, [c, b])$, it follows that $L' = U'$. From this it follows that $L(g, F, I) = U(g, F, I)$, as the supremum and infimum here are taken over a larger class of partitions, and so $g \in \mathcal{R}_1(F, I)$.

For the second statement, consider Example 1, where $g \in \mathcal{R}_2(F, [0, 1])$ and $g \in \mathcal{R}_2(F, [0, 2])$.

- **Application - Probability Theory:**

A cumulative distribution function (CDF) is a non-negative, non-decreasing function F defined on the entire real line R with the properties as RHS.

$$\lim_{x\to-\infty} F_X(x) = 0$$

$$\lim_{x\to+\infty} F_X(x) = 1$$

F is the CDF for a random variable X if Pr(X ≤ x) = F(x).

Undergraduate probability texts usually recognize two types of random variable: discrete and continuous. A discrete random variable takes at most countably many values $x_1, x_2, \ldots$. If we let $p_i = Pr(X = x_i)$ for each i, then $F(x) = \sum_{x_i \le x} p_i.$

A random variable has continuous distribution if it takes all the values in some (bounded or unbounded) interval in the real line, and Pr(X = x) = 0 for every $x \in R$. In practice, the textbooks consider only random variables with CDFs that are differentiable (except possibly at finitely many points).

The probability density function(pdf) is then defined by $f(x) = F'(x)$. The expectation of a random variable g(X) is defined respectively.

Using the RS integral, we may say that in either case. One must first make the obvious extension to an improper RS integral (Stieltjes himself considered the case b = ∞ in his 1894 paper). Then for discrete random variables, this equation is clearly valid. For continuous random variables, one must use the result that F is differentiable and $f = F'$.

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)dF(x)$$

A random variable is said to have a mixed-type distribution if its range is uncountable and yet there are points with Pr(X = x) > 0. Such random variables arise naturally, but the second equation are beyond the scope of most undergraduate texts. The RS integral allows such random variables to be dealt with in the same fashion as discrete and continuous ones. In particular, the expectation E(g(X)) is defined by equation 2 in this case as well.

1.10 Important Inequalities

In this section, we obtain the proofs of three famous inequalities involving expectations. We shall make use of these inequalities in the remainder of the text. We begin with a useful result.

Theorem 1.10.1. *Let X be a random variable and let m be a positive integer. Suppose $E[X^m]$ exists. If k is an integer and $k \leq m$, then $E[X^k]$ exists.*

Proof: We shall prove it for the continuous case; but the proof is similar for the discrete case if we replace integrals by sums. Let $f(x)$ be the pdf of X. Then

$$\begin{aligned}\int_{-\infty}^{\infty} |x|^k f(x)\,dx &= \int_{|x|\leq 1} |x|^k f(x)\,dx + \int_{|x|>1} |x|^k f(x)\,dx \\ &\leq \int_{|x|\leq 1} f(x)\,dx + \int_{|x|>1} |x|^m f(x)\,dx \\ &\leq \int_{-\infty}^{\infty} f(x)\,dx + \int_{-\infty}^{\infty} |x|^m f(x)\,dx \\ &\leq 1 + E[|X|^m] < \infty, \end{aligned} \tag{1.10.1}$$

which is the the desired result. ■

Techniques:
{ ☆ classfication (key)
makes them larger

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,*

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

pay attention to!
also the tail probabilities
Just know the meaning.

Proof. The proof is given when the random variable X is of the continuous type; but the proof can be adapted to the discrete case if we replace integrals by sums. Let $A = \{x : u(x) \geq c\}$ and let $f(x)$ denote the pdf of X. Then

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)\,dx = \int_A u(x)f(x)\,dx + \int_{A^c} u(x)f(x)\,dx.$$

Since each of the integrals in the extreme right-hand member of the preceding equation is nonnegative, the left-hand member is greater than or equal to either of them. In particular,

$$E[u(X)] \geq \int_A u(x)f(x)\,dx.$$

the proof is not so hard.

However, if $x \in A$, then $u(x) \geq c$; accordingly, the right-hand member of the preceding inequality is not increased if we replace $u(x)$ by c. Thus

$$E[u(X)] \geq c \int_A f(x)\,dx.$$

Since

$$\int_A f(x)\,dx = P(X \in A) = P[u(X) \geq c],$$

it follows that

$$E[u(X)] \geq cP[u(X) \geq c],$$

which is the desired result. ■

The preceding theorem is a generalization of an inequality that is often called *Chebyshev's inequality*. This inequality will now be established.

Theorem 1.10.3 (Chebyshev's Inequality). *Let the random variable X have a distribution of probability about which we assume only that there is a finite variance σ^2, (by Theorem 1.10.1 this implies the mean $\mu = E(X)$ exists). Then for every $k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \tag{1.10.2}$$

or, equivalently,

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

Proof. In Theorem 1.10.2 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 member of the preceding inequality is σ^2, the inequality may be written

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

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

A convenient form of Chebyshev's Inequality is found by taking $k\sigma = \epsilon$ for $\epsilon > 0$. Then equation (1.10.2) becomes

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}, \quad \text{for all } \epsilon > 0. \tag{1.10.3}$$

Hence, the number $1/k^2$ is an upper bound for the probability $P(|X - \mu| \geq k\sigma)$. In the following example this upper bound and the exact value of the probability are compared in special instances.

Example 1.10.1. Let X have the pdf

$$f(x) = \begin{cases} \frac{1}{2\sqrt{3}} & -\sqrt{3} < x < \sqrt{3} \\ 0 & \text{elsewhere.} \end{cases}$$

Here $\mu = 0$ and $\sigma^2 = 1$. If $k = \frac{3}{2}$, we have the exact probability

$$P(|X - \mu| \geq k\sigma) = P\left(|X| \geq \frac{3}{2}\right) = 1 - \int_{-3/2}^{3/2} \frac{1}{2\sqrt{3}}\, dx = 1 - \frac{\sqrt{3}}{2}.$$

By Chebyshev's inequality, this probability has the upper bound $1/k^2 = \frac{4}{9}$. Since $1 - \sqrt{3}/2 = 0.134$, approximately, the exact probability in this case is considerably less than the upper bound $\frac{4}{9}$. If we take $k = 2$, we have the exact probability $P(|X - \mu| \geq 2\sigma) = P(|X| \geq 2) = 0$. This again is considerably less than the upper bound $1/k^2 = \frac{1}{4}$ provided by Chebyshev's inequality. □

In each of the instances in the preceding example, the probability $P(|X - \mu| \geq k\sigma)$ and its upper bound $1/k^2$ differ considerably. This suggests that this inequality might be made sharper. However, if we want an inequality that holds for every $k > 0$ and holds for all random variables having a finite variance, such an improvement is impossible, as is shown by the following example. the optimal bound considering all R

Example 1.10.2. Let the random variable X of the discrete type have probabilities $\frac{1}{8}, \frac{6}{8}, \frac{1}{8}$ at the points $x = -1, 0, 1$, respectively. Here $\mu = 0$ and $\sigma^2 = \frac{1}{4}$. If $k = 2$, then $1/k^2 = \frac{1}{4}$ and $P(|X - \mu| \geq k\sigma) = P(|X| \geq 1) = \frac{1}{4}$. That is, the probability $P(|X - \mu| \geq k\sigma)$ here attains the upper bound $1/k^2 = \frac{1}{4}$. Hence the inequality cannot be improved without further assumptions about the distribution of X. □

Definition 1.10.1.

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

$$\phi[\gamma x + (1 - \gamma)y] \leq \gamma\phi(x) + (1 - \gamma)\phi(y). \tag{1.10.4}$$

convex

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

Depending on existence of first or second derivatives of ϕ, the following theorem can be proved.

Theorem 1.10.4. *If ϕ is differentiable on (a, b) then*

(a) ϕ is convex if and only if $\phi'(x) \leq \phi'(y)$, for all $a < x < y < b$,

(b) ϕ is strictly convex if and only if $\phi'(x) < \phi'(y)$, for all $a < x < y < b$.

If ϕ is twice differentiable on (a, b) then

(a) ϕ is convex if and only if $\phi''(x) \geq 0$, for all $a < x < b$,

(b) ϕ is strictly convex if $\phi''(x) > 0$, for all $a < x < b$.

Of course the second part of this theorem follows immediately from the first part. While the first part appeals to one's intuition, the proof of it can be found in most analysis books; see, for instance, Hewitt and Stromberg (1965). A very useful probability inequality follows from convexity.

Theorem 1.10.5 (Jensen's Inequality). *If ϕ is convex on an open interval I and X is a random variable whose support is contained in I and has finite expectation, then*

$$\phi[E(X)] \leq E[\phi(X)]. \tag{1.10.5}$$

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

simple proof on the first page of important inequalities (back)

more rigurous proof (back of this page)

- Expand $\phi(x)$ as Talor expansion about $\mu = E(X)$ of order two:

$$\phi(x) = \phi(\mu) + \phi'(\mu)(x-\mu) + \phi''(\xi)\frac{(x-\mu)^2}{2!} \geq \phi(\mu) + \phi'(\mu)(x-\mu)$$

$(\phi''(\xi) \geq 0)$

Taking expectation of both sides $E(\phi(X)) \geq \phi(E(X)) + E(\phi'(\mu)(x-\mu))$

$\phi'(\mu)(E(X) - E(X)) = 0$

Get the right answer.

(Proof of Jensen inequality)

- Another Proof of Chebyshev's Inequality:

$A = \{x \mid |x-\mu| \geq k\delta\}$. $E(X-\mu)^2 = \sum_{x \in A} (x-\mu)^2 f(x) + \sum_{x \in A'} (x-\mu)^2 f(x)$

δ^2 $\geq \sum_{x \in A} (x-\mu)^2 f(x)$

$\Rightarrow k^2\delta^2 \geq \sum_{x \in A} (x-\mu)^2 k^2 f(x)$ or $\delta^2 \geq \sum_{x \in A} (x-\mu)^2 k^2 f(x) \geq k^2\delta^2 \sum_{x \in A} f(x)$

definition of A $\rightarrow$ $\leq (x-\mu)^2$ for any x in A $\Rightarrow P = \sum_{x \in A} f(x) \leq \frac{1}{k^2}$

$\Rightarrow \sum_{x \in A} k^2 f(x) \leq 1$

(Discrete Form)

STA2001 Section THREE Continuous Distributions

1. Random Variable:

We consider RVs with *a continuous range of possible values*, i.e., ***unions of intervals***, which are quite common (e.g., velocity of a vehicle traveling along the high way).

Now there are many probability density functions that could describe probabilities associated with a random variable X. Thus, we say that the **probability density function (pdf)** of a random variable X of the **continuous type**, with space $\overline{S}$ that is an interval or union of intervals, is an integrable function $f(x)$ satisfying the following conditions:

(a) $f(x) \geq 0, \qquad x \in \overline{S}.$

(b) $\int_{\overline{S}} f(x)\,dx = 1.$

(c) If $(a,b) \subseteq \overline{S}$, then the probability of the event $\{a < X < b\}$ is

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

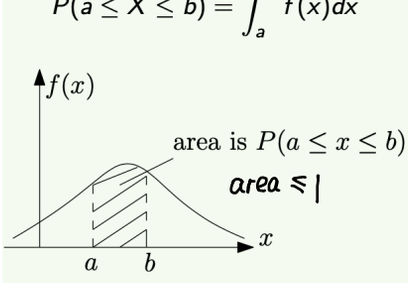
The corresponding distribution of probability is said to be of the continuous type.

2. Discrete RV vs. Continuous RV: (Differences)

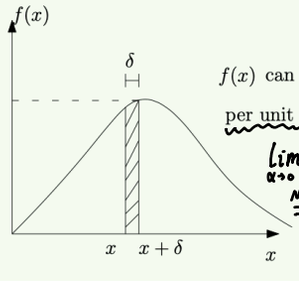
Discrete RV:	Continuous RV:
pmf $f(x) : \overline{S} \to (0, 1]$	pdf $f(x) : \overline{S} \to (0, \infty)$
① Domain: $f(x) = P(X = x) \leq 1$	pdf expresses *density* not probability
② Range	
1. $f(x) > 0$	1. $f(x) > 0$
③ marks	
2. $\sum_{x \in \overline{S}} f(x) = 1$	2. $\int_{\overline{S}} f(x)dx = 1$
3. $P(X \in A) = \sum_{x \in A} f(x)$	3. $P(X \in A) = \int_A f(x)dx$

3. Interpretation of pdf:

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



$$P(x \leq X \leq x+\delta) = \int_x^{x+\delta} f(t)dt \approx f(x)\delta$$



$f(x)$ can be viewed as the probability mass per unit length near x

$\lim_{a \to 0} P(x \leq X \leq x+a) = \lim_{a \to 0} \int_x^{x+a} f(t)dt$

MVT $\lim_{a \to 0} a f(c)$ $c \in (x, x+a)$

$\lim_{a \to 0} c = x$

$\lim_{a \to 0} a f(x)$ density

Remarks:

(1)We often ***extend the domain*** of f (x) ***from $\bar{S}$ to R*** and let f (x) = 0, $x \notin \bar{S}$. In this case, f (x) : R $\rightarrow$ [0, ∞) and S is called the support of X.

(2)For any single value a, P (X = a) = $\int_a^a$ f (x)dx = 0. Therefore, ***including or excluding*** the end points of an interval ***has no effect*** on its probability:

P(a$\leqslant$ X $\leqslant$b) = P(a< X $\leqslant$b) = P(a $\leqslant$X <b) = P(a< X <b)

(3). pdf needs to be neither bounded nor continuous.

$$\begin{cases} f(x) \geq 0, & x \in R \\ \int_{-\infty}^{\infty} f(x)dx = 1 \\ P(a \leq X \leq b) = \int_a^b f(x)dx \end{cases}$$

4. Cumulative distribution function (CDF):

The **cumulative distribution function** (**cdf**) or **distribution function** of a random variable X of the continuous type, defined in terms of the pdf of X, is given by

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

(1)F(x) is nondecreasing

(2)Relation between the probability function and the cdf: P(a$\leqslant$X $\leqslant$b) = F(b) $-$ F(a)

(3)Relation between the pdf and the cdf: f (x) = F′(x), for those values of x at which F(x) is differentiable.

5. Uniform Distribution:

The random variable X has a **uniform distribution** if its pdf is equal to a constant on its support. In particular, if the support is the interval $[a, b]$, then

$$f(x) = \frac{1}{b-a}, \qquad a \leq x \leq b. \qquad F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \leq x < b, \\ 1, & b \leq x, \end{cases}$$

6. Expectations:

Theorem 1.8.1. *Let X be a random variable and let $Y = g(X)$ for some function g.*

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

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx. \tag{1.8.2}$$

(b). *Suppose X is discrete with pmf $p_X(x)$. Suppose the support of X is denoted by S_X. If $\sum_{x \in S_X} |g(x)| p_X(x) < \infty$, then the expectation of Y exists and it is given by*

$$E(Y) = \sum_{x \in S_X} g(x) p_X(x). \tag{1.8.3}$$

need absolutely converges

Definition 1.9.3 (Moment Generating Function (mgf)). *Let X be a random variable such that for some $h > 0$, the expectation of e^{tX} exists for $-h < t < h$. The* **moment generating function** *of X is defined to be the function $M(t) = E(e^{tX})$, for $-h < t < h$. We will use the abbreviation* **mgf** *to denote moment generating function of a random variable.*

7. The (100p)th percentile:

It is a number π_p such that the area ***under f (x) to the left of*** π_p is p. That is,

$$p = \int_{-\infty}^{\pi_p} f(x)dx = F(\pi_p)$$

The 50th percentile is called ***the median***. We let $m = \pi_{0.5}$. The 25th and 75th percentiles are called the first and third quartiles, respectively, and are denoted by $q_1 = \pi_{0.25}$ and $q_3 = \pi_{0.75}$. Of course, the median $m = \pi_{0.5} = q_2$ is also called the second quartile.

8. Exponential Distribution:

Description, Definition & Expectations

For an interval with length T, the number of occurrences X has $E[X] = \lambda T$ and thus its pmf is

$$f(x) = \frac{(\lambda T)^x e^{-\lambda T}}{x!}, \quad x = 0, 1, \cdots$$

$P(X = 0) = e^{-\lambda T} = P$(no occurrence in the interval with length T)

New random experiment

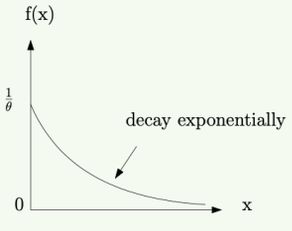
We are interested in the waiting time until the first occurrence for the APP, which is denoted by W.

Definition

A RV X has an exponential distribution if its pdf is

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \quad x \geq 0, \theta > 0$$

Accordingly, the waiting time until the first occurrence for an APP has an exponential distribution with $\theta = \frac{1}{\lambda}$ [λ: the number of occurrences per unit time]



1) Derive CDF of W, F(w)
2) f (w) = F′(w)

F(w)=P(W ⩽w)

Assume that the waiting time is nonnegative. Then,

For w ⩾ 0,

F(w) = 0, for w < 0.

F(w) = P(W ⩽ w) = 1 − P(W >w)

Where P(W > w) = P(no occurrences in [0,w]) = $e^{-\lambda w}$

$$M(t) = \int_0^\infty e^{tx}\left(\frac{1}{\theta}\right)e^{-x/\theta}\,dx = \lim_{b\to\infty}\int_0^b \left(\frac{1}{\theta}\right)e^{-(1-\theta t)x/\theta}\,dx$$

$$= \lim_{b\to\infty}\left[-\frac{e^{-(1-\theta t)x/\theta}}{1-\theta t}\right]_0^b = \frac{1}{1-\theta t}, \qquad t < \frac{1}{\theta}.$$

$$M'(t) = \frac{\theta}{(1-\theta t)^2}$$

$$M''(t) = \frac{2\theta^2}{(1-\theta t)^3}.$$

r an exponential distribution, we have

$$\mu = M'(0) = \theta \qquad \text{and} \qquad \sigma^2 = M''(0) - [M'(0)]^2 = \theta^2.$$

9. Gamma Distribution: Description, Definition & Expectations

Description

Consider an APP with average number of occurence λ in an unit interval. We are interested in the waiting time, denoted by W, until the αth occurrence, $\alpha = 1, 2, \ldots$.

In the (approximate) Poisson process with mean λ, we have seen that the waiting time until the first occurrence has an exponential distribution. We now let W denote the waiting time until the αth occurrence and find the distribution of W.

The cdf of W when $w \geq 0$ is given by

$$\begin{aligned} F(w) &= P(W \leq w) = 1 - P(W > w) \\ &= 1 - P(\text{fewer than } \alpha \text{ occurrences in } [0, w]) \\ &= 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}}{k!}, \end{aligned} \qquad (3.2\text{-}1)$$

$$\begin{aligned} F'(w) &= \lambda e^{-\lambda w} - e^{-\lambda w} \sum_{k=1}^{\alpha-1} \left[\frac{k(\lambda w)^{k-1}\lambda}{k!} - \frac{(\lambda w)^k \lambda}{k!} \right] \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \left[\lambda - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} \right] \\ &= \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} e^{-\lambda w}. \end{aligned}$$

A pdf of this form is said to be one of the gamma type, and the random variable W is said to have a gamma distribution.

The gamma function is defined by:

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y}\, dy, \qquad 0 < t.$$

This integral is positive for $0 < t$ because the integrand is positive. Values of it are often given in a table of integrals. If $t > 1$, integration of the gamma function of t by parts yields

$$\begin{aligned} \Gamma(t) &= \left[-y^{t-1} e^{-y} \right]_0^\infty + \int_0^\infty (t-1) y^{t-2} e^{-y}\, dy \\ &= (t-1) \int_0^\infty y^{t-2} e^{-y}\, dy \; = \; (t-1)\Gamma(t-1) \end{aligned}$$

mgf (exercise 3.2-7)

$$M(t) = \frac{1}{(1-\theta t)^\alpha}, \quad t < \frac{1}{\theta}$$

$$E[x] = \alpha\theta, \quad Var[X] = \alpha\theta^2$$

Let us now formally define the pdf of the gamma distribution and find its characteristics. The random variable X has a **gamma distribution** if its pdf is defined by

$$f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}, \qquad 0 \leq x < \infty.$$

A special case: when α = 1, Gamma distribution reduces to exponential distribution.

10. Chi-square Distribution: Description, Definition & Expectations

Chi-square Distribution is widely used because it has close relationship with uniform distribution.

We now consider a special case of the gamma distribution that plays an important role in statistics. Let X have a gamma distribution with $\theta = 2$ and $\alpha = r/2$, where r is a positive integer. The pdf of X is

$$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}, \qquad 0 < x < \infty.$$

We say that X has a **chi-square distribution with r degrees of freedom**, which we abbreviate by saying that X is $\chi^2(r)$. The mean and the variance of this chi-square distribution are, respectively,

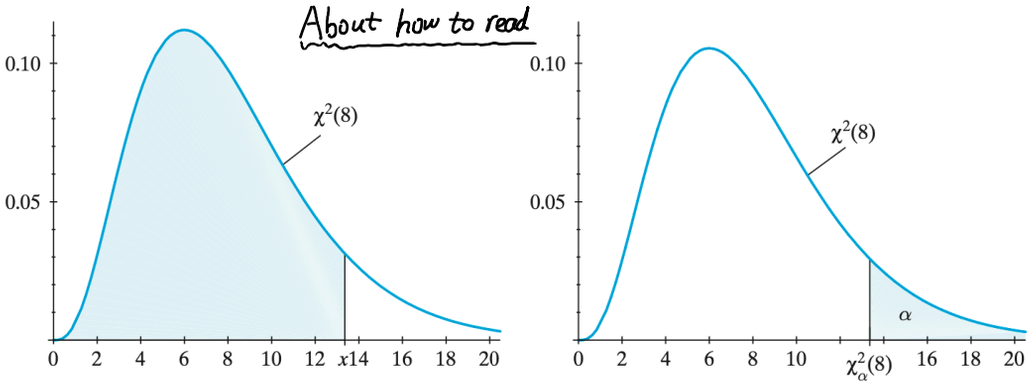
$$\mu = \alpha\theta = \left(\frac{r}{2}\right)2 = r \qquad \text{and} \qquad \sigma^2 = \alpha\theta^2 = \left(\frac{r}{2}\right)2^2 = 2r.$$

That is, the mean equals the number of degrees of freedom, and the variance equals twice the number of degrees of freedom. An explanation of "number of degrees of freedom" is given later. From the results concerning the more general gamma distribution, we see that its moment-generating function is

$$M(t) = (1 - 2t)^{-r/2}, \qquad t < \frac{1}{2}.$$

Table IV The Chi-Square Distribution

About how to read



$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

probability

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
			1.635	2.204	10.64	12.59	14.45	16.81

11. Normal Distribution (most important)

Definition & Expectations

Definition 3.4.1 (Normal Distribution). *We say a random variable X has a* **normal distribution** *if its pdf is*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad \text{for } -\infty < x < \infty. \tag{3.4.6}$$

The parameters μ and σ^2 are the mean and variance of X, respectively. We will often write that X has a $N(\mu, \sigma^2)$ distribution.

(prove that it is a pdf by hand)

(i) $f(x) > 0$, absolutely

(ii) $\int_{-\infty}^{\infty} f(x)dx \overset{z=\frac{x-\mu}{\sigma}}{=\!=\!=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{z^2}{2}\} dz \triangleq I$, then $I^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\{-\frac{z^2}{2}\}dz \int_{-\infty}^{\infty} \exp\{-\frac{z^2}{2}\}dz$

$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-\frac{x^2+y^2}{2}\}dxdy$

$\Rightarrow x = r\cos\theta$ & $y = r\sin\theta$

$I^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{\cdots\}dxdy = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r\,dr\,d\theta$

(calculate the mgf, mean and variance) $= \frac{1}{2\pi}\int_0^{2\pi} 1\,d\theta \int_0^{\infty} e^{-\frac{r^2}{2}} d\frac{r^2}{2} = 1$

mgf: $M(t) = \int_{-\infty}^{\infty} e^{tx} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \frac{1}{\sqrt{2\pi}\sigma} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x^2 - (2\sigma^2 t + 2\mu)x + \mu^2)\}$

$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu-\sigma^2 t)^2}{2\sigma^2}\} e^{\mu t + \frac{1}{2}\sigma^2 t^2} dx = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$N(\mu+\sigma^2 t, \sigma^2)$

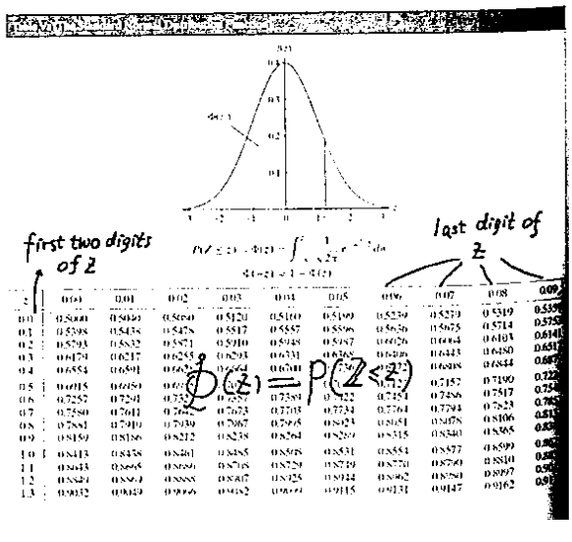
$M'(0) = \mu$, $M''(0) - M'(0) = \sigma^2$

12. Standard Normal Distribution:

If Z is $N(0,1)$, we shall say that Z has a **standard normal distribution**. Moreover, the cdf of Z is

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw.$$

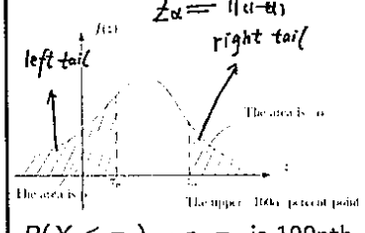
Because of the symmetry of the standard normal pdf, it is true that $\Phi(-z) = 1 - \Phi(z)$ for all real z.



13. The upper 100α percent point:

Definition

The number z_α such that $P(Z \ge z_\alpha) = \alpha$.



$P(X \le \pi_p) = p$, π_p is 100pth percentile.

Note:

$$P(Z < z_\alpha) = 1 - P(Z \ge z_\alpha) = 1 - \alpha$$

So z_α is the $100(1-\alpha)$th percentile

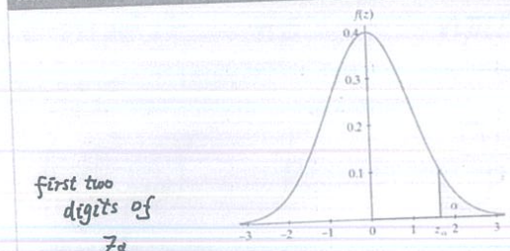
Theorem

If Y is $N(\mu, \sigma^2)$, then $X = \frac{Y-\mu}{\sigma}$ is $N(0,1)$

Use CDF techinges, we can show this.

14. Relation between normal and χ^2 distribution:
(proof by hand)

Table Vb The Standard Normal Right-Tail Probabilities



first two digits of z_α — the last digit of z_α

$P(Z > z_\alpha) = \alpha$

$P(Z > z) = 1 - \Phi(z) = \Phi(-z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3633	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823

Theorem 3.3-2 If the random variable X is $N(\mu, \sigma^2)$, $\sigma^2 > 0$, then the random variable $V = (X-\mu)^2/\sigma^2 = Z^2$ is $\chi^2(1)$.

(i) CDF technique:

$$G(v) = P(V \le v) = P(-\sqrt{v} \le Z \le \sqrt{v}) = \int_{-\sqrt{v}}^{\sqrt{v}} f(z)\,dz = 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,dz$$

pdf of V: $g(v) = G'(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v} \cdot \frac{1}{2\sqrt{v}} \times 2 = \frac{1}{\sqrt{2\pi v}} e^{-\frac{1}{2}v}$, $v > 0$

We know for Chi-square distribution: pdf of $\chi^2(1) = \frac{1}{\Gamma(\frac{1}{2})\sqrt{2}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$

$$\int_{-\infty}^{\infty} g(v)\,dv = \frac{1}{\sqrt{\pi}} \int_0^{\infty} x^{\frac{1}{2}-1} e^{-x}\,dx \overset{\text{definition}}{=} \frac{1}{\sqrt{\pi}} \Gamma(\tfrac{1}{2}) = 1. \quad \text{so } \Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

$\Rightarrow V \sim \chi^2(1)$

- $\delta^2 \downarrow$ — curve more peaked (normal distribution) (shaper)

(ii) mgf technique:

$$M(t) = e^{tV^2} = \int_0^{\infty} e^{tv^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\,dv = \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{(t-\frac{1}{2})v^2}\,dv, \text{ for } t < \tfrac{1}{2}, \text{ we have}$$

$$\overset{\text{original}}{=} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)v^2}\,dv = (\tfrac{1}{1-2t})^{\frac{1}{2}} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{1-2t}\,v)^2} d(\sqrt{1-2t}\,v) = (\tfrac{1}{1-2t})^{\frac{1}{2}}, \ t < \tfrac{1}{2}$$

which is mgf of $\chi^2(1)$.

STA2001 Section FOUR Bivariate Distributions

Topic I Bivariate Distributions

1. **Univariate Random Variables:** a random experiment whose outcome is one thing of interest.

Bivariate Distributions: two random experiments jointly each of whose outcome is one thing of interest or a random experiment whose outcome is a pair of two things of interest

Definition 2.1.1 (Random Vector). *Given a random experiment with a sample space C. Consider two random variables X_1 and X_2, which assign to each element c of C one and only one ordered pair of numbers $X_1(c) = x_1$, $X_2(c) = x_2$. Then we say that (X_1, X_2) is a* **random vector**. *The* space *of (X_1, X_2) is the set of ordered pairs $\mathcal{D} = \{(x_1, x_2) : x_1 = X_1(c), x_2 = X_2(c), c \in C\}$.*

Then, it holds that

$$\overline{S} \subseteq \overline{S_X} \times \overline{S_Y} = \{(x, y) | x \in \overline{S_X}, y \in \overline{S_Y}\}$$

2. **Five Basic Sample Space:**

$\overline{S} = \{\text{all possible values of } (X, Y)\}$

$\overline{S_X} = \{\text{all possible values of } X\} = \{x | (x, y) \in \overline{S}, y \in \overline{S_Y}\}$

$\overline{S_Y} = \{\text{all possible values of } Y\} = \{y | (x, y) \in \overline{S}, x \in \overline{S_X}\}$

$S_X(y) = \{x | (x, y) \in S\}$ for a given $y \in S_Y$

$\overline{S_Y}(x) = \{y | (x, y) \in \overline{S}\}$ for a given $x \in \overline{S_X}$

Topic II Discrete and Continuous Type of Bivariate Distributions

1. Joint Probability Mass Function of Discrete Type:

Let X and Y be two random variables defined on a discrete space. Let S denote the corresponding two-dimensional space of X and Y, the two random variables of the discrete type. The probability that $X = x$ and $Y = y$ is denoted by $f(x, y) = P(X = x, Y = y)$. The function $f(x, y)$ is called the **joint probability mass function** (joint pmf) of X and Y and has the following properties:

(a) $0 \leq f(x, y) \leq 1$.

(b) $\sum\sum_{(x,y)\in S} f(x, y) = 1$.

(c) $P[(X, Y) \in A] = \sum\sum_{(x,y)\in A} f(x, y)$, where A is a subset of the space S.

2. Joint Probability Mass Function of Continuous Type:

The **joint probability density function** (joint pdf) of two continuous-type random variables is an integrable function $f(x,y)$ with the following properties:

(a) $f(x,y) \geq 0$, where $f(x,y) = 0$ when (x,y) is not in the support (space) S of X and Y.

(b) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$. ($\iint_S f(x,y)\,dx\,dy = 1$)

(c) $P[(X,Y) \in A] = \iint_A f(x,y)\,dx\,dy$, where $\{(X,Y) \in A\}$ is an event defined in the plane.

3. Marginal pmfs & pdfs:

Let X and Y have the joint probability mass function $f(x,y)$ with space S. The probability mass function of X alone, which is called the **marginal probability mass function of** X, is defined by

$$f_X(x) = \sum_{y \in \bar{S}_Y(x)} f(x,y) = P(X = x), \qquad x \in S_X,$$

where the summation is taken over all possible y values for each given x in the x space S_X. That is, the summation is over all (x,y) in S with a given x value. Similarly, the **marginal probability mass function of** Y is defined by

$$f_Y(y) = \sum_{x \in \bar{S}_X(y)} f(x,y) = P(Y = y), \qquad y \in S_Y,$$

The respective **marginal pdfs** of continuous-type random variables X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy, \qquad x \in S_X,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx, \qquad y \in S_Y,$$

4. Mathematical Expectations

Let $Y = g(X_1, X_2)$ for some real valued function.

For continuous type:

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1,x_2) f_{X_1,X_2}(x_1,x_2)\,dx_1 dx_2. \qquad (2.1.10)$$

For discrete type:

$$E(Y) = \sum_{x_1}\sum_{x_2} g(x_1,x_2) p_{X_1,X_2}(x_1,x_2). \qquad (2.1.11)$$

When g (X, Y) = X, E[X] is the mean of X

When g (X , Y) = (X - EX)^2, E[(X - EX)^2] is the variance of X

(Two ways to calculate EX)

Discrete Type:

$$EX = \sum_x \sum_y x f(x,y) = \begin{cases} \sum_{(x,y)\in\bar{S}} x f(x,y) \\ \sum_{x \in \bar{S}_X} x f_X(x) \end{cases}$$

Continous Type

$$EX = \int_{\bar{S}_X}\int_{\bar{S}_Y} x f(x,y)\,dy\,dx = \begin{cases} \iint_{\bar{S}} x f(x,y)\,dy\,dx \\ \int_{\bar{S}_X} x f_X(x)\,dx \quad \text{(easy to compute)} \end{cases}$$

Topic III Covariance and Correlation Coefficient

1. Covariance:

A special mathematical expectation (Take g (X, Y) as (X-EX)*(Y-EY))

(a) If $u(X,Y) = (X-\mu_X)(Y-\mu_Y)$, then

$$E[u(X,Y)] = E[(X-\mu_X)(Y-\mu_Y)] = \sigma_{XY} = \mathrm{Cov}(X,Y)$$

is called the **covariance** of X and Y.

$$\mathrm{Cov}(X,Y) = \sum_{(x,y)\in S} (x-E(X))(y-E(Y))f(x,y) = E(XY) - E(X)E(Y)$$

When Cov (X, Y) = 0, X and Y are uncorrelated.
When Cov (X, Y) > 0, X and Y are positively correlated.
When Cov (X, Y) < 0, X and Y are negatively correlated.

Independence $\Rightarrow$ Uncorrelation: Independence $\Rightarrow f(x,y) = f_X(x)f_Y(y)$ (also true for continuous type)
$\Rightarrow \mathrm{cov}(X,Y) = E(XY) - E(X)E(Y) = \sum_x \sum_y xy f_X(x) f_Y(y) - E(X)E(Y) = 0 \Rightarrow$ Uncorrelation
Uncorrelation $\not\Rightarrow$ Independence: (e.g. Let X and Y be RVs that take values (1, 0), (0, 1), (1, 0), (0, 1), each with probability ¼.)

2. Correlation Coefficient:

(b) If the standard deviations σ_X and σ_Y are positive, then

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

is called the **correlation coefficient** of X and Y.

$E((V+tW)^2) = 0 \Leftrightarrow \mathrm{Var}(V+tW) = 0$
$\Leftrightarrow V+tW = 0$
Chebyshev's inequality $P(|X-\mu| \geq \varepsilon^2) \leq \frac{\delta^2}{\varepsilon^2}$ for $\varepsilon > 0$ (when $\delta = 0 \Rightarrow X = \mu$)

(1) It is a normalized version of Cov (X, Y) and in fact $-1 =< \rho(X, Y) =< 1$ and the size of $|\rho|$ provides a normalized measure of the extent to which this is true.

(2) $\rho = 1$ or ($\rho = -1$) if and only if there exists a positive (or negative respectively) constant c such that Y - E (Y) = c*(X - E (X))

(Proof) Consider $E([(X-EX)\cdot t + (Y-EY)]^2) \geq 0$

However, $E = t^2 \mathrm{Var}(X) + \mathrm{Var}(Y) + 2t\,\mathrm{COV}(X,Y)$, a function regarding t

So $\Delta = 4\mathrm{cov}^2(X,Y) - 4\mathrm{Var}(X)\mathrm{Var}(Y) \leq 0$, $\rho^2 = \frac{\mathrm{cov}^2(X,Y)}{\mathrm{Var}(X)\mathrm{Var}(Y)} \leq 1$

$\therefore |\rho| \leq 1$. $\Delta = 0$, $\rho = \pm 1$ if & only if $(Y-EY) + t(X-EX) = 0$.

Because $t = -\frac{\mathrm{cov}(X,Y)}{\mathrm{Var}(X)} = -\frac{\rho\sigma_Y}{\sigma_X}$, we get least square regression line:

$$y - \mu_Y = \frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X)$$ (meaning E(error²) minimized ("0"))

Topic IV Independent Random Variables

Independent RVs:

where the summation is taken over all possible x values for each given y in the y space S_Y. The random variables X and Y are **independent** if and only if, for every $x \in S_X$ and every $y \in S_Y$,

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or, equivalently,

$$f(x, y) = f_X(x)f_Y(y);$$

otherwise, X and Y are said to be **dependent**.

Topic V Conditional pmfs and pdfs

1. Conditional pmfs & pdfs:

Definition 4.3-1 {density

The **conditional probability mass function of** X, given that $Y = y$, is defined by

$$g(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \qquad \text{provided that } f_Y(y) > 0.$$

Similarly, the **conditional probability mass function of** Y, given that $X = x$, is defined by {density

$$h(y \mid x) = \frac{f(x, y)}{f_X(x)}, \qquad \text{provided that } f_X(x) > 0.$$

$$E(Y|X = x) = \int_{S_Y(x)} yh(y|x)dy$$

$$\begin{aligned} Var(Y|X = x) &= E\{[Y - E(Y|X = x)]^2|X = x\} \\ &= \int_{S_Y(x)} [y - E(Y|X = x)]^2 h(y|x)dy \\ &= E[Y^2|X = x] - [E(Y|X = x)]^2 \end{aligned}$$

$$E(Y|X = x) = \sum_{y \in S_Y(x)} yh(y|x) \rightarrow \text{conditional mean}$$

$$\begin{aligned} Var(Y|X = x) &\triangleq E\{[Y - E(Y|X = x)]^2|X = x\} \\ &= \sum_{y \in S_Y(x)} [y - E(Y|X = x)]^2 h(y|x) \\ &= E(Y^2|X = x) - [E(Y|X = x)]^2 \end{aligned}$$

Additions about Conditional pmfs

Suppose that the latter conditional mean is a linear function of x; that is, $E(Y \mid x) = a + bx$. Let us find the constants a and b in terms of characteristics μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ. This development will shed additional light on the correlation coefficient ρ; accordingly, we assume that the respective standard deviations σ_X and σ_Y are both positive, so that the correlation coefficient will exist.

It is given that

$$\sum_y y\,h(y \mid x) = \sum_y y \frac{f(x,y)}{f_X(x)} = a + bx, \qquad \text{for } x \in S_X,$$

where S_X is the space of X and S_Y is the space of Y. Hence,

$$\sum_y y f(x,y) = (a + bx) f_X(x), \qquad \text{for } x \in S_X, \tag{4.3-1}$$

and

$$\sum_{x \in S_X} \sum_y y f(x,y) = \sum_{x \in S_X} (a + bx) f_X(x).$$

That is, with μ_X and μ_Y representing the respective means, we have

$$\mu_Y = a + b\mu_X. \tag{4.3-2}$$

In addition, if we multiply both members of Equation 4.3-1 by x and sum the resulting products, we obtain

$$\sum_{x \in S_X} \sum_y xy f(x,y) = \sum_{x \in S_X} (ax + bx^2) f_X(x).$$

That is,

$$E(XY) = aE(X) + bE(X^2)$$

or, equivalently,

$$\mu_X \mu_Y + \rho \sigma_X \sigma_Y = a\mu_X + b(\mu_X^2 + \sigma_X^2). \tag{4.3-3}$$

The solution of Equations 4.3-2 and 4.3-3 is

$$a = \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \mu_X \qquad \text{and} \qquad b = \rho \frac{\sigma_Y}{\sigma_X},$$

which implies that if $E(Y \mid x)$ is linear, it is given by

$$E(Y \mid x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X).$$

So if the conditional mean of Y, given that $X = x$, is linear, it is exactly the same as the best-fitting line (least squares regression line) considered in Section 4.2.

By symmetry, if the conditional mean of X, given that $Y = y$, is linear, then

$$E(X \mid y) = \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y).$$

We see that the point $[x = \mu_X, E(Y \mid X = \mu_X) = \mu_Y]$ satisfies the expression for $E(Y \mid x)$ and $[E(X \mid Y = \mu_Y) = \mu_X, y = \mu_Y]$ satisfies the expression for $E(X \mid y)$. That is, the point (μ_X, μ_Y) is on each of the two lines. In addition, we note that the product of the coefficient of x in $E(Y \mid x)$ and the coefficient of y in $E(X \mid y)$ equals ρ^2 and the ratio of these two coefficients equals σ_Y^2/σ_X^2. These observations sometimes prove useful in particular problems.

Topic VII Bivariate Normal Distributions

1. Definition and pdf

Definition

Let X and Y be 2 continuous RVs and have the joint pdf

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp[-\frac{1}{2}q(x,y)], x \in \mathbb{R}, y \in \mathbb{R},$$

$$q(x,y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$$

where $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y > 0$ and $|\rho| < 1$. Then X and Y are said to be bivariate normally distributed.

2. Properties:

1) Marginal pdf of X and Y are normal with
$X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$

2) Conditional pdf of X given that Y = y is normal with mean $\mu_X + (\sigma_X/\sigma_Y)\rho(y - \mu_Y)$ and variance $(1 - \rho^2)\sigma_X^2$

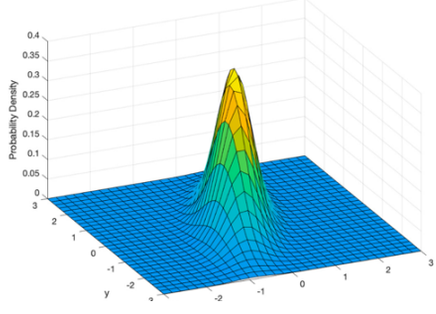
$$X|Y = \mathcal{X} \sim N\left(\mu_X + \frac{\sigma_X}{\sigma_Y}\rho(y-\mu_Y), (1-\rho^2)\sigma_X^2\right)$$

Moreover,

$$Y|X = y \sim N\left(\mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x-\mu_X), (1-\rho^2)\sigma_Y^2\right)$$

3) Independence $\Longleftrightarrow$ Uncorrelation under this circumstance

3. Geometric interpretation



1) If we intersect this surface with planes parallel to the yz-plane (i.e., with $x = x_0$), we have
$f(x_0, y) = f_X(x_0)h(y \mid x_0)$. (Ball-Shaped Curve)

2) If we intersect this surface with planes parallel to the xz-plane (i.e., with $y = y_0$), we have
$f(x, y_0) = f_Y(y_0)g(x \mid y_0)$. (Ball-Shaped Curve)

3) If we intersect this surface with planes parallel to the xy-plane (i.e., with $z = z_0$) with $0 < z_0 < \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$

Then we have an ellipse:
$z_0 2\pi\sigma_X\sigma_Y(1-\rho^2)^{0.5} = \exp\{-0.5q(x,y)\}$.

STA2001 Section FIVE Distributions of Functions of Random Variables

Topic I Function of One Random Variable

1. **Function of Random Variables:** Let X be a RV of either discrete or continuous type. Consider a function of X, say Y = u(X). Then Y is also a RV and has its pmf or pdf.

2. **Change-of-variable technique:**

(i) Discrete case: Let pmf of X be f(x): $S_X \to (0,1]$, and

Y = u(X) be ***a one-to-one mapping*** with inverse $X = u^{-1}(Y)$.

Then the pmf of Y: g(y) = P(Y = y)= P(u(X) = y) = $P(X= u^{-1}(Y))$

For $y \in S_Y=\{y|y=u(x),x\in S_X\}$. ***Then, $g(y)=f[u^{-1}(Y)]$ for $y\in S_Y$***

(ii) Continuous Case: For a continuous RV X, generally first calculate the cdf F(x), then the corresponding pdf f(x). $$F(x) = \int_0^x f(t)dt \Rightarrow F'(x) = \frac{dF(x)}{dx} = f(x)$$

Conditions: Y = u(X) is continuous, ***strictly increasing or decreasing***, has inverse function $X = u^{-1}(Y)$, whose derivative $du^{-1}(Y)$ exists.

(Calculate by the above idea) $$g(y) = f(u^{-1}(y))\left|\frac{du^{-1}(y)}{dy}\right|$$

Topic II Random Number Generator

1. Random Number Generator:

Theorem 5.1-1 Let Y have a distribution that is $U(0,1)$. Let $F(x)$ have the properties of a cdf of the continuous type with $F(a) = 0$, $F(b) = 1$, and suppose that $F(x)$ is strictly increasing on the support $a < x < b$, where a and b could be $-\infty$ and ∞, respectively. Then the random variable X defined by $X = F^{-1}(Y)$ is a continuous-type random variable with cdf $F(x)$.

Proof The cdf of X is

$$P(X \le x) = P[F^{-1}(Y) \le x], \qquad a < x < b.$$

Since $F(x)$ is strictly increasing, $\{F^{-1}(Y) \le x\}$ is equivalent to $\{Y \le F(x)\}$. It follows that

$$P(X \le x) = P[Y \le F(x)], \qquad a < x < b.$$

But Y is $U(0,1)$; so $P(Y \le y) = y$ for $0 < y < 1$, and accordingly,

$$P(X \le x) = P[Y \le F(x)] = F(x), \qquad 0 < F(x) < 1.$$

That is, the cdf of X is $F(x)$. □

2. Random Number Generator (Inverse Theorem):

Theorem 5.1-2 Let X have the cdf $F(x)$ of the continuous type that is strictly increasing on the support $a < x < b$. Then the random variable Y, defined by $Y = F(X)$, has a distribution that is $U(0,1)$.

Proof Since $F(a) = 0$ and $F(b) = 1$, the cdf of Y is

$$P(Y \le y) = P[F(X) \le y], \qquad 0 < y < 1.$$

However, $\{F(X) \le y\}$ is equivalent to $\{X \le F^{-1}(y)\}$; thus,

$$P(Y \le y) = P[X \le F^{-1}(y)], \qquad 0 < y < 1.$$

Since $P(X \le x) = F(x)$, we have

$$P(Y \le y) = P[X \le F^{-1}(y)] = F[F^{-1}(y)] = y, \qquad 0 < y < 1,$$

which is the cdf of a $U(0,1)$ random variable. □

3. Random Number Generator (Inverse Theorem):

[Random number generator from arbitrary distribution]

(i) generator a random number y from U(0, 1)

(ii) Take x = $F^{-1}(y)$, then x is a random number generated from the distribution or RV with cdf F(x).

(iii) When it is not one-to-one mapping, no general solutions.

4. ***Histogram*** (直方圖) for continuous distribution:

The simplest form of a histogram is constructed as follows:

1. Divide (or "bin") the sample space of the distribution into a sequence of adjacent, non-overlapping and equally spaced subintervals.
2. Treat each subinterval as an event, then count how many observed numerical outcomes fall into each subinterval and calculate the relative frequency
3. Draw a rectangle erected over the bin ***with height equal to the relative frequency divided by the width of each subinterval.***

Remark:

*Note that ***the area of the histogram is equal to 1,*** thus histogram gives an ***approximation of the probability density function*** of the underlying random variable.

Topic III Several Random Variables (Multivariate RVs)

1. Derivation of Multivariate RVs:

The multivariate RVs can arise in many different ways.

(e.g. We can perform a random experiment n times and let X_i, i = 1, ⋯ , n denote the RV for the i^{th} repetition of the random experiment. Then (X_1, ⋯ , X_n) is a multivariate RV.)

*

2. ***N independent RVs:***

The n RVs $X_1, \cdots, X_n$ are said to be (mutually) independent if $f(x_1, \cdots, x_n) = f_{X1}(x_1) \cdot \cdots \cdot f_{Xn}(x_n)$, where $f(x_1, \cdots, x_n)$ is the joint pmf or pdf of $X_1 \cdots X_n$, and $f_{Xi}(x_i)$ is the marginal pmf or pdf of X_i, $i = 1, \cdots, n$.

Remark: If $X_1, \cdots, X_n$ are independent, then any pair of them, any triple of them, $\cdots$, any $(n-1)$ of them are also independent.

3. i.i.d.:

Independently and identically distributed (i.i.d.) RVs $X_1, \cdots, X_n$, are also called ***random sample of size n from a common distribution.***

4. **Mathematical Expectation of Independent RVs:**

Theorem 5.3-1 Say $X_1, X_2, \ldots, X_n$ are independent random variables and $Y = u_1(X_1)\,u_2(X_2)\cdots u_n(X_n)$. If $E[u_i(X_i)]$, $i = 1, 2, \ldots, n$, exist, then

$$E(Y) = E[u_1(X_1)\,u_2(X_2)\cdots u_n(X_n)] = E[u_1(X_1)]E[u_2(X_2)]\cdots E[u_n(X_n)].$$

Proof In the discrete case, we have

$$\begin{aligned} &E[u_1(X_1)u_2(X_2)\cdots u_n(X_n)] \\ &= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} u_1(x_1)\,u_2(x_2)\cdots u_n(x_n)f_1(x_1)f_2(x_2)\cdots f_n(x_n) \\ &= \sum_{x_1} u_1(x_1)f_1(x_1)\sum_{x_2} u_2(x_2)f_2(x_2)\cdots\sum_{x_n} u_n(x_n)f_n(x_n) \\ &= E[u_1(X_1)]E[u_2(x_2)]\cdots E[u_n(X_n)]. \end{aligned}$$

In the proof of the continuous case, obvious changes are made; in particular, integrals replace summations.

Theorem 5.3-2 If $X_1, X_2, \ldots, X_n$ are n independent random variables with respective means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then the mean and the variance of $Y = \sum_{i=1}^n a_i X_i$, where $a_1, a_2, \ldots, a_n$ are real constants, are, respectively,

$$\mu_Y = \sum_{i=1}^n a_i\mu_i \qquad \text{and} \qquad \sigma_Y^2 = \sum_{i=1}^n a_i^2\,\sigma_i^2.$$

(proof by hand)

Mean: by the property of mathematical expectation

$$\mu_Y = E(Y) = E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i) = \sum_{i=1}^n a_i\mu_i$$

Variance: $\sigma_Y^2 = E((Y-EY)^2) = E\left(\left(\sum_{i=1}^n a_i X_i - \sum_{i=1}^n a_i\mu_i\right)^2\right) = \sum_{i=1}^n a_i^2 E\left((X_i-\mu_i)^2\right) + \sum_{i=1}^n \sum_{j\neq i}^n a_i a_j E\left((X_i-\mu_i)(X_j-\mu_j)\right)$

Because $E\left((X_i-\mu_i)(X_j-\mu_j)\right) = COV(X_i, X_j) = 0$, ($X_i$, X_j are independent)

$$\sigma_Y^2 = \sum_{i=1}^n a_i^2 E\left((X_i-\mu_i)^2\right) = \sum_{i=1}^n a_i^2\sigma_i^2 \quad . \square$$

5. **Mean of Random Sample/Sample mean (a type of statistics/a statistic):**

Now consider the **mean of a random sample**, $X_1, X_2, \ldots, X_n$, from a distribution with mean μ and variance σ^2, namely,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

which is a linear function with each $a_i = 1/n$. Then

$$\mu_{\overline{X}} = \sum_{i=1}^{n} \left(\frac{1}{n}\right)\mu = \mu \qquad \text{and} \qquad \sigma^2_{\overline{X}} = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

Topic IV Moment Generating Function Technique

1. mgf of n independent RVs:

Theorem 5.4-1 If $X_1, X_2, \ldots, X_n$ are independent random variables with respective moment-generating functions $M_{X_i}(t)$, $i = 1,2,3,\ldots,n$, where $-h_i < t < h_i, i = 1,2,\ldots,n$, for positive numbers $h_i, i = 1,2,\ldots,n$, then the moment-generating function of $Y = \sum_{i=1}^{n} a_i X_i$ is

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t), \text{ where } -h_i < a_i t < h_i, \ i = 1,2,\ldots,n.$$

(proof by hand)

$M_Y(t) = E(e^{Yt}) = E(e^{\sum_{i=1}^{n} a_i X_i t}) = E(\prod_{i=1}^{n} e^{a_i X_i t})$

By Theorem 5.3-1, because $X_1, X_2, \cdots X_n$ are independent,

$E(\prod_{i=1}^{n} e^{a_i X_i t}) = \prod_{i=1}^{n} E(e^{a_i X_i t}) = \prod_{i=1}^{n} M_{X_i}(a_i t)$,

where $-h_i < a_i t < h_i$, $i = 1,2,3,\cdots,n$

• □

2. Corollaries:

Corollary 5.4-1 If $X_1, X_2, \ldots, X_n$ are observations of a random sample from a distribution with moment-generating function $M(t)$, where $-h < t < h$, then

(a) the moment-generating function of $Y = \sum_{i=1}^{n} X_i$ is

$$M_Y(t) = \prod_{i=1}^{n} M(t) = [M(t)]^n, \quad -h < t < h;$$

(b) the moment-generating function of $\overline{X} = \sum_{i=1}^{n} (1/n) X_i$ is

$$M_{\overline{X}}(t) = \prod_{i=1}^{n} M\left(\frac{t}{n}\right) = \left[M\left(\frac{t}{n}\right)\right]^n, \quad -h < \frac{t}{n} < h.$$

3. Applied to Chi-square distribution:

Theorem 5.4-2 Let $X_1, X_2, \ldots, X_n$ be independent chi-square random variables with $r_1, r_2, \ldots, r_n$ degrees of freedom, respectively. Then $Y = X_1 + X_2 + \cdots + X_n$ is $\chi^2(r_1 + r_2 + \cdots + r_n)$.

Proof By Theorem 5.4-1 with each $a = 1$, the mgf of Y is

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = (1-2t)^{-r_1/2}(1-2t)^{-r_2/2}\cdots(1-2t)^{-r_n/2}$$
$$= (1-2t)^{-\Sigma r_i/2}, \qquad \text{with } t < 1/2,$$

which is the mgf of a $\chi^2(r_1 + r_2 + \cdots + r_n)$. Thus, Y is $\chi^2(r_1 + r_2 + \cdots + r_n)$.

(what if for $X_1 - X_2 + X_3 - \ldots$ and $2X_1$?)

1) $M_Y(t) = M_{X_1}(t) M_{X_2}(-t) \cdots$, not χ^2 distribution.

2) $M_Y(t) = M_{X_1}(2t) = \frac{1}{(1-4t)^{\frac{n}{2}}}$, still not χ^2 distribution.

Topic V Random function associated with normal distribution

1. Random function associated with n normal distribution:

Theorem 5.5-1 If $X_1, X_2, \ldots, X_n$ are n mutually independent normal variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively, then the linear function

$$Y = \sum_{i=1}^{n} c_i X_i$$

has the normal distribution

$$N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right).$$

2. Corollaries:

Corollary 5.5-1 If $X_1, X_2, \ldots, X_n$ are observations of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$, then the distribution of the sample mean $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ is $N(\mu, \sigma^2/n)$.

(which means $(\overline{X} - \mu)/(\sigma/n^{0.5})$ follows standard normal distribution N(0, 1).)

3. *Sample Mean and Sample Variance:*

Example 2.8.1 (Sample Mean). Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with common mean μ and variance σ^2. The **sample mean** is defined by $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$. This is a linear combination of the sample observations with $a_i \equiv n^{-1}$; hence, by Theorem 2.8.1 and Corollary 2.8.2, we have

$$E(\overline{X}) = \mu \text{ and } \text{Var}(\overline{X}) = \frac{\sigma^2}{n}. \tag{2.8.3}$$

By Definition 4.1.3 of Chapter 4, we say that $\overline{X}$ is an **unbiased estimator** of μ. ■

Example 2.8.2 (Sample Variance). Define the **sample variance** by

$$S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)^{-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right). \tag{2.8.4}$$

where the second equality follows after some algebra; see Exercise 2.8.1. Using the above theorems, the results of the last example, and the fact that $E(X^2) = \sigma^2 + \mu^2$, we have the following:

$$\begin{aligned} E(S^2) &= (n-1)^{-1}\left(\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right) \\ &= (n-1)^{-1}\{n\sigma^2 + n\mu^2 - n[(\sigma^2/n) + \mu^2]\} \\ &= \sigma^2. \end{aligned} \tag{2.8.5}$$

Hence, S^2 is an unbiased estimator of σ^2. ■

4. Student's t distribution:

Theorem 5.5-3 **(Student's t distribution)** Let

$$T = \frac{Z}{\sqrt{U/r}},$$

where Z is a random variable that is $N(0,1)$, U is a random variable that is $\chi^2(r)$, and Z and U are independent. Then T has a t distribution with pdf

$$f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{\pi r}\,\Gamma(r/2)} \frac{1}{(1+t^2/r)^{(r+1)/2}}, \qquad -\infty < t < \infty.$$

(proof by hand)

Property: Student's t Distribution is ***a heavy tailed distribution*** compared with Standard Normal distribution (Its tails are very fat.).

- Joint pdf of z & U: (because of independence)
$f_{z,u}(z,u) = f_z(z)\cdot f_u(u) = \frac{1}{\sqrt{2\pi}\,\Gamma(\frac{r}{2})\cdot 2^{\frac{r}{2}}}\, u^{\frac{r}{2}-1} e^{-(\frac{z^2}{2}+\frac{u}{2})}$

CDF of T satisfies:
$F(t) = P(T\le t) = P(z\le\sqrt{\frac{u}{r}}t) = \int_0^{\infty}\int_{-\infty}^{\sqrt{\frac{u}{r}}t} f_{z,u}(z,u)\,dz\,du$

By Fundamental Theorem of calculus:
$f(t) = F'(t) = \int_0^{\infty} \frac{1}{\sqrt{2\pi r}\cdot\Gamma(\frac{r}{2})}\left(\frac{r}{r+t^2}\right)^{\frac{r+1}{2}} \cdot \left(\frac{u(r+t^2)}{2r}\right)^{\frac{r+1}{2}-1} \cdot e^{-\frac{(r+t^2)u}{2r}} \cdot \frac{\Gamma(\frac{r+1}{2})}{\sqrt{2r}\,\Gamma(\frac{r}{2})}\cdot\left(\frac{r}{r+t^2}\right)^{\frac{r+1}{2}}$ $d\frac{(r+t^2)u}{2r}$. □

- Another way: Let $v = \sqrt{\frac{u}{r}}$, $v = u$

$f_{T,V}(t,v) = f_{z,u}(t\sqrt{\tfrac{v}{r}}, v^2)\cdot\left|\frac{\partial(z,u)}{\partial(t,v)}\right|$

$= \begin{cases} \frac{1}{\sqrt{2\pi}\,\Gamma(\frac{r}{2})2^{\frac{r}{2}}} v^{r-1}\exp\left[-\frac{v^2}{2}(1+\frac{t^2}{r})\right]\sqrt{\frac{v}{r}}, & v>0 \\ 0, & \text{elsewhere} \end{cases}$

$f_T(t) = \int_0^{\infty} \frac{v^{\frac{r}{2}-1}}{\sqrt{2\pi}\,\Gamma(\frac{r}{2})2^{\frac{r}{2}}}\exp\left[-\frac{v}{2}(1+\frac{t^2}{r})\right]\sqrt{\frac{v}{r}}\,dv$

$= \cdots$ like the left. □

5. *Student's Theorem*

(show the proof by hand)

Theorem 3.6.1. *Let $X_1,\ldots,X_n$ be iid random variables each having a normal distribution with mean μ and variance σ^2. Define the random variables,*

$$\overline{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i \text{ and } S^2 = \tfrac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2.$$

Then,

(a). $\overline{X}$ *has a* $N\left(\mu, \frac{\sigma^2}{n}\right)$ *distribution.*

(b). $\overline{X}$ *and* S^2 *are independent.* (proof is omitted for (b))

(c). $(n-1)S^2/\sigma^2$ *has a* $\chi^2(n-1)$ *distribution.*

(d). *The random variable*

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}}, \qquad (3.6.8)$$

has a Student t-distribution with $n-1$ degrees of freedom.

(d) By (a) & (c) & Theorem 5.5-3

T is a t-distribution $t(n-1)$

by $\dfrac{\frac{\overline{X}-\mu}{\delta/\sqrt{n}} = Z}{\sqrt{\frac{(n-1)S^2}{\delta^2}/n-1}} = \dfrac{\overline{X}-\mu}{S/\sqrt{n}}$. □

(a). By mgf technique, Mgf of $\overline{X}$: $M_{\overline{X}}(t) = \exp\left\{\mu t + \frac{\sigma^2}{2n}t^2\right\}$

(c) $\frac{(n-1)S^2}{\delta^2} = \sum_{i=1}^n \left(\frac{X_i-\overline{X}}{\delta}\right)^2$. Because $\left(\frac{X_i-\mu}{\delta}\right)^2 \sim \chi^2(1)$ (proved before)

$\Rightarrow \sum_{i=1}^n \left(\frac{X_i-\mu}{\delta}\right)^2 = \sum_{i=1}^n \left(\frac{X_i-\overline{X}}{\delta}+\frac{\overline{X}-\mu}{\delta}\right)^2 \sim \chi^2(n)$, $\therefore \sum_{i=1}^n \left(\frac{X_i-\overline{X}}{\delta}\right)^2 + n\left(\frac{\overline{X}-\mu}{\delta}\right)^2 + 2\sum_{i=1}^n \left(\frac{X_i-\overline{X}}{\delta}\right)\left(\frac{\overline{X}-\mu}{\delta}\right) \sim \chi^2(n)$

$\therefore \frac{(n-1)S^2}{\delta^2} + \left(\frac{\overline{X}-\mu}{\delta/\sqrt{n}}\right)^2 \sim \chi^2(n)$. By (b) (the cross term $=0$)

$\Rightarrow M_{\frac{(n-1)S^2}{\delta^2}}(t)\cdot M_{\left(\frac{\overline{X}-\mu}{\delta/\sqrt{n}}\right)^2}(t) = (1-2t)^{-\frac{n}{2}} \Rightarrow M_{\frac{(n-1)S^2}{\delta^2}}(t) = (1-2t)^{-\frac{n-1}{2}}$

$(1-2t)^{-\frac{1}{2}} \sim \chi(1)$ $\Rightarrow \frac{(n-1)S^2}{\delta^2} \sim \chi^2(n-1)$

Topic VI Convergence in Random Variables

1. Convergence in Distribution/ converge weakly/ converge in law to a random variable X:

Definition 4.3.1 (Convergence in Distribution). *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be, respectively, the cdfs of X_n and X. Let $C(F_X)$ denote the set of all points where F_X is continuous. We say that X_n* **converges in distribution** *to X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \text{for all } x \in C(F_X).$$

We denote this convergence by

$$X_n \xrightarrow{D} X.$$

(e.g.) Proof that a sequence of t-distribution with degrees of freedom n has limits as a standard normal distribution.

$$\lim_{n\to\infty} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\frac{n}{2}}\,\Gamma(\frac{n}{2})} = \lim_{n\to\infty} \left(\left(1+\frac{1}{n}\right)^{\frac{n}{2}}\right)^{\frac{n+2}{2n}} \cdot e^{-\frac{1}{2}}$$

$$= \lim_{n\to\infty} e^{-\frac{1}{2}} \cdot e^{\frac{n+2}{2n}} = \lim_{n\to\infty} e^{\frac{1}{n}} = 1$$

By Lebesgue Dominated Convergence Theorem,

$$\lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} \int_{-\infty}^{t} f_n(y)\,dy = \int_{-\infty}^{t} \lim_{n\to\infty} f_n(y)\,dy$$

$$\lim_{n\to\infty} f_n(y) = \lim_{n\to\infty} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\frac{n}{2}}\,\Gamma(\frac{n}{2})} \cdot \lim_{n\to\infty} \left(\frac{1}{1+\frac{y^2}{n}}\right)^{\frac{1}{2}} \cdot \lim_{n\to\infty} \left(\frac{1}{\sqrt{2\pi}}\left(1+\frac{y^2}{n}\right)^{-\frac{n}{2}}\right)$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \sim N(0,1)$$

According to Stirling's Formula, $\Gamma(k+1) \overset{k \text{ large}}{\approx} \sqrt{2\pi}\, k^{\frac{k+1}{2}} e^{-k}$

2. Convergence in Probability:

Definition 4.2.1. *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable defined on a sample space. We say that X_n* **converges in probability** *to X if for all $\epsilon > 0$*

$$\lim_{n\to\infty} P[|X_n - X| \geq \epsilon] = 0,$$

or equivalently,

$$\lim_{n\to\infty} P[|X_n - X| < \epsilon] = 1.$$

If so, we write

$$X_n \xrightarrow{P} X.$$

Topic VII Central Limit Theorem (CLT)

1. Central Limit Theorem:

Theorem 4.4.1. *Let $X_1, X_2, \ldots, X_n$ denote the observations of a random sample from a distribution that has mean μ and positive variance σ^2. Then the random variable $Y_n = (\sum_1^n X_i - n\mu)/\sqrt{n}\sigma = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ converges in distribution to a random variable which has a normal distribution with mean zero and variance 1.*

(show the proof by hand)

It can be showed using Limit mgf tech.

$\frac{\bar{X}-\mu}{\delta/\sqrt{n}} = \sum_{i=1}^{n} \frac{X_i-\mu}{\sqrt{n}\delta} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \cdot \frac{X_i-\mu}{\delta}$, Let $Z_i = \frac{X_i-\mu}{\delta}$, then Z_i has distribution with $\mu=0$, $\delta^2=1$

$M_{\frac{\bar{x}-\mu}{\delta/\sqrt{n}}}(t) = \prod_{i=1}^{n} M_{Z_i}\left(\frac{t}{\sqrt{n}}\right) \xrightarrow{\text{Taylor's Theorem}} \prod_{i=1}^{n} \left(1+\frac{M''_{Z_i}(\xi)}{2n}t^2\right)$ $\left(M_{Z_i}(t)=1+\frac{M''_{Z_i}(\xi)}{2}t^2\right)$

$\overset{\text{i.i.d.}}{=} \left(1+\frac{M''(\xi)}{2n}t^2\right)^n$. Because $-\frac{t}{\sqrt{n}} < \xi < \frac{t}{\sqrt{n}}$, $\lim_{n\to\infty} M''(\xi) = M''(0) = \mu^2+\delta^2 = 1$

$\therefore \lim_{n\to\infty} M_{\frac{\bar{x}-\mu}{\delta/\sqrt{n}}}(t) = \lim_{n\to\infty} \left(1+\frac{M''(\xi)t^2}{2n}\right)^n = \lim_{u\to\infty} \left(1+\frac{t^2}{2n}\right)^n = e^{\frac{1}{2}t^2}$

$\therefore \frac{\bar{X}-\mu}{\delta/\sqrt{n}}$ converges to $N(0,1)$ when $n\to\infty$. □

3. ***Usage of CLT:***

CLT can be used to estimate: (when n is sufficiently large, which depends.)

(i) $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ as a standard normal distribution N (0, 1);

(ii) $\bar{X}$ as a normal distribution N (μ, σ^2/n);

(iii) $\sum_{i=1}^{n} X_i$ as a normal distribution N ($n\mu$, $n\sigma^2$);

Topic VIII Approximations for Discrete Distributions

1. ***Histogram for Discrete Distribution:***

Consider a discrete RV Y with pmf f(y): S→(0,1] with S = {0,1,⋯ ,n}. Then the histogram for Y:

h(y) = f(k), y∈((k−1)/2, (k+1)/2), k = 0, 1, ⋯, n

For k=0, 1,⋯, n, P(Y = k) = f(k) corresponds to *the area of the rectangle with a height* of P(Y = k) and a base of ***length 1 centered at k.***

2. ***Half-unit correction for continuity:***

In using the normal distribution to approximate probabilities for the **discrete distributions**, areas under the pdf for the normal distribution will be used to approximate areas of rectangles in the probability histogram for the binomial distribution.

Since these rectangles have **unit base** centered at the integers, this is called a **half-unit correction** for continuity.

Note that, for an integer k, **P(Y = k) = P(k−1/2 < Y < k+1/2)**

Topic IX Bounds for Probability of RVs

1. ***Chebyshev Inequality & Markov's Inequality for tail probabilities (omitted)***
2. ***Weak Law of Large Numbers:***

Let X be the sample mean of a random sample $X_1, X_2, \cdots, X_n$ from a distribution with finite nonzero variance. Then X converges in probability to μ, i.e.

$$\lim_{n\to\infty} P\left(|\overline{X} - \mu| \geq \varepsilon\right) = 0$$

Topic X Limiting Moment Generating Function Technique

1. *Limiting Mgf Tech:*

Theorem 5.9-1 If a sequence of mgfs approaches a certain mgf, say, $M(t)$, for t in an open interval around 0, then the limit of the corresponding distributions must be the distribution corresponding to $M(t)$.

(Measured by CDF, i.e., convergence in probability.)

2. *Estimate Binomial Distribution as Poisson Distribution*

(proof by hand) Consider $X \sim b(n,p)$, assume $\lambda = np$ holds constant.

$$M_X(t) = (1-p+pe^t)^n = (1+p(e^t-1))^n = (1+\frac{\lambda(e^t-1)}{n})^n$$

$$\lim_{n\to\infty} M_X(t) = \lim_{n\to\infty}(1+\frac{\lambda(e^t-1)}{n})^n = e^{\lambda(e^t-1)}, \text{ which is mgf of Poisson}(np).$$

Under this circumstance, λ holds but $n\to\infty$, therefore, we need small p & large n to give a more precise estimation. □

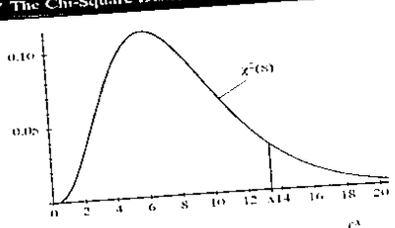
Probability and Statistics I

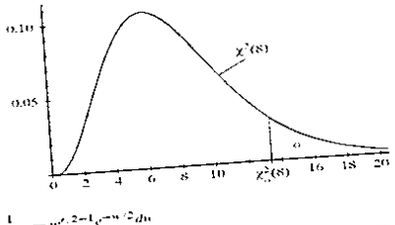
Mid-term Examination
SDS, CUHK(SZ)

March 20, 2021

Answer the multiple choice questions (Section I) in the Answer Card, and answer the regular questions (Section II) in the Answer Book.

Table IV The Chi-Square Distribution





$$P(X \le \lambda) = \int_0^\lambda \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

	$P(X \le \lambda)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21
11	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72
12	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22
13	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69
14	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14
15	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58
16	5.812	6.908	7.962	9.312	23.54	26.30	28.84	32.00
17	6.408	7.564	8.672	10.08	24.77	27.59	30.19	33.41
18	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.80
19	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19
20	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57
21	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93
22	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64
24	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31
26	12.20	13.84	15.38	17.29	35.56	38.88	41.92	45.64
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96
28	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
40	22.16	24.43	26.51	29.05	51.80	55.76	59.34	63.69
50	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
60	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
70	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4
80	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3

This table is abridged and adapted from Table III in *Biometrika Tables for Statisticians*, edited by E.S.Pearson and H.O.Hartley.

Multiple Choices (72 points)

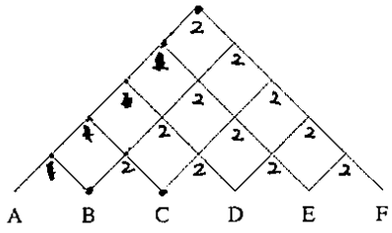
- **3 points for each correct answer; -1 point for each incorrect answer; 0 points for no answer**
- **For each question, only choose (at most) one out of four given choices (A,B,C and D). If you choose more than one choice in one question, your answer will be incorrect and 1 point will be deduced.**

1. Let A and B be two events. Suppose $P(A) = 0.4$, $P(B) = 0.5$.
   (a) $P(A \cap B') = 0.2$ if $P(A \cap B) = 0.2$ ✓
   (b) $P(A \mid B) = 0.4$ if A and B are mutually exclusive ✗
   (c) $P(A \cup B) = 0.7$ if A and B are independent ✓
   (d) $P(A \cap B \mid A \cup B) = 0.5$ if $P(A \mid B) = 0.6$. ✓

   Find which one of the following statements is correct.
   A. (a)(c)(d)
   B. (c)(d)

C. (b)(d)
D. (a)(c)

2. Skiers at the top of the mountain have a variety of choices as they head down the trails. Assume that at each intersection, a skier is equally likely to go left or right. Find the percent of skiers who end up at B and C, respectively.



A. (a) $\frac{5}{32}$, (b) $\frac{6}{16}$
B. (a) $\frac{5}{32}$, (b) $\frac{10}{32}$
C. (a) $\frac{4}{16}$, (b) $\frac{10}{32}$
D. (a) $\frac{4}{16}$, (b) $\frac{6}{16}$

3. In an urn are 7 blue and 3 red marbles.

(a) If you draw 3 marbles without replacement, what is the probability that less than 2 will be red? 0 red or 1 red

(b) If you draw the marbles one by one without replacement, and STOP when only one color of marbles are left, what is the probability that the first drawn marble is red and the last drawn marble is blue?

Find which one of the following statements is correct.

A. (a) $\frac{98}{120}$ (b) $\frac{1}{10}$
B. (a) $\frac{88}{120}$ (b) $\frac{1}{10}$
C. (a) $\frac{98}{120}$ (b) $\frac{1}{15}$
D. (a) $\frac{88}{120}$ (b) $\frac{1}{15}$

4. Let A, B and C be three events.

(a) If the events A and B are mutually exclusive, then A and B are NOT independent.
(b) If $A \subseteq B$, then A and B are NOT independent.
(c) If $P(A) = 0.4$, $P(B) = 0.6$ and $P(A \cup B) = 0.76$, then A and B are independent.
(d) Suppose A, B, C are pairwise independent. In addition, suppose A and $B \cup C$ are independent. Then A, B, C are mutually independent.

How many of the above four statements are TRUE?

A. 1
B. 2
C. 3
D. 4.

5. I have in my pocket five coins. Four of them are ordinary coins with equal chances of coming up head and tail when tossed, and the fifth has two heads.
(a) If I take one of the coins at random from my pocket and toss it, what is the probability that it comes up head?
(b) If I toss a randomly taken coin and it comes up head, what is the probability that it is the coin with two heads?

Find which one of the following statements is correct.

A. (a) $\frac{3}{5}$, (b) $\frac{1}{3}$
B. (a) $\frac{3}{5}$, (b) $\frac{1}{4}$
C. (a) $\frac{4}{7}$, (b) $\frac{1}{3}$
D. (a) $\frac{4}{7}$, (b) $\frac{1}{4}$

6. Suppose that for three dice of the standard type all 216 outcomes of a throw are equally likely. Denote the obtained scores of the three dice by X_1, X_2 and X_3, respectively.
(a) $P(X_1 + X_2 + X_3 \leq 5) = 10/216$,
(b) $P(\min\{X_1, X_2, X_3\} \geq 2) = 125/216$.
(c) $P(X_1 + X_2 < (X_3)^2) = 137/216$.

Find which one of the following statements is correct.

A. (a)(b)
B. (a)(c)
C. (b)(c)
D. (a)(b)(c)

7. A small plane went down and was missing, and the search was organized into three regions. Starting with the likeliest, they are:

Region	Initial chance the plane is there	Chance of being overlooked in the search
Mountains	0.5	0.3
Prairie	0.3	0.2
Sea	0.2	0.9

The last column gives the chance that if the plane is there, it will not be found. For example, if it went down at sea, there is 90% chance it will have disappeared, or otherwise not be found. Since the pilot is not equipped to long survive a crash in the mountains, it is particularly important to determine the chance that the plane went down in the mountains.
(a) Before any search is started, what is the chance that the plane is in the mountains?
(b) The initial search was in the mountains, and the plane was not found. Now what is the chance the plane is nevertheless in the mountains?
(c) The search was continued over the other two regions, and unfortunately the plane was not found anywhere. Finally now what is the chance that the plane is in the mountains?

Find which one of the following statements is correct.

A. (a) 0.50, (b) 0.23, (c) 0.38

B. (a) 0.50, (b) 0.28, (c) 0.30
C. (a) 0.20, (b) 0.23, (c) 0.38
D. (a) 0.20, (b) 0.28, (c) 0.30

8. Two balls are chosen randomly from an urn containing 7 white, 4 black, and 1 orange balls. Suppose that we win \$1 for each white ball drawn and we lose \$1 for each orange ball drawn. Denote $\$X$ as the amount that we can win. Determine the probability mass function $p(x)$ for (a) $x = 0$ and (b) $x = 2$, and determine the cumulative distribution function $F(x)$ for (c) $x = 0$ and (d) $x = 2$.

   Find which one of the following statements is correct.
   A. (a) $\frac{13}{66}$, (b) $\frac{7}{22}$, (c) $\frac{17}{66}$, (d) 1
   B. (a) $\frac{13}{66}$, (b) $\frac{7}{22}$, (c) $\frac{2}{33}$, (d) $\frac{15}{22}$
   C. (a) $\frac{2}{33}$, (b) $\frac{14}{33}$, (c) $\frac{15}{66}$, (d) 1
   D. (a) $\frac{2}{33}$, (b) $\frac{14}{33}$, (c) $\frac{2}{33}$, (d) $\frac{15}{22}$

9. Consider a random experiment of "tossing a coin (not necessarily a fair one) i
   initely until it turns out a head, and the number of dollars you will win is equ
   the number of tosses it takes to see the first head turning out." You are interest
   how many dollars you will win after the experiment. Assume $\mathrm{P}(\{H\}) = p$.
   (a) What is the expected amount of money you will win?
   (b) What is the probability that you will win at least 3 dollars?

   Find which one of the following statements is correct.
   A. (a) $\frac{1}{p}$, (b) $1 - p$
   B. (a) $\frac{1}{1-p}$, (b) p^2
   C. (a) p, (b) p^2
   D. (a) $\frac{1}{p}$, (b) $(1-p)^2$.

10. Consider a two-engine plane and a four-engine plane. Suppose that each engir
    each plane will fail independently with the same probability p (where $0 < p < 1$
    that each plane will make a safe flight if at least half of the engines remain opera
    (a) Flying with a four-engine plane is safer than flying with a two-engine plane;
    (b) Flying with a two-engine plane is safer than flying with a four-engine plane
    (c) Flying with a two-engine plane may be safer than flying with a four-engine p
    (d) Flying with a two-engine plane is the same as flying with a four-engine plar

    Find which one of the following statements is correct.
    A. (a)
    B. (d)
    C. (b) and (d)
    D. (c) .

11. Let X follow a discrete uniform distribution on $\{a, \ldots, b\}$, where a and b are int

with $a \leq b$. The pmf of X is

$$\mathrm{P}(X = x) = p(x) = \begin{cases} \frac{1}{b-a+1}, & \text{for } x \in \{a, \ldots, b\} \\ 0, & \text{otherwise} \end{cases}$$

Let $E(X)$, $Var(X)$ and $M_X(t)$ denote the mean, variance and moment generating function of X, respectively.

Find which one of the following statements is correct.

A. (a) $E(X) = \frac{a+b}{2}$, (b) $Var(X) = \frac{(b-a+1)^2+1}{12}$, (c) $M_X(1) = \frac{e^a - e^b}{(b-a+1)(1-e)}$

(a) $E(X) = \frac{b-a}{2}$, (b) $Var(X) = \frac{(b-a+1)^2-1}{12}$, (c) $M_X(1) = \frac{e^a - e^{(b-1)}}{(b-a+1)}$

(a) $E(X) = \frac{a+b}{2}$, (b) $Var(X) = \frac{(b-a+1)^2-1}{12}$, (c) $M_X(1) = \frac{e^a - e^{(b+1)}}{(b-a+1)(1-e)}$

(a) $E(X) = \frac{b-a}{2}$, (b) $Var(X) = \frac{(b-a+1)^2+1}{12}$, (c) $M_X(1) = \frac{e^a - e^b}{(b-a+1)}$

12. The number of times that an individual contracts a cold in a given year is a Poisson random variable with mean $\theta = 6$. Suppose a new wonder drug (based on large quantities of vitamin C) has just been marketed that reduces the Poisson mean to $\theta = 4$ for 60 percent of the population. For the other 40 percent of the population the drug has no appreciable effect on colds. If an individual tries the drug for a year and has 3 colds in that time, how likely is it that the drug is beneficial for him/her?

A. 0.82
B. 0.77
C. 0.56
D. 0.69 .

13. Ten percent of all trucks undergoing a certain inspection will fail the inspection. Assume that trucks are independently undergoing this inspection one at a time. The expected number of trucks inspected before a truck fails inspection is

A. 1 B. 5 C. 10 D. 20

14. A discrete random variable X has a pmf such that $P(X = k+1) = aP(X = k)$, where $1 > a > 0$ and $k = 0, 1, 2 \ldots$. Compute the probability $P(X \geq 11)$.

A. a^{11} B. $1 - a^{11}$ C. a^{10} D. a^{12}

15. Consider a sequence of N Bernoulli trials with probability of success being equal to p. It is observed that two successes occurred in these N trials, where $2 < N$ and N is even. What is the probability that one success occurred in the first $N/2$ trials?

A. $\frac{N/2}{\binom{N}{2}p(1-p)^{(N-2)}}$ B. $\frac{N}{2(N-1)}$ C. $\frac{(N/2)p^{(N/2)}}{\binom{N}{2}p(1-p)^{(N-2)}}$ D. $\frac{(N/2)p^{(N-2)}(1-p)^{(N/2-1)}}{\binom{N}{2}p(1-p)^{(N-2)}}$

16. A telephone company employs 5 operators who receive requests independently of one another. The number of the requests received by each operator has a Poisson

distribution, and each operator receives on average 1 request every 30 minutes. What is the probability that during a given 2 hour period, exactly 4 of the 5 operators receive no requests?

A. $5(e^{-8} - e^{-10})$.

B. $4(e^{-10} - e^{-16})$.

C. $5(e^{-16} - e^{-20})$.

D. $5(e^{-10} - e^{-16})$.

17. Suppose that X is a continuous random variable with the pdf

$$f(x) = \begin{cases} 1 + x, & -1 \leq x < 0, \\ 1 - x, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Define a random variable $Y = X^2$ [illegible] ne question is what is the value of $P(\frac{5}{4} < Y \leq \frac{7}{4})$.

A. $\sqrt{3} - \frac{3}{2}$ B. $2\sqrt{3} - 3$. C. $2 - \sqrt{3}$. D. $3 - 2\sqrt{2}$.

18. Suppose that $f_1(x)$ is the pdf of the standard normal distribution, and $f_2(x)$ is the pdf of the uniform distribution over $[-1, 3]$.Let $f(x)$ be a function defined as

$$f(x) = \begin{cases} af_1(x), & x \leq 0, \\ bf_2(x), & x > 0, \end{cases}$$

where $a > 0$ and $b > 0$. If $f(x)$ is a pdf, then which one of the following equalities must hold?

A. $2a + 3b = 4$.

B. $3a + 2b = 4$.

C. $a + b = 1$.

D. $a + b = 2$.

19. Let the random variable X have a distribution with probability density function

$$f(x) = \frac{1}{\theta} e^{-(x-\delta)/\theta}, \quad \delta < x < \infty.$$

What is the mean and variance of X?

A. $E(X) = \frac{\delta}{\theta}, Var(X) = (\frac{\delta}{\theta})^2$.

B. $E(X) = \theta + \delta, Var(X) = (\theta + \delta)^2$.

C. $E(X) = \theta - \delta, Var(X) = \theta^2$.

D. $E(X) = \theta + \delta, Var(X) = \theta^2$.

20. Consider the nonnegative random variable X with the pdf

$$f_X(x) = \alpha x e^{-x^2} + \beta I(0,1)$$

Here, α and β are constants to be determined. $I(0,1)$ is the indicator function given by

$$I(0,1) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{elsewhere} \end{cases}$$

Find α and β such that the 80^{th} percentile is at the point $\pi_{0.8} = 1$.

A. $\alpha = 0.4e, \beta = 1 - 0.2e$
B. $\alpha = 1 - 0.2e, \beta = 0.4e$
C. $\alpha = 1 - 0.4e, \beta = 1 - 0.2e$
D. $\alpha = 0.4e, \beta = 0.2e$

$Y \sim \text{Poisson}(np)$

Given $n = C\hat{x}^2$, $p = \frac{1}{365}$, $P(Y \ge 1) \ge 0.5$, find smallest n.

21. For a class of students, consider the probability that at least two of them have their birthdays on the same day. What's the minimum size of class, (i.e., the minimum number of students in the class,) such that the probability is larger than 0.5? (For simplicity, suppose each year has 365 days.)

Hint: The probability of a binomial distribution $b(n,p)$ can be approximated by the probability of a Poisson distribution with $\lambda = np$ for $n \ge 20$ and $p \le 0.05$, and moreover, $\ln 2 = 0.693$.

A. 22 B. 23 C. 24 D. 25

A famous problem — birthday paradox

$$1 - \frac{A_N^n}{N^n} \ge 0.5 \iff 1 - \frac{N!}{N^n (N-n)!} = 1 - \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{n-1}{N}\right)$$

$$\approx 1 - \left(1 - \frac{1}{N}\right)\left(1 - \frac{1}{N}\right)^2 \cdots \left(1 - \frac{1}{N}\right)^{n-1} = 1 - \left(1 - \frac{1}{N}\right)^{\frac{n(n-1)}{2}}$$

$$\approx 1 - e^{-\frac{n(n-1)}{2N}} \ge 0.5$$

answers are wrong / not this problem

22. In a coin tossing game, the gambler has to pay the casino $\$X$ for each toss of the coin which has a probability p of a head and probability $q = 1 - p$ of a tail. The game stops when either a head shows up at which point he gets a reward of $\$Y$, or no head occurs after K tosses at which point he gets no reward. Find the expected net gain of the casino from the game.

A. $KX(1-p)^{K+1} + X\left(\frac{1}{p}\right)\left((1-q^K) - (K+1)pq^K\right) - Y$
B. $KX(1-p)^K + X\left((1-q^{K+1}) - Kpq^K\right) - Y$
C. $(K+1)X(1-p)^{K+1} + X\left(\frac{1}{p}\right)\left((1-q^K) - (K+1)pq^K\right) - Y$
D. $KX(1-p)^K + X\left(\frac{q}{p}\right)(1-q^K) + X(1-(K+1)q^K) - Y$

23. Consider the following statements:

(i) Let $X \sim N(\mu, 1)$ with $\mu \ge 0$. The largest possible value of c such that the inequality $P(|X| \le c) \le 0.8064$ holds is $c = 1.3$.

for any $\mu \ge 0$, therefore it should satisfy $\mu = 0$ (when)

(ii) Let $Z \sim N(0,1)$, then the probability that the quadratic equation $0.1x^2 + Zx + 0.04 = 0$ has real roots is 0.9.

(iii) If $X \sim N(0,3)$, then $\frac{E(X^6)}{E(X^4)} = 3$. ✓

Find which one of the following statements is correct.

A. All statements are ture. B. Only (i), (ii) and (iii) are ture. C. Only (ii) and (iii) are ture. D. Only (i) and (iii) are ture.

24. The number of cars passing a speed camera follows a Poisson distribution, and the average number of cars passing the camera in 1 minute is 2. Suppose current time is $t = 0$ and the unit is minute, and let the random variable Y be the time that the 6th car passes the speed camera and let the random variable Z be the time that the 1st car passes the speed camera.

Consider the following statements:

(i) The probability of at least 3 cars passing the speed camera in the first 2 minute is 0.7619 approximately.

(ii) $E(Y) = 3$, $\text{Var}(Y) = 1.5$, $E(Y^4) = 189$

(iii) for any $t, s > 0$, $P(Z > s + t | Z > t) = P(Z > s)$.

Find which one of the following statements is correct.

A. Only (i) is true. B. Only (i) and (ii) are true.
C. Only (ii) and (iii) are true. D. All statements are true.

II Regular Questions (28 points)

25. (12 points) Consider a sequence of n Bernoulli trials over the interval $[0, 1]$ which is divided into a large number n of subintervals each of width $1/n$ such that each trial occurs within a subinterval. The probability of a success in each trial is p, and the probability of two or more successes in each subinterval is practically zero. **Note: all details should be included!**

(a) (2 points) Supposing X is the number of successes in the interval $[0, 1]$, write an expression for $P(X = k)$ where $k = 0, 1, 2, \ldots, n$ is an integer.

(b) (1 point) Derive an expression for the ratio $\frac{P(X=k+1)}{P(X=k)}$.

(c) (4 points) Let $n \to \infty$ and the probability of success $p \to 0$ in such a way that the product np remains a constant μ. Derive an approximate expression for the limiting value of this ratio $\frac{P(X=k+1)}{P(X=k)}$ for a given fixed value of k. You must present your reasons in full.

(d) (4 points) Using the result in (c), derive the expression for $P(X = 0)$.

(e) (1 point) Finally, derive the expression for $P(X = k)$ for an arbitrary integer k.

26. (16 points) Let X be a continuous random variable with probability density function (pdf) defined as follows

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

Note: all details should be included!

(a) (4 points) Derive the moment generating function for X and show that $\text{E}(X) = 0$, $\text{Var}(X) = 1$.

(b) (3 points) Define a random variable $Y = aX + b$. What is the distribution of Y? Show the proof in details.

(c) (2 points) What is $E((Y - b)^{177})$?

(d) (7 points) Prove that $\text{Var}(X) = 1$ without using the moment generating function.

Probability and Statistics I

Mid-term Sample
SSE, CUHK(SZ)

Name: _____ Student ID: _____

Answer the multiple choice questions (Section I) in the Answer Card, and answer the regular questions (Section II) in the Answer Book.

I Multiple Choices (72 points)

- **3 points for each correct answer; -1 point for each incorrect answer; 0 points for no answer**
- **For each question, only choose (at most) one out of four given choices (A,B,C and D). If you choose more than one choice in one question, your answer will be incorrect and 1 point will be deduced.**

1. From a group of 3 first-year students, 4 sophomores, 4 juniors, and 3 seniors, a committee of size 4 is randomly selected. Find the probability that the committee will consist of (a) 1 from each class; (b) 2 sophomores and 2 juniors; (c) only sophomores or juniors.

   A. (a) 0.1439, (b) 0.0360, (c) 0.0699
   B. (a) 0.2510, (b) 0.0310, (c) 0.0952
   C. (a) 0.1539, (b) 0.4601, (c) 0.6905
   D. (a) 0.0438, (b) 0.3646, (c) 0.0799

Solution: (A)
(a) $\frac{(3\cdot4\cdot4\cdot3)}{\binom{14}{4}} = 0.1439$
(b) $\frac{\binom{4}{2}\binom{4}{2}}{\binom{14}{4}} = 0.0360$
(c) $\frac{\binom{8}{4}}{\binom{14}{4}} = 0.0699$

2. You ask your neighbor to water a sick plant while you are on vacation. Without water, it will die with probability 0.8; with water, it will die with probability 0.15. You are 90 percent certain that your neighbor will remember to water the plant. If the plant is dead upon your return, what is the probability that your neighbor forgot to water it?
   A. 0.372 B. 0.420 C. 0.101 D. 0.265

Solution: (A)
Let$A$ denote the event that the plant is alive and let W be the event that it was watered.

$P(A) = P(A \mid W)P(W) + P(A \mid W')P(W')$
$= (0.85)(0.9) + (0.2)(0.1) = 0.785$
$P(W' \mid A') = \frac{P(A'|W')P(W')}{P(A')} = \frac{(0.8)(0.1)}{1-0.785} = \frac{0.08}{0.215} = 0.372$

3. The random variable X has the probability density function

$$f(x) = \begin{cases} ax + bx^2, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases}$$

If $E(X) = 0.6$, find (a)$P(X < \frac{1}{2})$ and (b) $\text{Var}(X)$.

A. $P(X < \frac{1}{2}) = 0.4$, $\text{Var}(X) = 0.06$
B. $P(X < \frac{1}{2}) = 0.35$, $\text{Var}(X) = 0.06$
C. $P(X < \frac{1}{2}) = 0.35$, $\text{Var}(X) = 0.09$
D. $P(X < \frac{1}{2}) = 0.4$, $\text{Var}(X) = 0.09$

Solution: (B)
Since $1 = \int_0^1 (ax + bx^2)dx = \frac{a}{2} + \frac{b}{3}$
$0.6 = \int_0^1 (ax^2 + bx^3)dx = \frac{a}{3} + \frac{b}{4}$
we obtain $a = 3.6, b = -2.4$.Hence,
(a) $P(X < \frac{1}{2}) = \int_0^{1/2} (3.6x - 2.4x^2)dx = (1.8x^2 - 0.8x^3) \mid_0^{1/2} = 0.35$
(b) $E[X^2] = \int_0^1 (3.6x^3 - 2.4x^4)dx = 0.42$, so $\text{Var}(X) = 0.42 - 0.36 = 0.06$.

4. Suppose that the length of a phone call in minutes is an exponential random variable with parameter $\theta = 10$. If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait
(a) more than 10 minutes;
(b) between 10 and 20 minutes.

A. (a) e, (b) $e^2 - e$
B. (a) e^{-10}, (b) $e^{-10} - e^{-20}$
C. (a) e, (b) $e^{-1} - e^{-2}$
D. (a) e^{-1}, (b) $e^{-1} - e^{-2}$

Solution: (D)
Let X denote the length of the call made by the person in the booth. $F(x)$ is the cdf at x in the solution. Then the desired probabilities are
(a) $P(X > 10) = 1 - F(10) = 1 - \int_0^{10} \frac{1}{10} e^{-\frac{x}{10}} dx = 1 - (-e^{-\frac{x}{10}}) \,|_0^{10} = e^{-1}$.
(b) $P(10 < X < 20) = F(20) - F(10) = e^{-1} - e^{-2}$

5. The following gambling game, known as the wheel of fortune (or chuck-a-luck), is quite popular at many carnivals and gambling casinos: A player bets on one of the numbers 1 through 6. Three dice are then rolled, and if the number bet by the player appears i times, $i = 1, 2, 3$, then the player wins i units; if the number bet by the player does not appear on any of the dice, then the player loses 1 unit. Let X denote the player's winnings in the game. (Actually, the game is played by spinning a wheel that comes to rest on a slot labeled by three of the numbers 1 through 6, but this variant is mathematically equivalent to the dice version.) Consider the following five statements:
(i) The random variable X follows binomial distribution.
(ii) $E(X) = \frac{17}{216}$
(iii) $P(X = 2) = \frac{15}{216}$
(iv) $P(X = 3) = \frac{1}{216}$
(v) This is an unfair game for the player. That is, $E(X) < 0$.

A. Only (i), (ii), (iii) and (iv) are true.
B. Only (iii), (iv) and (v) are true.
C. Only (ii), (iii) and (iv) are true.
D. Only (i), (iv) and (v) are true.

Solution: (B)
If we assume that the dice are fair and act independently of one another, then the number of times that the number bet appears is a binomial Hence, letting X denote the player's winnings in the game, we have $P(X = -1) = \binom{3}{0}(\frac{1}{6})^0(\frac{5}{6})^3 = \frac{125}{216}$
$P(X = 1) = \binom{3}{1}(\frac{1}{6})^1(\frac{5}{6})^2 = \frac{75}{216}$
$P(X = 2) = \binom{3}{2}(\frac{1}{6})^2(\frac{5}{6})^1 = \frac{15}{216}$

$P(X=3)=\binom{3}{3}(\frac{1}{6})^3(\frac{5}{6})^0=\frac{1}{216}$
In order to determine whether or not this is a fair game for the player, let us calculate $E[X]$. From the preceding probabilities, we obtain
$E(X)=\frac{-125+75+30+3}{216}=\frac{-17}{216}$
Hence, in the long run, the player will lose 17 units per every 216 games he plays.

6. Find $P(X=4)$ if X has a Poisson distribution such that $3P(X=1)=P(X=2)$.

   A. 0.285 B. 0.313 C. 0.238 D. 0.134

Solution: (D)
$3\frac{\lambda^1 e^{-\lambda}}{1!}=\frac{\lambda^2 e^{-\lambda}}{2!}$
$e^{-\lambda}\lambda(\lambda-6)=0$
$\lambda=6$
Thus $P(X=4)=P(X\leq 4)-P(X\leq 3)=0.285-0.151=0.134$

7. Consider the moment generating function of X

$$M_X(t)=e^{3t+9t^2}.$$

   And Y has the moment generating function, for $t<3$,

$$M_Y(t)=\frac{3e^{2t}}{(3-t)}.$$

   Statement 1 : Y has $E[Y]$ of $\frac{7}{9}$
   Statement 2 : Y has $\mathrm{Var}(Y)$ of $\frac{1}{3}$
   Statement 3 : X follows normal distribution $N(3,9)$

   How many statements are correct?

   A. 0 B. 1 C. 2 D. 3

Solution: (A)
$E[Y] = \frac{\partial}{\partial t} M_Y(t) = \frac{21e^{2t} - 6te^{2t}}{(3-t)^2}|_{t=0} = \frac{7}{3}$ Take the second derivative to get $E[Y^2]$ and then $Var(Y) = \frac{1}{9}$.

The moment generating function of a normal distribution is $M(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$, so that $X \sim N(3, 18)$.

8. CUHKSZ students sometimes delay laundry for a few days.
A busy student must complete 3 problem sets before doing laundry. Each problem set requires 1 day with probability 2/3 and 2 days with probability 1/3. Let B be the number of days a busy student delays laundry. What is $E[B]$?

A. 3 B. 4 C. $\frac{3}{4}$ D. $\frac{4}{3}$

Solution: (B)
The expected time to complete a problem set is:

$1 \times \frac{2}{3} + 2 \times \frac{1}{3} = \frac{4}{3}$ Therefore, the expected time to complete all three problem sets is: $E[B] = E[pset1] + E[pset2] + E[pset3] = \frac{4}{3} + \frac{4}{3} + \frac{4}{3} = 4$

9. CUHKSZ students sometimes delay laundry for a few days.
A relaxed student rolls a fair, 6-sided die in the morning. If he rolls a 1, then he does his laundry immediately (with zero days of delay). Otherwise, he delays for one day and repeats the experiment the following morning. Let R be the number of days a relaxed student delays laundry. What is $E[R]$?

A. 5 B. 6 C. $\frac{5}{6}$ D. $\frac{6}{5}$

Solution: (A)
If we regard doing laundry as a failure, then the mean time to failure is $1/(1/6) = 6$. However, this counts the day laundry is done, so the number of days delay is 6-1 = 5. Alternatively, we could derive the answer as follows:

$E[R] = \sum_{k=0}^{\infty} P(R > k) = \frac{5}{6} + (\frac{5}{6})^2 + (\frac{5}{6})^3 + ... = \frac{5}{6} \times (1 + \frac{5}{6} + (\frac{5}{6})^2 + ...) = \frac{5}{6} \times \frac{1}{1-\frac{5}{6}} = 5$

10. Suppose that Pandora Restaurant everyday (24 hours) receives 720 complaints on average. It is assumed that the number of complaints received follow an approximate Poisson process. Suppose that X (in month) is the time it takes for Pandora to reply to a complaint, and X follows a gamma distribution with mean $\frac{3}{2}$ months and standard deviation $\sqrt{\frac{3}{4}}$ months.

    Statement 1 : The probability that Pandora will have to wait longer than 21.06 minutes for the first 7th complain is 0.05.
    Statement 2 : X follows the gamma distribution with $\alpha = 4$.
    Statement 3 : Let $Y = 3X$. Then Y has a gamma distribution with the moment generating function $(\frac{\frac{2}{3}}{\frac{2}{3}-t})^3$

    Which statement is true?

    A. Statements 1 and 3 only
    B. Statements 2 and 3 only
    C. Statements 3 only
    D. None of them

Solution: (C)
1: The mean rate of complains per minute is $\lambda = \frac{1}{2}$. Thus $\theta = 2$ and $\alpha = \frac{r}{2} = 7$. If Z denotes the waiting time until the 7th complain, then Z is $\chi^2(14)$. $P(Z > 21.06) = 0.1$.
2: $\mu = \alpha\theta = \frac{3}{2}$, $\sigma^2 = \alpha\theta^2 = \frac{3}{4}$. $\alpha = 3$.
3: $M_X(t) = (\frac{2}{2-t})^3$, $M_Y(t) = E[e^{tY}] = E[^{t3X}] = M_X(3t) = (\frac{2}{2-3t})^3$

11. Consider these 3 statements. We denote A' and B' to be the complement of the set A and B respectively.

Statement 1: If events A and B are mutually exclusive and exhaustive, A' and B' are mutually exclusive.
Statement 2: If events A and B are mutually exclusive but not exhaustive, A' and B' are exhaustive.
Statement 3: If events A and B are exhaustive but not mutually exclusive, A' and B' are exhaustive.

Choose the correct option.

A. Statement1–False, Statement2–True, Statement3–True
B. Statement1–True, Statement2–True, Statement3–False
C. Statement1–True, Statement2–False, Statement3–False
D. Statement1–False, Statement2–True, Statement3–False

Solution: (B)
S1: $A' \cap B' = (A \cup B)' = S' = \emptyset$. Thus the events A' and B' are mutually exclusive.
S2: Let $C = (A' \cup B')'$, that is the part that is not contained in $A' \cup B'$. Using De Morgan's Law $C = A \cap B = \emptyset$. Thus, there is nothing that is not a part of A' or B'. Hence, A' and B' are mutually exhaustive.
S3: As in previous part, let $C = (A' \cup B')' = A \cap B$ which is not null. Thus, A' and B' are not mutually exhaustive.

12. Suppose we roll two fair six-sided dice, one red and one blue. Let A be the event that the two dice show the same value. Let B be the event that the sum of the two dice is equal to 12. Let C be the event that the red die shows 4. Let D be the event that the blue die shows 4. Consider the following 5 statements.

    1. A and B are independent.
    2. A and C are independent.
    3. A and D are independent.
    4. C and D are independent.
    5. A, C and D are mutually independent.

    Which statement is true?

A. 2 only
B. 2 and 3 only
C. 2 and 3 and 4 only
D. All of them

Solution: (C)

13. Consider following 4 statements:

(i) If X follows a Gamma distribution with parameter $\alpha = 2$ and $\theta = 1$, then $E[X^4] = 120$
(ii) Consider the Gamma function with a positive integer α, then we could write it as $\Gamma(\alpha) = (\alpha - 1)!$
(iii) If $X \sim N(3, 1)$, then $E[X^3]$=27
(iv) If X follows a Chi-square distribution with 1 degree of freedom, then $P(X \geq 2.706) = 0.9$

A. All statements are ture.
B. Only (i), (ii) and (iii) are ture.
C. Only (iii) and (iv) are ture.
D. Only (i) and (ii) are ture.

Solution: (D)
(i). Using the moment generating function of gamma distribution, we have $E[X^4]$=$\alpha(\alpha+1)(\alpha+2)(\alpha+3)\theta^4 = 120$.
(ii). This a property of Gamma function and it can be proved by using integration by part.
(iii). Let $X = Z + 3$, where $Z \sim N(0, 1)$. Notice that $E[Z] = 0$, $E[Z^2] = 1$ and $E[Z^3] = 0$, we have

$$\mathbb{E}[X^3] = \mathbb{E}[Z^3 + 9Z^2 + 27Z + 27] = 36$$

(iv). This can be checked through the table. The correct one is $P(X \leq 2.706) = 0.9$
Therefore, only (i) and (ii) are true.

14. The weekly amount of downtime X (in hours) for a certain industrial machine has following probability density function

$$f(x) = \begin{cases} \frac{1}{16}x^2e^{-x/2}, & x \geq 0, \\ 0, & \text{otherwise} \end{cases}$$

Consider the following statements:

(i) The random variable X follows a exponential distribution.
(ii) $E(X) = 1/2$, Var(X)=1/4
(iii) The 10th percentile of the random variable X is 2.204.
(iv) The inequality $P(X \geq 2.204) < \frac{E(X)}{2.204}$ holds.

A. Only (i) and (ii) are true
B. Only (i) and (iii) are true
C. Only (ii) and (iv) are true
D. Only (iii) and (iv) are true

Solution: (D)
From the expression of probability density function we find that X actually follows a Gamma distribution with parameters $\alpha = 3$, $\theta = 2$, and equivalently X follows a Chi-square distribution with degree of freedom $r = 6$ (since $\theta = 2$ and $\alpha = 3$). Therefore,
(i) is wrong since $\alpha \neq 1$.
(ii) is wrong, since $E(X) = \alpha\theta = 6$, $Var(X) = \alpha\theta^2 = 12$.
(iii) is correct. The 10th percentile for the random variable X is a number such that $P\left(X \leq \chi^2_{0.9}(6)\right) = 0.1$, where $\alpha = 0.1$ in this case. The value of $\chi^2_{0.9}(6)$ could be obtained by checking the table of Chi-square distribution.
(iv) is correct. We find that $P(X \geq 2.204) = 1 - 0.1 = 0.9$ follows from (iii), and since $E(X) = 6$, the inequality holds(the inequality is known as the Markov inequality but we actually do not need to know it to find the answer).

15. Given a random variable X with following probability density function

$$f(x) = \begin{cases} 4x^3, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases}$$

What's the median of this distribution?

A. 2^{-1} B. 2^{-2} C. $2^{-1/2}$ D. $2^{-1/4}$

Solution: (D)
Let m be the value of median. We have $\int_0^m 4x^3\ dx = 1/2$, and yields that $m = 2^{-1/4}$.

16. Suppose X follows a uniform distribution on the interval $[-2, 1]$, that is, $X \sim U[-2, 1]$. Let $Y = X^2$. Which one of the following is the cumulative distribution function of Y?

A. $F_Y(y) = \begin{cases} 2\sqrt{y}/3, & 0 \le y \le 1, \\ \left(1+\sqrt{y}\right)/3, & 1 < y \le 4, \\ 0, & y < 0, \\ 1, & y > 4 \end{cases}$

B. $F_Y(y) = \begin{cases} \left(1+\sqrt{y}\right)/3, & 0 \le y \le 1 \\ \sqrt{y}/2, & 1 < y \le 4 \\ 0, & y < 0, \\ 1, & y > 4 \end{cases}$

C. $F_Y(y) = \begin{cases} (1+\sqrt{y})/3, & 0 \le y \le 4 \\ 0, & y < 0, \\ 1, & y > 4 \end{cases}$

D. $F_Y(y) = \begin{cases} 2\sqrt{y}/3, & 0 \le y \le 4 \\ 0, & y < 0, \\ 1, & y > 4 \end{cases}$

Solution: (A)
The cdf of X is given by $F_X(x) = (x+2)/3$, where $-2 \le x \le 1$
Consider the cdf of Y for $0 \le y \le 1$ (therefore $-1 \le x \le 1$), we have

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(X^2 \le y) \\ &= P(-\sqrt{y} \le X \le \sqrt{y}) \end{aligned}$$

$$
\begin{aligned}
&= P(X \le \sqrt{y}) - P(X \le -\sqrt{y}) \\
&= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\
&= 2\sqrt{y}/3
\end{aligned}
$$

Then, we consider the case $1 < y \le 4$, therefore $-2 \le x \le -1$ and we have

$$
\begin{aligned}
F_Y(y) = P(Y \le y) &= P(Y < 1) + P(1 \le Y \le y) \\
&= \frac{2}{3} + P(1 \le X^2 \le y) \\
&= \frac{2}{3} + P(-\sqrt{y} \le X \le -1) \\
&= \frac{2}{3} + P(X \le -1) - P(X \le -\sqrt{y}) \\
&= \frac{2}{3} + F_X(-1) - F_X(-\sqrt{y}) \\
&= \left(1 + \sqrt{y}\right)/3
\end{aligned}
$$

Combine them we have

$$
F_Y(y) = \begin{cases} 2\sqrt{y}/3, & 0 \le y \le 1, \\ \left(1 + \sqrt{y}\right)/3, & 1 < y \le 4, \\ 0, & y < 0, \\ 1, & y > 4 \end{cases}
$$

17. Suppose there are two well defined events, event A and event B. The probability that only one of them occurs is 0.3, and $P(A)+P(B) = 0.5$. What's the probability that at least one of them not occur?
A. 0.3 B. 0.5 C. 0.6 D. 0.9

Solution: (D)
"Only one of them occurs" means, $P\left((A \cap B') \cup (A' \cap B)\right) = 0.3$. That is,

$$
\begin{aligned}
0.3 &= P\left((A \cap B') \cup (A' \cap B)\right) \\
&= P(A) - P(A \cap B) + P(B) - P(A \cap B) \\
&= 0.5 - 2P(A \cap B)
\end{aligned}
$$

Therefore,

$$P(A \cap B) = 0.1$$

$$P(A' \cup B') = 1 - P(A \cap B) = 0.9.$$

18. Let X be a random variable with following probability mass function

$$f(x) = \begin{cases} \frac{c}{x!}, & \text{for } x=0,1,2,\ldots, \\ 0, & \text{otherwise} \end{cases}$$

where c is a constant.
Determine the value of c and compute the expectation of X.
A. $c = e^{-2},\ E[X] = e^{-1}$
B. $c = e^{-1},\ E[X] = 1$
C. $c = 1,\ E[X] = e$
D. $c = e,\ E[X] = e^2$

Solution: (B)
By the power series of e and the properties of probability mass function, we have

$$\sum_{x=0}^{\infty} \frac{c}{x!} = c(\sum_{x=0}^{\infty} \frac{1}{x!}) = ce = 1.$$

Therefore $c = 1/e$, and $E[X]$ is given by

$$E[X] = \sum_{x=0}^{\infty} x\frac{c}{x!} = c\sum_{x=1}^{\infty} \frac{1}{(x-1)!} = ce = 1.$$

19. Suppose $P(A|B) = P(B|A) = \frac{1}{4}, P(A') = \frac{2}{3}$, Which one of the following claims is true about these statements?
A. A and B are independent, and $P(A \cap B) = \frac{5}{12}$.
B. A and B are independent, and $P(A) = P(B)$.

C. A and B are dependent, and $P(A \cup B) = \frac{7}{12}$.
D. A and B are dependent, and $P(A|B') = P(A|B)$

Solution: (C)
Since

$$P(A|B) = P(B|A) = \frac{1}{4}$$

That is,

$$\frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A)} = \frac{1}{4}.$$

Therefore

$$P(A \cap B) \neq P(A)P(B)$$

Then,

$$P(A) = 1 - P(A') = \frac{1}{3}, P(B) = \frac{1}{3}, P(A \cap B) = \frac{1}{4}P(A) = \frac{1}{12}$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{3} + \frac{1}{3} - \frac{1}{12} = \frac{7}{12}.$$

20. Products produced by a machine has a 3% defective rate. The first 10 inspections have been found to be free of defectives. What is the probability that the first defective will occur on the 15th inspection?

A. 0.97^4
B. $0.97^5 \times 0.03$
C. $1 - 0.97^4 \times 0.03$
D. $0.97^4 \times 0.03$

Solution: (D)
Let X be the number of products need to be checked to detect the first defective product. X has geometric distribution with $p = 0.03, q = 0.97$.

$$P(X = 15|X > 10) = \frac{P(X = 15 \cap X > 10)}{P(X > 10)}$$

$$= \frac{P(X=15)}{P(X>10)}$$
$$= \frac{q^{14}p}{\sum_{k=11}^{\infty} q^{k-1}p} = \frac{q^{14}p}{q^{10}} = q^4 p$$

Since $q = 0.97$. Therefore it should be D.

21. Suppose that $X \sim N(3, 2^2)$, find the largest d such that $P(X > d) \geq 0.9$.
A. 5.56 B. 4.68 C. 2.08 D. 0.44

Solution: (D)

$$P(X > d) = 1 - P(X \leq d) = 1 - P(\frac{X-3}{2} \leq \frac{d-3}{2})$$
$$= 1 - \Phi(\frac{d-3}{2}) \geq 0.9$$
$$\Phi(\frac{d-3}{2}) \leq 0.1$$

Check that table we know that $\Phi(1.28) = 0.9$.And $\Phi(-1.28) = 1 - \Phi(1.28) = 0.1$ Therefore $\frac{d-3}{2} = -1.28$, that is $d \leq 0.44$.

22. Suppose that the moment-generating function $M_X(t)$ of the continuous random variable X has the property $M_X(t) = e^t M_X(-t)$ for all t. What is $E(X)$?
A. $\frac{1}{4}$ B. $\frac{1}{2}$ C. 2 D. 4

Solution: (B)
Since $M_X(t) = e^t M_X(-t)$ can be written as $E(e^{tX}) = e^t E(e^{-tX}) = E(e^{t(1-X)})$ and thus $M_X(t) = M_{1-X}(t)$ for all t. Since the moment-generating function determines uniquely the probability distribution, it follows that the random variable X has the same distribu-

tion as the random variable $1 - X$. Hence $E(X) = E(1 - X)$ and so $E(X) = \frac{1}{2}$.

23. Let X be lifetime (measured in hours) of a cetain type of electronic device, and its probabiltiy density function given by

$$f(x) = \begin{cases} \frac{10}{x^2}, & x > 10 \\ 0, & x \leq 10 \end{cases}$$

What is the probability that at least 2 of 4 such types of devices will function for at least 15 hours?
A. $\frac{57}{81}$ B. $\frac{59}{81}$ C. $\frac{69}{81}$ D. $\frac{72}{81}$

Solution: (D)
Let X be the lifetime (measured in hours) of a certain type of device. Then,

$$\begin{aligned} P(X \geq 15) &= \int_{15}^{\infty} \frac{10}{x^2} dx = [-\frac{10}{x}]_{15}^{\infty} \\ &= \lim_{x \to \infty} -\frac{10}{x} - (-\frac{10}{15}) \\ &= -\lim_{x \to \infty} \frac{10}{x} + \frac{2}{3} = \frac{2}{3} \end{aligned}$$

Let Y be the number of devices that function at least 15 hours. Then,

$$P(X \geq 2) = 1 - \binom{4}{0}(\frac{2}{3})^0(\frac{1}{3})^4 - \binom{4}{1}(\frac{2}{3})^1(\frac{1}{3})^3 = \frac{72}{81}$$

24. Consider 3 urns. Urn A contains 2 white and 4 red balls; urn B contains 8 white and 4 red balls; and urn C contains 1 white and 3 red balls. If 1 ball is selected from each urn, what is the probability that the ball chosen from urn A was white, given that exactly 2 white balls were selected?
A. $\frac{6}{11}$ B. $\frac{9}{11}$ C. $\frac{8}{11}$ D. $\frac{7}{11}$

Solution: (D)
The probability in the question should be

$$P(\text{Ball from A white}|\text{2 white balls selected})$$
$$= \frac{P(\text{Ball from A white} \cap \text{2 white balls selected})}{P(\text{2 white balls selected})}$$

First, consider $P(\text{Ball from } A \text{ white} \cap \text{2 white balls selected})$. This means that the ball chosen from A must be white and either the ball from B or C is white and the other one is not. The probability of drawing a white ball from A is $\frac{2}{6}$. Likewise, the probability of drawing a white ball from B, C is $\frac{8}{12}$ and $\frac{1}{4}$ respectively. The probability of not drawing the white ball from B, C is $\frac{4}{12}$ and $\frac{3}{4}$ respectively. So,

$$P(\text{Ball from A white} \cap \text{2 white balls selected}) = \frac{2}{6}(\frac{8}{12} \cdot \frac{3}{4} + \frac{1}{4} \cdot \frac{4}{12})$$

Now consider $P(\text{2 white balls selected})$. There are 3 ways we can choose the two white balls: choose a white ball from A and B and a red ball from C, choose a white ball from B and C and a red ball from A and choose a white ball from A and C and a red ball from B. So the denominator is

$$\frac{2}{6} \cdot \frac{8}{12} \cdot \frac{3}{4} + \frac{4}{6} \cdot \frac{8}{12} \cdot \frac{1}{4} + \frac{2}{6} \cdot \frac{4}{12} \cdot \frac{1}{4}$$

Therefore the probability is $\frac{7}{11}$.

II Regular Questions (28 points)

25. (a) A random variable X with parameter $\theta \in \mathbb{R}$ belongs to the exponential family if its probability mass function or probability density function can be written as

$$f(x) = h(x)e^{\theta x - A(\theta)},$$

where $h(x)$ and $A(\theta)$ are some known functions. Note that $h(x)$ does not depend on θ and $A(\theta)$ does not depend on x.

(a1) (2 points) Show that a Poisson distribution with mean $\lambda > 0$

belongs to the exponential family by identifying appropriate $h(x)$, θ and $A(\theta)$.

(a2) (3 points) Assume that X is a continuous random variable that belongs to the exponential family. Show that the moment generating function $M(t)$ of X can be written as

$$M(t) = E(e^{tX}) = e^{A(\theta+t)-A(\theta)}.$$

(a3) (3 points) Continuing part (a2), and assume that $A(\theta)$ is twice differentiable so that its first and second derivative with respect to θ exist. Show that the mean and variance of X are

$$E(X) = \frac{d}{d\theta}A(\theta), \quad \text{Var}(X) = \frac{d^2}{d\theta^2}A(\theta).$$

(b) Let n be a fixed positive integer, and X be a U(0, n) random variable, that is, a continuous uniform distribution on the interval $(0, n)$. Define $Y = \lfloor X \rfloor$, where for a real number x, $\lfloor x \rfloor$ is x rounded down to the nearest integer (e.g., $\lfloor 5.7 \rfloor = 5$).

(b1) (3 points) Find $f(y)$, the probability mass function (pmf) of Y.

(b2) (2 points) Find $F(y)$, the cumulative distribution function (cdf) of Y.

(b3) (1 points) Find $E(Y)$, the mean of Y.

Solution: Part (a1). The probability mass function of Poisson random variable with mean λ is given by

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{1}{x!}e^{(\ln\lambda)x}e^{-\lambda} = h(x)e^{\theta x - A(\theta)},$$

where we take

$$\begin{aligned} h(x) &= \frac{1}{x!} \\ \theta &= \ln\lambda \\ A(\theta) &= \lambda. \end{aligned}$$

Part (a2).

$$\begin{aligned} E(e^{tX}) &= \int e^{tx} h(x) e^{\theta x - A(\theta)}\, dx \\ &= e^{-A(\theta)} \int h(x) e^{(\theta+t)x}\, dx \\ &= e^{A(\theta+t)-A(\theta)} \int h(x) e^{(\theta+t)x - A(\theta+t)}\, dx \\ &= e^{A(\theta+t)-A(\theta)}, \end{aligned}$$

where in the last equality we use $\int h(x) e^{(\theta+t)x - A(\theta+t)}\, dx = 1$.

Part (a3).

$$\begin{aligned} \frac{d}{dt} E(e^{tX}) &= e^{A(\theta+t)-A(\theta)} \frac{d}{dt} A(\theta+t) \\ \frac{d^2}{dt^2} E(e^{tX}) &= e^{A(\theta+t)-A(\theta)} \frac{d^2}{dt^2} A(\theta+t) + e^{A(\theta+t)-A(\theta)} \left(\frac{d}{dt} A(\theta+t) \right)^2. \end{aligned}$$

As a result, by taking $t = 0$ above, we have

$$\begin{aligned} E(X) &= \frac{d}{dt} E(e^{tX}) \bigg|_{t=0} = \frac{d}{d\theta} A(\theta), \\ E(X^2) &= \frac{d^2}{dt^2} E(e^{tX}) \bigg|_{t=0} = \frac{d^2}{d\theta^2} A(\theta) + \left(\frac{d}{d\theta} A(\theta) \right)^2, \\ \mathrm{Var}(X) &= E(X^2) - E(X)^2 = \frac{d^2}{d\theta^2} A(\theta). \end{aligned}$$

Part (b1). Y can take values in the set $\bar{S} = \{0, 1, 2, \ldots, n-1\}$. For $y \in \{0, 1, 2, \ldots, n-1\}$,

$$f(y) = P(Y = y) = P(y \le X < y + 1) = \int_y^{y+1} \frac{1}{n}\, dx = \frac{1}{n}.$$

As a result, we have

$$f(y) = \begin{cases} \dfrac{1}{n}, & y \in \{0, 1, 2, \ldots, n-1\}, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, Y has a discrete uniform distribution on $\bar{S}$.

Part (b2).

$$F(y) = \begin{cases} 0, & y < 0, \\ \frac{1}{n}, & 0 \le y < 1, \\ \frac{2}{n}, & 1 \le y < 2, \\ \vdots & \vdots \\ \frac{n-1}{n}, & n-2 \le y < n-1, \\ 1, & n-1 \le y. \end{cases}$$

Part (b3).

$$E(Y) = \frac{1}{n}\sum_{y=1}^{n-1} y = \frac{n-1}{2}.$$

26. Consider a normal random variable $X \sim N(\mu, \sigma^2)$ with $\sigma > 0$.

(a1) (4 points) Show that for any $a < b$,

$$P(-b \le X \le -a) = P(a + 2\mu \le X \le b + 2\mu).$$

(a2) (8 points) **By using the moment generating function technique**, show that $(X^2 - 2\mu X + \mu^2)/\sigma^2$ has a Chi-square distribution with degrees of freedom 1, i.e.,

$$(X^2 - 2\mu X + \mu^2)/\sigma^2 \sim \chi^2(1)$$

(a3) (2 points) What is $E\left(\frac{X^2 - 2\mu X + \mu^2}{\sigma^2}\right)^2$?

Solution: Part (a). There are at least two possible solutions. One is based on the property of the cdf of $N(0,1)$, i.e., $\Phi(x)$ that $\Phi(-x) = 1 - \Phi(x)$. The other one is based on the coordinate change.

The first solution is

$$\begin{aligned}
P(-b \le X \le -a) &= P\left(\frac{-b-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{-a-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{-a-\mu}{\sigma}\right) - \Phi\left(\frac{-b-\mu}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{a+\mu}{\sigma}\right) - 1 + \Phi\left(\frac{b+\mu}{\sigma}\right) \\
&= \Phi\left(\frac{b+\mu}{\sigma}\right) - \Phi\left(\frac{a+\mu}{\sigma}\right) \\
&= P\left(\frac{a+\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b+\mu}{\sigma}\right) \\
&= P(a+2\mu \le X \le b+2\mu)
\end{aligned}$$

Let $f(x)$ be the pdf of X. Then the second solution is simply to show that

$$\int_{-b}^{-a} f(x)dx = \int_{a+2\mu}^{b+2\mu} f(z)dz.$$

This is true by noting the pdf of X and taking the coordinate change

$$z = -x + 2\mu.$$

Part (b). The question is equivalent to show by moment generating function technique that if $Y \sim N(0,1)$, $Y^2 \sim \chi^2(1)$. By the definition of moment generating function, we have

$$M(t) = Ee^{tY^2} = \int_{-\infty}^{\infty} e^{ty^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{(t-\frac{1}{2})y^2} dy$$

For $t < 1/2$, we further have

$$\begin{aligned}
M(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)y^2} dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y}{\frac{1}{(1-2t)^{\frac{1}{2}}}}\right)^2} dy \\
&= \frac{1}{(1-2t)^{\frac{1}{2}}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\frac{1}{(1-2t)^{\frac{1}{2}}}} e^{-\frac{1}{2}\left(\frac{y}{\frac{1}{(1-2t)^{\frac{1}{2}}}}\right)^2} dy
\end{aligned}$$

Since

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\frac{1}{(1-2t)^{\frac{1}{2}}}} e^{-\frac{1}{2}\left(\frac{\frac{y}{1}}{\frac{1}{(1-2t)^{\frac{1}{2}}}}\right)^2} dy = 1$$

we have

$$M(t) = \frac{1}{(1-2t)^{\frac{1}{2}}}, \quad t < \frac{1}{2},$$

which is the moment generating function of $\chi^2(1)$. This completes the proof.

Part (c). Let $Z \sim \chi^2(1)$. Then the question is to calculate $E(Z^2)$. Since

$$E(Z) = 1, Var(Z) = 2,$$

$$E(Z^2) = Var(Z) + (E(Z))^2 = 2 + 1 = 3.$$

The same result can be obtained by calculating $M''(0) = 3$.

Probability and Statistics I

Final Examination
SSE, CUHK(SZ)

May 15, 2017

Name: _____ Student ID: _____

Answer all the questions in the Answer Book.
This page has no questions.

1. (8 points) Let Y_1, Y_2, Y_3 be independent random variables which have the Bernoulli distribution with the probability of success p.
   (a) (1 point) Define a new random variable $Z = Y_1 + Y_2 + Y_3$. Find the distribution of Z. Provide the details of the derivation.
   (b) (7 points) For $k = 1, 2$, let

$$X_k = \begin{cases} 1, & Y_1 + Y_2 + Y_3 = k, \\ -1 & Y_1 + Y_2 + Y_3 \neq k. \end{cases}$$

   i. (2 points) Find the joint pmf of X_1, X_2.
   ii. (2 points) Find the marginal pmfs of X_1 and X_2, respectively.
   iii. (2 points) Find the value of the success probability p that minimizes $\mathrm{E}(X_1 X_2)$.
   iv. (1 point) Compute $\mathrm{Cov}(X_1 - X_2, X_2)$.

Solution:

(a)

$$\mathrm{E}(e^{tZ}) = \mathrm{E}(e^{t(Y_1+Y_2+Y_3)}) = \mathrm{E}(e^{tY_1})\mathrm{E}(e^{tY_2})\mathrm{E}(e^{tY_3}) = (1 - p + pe^t)^3$$

Hence Z has a binomial distribution $Z = Y_1 + Y_2 + Y_3 \sim b(3, p)$

(b) i. Let $f(X_1, X_2)$ be the joint pmf of X_1 and X_2,then we have

$$\begin{aligned} f(-1,-1) &= P\{X_1 = -1, X_2 = -1\} \\ &= P\{Y_1 + Y_2 + Y_3 \neq 1, Y_1 + Y_2 + Y_3 \neq 2\} \\ &= P\{Y_1 + Y_2 + Y_3 = 0\} + P\{Y_1 + Y_2 + Y_3 = 3\} \\ &= (1-p)^3 + p^3 \end{aligned}$$

$$\begin{aligned} f(-1,1) &= P\{X_1 = -1, X_2 = 1\} \\ &= P\{Y_1 + Y_2 + Y_3 \neq 1, Y_1 + Y_2 + Y_3 = 2\} \\ &= \binom{3}{2} p^2(1-p) \\ &= 3p^2(1-p) \end{aligned}$$

$$\begin{aligned} f(1,-1) &= P\{X_1 = 1, X_2 = -1\} \\ &= P\{Y_1 + Y_2 + Y_3 = 1, Y_1 + Y_2 + Y_3 \neq 2\} \\ &= P\{Y_1 + Y_2 + Y_3 = 1\} \\ &= \binom{3}{1} p(1-p)^2 = 3p(1-p)^2 \end{aligned}$$

$$\begin{aligned} f(1,1) &= P\{X_1 = 1, X_2 = 1\} \\ &= P\{Y_1 + Y_2 + Y_3 = 1, Y_1 + Y_2 + Y_3 = 2\} = 0 \end{aligned}$$

ii. Then we could get the marginal pmf of X_k as follows:

$$f_{X1}(x) = \begin{cases} 1 - 3p + 6p^2 - 3p^3, & X_1 = -1, \\ 3p(1-p)^2, & X_1 = 1. \end{cases}$$

$$f_{X2}(x) = \begin{cases} 1 - 3p^2 + 3p^3, & X_2 = -1, \\ 3p^2(1-p), & X_2 = 1. \end{cases}$$

iii.

$$\begin{aligned} &\mathrm{E}(X_1X_2) = 1 \times P\{X_1 = -1, X_2 = -1\} + (-1) \times P\{X_1 = -1, X_2 = 1\} \\ &+ (-1) \times P\{X_1 = 1, X_2 = -1\} + 1 \times P\{X_1 = 1, X_2 = 1\} = 1 - 6p + 6p^2 \end{aligned}$$

When $p = \frac{1}{2}$, $\mathrm{E}(X_1X_2)$ gets the minimum value $-\frac{1}{2}$.

iv.

$$\begin{aligned} \mathrm{Cov}(X_1 - X_2, X_2) &= \mathrm{Cov}(X_1, X_2) - \mathrm{Cov}(X_2, X_2) \\ &= \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\mathrm{E}(X_2) - \mathrm{Var}(X_2) \\ &= -12p^2 - 24p^3 + 144p^4 - 180p^5 + 72p^6 \end{aligned}$$

2. (8 points) Let T have a student's t distribution with r degrees of freedom.

$$T = \frac{Z}{\sqrt{\frac{U}{r}}}$$

where Z has a standard normal distribution, that is $Z \sim N(0,1)$, and U has a chi-square distribution with degrees of freedom r, that is $U \sim \chi^2(r)$, and Z and U are independent.

(a) (2 points) Find $\mathrm{E}(Z)$ and $\mathrm{E}(Z^2)$.

(b) (4 points) Find $\mathrm{E}(\frac{1}{\sqrt{U}})$ and $\mathrm{E}(\frac{1}{U})$.

(c) (1 point) Show that $\mathrm{E}(T) = 0$ provided that $r \geq 2$.

(d) (1 point) Show that $\mathrm{Var}(T) = \frac{r}{r-2}$ provided that $r \geq 3$.

Solution: Please refer to Exercise 5.5-14.

(a) Because Z is $N(0,1)$, $\mathrm{E}(Z) = 0$ and $\mathrm{E}(Z^2) = 1$.

(b) Since $U \sim \chi^2(r)$ so it follows that

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{\sqrt{U}}\right] &= \int_0^\infty \frac{1}{\sqrt{u}} \frac{1}{\gamma(r/2)2^{r/2}} u^{r/2-1} e^{-u/2} du \\
&= \frac{\gamma[(r-1)/2]}{\sqrt{2}\gamma(r/2)} \int_0^\infty \frac{1}{\gamma[(r-1)/2]2^{(r-1)/2}} u^{(r-1)/2-1} e^{-u/2} du \\
&= \frac{\gamma[(r-1)/2]}{\sqrt{2}\gamma(r/2)}.
\end{aligned}$$

Note that the last integral is equal to one because the integrand is the pdf of a $\chi^2(r-1)$ random variable.

To find $E[\frac{1}{U}]$ we have

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{U}\right] &= \int_0^\infty \frac{1}{u} \frac{1}{\gamma(r/2)2^{r/2}} u^{r/2-1} e^{-u/2} du \\
&= \frac{\gamma[(r-2/2)]}{2\gamma(r/2)} \int_0^\infty \frac{1}{\gamma[(r-2)/2]2^{(r-2)/2}} u^{(r-2)/2-1} e^{-u/2} du \\
&= \frac{\gamma(r/2-1)}{2\gamma(r/2-1)(r/2-1)} = \frac{1}{r-2}.
\end{aligned}$$

Note that the last integral is equal to one because the integrand is the pdf of a $\chi^2(r-2)$ random variable.

(c)

$$\mathrm{E}[T] = \mathrm{E}\left[\frac{Z}{\sqrt{U/r}}\right] = \mathrm{E}(Z)\mathrm{E}\left[\frac{1}{\sqrt{U/r}}\right] = 0[\frac{\sqrt{r}\gamma[(r-1)/2]}{\sqrt{2}\gamma(r/2)}] = 0,$$

provided $r \geq 2$;

(d)

$$\mathrm{Var}(T) = \mathrm{E}(T^2) - 0^2 = \mathrm{E}[Z^2]\mathrm{E}[r/U] = \frac{r}{r-2},$$

provided $r \geq 3$.

3. (9 points) Let X and Y be two random variables with the joint pdf

$$f(x,y) = \frac{1}{8}, \quad 0 \leq y \leq 4,\ y \leq x \leq y+2.$$

(a) (2 points) Find $f_X(x)$, the marginal pdf of X.

(b) (1 point) Find $f_Y(y)$, the marginal pdf of Y.

(c) (2 points) Determine $h(y|x)$, the conditional pdf of Y, given that $X = x$.

(d) (1 point) Determine $g(x|y)$, the conditional pdf of X, given that $Y = y$.

(e) (2 points) Compute $\mathrm{E}(Y|x)$, the conditional mean of Y, given that $X = x$.

(f) (1 point) Compute $\mathrm{E}(X|y)$, the conditional mean of X, given that $Y = y$.

Solution: Please refer to Exercise 4.4-18.

(a)

$$f_X(x) = \begin{cases} \int_0^x \frac{1}{8} dy = \frac{x}{8}, & 0 \le x \le 2, \\ \int_{x-2}^x \frac{1}{8} dy = \frac{1}{4}, & 2 < x < 4, \\ \int_{x-2}^4 \frac{1}{8} dy = \frac{6-x}{8}, & 4 \le x \le 6. \end{cases}$$

(b)

$$f_Y(y) = \int_y^{y+2} \frac{1}{8} dx = \frac{1}{4}, \quad 0 \le y \le 4.$$

(c)

$$h(y|x) = \begin{cases} \frac{1}{x}, & 0 \le y \le x, \quad 0 \le x \le 2, \\ \frac{1}{2}, & x-2 < y < x, \quad 2 < x < 4, \\ \frac{1}{(6-x)}, & x-2 \le y \le 4, \quad 4 \le x \le 6. \end{cases}$$

(d)

$$g(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{1}{2}, \quad y \le x \le y+2, \quad 0 \le y \le 4.$$

(e)

$$E(Y|x) = \begin{cases} \int_0^x y \frac{1}{x} dy = \frac{x}{2}, & 0 \le x \le 2, \\ \int_{x-2}^x y \frac{1}{2} dy = x - 1, & 2 < x < 4, \\ \int_{x-2}^4 y \frac{1}{6-x} dy = \frac{x+2}{2}, & 4 \le x \le 6. \end{cases}$$

(f)

$$E(X|y) = \int_y^{y+2} x \frac{1}{2} dx = y + 1, \quad 0 \le y \le 4.$$

4. (10 points) Let X and Y have a bivariate normal distribution with $\mu_X = -3, \mu_Y = 10, \sigma_X^2 = 25, \sigma_Y^2 = 9$, and $\rho = 3/5$. Answer the following questions (if necessary, make use of the standard normal table shown in Figure 1).

(a) (1 point) Compute $P(-4 < X < 4)$.

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

Figure 1: **The standard normal table**. The entries in this table provide the numerical values of $\Phi(x) = P(X \leq x)$, where X is a standard normal random variable, for x between 0 and 1.49. For example, to find $\Phi(0.72)$, we look at the row corresponding to 0.7 and the column corresponding to 0.02 so that $\Phi(0.72) = 0.7642$. When x is negative, the value of $\Phi(x)$ can be found using the formula $\Phi(x) = 1 - \Phi(-x)$.

(b) (3 points) Compute $P(-4 < X < 4|Y = 13)$.

(c) (3 points) Determine $\mathrm{E}(X^2|Y = 13)$ and $\mathrm{E}[X(X-1)|Y = 13]$.

(d) (3 points) Determine $\mathrm{E}(X+Y)$, $\mathrm{Var}(X+Y)$ and $\mathrm{E}(XY)$.

Solution: Please refer to Exercise 4.5-1.

(a)

$$P(-4 < X < 4) = \Phi(\frac{4-(-3)}{\sqrt{25}}) - \Phi(\frac{-4-(-3)}{\sqrt{25}})$$
$$= \Phi(1.4) - \Phi(-0.2) = \Phi(1.4) - (1 - \Phi(0.2)) = 0.4985$$

(b)

$$\mu_{X|Y=13} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(13 - \mu_Y) = -3 + (3/5)(5/3)(13-10) = 0.$$

$$\sigma^2_{X|Y=13} = \sigma^2_X(1-\rho^2) = 25(1-(3/5)^2) = 16.$$

$$P(-4 < X < 4|Y = 13) = \Phi(\frac{4-0}{\sqrt{16}}) - \Phi(\frac{-4-0}{\sqrt{16}}) = \Phi(1) - \Phi(-1)$$
$$= \Phi(1) - (1 - \Phi(1)) = 0.6826.$$

(c)

$$\mathrm{E}(X^2|Y=13) = \sigma^2_{X|Y=13} + \mu^2_{x|Y=13} = 16.$$

$$\mathrm{E}(X(X-1)|Y=13) = \mathrm{E}(X^2|Y=13) - \mathrm{E}(X|Y=13) = 16.$$

(d)

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y) = -3 + 10 = 7.$$

$$\mathrm{Var}(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y = 25 + 9 + 2 \times (3/5) \times 5 \times 3 = 52.$$

$$\mathrm{E}(XY) = \rho\sigma_X\sigma_Y + \mu_X\mu_Y = (3/5) \times 5 \times 3 + (-3) \times 10 = -21.$$

5. (10 points) Let X be a random variable of distribution $N(0,1)$, and define $Y = e^X$.
   (a) (3 points) Find the pdf of Y.
   (b) (2 points) Compute $P(1 < Y < 2)$, using $\ln 2 = 0.69$ and the standard normal table as shown in Figure 1.
   (c) (5 points) Find the mean and variance of Y.

Solution: Please refer to Exercises 3.3-14 and 5.1-13.

(a)

$$\begin{aligned} G(y) &= P(Y \le y) \\ &= P(e^X \le y) \\ &= P(X \le \ln y) \\ &= \int_{-\infty}^{\ln y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \end{aligned}$$

$$g(y) = G(y)' = \frac{1}{y\sqrt{2\pi}} e^{-(\ln y)^2/2}.$$

(b)

$$\begin{aligned} P(1 < Y < 2) &= P(\ln 1 < X < \ln 2) \\ &= \Phi(0.69) - \Phi(0) \\ &= 0.7549 - 0.5 = 0.2549 \end{aligned}$$

(c)

$$M_X(t) = \mathrm{E}(e^{tX}) = e^{t^2/2}.$$

$$\mathrm{E}(Y) = \mathrm{E}(e^X) = M_X(1) = e^{0.5}.$$

$$\mathrm{E}(Y^2) = \mathrm{E}(e^{2X}) = M_X(2) = e^2.$$

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 = e^2 - e^1.$$

6. (10 points) Let X_1 and X_2 be two independent continuous random variables and let $Y = X_1 + X_2$.
   (a) (5 points) Denote the pdf of X_1 as $f_{X_1}(x_1), x_1 \in (-\infty, \infty)$, and the pdf of X_2 as $f_{X_2}(x_2), x_2 \in (-\infty, \infty)$. Express the pdf of Y in terms of the pdfs of X_1 and X_2.
   (b) (5 points) Assume that X_1 and X_2 have the same pdf: $f(x) = e^{-x}, 0 < x < \infty$.
      i. (2 points) Use the moment-generating function technique to find the pdf of Y.
      ii. (3 points) Use your conclusion in part (a) to find the pdf of Y.

Solution:

(a)

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(X_1 + X_2 \le y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_1} f_{X_1}(x_1) f_{X_2}(x_2) dx_2 dx_1 \\ &= \int_{-\infty}^{\infty} f_{X_1}(x_1) [\int_{-\infty}^{y-x_1} f_{X_2}(x_2) dx_2] dx_1 \\ &= \int_{-\infty}^{\infty} f_{X_1}(x_1) F_{X_2}(y - x_1) dx_1 \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= \frac{d}{dy} \int_{-\infty}^{\infty} f_{X_1}(x_1) F_{X_2}(y - x_1) dx_1 \\ &= \int_{-\infty}^{\infty} f_{X_1}(x_1) \frac{dF_{X_2}}{y}(y - x_1) dx_1 \end{aligned}$$

$$= \int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(y - x_1) dx_1, \quad y \in (-\infty, \infty)$$

(b) i. X_1 and X_2 have a gamma distribution with $\alpha = 1, \theta = 1$.

$$M_{X_1}(t) = M_{X_2}(t) = \frac{1}{1-t}, \quad t < 1.$$

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) = \frac{1}{(1-t)^2}, \quad t < 1.$$

so that Y has a gamma distribution with $\alpha = 2, \theta = 1$.

$$f_Y(y) = ye^{-y}, \quad 0 < y < \infty.$$

ii. $0 < x_1 < \infty, 0 < x_2 = y - x_1 < \infty$, so that $0 < x_1 < y$.

$$\begin{aligned} f_Y(y) &= \int_0^y f_{X_1}(x_1) f_{X_2}(y - x_1) dx_1 \\ &= \int_0^y e^{-x_1} e^{-(y-x_1)} dx_1 \\ &= \int_0^y e^{-y} dx_1 \\ &= ye^{-y}, \quad 0 < y < \infty. \end{aligned}$$

7. (10 points) Two components operate in parallel in a device, so the device fails when and only when both components fail. The lifetimes, X_1 and X_2, of the respective components are independent and identically distributed with an exponential distribution with $\theta = 2$ (i.e., the mean value is 2). The cost of operating the device is $Z = 2Y_1 + Y_2$, where $Y_1 = \min(X_1, X_2)$ and $Y_2 = \max(X_1, X_2)$.
   (a) (2 points) Show that $P(X_1 > a + b | X_1 > a) = P(X_1 > b)$ for $a \geq 0, b \geq 0$.
   (b) (2 points) Show that $W = X_1 + X_2$ has a gamma distribution with parameters $\alpha = 2$ and $\theta = 2$.
   (c) (4 points) Compute the pdf of Y_1 and the pdf of Y_2.
   (d) (2 points) Compute E(Z).

Solution:

(a) Please refer to Exercise 3.2-3.
Since X_1 has an exponential distribution with parameter $\theta = 2$, $X_1 \geq 0$.
If $a \geq 0$,$b \geq 0$, $P(X_1 > a + b | X_1 > a) = \frac{P(X_1 > a+b)}{P(X_1 > a)} = \frac{e^{-(a+b)/2}}{e^{-a/2}} = e^{-b/2} = P(X_1 > b)$.

(b) Please refer to Exercise 5.4-8.
The mgf of W is

$$\mathrm{E}(e^{tW}) = \mathrm{E}(e^{t(X_1+X_2)}) = \mathrm{E}(e^{tX_1})E(e^{tX_2}) = \frac{1}{1-2t}\cdot\frac{1}{1-2t} = \frac{1}{(1-2t)^2}, t < \frac{1}{2}$$

Hence, W follows a gamma distribution with parameters $\alpha = 2$ and $\theta = 2$.

(c) Please refer to Exercise 5.3-19.

$$\begin{aligned}
F_{Y_1}(y_1) = P(Y_1 \le y_1) &= 1 - P(\min(X_1, X_2) > y_1) \\
&= 1 - P(X_1 > y_1)P(X_2 > y_1) \\
&= 1 - (1 - F_{X_1}(y_1))(1 - F_{X_2}(y_1)) \\
&= 1 - (e^{-\frac{1}{2}y_1})^2 = 1 - e^{-y_1}, \quad y_1 \ge 0 \\
\text{Hence,} \quad f_{Y_1}(y_1) &= e^{-y_1}, \quad y_1 \ge 0 \\
F_{Y_2}(y_2) = P(Y_2 \le y_2) &= P(\max(X_1, X_2) \le y_2) \\
&= P(X_1 \le y_2)P(X_2 \le y_2) = F_{X_1}(y_2)F_{X_2}(y_2) \\
&= (1 - e^{-\frac{1}{2}y_2})^2 = 1 + e^{-y_2} - 2e^{-\frac{1}{2}y_2}, \quad y_2 \ge 0 \\
\text{Hence,} \quad f_{Y_2}(y_2) &= -e^{-y_2} + e^{-\frac{1}{2}y_2}, \quad y_2 \ge 0
\end{aligned}$$

(d) Please also refer to Exercise 5.3-19.

$$\begin{aligned}
\mathrm{E}(Y_1) &= 1 \\
\mathrm{E}(Y_2) &= \int_0^{+\infty} -ye^{-y} + ye^{\frac{1}{2}y}dy \\
&= \lim_{a\to+\infty}[ye^{-y} + e^{-y} - 2ye^{-\frac{1}{2}y} - 4e^{-\frac{1}{2}y}]\Big|_0^a \\
&= 3 \\
\text{Hence,} \quad \mathrm{E}(Z) &= E(2Y_1 + Y_2) = 2 \times 1 + 3 = 5
\end{aligned}$$

Or, obviously, $\mathrm{E}(Z) = \mathrm{E}(Y_1 + Y_2) + \mathrm{E}(Y_1) = \mathrm{E}(X_1 + X_2) + \mathrm{E}(Y_1)$, and by using the conclusion of (b),we get $\mathrm{E}(Z) = 2 \times 2 + 1 = 5$.

8. (10 points) Suppose that the number of customers visiting a fast food restaurant in a given day is K, which follows a Poisson distribution with parameter λ. Assume that each customer purchases a drink with probability p, independently from other customers, and moreover, the value of p is unchanged for different value of K. Let X be the number of customers who purchase drinks, and let Y be the number of customers who do not purchase drinks. So $X + Y = K$.

   (a) (2 points) Given that $K = k$, what is the expectation of X, namely $\mathrm{E}(X|K = k)$ and the variance of X, namely $\mathrm{Var}(X|K = k)$?

   (b) (5 points) What is the pmf of X and the pmf of Y? Show that X and Y are

independent. Hint: make use of the formula shown below

$$e^x = \sum_{t=0}^{\infty} \frac{x^t}{t!} \tag{1}$$

(c) (3 points) Use the normal distribution, Central Limit Theorem (CLT) and half-unit correction for continuity to approximate $P(a \leq \overline{X} \leq b)$, where a and b are integers and $0 < a < b$, and $\overline{X}$ is the mean of a random sample $X_1, X_2, X_3, \ldots, X_{100}$ of size 100 from the distribution of X, i.e. $\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$. Express the answer in terms of the cdf Φ of the standard normal distribution $N(0,1)$.

Solution:

(a) Given that $K = k$, X follows a binomial distribution with parameters k and p, so $\mathrm{E}(X|K=k) = kp$ and $\mathrm{Var}(X|K=k) = kp(1-p)$.

(b) Let $f_K(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ be the pmf of K, which has a Poisson distribution with parameter λ.
Let $q = 1-p$, and the pmf of X, $f_X(x)$ is given by

$$\begin{aligned}
f_X(x) &= \sum_{k=0}^{\infty} P(X=x, K=k) \\
&= \sum_{k=0}^{\infty} P(X=x|K=k)P(K=k) = \sum_{k=0}^{\infty} P(X=x|K=k)f_K(k) \\
&= \sum_{k=x}^{\infty} \binom{k}{x} p^x q^{k-x} \exp(-\lambda)\frac{\lambda^k}{k!} \\
&= \frac{\exp(-\lambda)(\lambda p)^x}{x!} \sum_{k=x}^{\infty} \frac{(\lambda q)^{k-x}}{(k-x)!} \\
&= \frac{\exp(-\lambda)(\lambda p)^x}{x!} \exp(\lambda q) \qquad \text{given by the formula (1)} \\
&= \frac{\exp(-\lambda p)(\lambda p)^x}{x!} \quad x = 0, 1, 2 \ldots
\end{aligned}$$

Thus, we conclude that X has a poisson distribution with parameter λp, and similarly, Y has a poisson distribution with parameter $\lambda q = \lambda(1-p)$, i.e.

$$f_Y(y) = \frac{\exp(-\lambda q)(\lambda q)^y}{y!} \quad y = 0, 1, 2 \ldots$$

To find the joint pmf of X and Y,$f(x,y)$, we can use the law of total probability:

$$f(x,y) = \sum_{k=0}^{\infty} P(X=x, Y=y|K=k)f_K(k) \qquad x = 0, 1, 2 \ldots \quad y = 0, 1, 2 \ldots$$

But note that$P(X = x, Y = y|K = k) = 0$ if $k \neq x + y$

$$\begin{aligned} f(x,y) &= P(X = x, Y = y|K = x + y) f_K(x + y) \\ &= P(X = x|K = x + y) f_K(x + y) \\ &= \binom{x+y}{x} p^x q^y \exp(-\lambda) \frac{\lambda^{x+y}}{(x+y)!} \\ &= \frac{\exp(-\lambda p)(\lambda p)^x}{x!} \cdot \frac{\exp(-\lambda q)(\lambda q)^y}{y!} \\ &= f_X(x) f_Y(y) \end{aligned}$$

X and Y are independent, since as we saw above

$$f(x,y) = f_X(x) f_Y(y).$$

(c) Please refer to a similar question in Exercise 5.7-13.

Since X follows a Poisson distribution with parameter λp, we have $\mathrm{E}(X_i) = \lambda p$ and $\mathrm{Var}(X_i) = \lambda p$. By applying the Central Limit Theorem, we can use $N(100\lambda p, 100\lambda p)$ to approximate the distribution of $\sum_{i=1}^{100} X_i$.

$$\begin{aligned} P(a \leq \overline{X} \leq b) &= P(100a \leq \sum_{i=1}^{100} X_i \leq 100b) \\ &\approx P(\frac{100a - 100\lambda p - \frac{1}{2}}{\sqrt{100\lambda p}} \leq Z \leq \frac{100b - 100\lambda p + \frac{1}{2}}{\sqrt{100\lambda p}}) \\ &= \Phi(\frac{100b - 100\lambda p + \frac{1}{2}}{10\sqrt{\lambda p}}) - \Phi(\frac{100a - 100\lambda p - \frac{1}{2}}{10\sqrt{\lambda p}}) \end{aligned}$$

9. (15 points) This question considers a communication system, which comprises three components: an encoder, a noisy channel and a decoder. In your solution of this question, you may need the pdf of the standard normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

and the cdf of the standard normal distribution

$$\Phi(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

(a) (4 points) The noisy channel has an input and an output. The input of the channel is a discrete random variable X with the uniform distribution over $\{1, -1\}$, which is generated by the encoder. The output of the channel is a continuous random variable Y defined as

$$Y = X + Z,$$

where Z has the normal distribution $N(0, \sigma^2)$, $\sigma > 0$, and X and Z are independent.

i. What is the distribution of Y given that $X = 1$? Determine the conditional pdf of Y given $X = 1$.

ii. What is the distribution of Y given that $X = -1$? Determine the conditional pdf of Y given $X = -1$.

iii. Give the pdf of Y.

(b) (6 points) After getting the output Y of the channel, the decoder makes a *decision* about the input X of the channel with respect to a *threshold* y: If $Y > y$, $\hat{X} = 1$; otherwise, $\hat{X} = -1$. The probability $P(\hat{X} \neq X)$ is called the decision error probability and is actually a function of y.

i. Find the decision error probability when $X = 1$, i.e., $P(\hat{X} \neq X|X = 1)$.

ii. Find the decision error probability when $X = -1$, i.e., $P(\hat{X} \neq X|X = -1)$.

iii. Find the decision error probability $P(\hat{X} \neq X)$.

iv. Determine the optimal value of y that minimizes the decision error probability. What is the minimum decision error probability?

(c) (5 points) Fix an integer $n > 0$. Suppose that the random variable X is transmitted n times through the channel. For the ith transmission, $i = 1, \ldots, n$, the channel takes X as the input and generates the output $Y_i = X + Z_i$, where Z_i has the normal distribution $N(0, \sigma^2)$. Moreover, $X, Z_1, Z_2, \ldots, Z_n$ are mutually independent. In this case, the decoder can use $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ to make the decision about the input X.

i. What are the mean and variance of $\overline{Y}$?

ii. Using Chebyshev's Inequality, show that $\overline{Y}$ converges to X in probability, i.e., for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{Y} - X| \geq \epsilon) = 0.$$

Solution:

(a) i. Given $X = 1$, Y is $N(1, \sigma^2)$, which has pdf $f((y-1)/\sigma)/\sigma$.

ii. Given $X = -1$, Y is $N(-1, \sigma^2)$, which has pdf $f((y+1)/\sigma)/\sigma$.

iii. $[f((y-1)/\sigma) + f((y+1)/\sigma)]/(2\sigma)$.

(b) i.

$$\begin{aligned} P(\hat{X} \neq X|X = 1) &= P(\hat{X} = -1|X = 1) \\ &= P(Y \leq y|X = 1) \\ &= \Phi((y-1)/\sigma). \end{aligned}$$

ii.

$$\begin{aligned} P(\hat{X} \neq X|X = -1) &= P(\hat{X} = 1|X = -1) \\ &= P(Y > y|X = -1) \\ &= 1 - \Phi((y+1)/\sigma). \end{aligned}$$

iii.

$$\begin{aligned} p_{\mathrm{e}}(y) &\triangleq P(\hat{X} \neq X) \\ &= P(\hat{X} \neq X|X=1)P(X=1) + P(\hat{X} \neq X|X=-1)P(X=-1) \\ &= [\Phi((y-1)/\sigma) + 1 - \Phi((y+1)/\sigma)]/2. \end{aligned}$$

iv. Taking the derivative with respect to y, we have

$$p'_{\mathrm{e}}(y) = \frac{1}{2\sigma}\left(f((y-1)/\sigma) - f((y+1)/\sigma)\right).$$

Using the symmetry, bell-shape of the standard normal pdf, we have that

- when $y < 0$, $p'_{\mathrm{e}}(y) < 0$;
- when $y = 0$, $p'_{\mathrm{e}}(y) = 0$;
- when $y > 0$, $p'_{\mathrm{e}}(y) > 0$.

Hence, $p_{\mathrm{e}}(y)$ is minimized at $y = 0$ and the optimal value is $[\Phi(-1/\sigma) + 1 - \Phi(1/\sigma)]/2 = \Phi(-1/\sigma)$.

(c) Write $\overline{Y} = X + \overline{Z}$.

i. $E(\overline{Y}) = E(X) + E(\overline{Z}) = 0$. $Var(\overline{Y}) = Var(X) + Var(\overline{Z}) = 1 + \sigma^2/n$.

ii. By Chebyshev's Inequality, $P(|\overline{Y} - X| \geq \epsilon) = P(|\overline{Z} - 0| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$. Taking the limit on both sides,

$$\lim_{n\to\infty} P(|\overline{Y} - X| \geq \epsilon) \leq 0.$$

Together with $P(|\overline{Y} - X| \geq \epsilon) \geq 0$, we have $\lim_{n\to\infty} P(|\overline{Y} - X| \geq \epsilon) = 0$.

10. (10 points) Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. Suppose that it has been known that the random sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the random sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

are independent.

(a) (8 points) Prove that

$$\frac{n-1}{\sigma^2}S^2$$

has a chi-square distribution with degrees of freedom $n-1$, i.e.,

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1).$$

(b) (2 points) Prove that

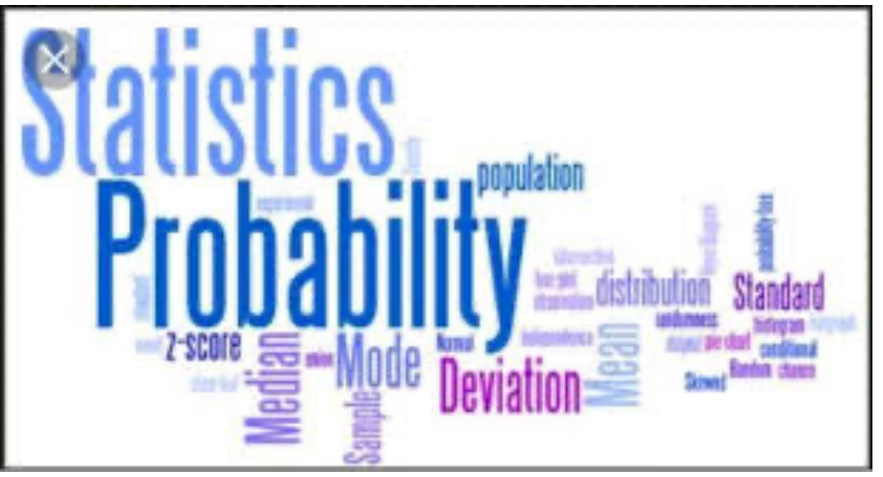
$$\mathrm{E}(S^2) = \sigma^2.$$

Solution:

(a) See the proof of Theorem 5.5-2 in the textbook on page 203.

(b) Note that $\mathrm{E}(X) = r$ for $X \sim \chi^2(r)$. Then it follows from $\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$ that,

$$\mathrm{E}\left(\frac{n-1}{\sigma^2}S^2\right) = n-1 \Longrightarrow \mathrm{E}(S^2) = \sigma^2.$$



Probability and Statistics (II)

NOTEBOOK

WANG Yuzhe (Youthy)

Spring, 2022

Section I – Review and Additions for STA 2001

WANG Yuzhe*

March 17, 2022

1 Regression Line

1.1 *Properties of Covariance*

Let X, Y be two random variables.

Property I For any constant a, b, c, d, $\mathrm{Cov}(aX + c, bY + d) = ac\ \mathrm{Cov}(X, Y)$.

Property II $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$.

(*proof.*) $\mathrm{Var}(X \pm Y) = E\left[(X \pm Y)^2\right] - E^2(X \pm Y) = E(X^2) - E^2(X) + E(Y^2) - E^2(Y) \pm 2\left[E(XY) - E(X)E(Y)\right] = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$.

Property III Let Z be a random variable. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$.

(*proof.*) Directly by definition.

1.2 *Linear Regression Line*

The two-variable series (X_i, Y_i), $1 \le i \le n$, are given.

Question I: Which line represents these data best?

Answer I: There are many choices. One option: $\left(\widehat{a}, \widehat{b}\right) = \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - b - aX_i)^2$.

Question II: Solve for $\widehat{a}$ and $\widehat{b}$.

Answer II: $\begin{cases} \widehat{a} = \mathrm{Corr}(X, Y)\dfrac{\mathrm{SD}(Y)}{\mathrm{SD}(X)} \\ \widehat{b} = \overline{Y} - \widehat{a}\overline{X} \end{cases}$.

(*proof.*) For any given slope a, the optimal b is $\overline{Y} - a\overline{X}$ (easy to verify). Thus, plug $b = \overline{Y} - a\overline{X}$ into $\sum_{i=1}^{n}(Y_i - b - aX_i)^2$ and use the property of the laughing parabola with variable a, we get $\widehat{a} = \dfrac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$.

Ouestion III: How can we measure errors?

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

Answer III: (*Defn.*) Residual $e_i := Y_i - \widehat{Y}_i, 1 \le i \le n$; $S_{xx} := \sum_{i=1}^{n}(X_i - \overline{X})^2 = n\text{Var}(X)$; Total variability $S_{yy} := \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = n\text{Var}(Y)$; $S_{xy} := \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = n\text{Cov}(X,Y)$; Unexplained variability $S_{ee} := \sum_{i=1}^{n} e_i^2 = n\text{Var}(Y)[1 - \text{Corr}^2(X,Y)]$.; Explained variability $S_{\widehat{y}\widehat{y}} := \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$; Coefficient of determination $R^2 = \text{Corr}^2(X,Y)$.

(properties) $\frac{S_{ee}}{S_{yy}} = 1 - R^2$; $S_{ee} + S_{\widehat{y}\widehat{y}} = S_{yy}$. (proof: $\sum_{i=0}^{n}(\widehat{Y}_i - \overline{Y})(\widehat{Y}_i - Y_i) = \sum_{i=1}^{n}(Y_i - \overline{Y} + \widehat{a}\overline{X} - \widehat{a}X_i)(\overline{Y} - \widehat{a}\overline{X} + \widehat{a}X_i - \overline{Y}) = n\widehat{a}\sum_{i=1}^{n}[\text{Cov}(X,Y) - \widehat{a}\text{Var}(X)] = 0$.)

Remark: Correlation is not causality. Had it been, we would land in a circular argument. There might be a third factor which explains the two.

2 Common Random Variables

2.1 Conditional Distribution

(*Thm I*) **Theorem of Total Expectation/ Law of Iterated expectations:**

$$E\left(E(X\middle|Y)\right) = E(X).$$

(proof.) $E(E(X|Y)) = \sum_y E(X|Y=y)P(Y=y) = \sum_y \left(\sum_x xP(X=x|Y=y)\right)P(Y=y) = \sum_y\left(\sum_x \frac{xP(X=x,Y=y)}{P(Y=y)}\right)P(Y=y) = \sum_x x\sum_y P(X=x,Y=y) = E(X)$.

(*Defn*) **Variance Within/ Unexplained Variance:**

$$E\left(\text{Var}(X\middle|Y)\right) = \sum_y \text{Var}(X\middle|Y=y)P(Y=y);$$

Variance Between/ Explained Variance:

$$\text{Var}\left(E(X\middle|Y)\right) = \sum_y \left[E(X\middle|Y=y) - E(E(X|Y))\right]^2 P(Y=y).$$

(*Thm II*) $\text{Var}(X) = E\left(\text{Var}(X\middle|Y)\right) + \text{Var}\left(E(X\middle|Y)\right)$. (Total Variance = Variance Within + Variance Between.)

(proof.) $\text{Var}(X) = \text{E}((X - \text{E}(X))^2) = \text{E}(\text{E}((X - \text{E}(X))^2)|Y) = \text{E}(\text{E}((X - \text{E}(X|Y) + \text{E}(X|Y) - \text{E}(X))^2)|Y)$

$= \text{E}(\text{E}((X - \text{E}(X|Y))^2)|Y) + 2\text{E}((\text{E}(X - \text{E}(X|Y))(\text{E}(X|Y) - \text{E}(X)))|Y) + \text{E}(\text{E}(X|Y) - \text{E}(X))^2)|Y) = \text{E}(\text{Var}(X|Y)) + 0 + \text{Var}(\text{E}(X|Y))$.

We next argue why the second term equals zero. Note that in fact it is twice $\text{Cov}(X - \text{E}(X|Y)), \text{E}(X|Y) - \text{E}(X)) = \text{Cov}(X - \text{E}(X|Y)), \text{E}(X|Y))$.

$\text{Cov}(\text{E}(X|Y), \text{E}(X|Y) - X) = \text{Var}(\text{E}(X|Y)) - \text{Cov}(X, \text{E}(X|Y))$, we need to show that the two terms in the right hand side are equal.

$\text{Cov}(X, \text{E}(X|Y)) = \text{E}(X\text{E}(X|Y)) - \text{E}(X)\text{E}(\text{E}(X|Y)) = \text{E}(\text{E}(X\text{E}(X|Y)|Y)) - \text{E}^2(X) = \text{E}(\text{E}(X|Y)\text{E}(X|Y)) - \text{E}^2(X)$

$= \text{E}((\text{E}(X|Y))^2) - \text{E}^2(X) = \text{E}((\text{E}(X|Y))^2) - \text{E}^2(\text{E}(X|Y)) = \text{Var}(\text{E}(X|Y))$,

2.2 Discrete Random Variables

2.2.1 *Poisson Distribution*

Suppose X and Y are two independent random variables, both following a Poisson distributions with parameters λ and μ, respectively. Then, $X + Y \sim Pois(\lambda + \mu)$.

(proof.)

$$\mathrm{P}(X+Y=k) = \sum_{i=\infty}^{\infty} \mathrm{P}(X=i, Y=k-i) = \sum_{i=\infty}^{\infty} \mathrm{P}(X=i)\mathrm{P}(Y=k-i)$$

$$= \sum_{i=0}^{k} \mathrm{P}(X=i)\mathrm{P}(Y=k-i) = \sum_{i=0}^{k} e^{-\lambda}\frac{\lambda^i}{i!}e^{-\mu}\frac{\mu^{k-i}}{(k-i)!}$$

$$e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}(\frac{\lambda}{\lambda+\mu})^i(\frac{\mu}{\lambda+\mu})^{k-i} = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!},$$

since the final summation is the sum of binomial probabilities with parameters k and $\lambda/(\lambda+\mu)$. **END**

2.2.2 *The Geometric Distribution*

The Hazard Function: $P(X = i|X \geq i) = \dfrac{p_i}{q_i}$ by $h_i, i \geq 1$. These probabilities are knows as the hazard probabilities, because they give the probability that one whose current age is i, namely has survived i years, $i \geq 1$, and this will be his final year.

In the area of actuary, the hazard is the single post important variable as it determined the yearly premium for life insurance. Clearly, $1 - h_i = \dfrac{q_{i+1}}{q_i}$; $p_i = \Pi_{j=1}^{i-1}(1-h_j)h_i \;\; i \geq 1$.

Meomeryless Property: The hazard probabilities in the case of some distributions are constant. This fact show the one does not age or rejuvenate when one's life spam follows these distributions. An alternative way to see this property is to note that

$$P(X > m + n|X > m) = P(X > n).$$

(How long people will live for another several years does NOT depend on the past.)

2.3 Continuous Random Variables

2.3.1 *Exponential Distribution*

Suppose $X \sim exp(\lambda)$. (kernel: $e^{-\lambda x}$, constant: λ, range: $[0, \infty)$).

Property I For any non-zero constant a, $aX \sim exp\left(\dfrac{\lambda}{a}\right)$.

(*proof.*) Use CDF of $exp(\lambda)$.

Property II n-th moment: $E(X^n) = \displaystyle\int_0^{\infty} \lambda x^n e^{-\lambda x} dx = \frac{n!}{\lambda^n}, \forall n \in \mathbb{N}^+$.

(*proof.*) Integration by part + Induction.

Property III It satisfies meomeryless property.

Property IV $\lceil X \rceil \sim geom(1 - e^{\lambda})$.

(*proof.*)

$$\mathrm{P}(\lceil X \rceil = i) = \int_{x=i-1}^{i} \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}|_{x=i-1}^{i} = -e^{\lambda i} + e^{\lambda(i-1)} = e^{(i-1)\lambda}(1-e^{-\lambda}),\; i \geq 1.$$

2.3.2 *Normal Distribution*

Suppose $X \sim N(\mu, \sigma^2)$. (kernel: $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, constant: $\dfrac{1}{\sqrt{2\pi}\sigma}$, range: $\mathbb{R}$).

(Defn) **Convolution Formula:** Let X and Y be two independent random variables with densities $f(x)$ and $g(y)$, respectively. The the density function of $X+Y$ at the point w, equals

$$f_{X+Y}(w) = \int_{-\infty}^{\infty} f(t)g(w-t)dt.$$

Property if $X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2$, are two independent random variables, $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2.3.3 *Chi-square Distribution*

Suppose $X \sim \chi^2_{(n)}$. (kernel: $x^{\frac{n}{2}-1}e^{-\frac{x}{2}}$, constant: $\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$, range: $[0, \infty)$).

Note that if $X \sim \chi^2_{(n)}$ and $Y \sim \chi^2_{(m)}$ are two independent random variables, then $X + Y \sim \chi^2_{(n+m)}$.

2.3.4 *Gamma Distribution*

Suppose $X \sim \Gamma(\alpha, \beta)$. (kernel: $x^{\alpha-1}e^{-\beta x}$, constant: $\dfrac{\beta^\alpha}{\Gamma(\alpha)}$, range: $[0, \infty)$.)

If $X \sim \Gamma(\alpha, \beta)$, then

$$\mathrm{E}(X^{-k}) = \frac{\Gamma(\alpha-k)}{\Gamma(\alpha)}\beta^k = \frac{\beta^k}{(\alpha-1)(\alpha-2)\cdots(\alpha-k)}, k < \alpha$$

Proof: Similar to Item 4. See home assignment no. 3. **End**

If $X \sim \Gamma(\alpha, \beta)$, then

$$\mathrm{E}(X^k) = \frac{\Gamma(\alpha+k)}{\Gamma(a)}\frac{1}{\beta^k} = \alpha(\alpha+1)\cdots(\alpha+k-1)\frac{1}{\beta^k},\ k \geq 1.$$

Proof:

$$\mathrm{E}(X^k) = \int_{x=0}^{\infty} x^k \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-x}\,dx = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\frac{1}{\beta^k}\int_{x=0}^{\infty}\frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)}x^{\alpha+k-1}e^{-\beta x}\,dx$$

Note that the value of the integral is 1 as it is the integral of a density function. **End**

if $X_i \sim \Gamma(\alpha_i, \beta)$, $i = 1, 2$, are two independent random variables, then $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \beta)$. Proof as above for the sum of two chi-square random variables.

Proof: Look at the kernel of $X_1 + X_2$ at the point x, using the convolution formula. It equals

$$\int_{t=0}^{x} t^{\alpha_1-1}e^{-\beta t}(x-t)^{\alpha_2-1}e^{-\beta(x-t)}\,dt = e^{-\beta x}\int_{t=0}^{x} t^{\alpha_1-1}(x-t)^{\alpha_2-1}\,dt$$

The rest is by change of variables as done above. **End**

2.3.5 *F Distribution & Student-t Distribution*

In summary,

$$t_{(n)} = \frac{Z}{\sqrt{\chi^2_{(n)}/n}}.$$

ratio between two independent chi-square random variables with m and n degrees of freedom, each divided by its number of degrees of freedom. In short,

$$F_{(m,n)} = \frac{\chi^2_{(m)}/m}{\chi^2_{(n)}/n}.$$

$$T|X = x \sim N(0, \frac{n}{x}) \quad \text{namely,} \quad f_{T|X=x}(t) = \frac{\sqrt{x}}{\sqrt{2\pi n}}e^{-\frac{xt^2}{2n}},$$

with penalty function quals $\sqrt{x}e^{-\frac{xt^2}{2n}}$. Of course,

$$\mathrm{E}(F_{(m,n)}) = \frac{n}{n-2},\ n > 2.$$

$$f_T(t) = \int_{x=0}^{\infty} f_{T|X=x}(t)f_X(x)\,dx$$

$$\mathrm{E}(1/F_{(m,n)}) = m/(m-2), \quad W \sim F(r_1, r_2) \xrightarrow{\text{yields}} \frac{1}{W} \sim F(r_2, r_1)$$

$$F_{1-\alpha}(r_1, r_2) = \frac{1}{F_\alpha(r_2, r_1)}$$

Properties of the Arithmetic Mean:

minimizes the sum of the squares of the residuals, namely the $f(x) = \sum_{i=1}^{n}(Y_i - x)^2$

Properties of the Geometric Mean:

The geometric mean minimizes this penalty function. $g(x) = \sum_{i=1}^{n}(\log Y_i - \log x)^2$

Properties of the harmonic mean, $H(Y)$:

The harmonic mean minimizes this penalty function $h(x)$. $h(x) = \sum_{i=1}^{n}\left(\frac{1}{Y_i} - \frac{1}{x}\right)^2.$

Properties of the Median:

Minimizing a loss function $h(x) = |Y_1 - x| + |Y_2 - x| + \cdots + |Y_n - x| = \sum_{i=1}^{n}|Y_i - x| = n \cdot \overline{|Y - x|}.$

2.4 Standard units and standardization

Definition: For Y_i, $1 \le i \le n$, let

$$Z_i = \frac{Y_i - \overline{Y}}{\mathrm{SD}(Y)},\ 1 \le i \le n$$

The linear transformation whch multiplies by $\frac{1}{\mathrm{SD}(Y)}$ and adds $-\frac{\overline{Y}}{\mathrm{SD}(Y)}$, is called **standardization.**

Regression to the Mean $\frac{y - \overline{Y}}{\mathrm{SD}(Y)} = \mathrm{Corr}(X, Y)\frac{x - \overline{X}}{\mathrm{SD}(X)}$

Additions:

(i) **Hypergeometric Distribution**

Consider a collection of $N = M + K$ similar objects, M of which belongs to one of two dichotomous classes and K of which belongs to the second class. A collection of n objects is selected from these N objects at random and without replacement. Find the probability that exactly x of these n objects belong to the first class and $n - x$ belong to the second.

$$f(x) = \frac{\binom{M}{x}\binom{K}{n-x}}{\binom{N}{n}} \qquad N = M + K$$

$X \sim Hypergeometric(N, M, n) \quad 0 \le x \le n, x \le M, n - x \le K$

$$\mu = E(X) = n\frac{M}{N};\ \ \sigma^2 = \mathrm{Var}(X) = n\frac{M}{N}\frac{K}{N}\frac{N-n}{N-1}$$

(ii) **Sample Proportion**:

the percentage of success $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\quad X \sim Bernoulli(p) \quad x = 0,1$

Section II – Point Estimation

WANG Yuzhe*

April 1, 2022

1 Generating Point Estimators

The point of departure in parameter estimation is that one assumes that a random variable under inspection belongs to some *parameter family of distributions* but the *actual value* of the parameter is unknown.

To learn what is the true parameter, one conducts a random sample, namely one inspects an independent series of some length, to be denoted by n, of i.i.d. random variables X_i, $1 \leq i \leq n$, all following the same target distribution. On the sample, one applies a function, itself a random variable, to be called in this context a **statistic**, estimating the parameter.

That is, we repeat the experiment n independent times, observe the sample, $X_1, X_2, \cdots, X_n$, and try to estimate the value of θ by using the observations $x_1, x_2, \cdots, x_n$. The function of $X_i, i = 1, \cdots, n$ used to estimate θ—say, the **statistic** $u(X_1, X_2, \cdots, X_n)$—is called an ***estimator*** of θ. We want it to be such that the computed ***estimate*** $u(x_1, x_2, \cdots, x_n)$ is usually close to θ. Since we are estimating one member of $\theta \in \Omega$, such an estimator is often called a ***point estimator***.

There are some criteria for judging how good an **estimator** is. Note that a similar question for an **estimate** is meaningless: Unless you know the actual value of the parameter which you estimate, there is no way to know the exact error committed.

1.1 *The Method of Moments*

To simply equate the first sample moment to the first theoretical moment. Next, if needed, the two second moments are equated, then the third moments, and so on, until we have enough equations to solve for the parameters.

Namely, *empirical distribution* is defined as the one who gets any of these n values with probability $\frac{1}{n}$. Then, replace $E(X^k)$ with $\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$.

Some illustrations are as follows.

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

1.1.1 Uniform Distributions

$X \sim U[a,b]$

$m_1 = (a+b)/2$ and $m_2 = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}$, we conclude that a pair of estimators for a and b, to be denote by $\hat{a}$ and $\hat{b}$,

$$\frac{\hat{a}+\hat{b}}{2} = \overline{X} \quad , \quad \frac{(\hat{b}-\hat{a})^2}{12} = \overline{X^2} - \overline{X}^2.$$

which needs to be solved for $\hat{a}$ and $\hat{b}$. As it turns out, $\hat{a} = (4\overline{X^2} - 3\overline{X}^2)/\overline{X}$ and $\hat{b} = 2\overline{X} - \hat{a}$.

1.1.2 Exponential Distributions

$X \sim exp(\lambda)$

$\mathrm{E}(X) = 1/\lambda$. Hence, replacing $\mathrm{E}(X)$ by the first empirical moment, namely by $\overline{X}$ leads to the estimator $\hat{\lambda} = 1/\overline{X}$. But this is not the only option, although it is the most natural one. As $\mathrm{E}(X^k) = k!/\lambda^k$, $k \geq 1$, replacing $\mathrm{E}(X^k)$ with $\overline{X^k}$ leads to the estimator

$$\hat{\lambda} = \left(\frac{k!}{\overline{X^k}}\right)^{\frac{1}{k}}, \ k \geq 1.$$

1.1.3 Normal Distributions

$X \sim N(\mu, \sigma^2)$

let $m_k = \mathrm{E}(X^k)$ be the k-th moment of $X \sim N(\mu, \sigma^2)$, $k \geq 1$.
Then, $m_1 = \mu$ and $m_2 = \sigma^2 + m_1^2$. Hence, $\mu = m_1$ and $\sigma^2 = m_2 - m_1^2$.
μ will be estimated by the estimator $\hat{\mu} = \overline{X}$ and σ^2 by the estimator $\hat{\sigma^2} = \overline{X^2} - \overline{X}^2$.

1.2 *Maximium Likelihood Estimator (MLE)*

One reasonable way to proceed toward finding a good estimate of θ is to regard the probability $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n; \theta)$ (or joint pmf) or product of pdf $\Pi_{i=1}^n f(x_i; \theta)$ as a function of θ and find the value of θ that maximizes it. That is, we find the θ value most likely to have produced these sample values. The joint pmf, when regarded as a function of θ, is frequently called the **likelihood function**, denoted as $L(X_1, \cdots, X_n; \theta), \theta \in \Omega \subset \mathbb{R}^m$.

The ***maximum likelihood estimator (MLE)*** is then defined as

$$\widehat{\theta} = \arg\max_{\theta \in \Omega} L(X_1, \cdots, X_n; \theta).$$

Some illustrations are as follows.

1.2.1 Bernoulli Distributions

$X \sim Ber(p)$

It is important to watch the technique presented here as it will be applicable for any other exponential family of distributions. First, the likelihood function is $L(X_1, \ldots, X_n; p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$.

$$\log L(X_1, \ldots, X_n; p) = \log p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} = \sum_{i=1}^n X_i \log p + (n - \sum_{i=1}^n X_i)\log(1-p).$$

Equating this derivative to zero, which is sufficient for maximization since the likelihood function is concave with p,

$$\hat{p} = \overline{X}.$$

Note that $\mathrm{E}(\hat{p}) = p$.

1.2.2 Exponential Distributions

$X \sim exp(\lambda)$

$L(X_1, \cdots, X_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$. Its logarithm equals $\log L(X_1, \cdots, X_n; \lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i$.

Taking derivative with respect to λ and equating it to zero, implies that $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$.

Since $\sum_{i=1}^{n} X_i \sim \Gamma(n, \lambda)$, we know that $E(\hat{\lambda}) = \mathrm{E}(\frac{1}{\overline{X}}) = n\mathrm{E}(\frac{1}{\sum_{i=1}^{n} X_i}) = \frac{n}{n-1}\lambda$.

1.2.3 Normal Distributions

$X \sim N(\mu, \sigma^2)$

the two parameters μ and σ^2 base on a random sample X_i, $1 \le i \le n$. Specifically, the likelihood function equals $\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = (\frac{1}{2\pi\sigma^2})^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$. Its logarithm, ignoring additive terms which are not a function of μ and/or σ^2, equals $-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$. Taking derivatives, first with respect to μ and then with respect to σ^2 The MLEs are derived l

$\hat{\mu} = \overline{X}$ and that $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$. Note the division here is by n and not by $n-1$.

had σ^2 been given, the MLE for μ would have be the same. had μ been given, the MLE for σ^2 would have been $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$.

1.2.4 Uniform Distributions

$X \sim U[0, \theta]$

$$L(X_1, \ldots, X_n; \theta) = \begin{cases} \frac{1}{\theta^n} & \max_{i=1}^{n} X_i \le \theta \\ 0 & \text{otherwise} \end{cases}$$

It is possible to see that the value which maximizes the likelihood function is $\ddot{\theta} = \max_{i=1}^{n} X_i$ and this will be the MLE for θ. $\mathrm{E}(\hat{\theta}) = \frac{n}{n+1}\theta$, and $\mathrm{Var}(\hat{\theta}) = \frac{n}{(n+1)^2(n+2)}\theta^2$.

(**Proof.** Assume that $\theta = 1$. Then, for $0 \le x \le 1$, $F_{\max_{i=1}^{n} X_i}(x) = \mathrm{P}(\max_{i=1}^{n} X_i \le x) = \mathrm{P}(X_i \le x, 1 \le i \le n) = \Pi_{i=1}^{n}\mathrm{P}(X_i \le n) = x^n$.)

2 Evaluating Point Estimators

2.1 *Mean Square Error and Unbiasedness*

(*Defn*) **Mean Square Error:**

$$f(\theta) = E\left(\widehat{\theta} - \theta\right)^2.$$

Decomposition of the MSE:

$$f(\theta) = \mathrm{Var}(\widehat{\theta}) + \left(E(\widehat{\theta}) - \theta\right)^2 = \text{variance} + \text{bias}^2.$$

(proof.) $E(\widehat{\theta}-\theta)^2 = E(\widehat{\theta}-E(\widehat{\theta})+E(\widehat{\theta})-\theta)^2 = E(\widehat{\theta}-E(\widehat{\theta}))^2 + 2E\left[(\widehat{\theta}-E(\widehat{\theta}))(E(\widehat{\theta})-\theta)\right] + (E(\widehat{\theta})-\theta)^2$. It is simple to get the middle term is 0, because $E\left[(\widehat{\theta}-E(\widehat{\theta}))(E(\widehat{\theta})-\theta)\right] =$

$$\left(E(\widehat{\theta}) - \theta\right) E(\widehat{\theta} - E(\widehat{\theta})) = 0.$$

The variance deals with the variation or the randomness of the estimator, while the bias deals with a consistency in its error. Of course, we like both parts to be as small as possible but at times there is a **trade-off** between the two.

An estimator for a parameter is said to be an **unbiased estimator (UBE)** in case that its mean value coincides with the parameter. In other words, its bias equal zero and hence its MSE coincides with its variance.

Note also that unbiasedness is NOT preserved under functions of the parameter. An exception is a linear function. Thus, it is possible, for example, that $E(\widehat{\theta}) = \theta$ but $E(\widehat{\theta^2}) \neq \theta^2$.

Some illustrations are as follows.

2.1.1 Normal Distributions

$X \sim N(\mu, \sigma^2)$

observe that $\mathrm{E}(\hat{\mu}) = \mu$ and that $\mathrm{E}(\frac{1}{n}\Sigma_{i=1}^{n}(X_i - \mu)^2) = \sigma^2$ $\left(\Sigma_{i=1}^{n}(X_i - \mu)^2)/\sigma^2 \sim \chi^2_{(n)}, \Sigma_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2 \sim \chi^2_{(n-1)}.\right)$

$\overline{X}$ is an unbiased estimator for μ, whose variance, and hence its MSE, equals σ^2/n. Both $\frac{1}{n}\Sigma_{i=1}^{n}(X_i - \mu)^2$ and S^2 are unbiased estimators for σ^2. we conclude that the variance of $\frac{1}{n}\Sigma_{i=1}^{n}(X_i - \mu)^2$ and of S^2 are, respectively, $\frac{1}{n^2}2n\sigma^4 = \frac{2}{n}\sigma^4$ and $\frac{1}{(n-1)^2}2(n-1)\sigma^4 = \frac{2}{n-1}\sigma^4$.

2.1.2 Exponential Distributions

$X \sim exp(\lambda)$

Recall that the MLE for λ in case of an exponential distribution is $1/\overline{X}$. it is not an unbiased estimator.

$\frac{n-1}{n}\frac{1}{\overline{X}}$ is an unbiased estimator for λ. What is the variance of this estimator?

We claim that it equals $\lambda^2/(n-2)$.

2.1.3 Bernoulli Distributions

$X \sim Ber(p)$

The sample mean, usually denoted in this context by $\hat{p}$ is an the MLE for p.

$\mathrm{MSE}(\hat{p}) = \mathrm{Var}(\hat{p}) = p(1-p)/n$. Consider an alternative estimator $\hat{p}_2 = \dfrac{n\hat{p} + 2}{n+4}$.

It is easy to see that this new estimator is a weighted average between $\hat{p}$ and $1/2$ with weights $n/(n+4)$ and $4/(n+4)$, respectively.

It is known as the "$2+2$" estimator as it as $\hat{p}$ but we artificially add two successes and two failures.

$$\mathrm{E}(\hat{p}_2) = \frac{np+2}{n+4} \quad \text{and} \quad \mathrm{Var}(\hat{p}_2) = \frac{n^2}{(n+4)^2}\frac{p(1-p)}{n}.$$

2.1.4 Uniform Distributions

$X \sim U[0, \theta]$

It was shown that its mean value equals $\frac{n}{n+1}\theta$, so it typically under shoots its target. In particular, it is biased. Clearly now, $\frac{n+1}{n}\hat{\theta}$ is unbiased. It is still not clear who among the two comes with a lower MSE.

among all estimators of the shape $C(n)\hat{\theta}$ for some function $C(n)$ the optimal choice is $C(n) = (n+2)/(n+1)$

Proof.

$\mathrm{E}((C(n)\hat{\theta}-\theta)^2) = \mathrm{Var}(C(n)\hat{\theta}) + \mathrm{E}(C(n)\hat{\theta}-\theta)^2 = C^2(n)\frac{n}{(n+1)^2(n+2)}\theta^2 + (C(n)\frac{n}{n+1}-1)^2\theta^2.$

Its minimum is indeed attained at $(n+2)/(n+1)$, as required. **End**

2.2 Consistency

Let $\widehat{\theta_n}$ be an estimator for θ which is based on a sample of size n. We say that the estimator is **consistent** if for any $\epsilon > 0$.

$$\lim_{n\to\infty} P_\theta\left(|\widehat{\theta_n} - \theta| < \epsilon\right) = 1.$$

If $X_1, \cdots, X_n$ is a random sample from a distribution with finite mean μ and variance σ^2 , then by the *Weak Law of Large Numbers*, the sample mean, $\overline{X}$, is a consistent estimator of μ.

(Remark.) It is possible to see that if $\lim_{n\to\infty} \mathrm{MSE}(\widehat{\theta_n}) = 0$, then consistency is guaranteed. What we further claim but without a proof is the fact that MLEs are consistent.

Consistency takes place when the series of estimators **converges in probability** to a degenerate random variable, namely a constant, which is the parameter itself.

(*Thm.*) **Slutsky's Theorem** Let X_i and Y_i, $i \geq 1$, be two series of random variables. Suppose X_i convergences in distribution to X and Y_i convergences in probability to a. Then X_iY_i convergences in distribution to aX and $X_i + Y_i$ convergences in distribution to $X + a$.

Claim: S^2/σ^2 converges in probability to 1.

Proof: $(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}$. Hence, $\mathrm{E}(S^2/\sigma^2) = 1$ and $\mathrm{Var}(S^2/\sigma^2) = 2/(n-1)$. Invoking Chebyshev inequality of the random variable S^2/σ^2, $\mathrm{P}(|\frac{S^2}{\sigma^2} - 1| > t) \leq \frac{2}{n-1}\frac{1}{t^2}$, which goes to zero when n goes to infinity. This concludes the proof.

2.3 Efficiency of Estimators

Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be two unbiased estimators (UBEs) for a parameter θ. Assume for some functions $C_1(n)$ and $C_2(n)$, $\mathrm{Var}(\widehat{\theta_1}) = C_1(n)$ and $\mathrm{Var}(\widehat{\theta_1}) = C_2(n)$, where n is the sample size.

(*Defn.*) **Relative efficiency** (of the first estimator in comparison with the second) is defined as $\dfrac{C_2(n)}{C_1(n)}$. In principle, this ratio can be a function of the parameter one estimates.

(*e.g.*) Let $X_i \sim N(\mu, \sigma^2)$, $1 \leq i \leq n$, be n independent random variables. The sample mean and the sample median are both UBEs for μ. Moreover, they are both consistent. The variance of the latter estimator does not have a close form expression. Yet, the limit of their relative efficiency when n goes to infinity equals $\dfrac{\pi}{2}$. Thus, we need an about 50 percent larger sample size in order to reach the same accuracy in sampling for μ if we use the sample median instead of the sample mean.

3 Improving Point Estimators

For simplicity, all estimators we choose for improvement are UBEs.

3.1 Method I: By Conditioning

(*Principle*: holding **unbiasedness** as well as reducing **variance**.)

For UBE $\widehat{\theta}$ and any random variables Y:

$$\begin{cases} E(E(\widehat{\theta}|Y)) = E(\widehat{\theta}) \\ \mathrm{Var}(E(\widehat{\theta}|Y)) = \mathrm{Var}(\widehat{\theta}) - E(\mathrm{Var}(\widehat{\theta}|Y)) \end{cases}.$$

Note: (i) Y is actually the random variable which is sampled and then, suppose Y turned out to equal y, the value of $E(\widehat{\theta}|Y=y)$ is printed, assuming it can be computed.

(ii) (problems) The computational issue may deter us from using $E(\widehat{\theta}|Y=y)$, and in many natural cases, no improvement takes place. The main problem is that $E(\widehat{\theta}|Y)$ may NOT be free of the parameter for some Y.

(e.g.1) Let U_i for $i=1,2$ be two independent [0,1]-uniform random variables. Define the random variable I as follows: $I = \begin{cases} 1 & U_1^2+U_2^2 \le 1 \\ 0 & otherwise \end{cases}$

Clearly, I is an unbiased estimator for $\pi/4$. with the variance of I $\frac{\pi}{4}(1-\frac{\pi}{4}) \approx 0.1686$. A reduced UBE for $\pi/4$ is $\mathrm{E}(I|U)$ which in fact means sampling only one of the two U_i's, $i=1,2$ and conditioning the original estimator based on that. Firstly,

$$\mathrm{E}(I|U=u) = \mathrm{P}(I=1|U=u) = \mathrm{P}(u^2+U_2^2 \le 1) = \mathrm{P}(U_2^2 \le 1-u^2) = \mathrm{P}(U_2 \le \sqrt{1-u^2}) = \sqrt{1-u^2}.$$

So the estimator is in fact $\sqrt{1-U^2}$. By definition, its expected value is $\pi/4$. its variance is $2/3-(\pi)^2/16 \approx 0.05$.

Of course, in case of an n-size random sample, the estimator is $\frac{1}{n}\sum_{i=1}^{n}\sqrt{1-U_i^2}$.

(e.g.2) **Poisson distribution.** Let $X_i \sim Pois(\lambda)$, $1 \le i \le n$.

Suppose there is an interest in estimating $e^{-\lambda}$ which is $\mathrm{P}(X_1=0)$. Define now the n random variables I_i, $1 \le i \le n$, via $I_i = 1$ of $X_i = 0$. Otherwise, $I_i = 0$. Clearly, $I_i \sim Ber(e^{-\lambda})$, $1 \le i \le n$. $\overline{I}$ is an UBE and its variance, which coincides with its MSE, equals $e^{-\lambda}(1-e^{\lambda})/n$. Let's now condition $\overline{I}$ on $\Sigma_{i=1}X_i$. First, observe that $\mathrm{E}(\overline{I}|\Sigma_{i=1}^n X_i) = \mathrm{E}(I_1|\Sigma_{i=1}^n X_i)$, The estimator is in fact $(\frac{n-1}{n})^{\sum_{i=1}^n X_i}$.

(**Proof.** The following is based on the fact that $\sum_{i=1}^n X_i \sim Pois(n\lambda)$. $\mathrm{E}(I_1|\sum_{i=1}^n X_i = k) = \mathrm{P}(X_1=0|\sum_{i=1}^n X_i = k) = \frac{e^{-\lambda}e^{(n-1)\lambda}\frac{((n-1)\lambda)^k}{k!}}{e^{-n\lambda}\frac{(n\lambda)^k}{k!}} = (\frac{n-1}{n})^k$)

In summary, the estimation of $e^{-\lambda}$ is done as follows: sampling takes place, $\Sigma_{i=1}^n X_i$ is computed, but what is typed is $((n-1)/n)^{\Sigma_{i=1}^n X_i}$.

Now for an unsuccessful example. $\mathrm{E}(\overline{I}|X_1)$is in fact $\frac{1}{n}(\mathrm{E}(I_1|X_1)) + (n-1)e^{-\lambda})$, which is not free of λ.

Hence, $\mathrm{E}(\overline{I}|X_1)$ is not a valid estimator.

3.2 Method II: By Mixing between Two Estimators

(*Principle*: holding **unbiasedness** as well as reducing **variance**.)

3.2.1 Independent UBEs

For two **independent** UBEs $\widehat{\theta}_1$ and $\widehat{\theta}_2$, and two real numbers λ_1, λ_2 s.t. $\lambda_1+\lambda_2=1$:

$$\begin{cases} E(\lambda_1\widehat{\theta}_1 + \lambda_2\widehat{\theta}_2) = \lambda_1 E(\widehat{\theta}_1) + \lambda_2 E(\widehat{\theta}_2) = \theta \\ \mathrm{Var}(\lambda_1\widehat{\theta}_1 + \lambda_2\widehat{\theta}_2)_{min} = \left[\frac{1}{\mathrm{Var}(\widehat{\theta}_1)} + \frac{1}{\mathrm{Var}(\widehat{\theta}_2)}\right]^{-1}, \quad \lambda_i \propto \frac{1}{\mathrm{Var}(\widehat{\theta}_i)}, i=1,2 \end{cases}.$$

Note: (i) $\lambda_i \propto \frac{1}{\mathrm{Var}(\widehat{\theta}_i)}, i=1,2$ gives the optimal solution, i.e., smallest variance. (The optimal weight of an estimator is proportional to the reciprocal of its variance.)

(ii) optimal case: variance = half of Harmonic Mean; sub-optimal case: variance = half of Arithmetic Mean.

3.2.2 Generalized 2 UBEs cases

For two NOT necessarily independent UBEs $\widehat{\theta_1}$ and $\widehat{\theta_2}$, and other conditions the same as above, then the optimal solutions are: $\lambda_1 = \dfrac{\mathrm{Var}(\widehat{\theta_2}) - \mathrm{Cov}(\widehat{\theta_1}, \widehat{\theta_2})}{\mathrm{Var}(\widehat{\theta_1}) + \mathrm{Var}(\widehat{\theta_2}) - 2\mathrm{Cov}(\widehat{\theta_1}, \widehat{\theta_2})}$, $\lambda_2 = 1 - \lambda_1$.

The minimum variance, $\mathrm{Var}(\lambda_1\widehat{\theta_1} + \lambda_2\widehat{\theta_2})_{min} = \dfrac{\mathrm{Var}(\widehat{\theta_1})\mathrm{Var}(\widehat{\theta_2}) - \mathrm{Cov}^2(\widehat{\theta_1}, \widehat{\theta_2})}{\mathrm{Var}(\widehat{\theta_1} - \widehat{\theta_2})}$.

Note: (i) How to give values to λ_1 and λ_2 depends on how $\widehat{\theta_1}$ and $\widehat{\theta_2}$ correlates with each other.

(ii) The more negatively $\widehat{\theta_1}$ and $\widehat{\theta_2}$ correlates, the better (i.e., the lower variance).

3.2.3 Generalized n independent UBEs cases

For n **independent** UBEs $\widehat{\theta_1}, \cdots, \widehat{\theta_n}$, and n real numbers $\omega_1, \cdots, \omega_n$ s.t. $\sum_{i=1}^{n} \omega_i = 1$:

$$\begin{cases} E(\sum_{i=1}^{n} \omega_i \widehat{\theta_i}) = \theta \\ \mathrm{Var}(\sum_{i=1}^{n} \omega_i \widehat{\theta_i})_{min} = \left[\sum_{i=1}^{n} \dfrac{1}{\mathrm{Var}(\widehat{\theta_i})}\right]^{-1}, \quad \omega_i \propto \dfrac{1}{\mathrm{Var}(\widehat{\theta_i})}, i = 1, 2 \end{cases}.$$

(*Proof.*) One option is to set the problem we face here as a standard optimization function, define the Lagrangean function, etc.

3.3 Method III: By Controling Variables

(*Principle*: holding **unbiasedness** as well as reducing **variance**.)

For any related pair of random variables X and Y, with avaiable mean of Y, denoted as μ_Y, where $\overline{X}$ is an UBE, $c \in \mathbb{R}$:

$$\begin{cases} E(\overline{X - c(Y - \mu_Y)}) = E(\overline{X}) = \theta \\ \mathrm{Var}(\overline{X - c(Y - \mu_Y)})_{min} = [1 - \mathrm{Corr}^2(X, Y)]\mathrm{Var}(\overline{X}), \quad c = \dfrac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} \end{cases}.$$

(e.g.1 cont'd) A reduced variance UBE for $\pi/4$ introduced there is $\sqrt{1 - U^2}$. It seems to be highly (negatively) correlated with U, which comes with $\mathrm{E}(U) = 0.5$ and $\mathrm{Var}(U) = 1/12$. Thus, a better UBE can be $\sqrt{1 - U^2} - c(U - 0.5)$ for an appropriate choice for c. What is the best choice for c? Towards this end, we need to find $\mathrm{Cov}(\sqrt{1 - U^2}, U)$.

For multi-parameter case, the thought of MLE methods becomes EM algorithm (see in notes of MAT3300 for details).

Section III – Interval Estimation

WANG Yuzhe*

April 26, 2022

1 The Idea of Interval Estimation

Usually the point estimation of the parameter and the true value are not the same. This bring us to **interval estimation**. Based on a sample, we compose a *random interval* and declare that the parameter is in this interval. Of course, we like this interval to be *short* in order that it will be informative.

When we say that a parameter is in an interval (a random one or not), we might be right or we might be wrong. There is no third way. We canNOT treat the parameter as a random variable and say that it belongs to the interval with some probability. What we can say is about times/ relative frequency.

Below is an example to illustrate interval estimation.

1.1 *An Example for Introduction*

Given a random sample $X_1, X_2, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$, we shall now consider the closeness of $\overline{X}$, the unbiased estimator of μ, to the unknown mean μ.

Firstly, recall that Z_p was defined via $P(Z \leq Z_p) = p$ when $Z \sim N(0,1)$. For the probability $1-\alpha$, $P\left(-Z_{1-\frac{\alpha}{2}} \leq \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$.

Then, by some simple algebra, $P\left(\overline{X} - \dfrac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \dfrac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$. So the probability that the random interval $\left[\overline{X} - \dfrac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}, \overline{X} + \dfrac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}\right]$, includes the unknown mean μ is $1-\alpha$. Once the sample is observed and the sample mean computed, the interval becomes known.

Since the probability that the random interval covers μ before the sample is drawn is equal to $1-\alpha$, we now call the computed interval above, a $100(1-\alpha)\%$ **confidence interval** for the unknown mean μ. The number $100(1-\alpha)\%$, or equivalently, $1-\alpha$, is called the **confidence coefficient**.

(Remark: (how to understand) Every time sample $X_1, \cdots, X_n$ to get $\overline{X}$, and go to both left and right by $\dfrac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}$ to generate a new interval. Judge whether μ is in that interval. Going to the fundamental interpretation for probability, it says that if this experiment is repeated large

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

amount of times, we believe that a fraction of $1-\alpha$ of the generated intervals will contain μ, while the rest will not.)

Choice of Sample Size: If $\overline{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error $|\overline{x}-\mu|$ will never exceed a specified amount E when the sample size $n \geq \left(\frac{Z_{\frac{1-\alpha}{2}}\sigma}{E}\right)^2$.

1.2 *Generalize a Confidence Interval*

Consider any arbitarty statistics X instead of $\overline{X}$. Note that the confidence interval is NOT necessarily symmetric. Any pair of positive numbers a and b, leads to $X-a \leq \mu \leq X+b$ being a confidence interval for μ with a confidence level of $1-\alpha$, if and only if $\phi(b)-\phi(a)=1-\alpha$. One extreme choice it to take $b=\infty$ and $a=Z_{1-\alpha}$. This choice is known as the **one-sided (right) confidence interval**.

In general, a confidence interval is based on sampling $X_1, \cdots, X_n$, constructing two statistics, $L(X_1, \cdots, X_n) \leq U(X_1, \cdots, X_n)$ and stating the θ dependent coverage probability

$$P_\theta(L(X_1, \cdots, X_n) \leq \theta \leq U(X_1, \cdots, X_n)).$$

2 Generating Confidence Interval (*Pivotal Quantities*)

It is desired that this probability be free of θ. Hence, the trick here is to state a statistic who is not free of the parameter but its distribution is. Followings are several examples.

2.1 *Confidence Interval for Means*

2.1.1 Mean of Normal Population when σ^2 is NOT given

Note that the case that mean of normal population when σ^2 is given is shown in above example. If σ^2 is unknown,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

Hence, we can generate a $100(1-\alpha)\%$ confidence interval

$$\overline{X} - \frac{S}{\sqrt{n}} t_{(n-1,1-\frac{\alpha}{2})} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{(n-1,1-\frac{\alpha}{2})}.$$

The width of the confidence interval is itself random, whether σ is given or not. Yet, $t_{(n-1,1-\frac{\alpha}{2})} \geq Z_{1-\frac{\alpha}{2}}$. For large sample sizes, this confidence interval holds without the normally assumption by applying a combination of the CLT and Slutsky's theorem.

2.1.2 Difference of Means (unpaired sample)

Suppose there are two independent normal populations, X population and Y population, with parameters μ_1 and σ_1^2, μ_2 and σ_2^2, respectively. Here is an interest in estimating the single parameter $\mu_2-\mu_1$. Towards this end, two independent random samples are conducted:

$X_i \sim N(\mu_1, \sigma_1^2), 1 \leq i \leq n_1$ and $Y_i \sim N(\mu_2, \sigma_2^2), 1 \leq i \leq n_2$. A sensible point estimator is

$$\overline{Y} - \overline{X} \sim N\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Hence, we can generate a $100(1-\alpha)\%$ confidence interval

$$\overline{Y} - \overline{X} - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} Z_{1-\frac{\alpha}{2}} \leq \mu_2 - \mu_1 \leq \overline{Y} - \overline{X} + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} Z_{1-\frac{\alpha}{2}}.$$

Suppose now, both variances are NOT in hand but $\sigma_2 = \sigma_1 = \sigma$, known as the equal variances assumption. In this case,

$$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \sim \chi^2_{(n_1+n_2-2)}$$

Set the **pooled sample variance** $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$, we get

$$\frac{\overline{Y} - \overline{X} - (\mu_2 - \mu_1)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

Then, the confidence interval can be obtained from this statistics.

2.1.3 Difference of Means (paired sample)

Suppose it is possible to conduct a sample of pairs $(X_i, Y_i), 1 \leq i \leq n$, where across pairs one can assume independence but not within a pair. Denote $Y_i - X_i$ by D_i, $1 \leq i \leq n$. We add a new assumption that D_i, $1 \leq i \leq n$ are normally distributed with mean $\mu_2 - \mu_1$, but exact variance is lost. Next, denote $\text{Var}(D_i)$ by σ_d^2 and let S_d^2 be the sample variance of the D_i's, $1 \leq i \leq n$. Then,

$$\frac{\overline{D} - (\mu_2 - \mu_1)}{S_d/\sqrt{n}} \sim t_{(n-1)}$$

We can then construct the confidence interval according to this statistics.

Our instincts can be lead us to think that unpaired sample is better than a paired one but this in not necessarily the case. If for a given i, Y_i and X_i are highly positively correlated, we would expect a small variance in D_i. Indeed, if Y_i overshoot its target, we would expect the same oto be the with X_i. These overshoots cancel each other when we consider D_i.

2.2 *Confidence Interval for Variances*

2.2.1 Variance of Normal Population

The point of departure now is that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Then a $(1-\alpha)$ confidence interval for σ^2 is $(n-1)\frac{S^2}{\chi^2_{(n-1,1-\frac{\alpha}{2})}} \leq \sigma^2 \leq (n-1)\frac{S^2}{\chi^2_{(n-1,\frac{\alpha}{2})}}$.

Remark. (i) For a confidence interval of σ rather than σ^2, all to do is to take the square roots in both hand sides of the above (due to monotone property of x^2). (ii) If μ is given, than S^2 can be replaced with $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ and the number of degrees of freedom goes up by one. (iii) Find the best CI $[q_1, q_2]$: because that the density function of a chi-square distribution, $f_{\chi^2_{(n-1)}}(x)$, is unimodel, the optimal pair satisfies $f_{\chi^2_{(n-1)}}(q_1) = f_{\chi^2_{(n-1)}}(q_2)$.

2.2.2 Ratio between Variances

Suppose $X_i \sim N(\mu_1, \sigma_1^2)$, $1 \leq i \leq n_1$ and $Y_i \sim N(\mu_2, \sigma_2^2)$, $1 \leq i \leq n_2$ and n_1+n_2 independent random variables. We are interested in constructing a confidence interval for the ratio $\frac{\sigma_2^2}{\sigma_1^2}$. Two cases include: (i) μ_1, μ_2 are in hand; (ii) μ_1, μ_2 are unknown. Just consider the second case,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(n_1-1,n_2-1)}$$

with sample variances S_1^2, S_2^2. And the $(1-\alpha)$ confidence interval is then

$$\left[\frac{S_2^2}{S_1^2}F_{(n_1-1,n_2-1,\frac{\alpha}{2})}, \frac{S_2^2}{S_1^2}F_{(n_1-1,n_2-1,1-\frac{\alpha}{2})}\right]$$

2.3 *Proportion in Population*

2.3.1 Proportion in one Population

Suppose $X \sim Bin(n,p)$. The value for n, the sample size, is given and one wishes to construct a confidence interval for p. Since $E(X) = np$ and $\mathrm{SD}(X) = \sqrt{np(1-p)}$, the normal approximation by central limit theorem (CLT) is that

$$\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$$

The larger n is, the more accurate is the approximation. Hence, a $(1-\alpha)$ CI is

$$\frac{X}{n} - \sqrt{\frac{p(1-p)}{n}}Z_{1-\frac{\alpha}{2}} \leq p \leq \frac{X}{n} + \sqrt{\frac{p(1-p)}{n}}Z_{1-\frac{\alpha}{2}}$$

Remark. (i) p should be replaced in the definition for the edges of the interval by $\widehat{p} = \frac{X}{n}$ itself, by the *weak law of large numbers* $\widehat{p}(1-\widehat{p})$ converges in probability to $p(1-p)$, then by Slutsky's theorem $\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}}$ converges in distribution to Z. Thus, replacing p by $\widehat{p}$ does not violate the approximation stated there. (ii) Another option is to replace $p(1-p)$ with 0.5 which is an upper bound on it. In this case we can say the confidence level is <u>at least</u> $1-\alpha$. (iii) The third option is an exact one, $(\widehat{p}-p)^2 \leq \frac{p(1-p)}{n}Z^2_{1-\frac{\alpha}{2}}$, let $a = 1 + \frac{Z^2_{1-\frac{\alpha}{2}}}{n}$, $b = \widehat{p} + \frac{Z^2_{1-\frac{\alpha}{2}}}{2n}$, $c = \widehat{p}^2$, we get the $(1-\alpha)$ CI as $\left[\frac{b-\sqrt{b^2-ac}}{a}, \frac{b+\sqrt{b^2-ac}}{a}\right]$. (iv) Choice of Sample Size: an

upper bound on the sample size making the error bounded by E is $\left(\frac{Z_{1-\frac{\alpha}{2}}}{2E}\right)^2$.

2.3.2 Difference in two Population Propotions

$X \sim Bin(n_1, p_1)$, and independent $Y \sim Bin(n_2, p_2)$. Now we are interested in $p_2 - p_1$. For large enough sample size n, we have approximately

$$\frac{\widehat{p_2} - \widehat{p_1} - (p_2 - p_1)}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}} + \sqrt{\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}} \sim N(0,1)$$

with $\widehat{p_1}, \widehat{p_2}$ defined like above.

2.4 *Uniform Distribution*

Suppose $X \sim U[0,\theta]$, then $\frac{X}{\theta} \sim U[0,1]$. The same can be said on the random variable $\max_{i=1}^n X_i$. Its CDF is x^n, $0 \le x \le 1$. Then, $\left[\frac{\max_{i=1}^n X_i}{\sqrt[n]{1-\frac{\alpha}{2}}}, \frac{\max_{i=1}^n X_i}{\sqrt[n]{\frac{\alpha}{2}}}\right]$ is a $(1-\alpha)$ CI for θ.

3 Prediction Intervals

Suppose one wishes to predict the value of X_{n+1} where $X_{n+1} \sim N(\mu, \sigma^2)$ where μ is not known to one. Two cases: the first is where σ^2 is in hand, the second where it is not. The point here to is to say something on one random variable, given the actual values of others. Yet, since the value of the others in random, the conclusion is also random. This resembles what we do in estimating parameter but there is a big difference: in prediction we deal with random variables while in estimation we deal with constants (parameters).

Get a random sample of X_i, $1 \le i \le n$, which is also independent of X_{n+1}. Our approach here is to use this sample in order to predict X_{n+1} through estimating μ (and σ^2 in case it is not given). A point prediction will be to predict X_{n+1} with $\overline{X}$, which is in fact using one's estimate of the mean and predict X_{n+1} with the estimate for its mean. Then,

$$X_{n+1} - \overline{X} \sim N\left(0, \left(1 + \frac{1}{n}\right)\sigma^2\right)$$

We can then derive a $100(1-\alpha)\%$ confidence interval $\left[\overline{X} - \sigma\sqrt{1+\frac{1}{n}}Z_{1-\frac{\alpha}{2}}, \overline{X} + \sigma\sqrt{1+\frac{1}{n}}Z_{1-\frac{\alpha}{2}}\right]$.

For the case where σ^2 is not given, as both X_{n+1} and X are both independent of S^2. Hence,

$$\frac{X_{n+1} - \overline{X}}{S\sqrt{1+\frac{1}{n}}} \sim t_{(n-1)}$$

then we can get the $(1-\alpha)$ interval for X_{n+1}.

Finally, note that the probabilistic statement on the correctness of the prediction interval is not with respect to X_{n+1} *given* $\overline{X}$ with respect to the joint distribution.

Section IV – Testing Hypothesis

WANG Yuzhe*

May 2, 2022

1 Simple Hypotheses and Tests Introduction

A ***statistical hypothesis*** is a statement about the parameters of one or more populations. ***Statistical hypothesis testing*** of parameters are the fundamental methods used at the data analysis stage of a comparative experiment. Below is an example of the simplest case of a test.

1.1 *Null and Alternative Hypotheses*

Suppose $X \sim N(\mu, 1)$, where μ is unknown. Samples X_i, $1 \leq i \leq n$ are i.i.d. and follow $N(\mu, 1)$. Assume that μ can get one of the following two numbers: μ_0 or μ_1. The former option is called the **null hypothesis** (representing *initial belief/claim/assumption*), denoted by H_0, while the latter is referred to as the **alternative hypothesis** (representing *competing belief/claim/assumption*), denoted by H_1. As the null and the alternative both come up with a single value, they are said to be **simple hypotheses**.

1.2 *Errors in Hypothesis Test and Rejection Region*

There are totally 2 type of errors in the test. The first, called **first type of error** is when H_0 is true but you have decided against it (or you have **reject**ed the null hypothesis). The second, called **second type of error**, is when you did not reject the null, while H_1 is true.

state \decision	"H_0"	"H_1"
H_0	true	type-I error
H_1	type-II error	true

Hypothesis-testing procedures rely on using the information in a *random sample from the population of interest*. For instance, suppose $\mu_1 > \mu_0$, "if $\overline{X} \geq \dfrac{\mu_0 + \mu_1}{2}$ reject H_0". This is a makes sense **decision rule**, but not sure how this intuitive justification is extended to more complicated cases. The key questions here are how to choose **detection statistic** and how to determine **threshold**.

Let's assume that the sample can be summarized by some statistic $T(X_1, \cdots, X_n)$ (which, as opposed to the confidence interval, can be a vector). A decision rule (or a test) will be

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

as follows: Partition the possible values that T can receive into two *exhaustive and mutually exclusive regions*. The first set will be called the **rejection region** and we will denoted it by C. So the test will be "reject H_0 if $T(X) \in C$."

1.3 *Significance and Power*

For judging the quality of a test, we judge the test, not the decision, namely not the test's results. There are hence two reasonable criteria. One is the probability that $T(X) \in C$ when H_0 is true. This can be written as $P_{H_0}[T(X_1, \cdots, X_n) \in C] \stackrel{\triangle}{=} \alpha$, which is the probability of the first type of error (when H_0 is true), and also referred to as the ***significance level*** of the test. Similarly, the probability of the second type of error $P_{H_1}[T(X_1, \cdots, X_n) \notin C]$ is denoted by β. From some historic reasons the focus here is on $1-\beta$, referred to as the ***power*** of the test. A good test will come with *a small significance level and high power*.

We can see that $T(X_1, \cdots, X_n) = \overline{X}$ and $C = \left[\frac{\mu_0 + \mu_1}{2}, \infty\right)$. In the case where the statistic is compared with a single value in order to make the decision, the latter is called the **critical value**. By changing the critical value, namely the lower bound of the rejection interval, we can decrease α or β but not both. In fact, these two criteria are antagonistic: improving one is always on the expense of the other.

Hence, what we can look for is for a given α, what is the *most powerful test* (with highest power), among all those tests whose significance level is at most α. Denote by $\beta(\alpha)$ the lowest possible β given $0 \leq \alpha \leq 1$. Drawing the function $\beta(\alpha)$ along the unit interval will get the **efficient frontier** (all points below are not achievable, and all points above it can be beat by a better test).

(*Claim.*) The efficient frontier $\beta(\alpha)$ is a convex function. The proof is as follows. Let S_i be a test whose significance level is α_i and power is $1-\beta(\alpha_i)$, $i = 1, 2$. Consider the following test. Plan to execute the two tests. For example, if they are based on two different statistics, then sample and compute both. In parallel, sample independently for I where $I = 1$ or 2 with probabilities p and $1-p$, respectively, for some p, $0 \leq p \leq 1$. Define the following test: reject H_0, if $I = i$ and the test statistic for S_i turns out to be in its rejection region, $i = 1, 2$. The significance level of this test and its probability of second type of error are clearly $p\alpha_1 + (1-p)\alpha_2$ and $p\beta(\alpha_1) + (1-p)\beta(\alpha_2)$. Thus, $\beta(p\alpha_1 + (1-p)\alpha_2) \leq p\beta(\alpha_1) + (1-p)\beta(\alpha_2)$.

2 Likelihood Ratio Test

2.1 *Neyman-Pearson Lemma*

Best Critical Region: as above shows, we need the critical region C to have type-I error as α and the largest power $1-\beta(\alpha)$, then that C is called the *best critical region of size* α.

(***Neyman-Pearson Lemma.***) Let $X_1, \cdots, X_n$ be a random variable of size n from a distribution with pdf/pmf $f(x;\theta)$. Define Likelihood function the same as in Section II, (i.e., $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$). If there is a positive constant k and a subset C of the sample space such that

(a) $P[(X_1, \cdots, X_n) \in C; \theta_0] = \alpha$;

(b) $\frac{L(\theta_0)}{L(\theta_1)} \leq k$, $\forall\ (x_1, \cdots, x_n) \in C$ and $\frac{L(\theta_0)}{L(\theta_1)} \geq k$, $\forall\ (x_1, \cdots, x_n) \in C^c$

Then, C is the best critical region of size α for testing the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$.

Some examples are as follows.

2.1.1 Simple Hypothesis for Normal Mean

Back to the example in Section 1.1, likelihood ratio for it is $\frac{\frac{1}{(2\pi)^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_0)^2}}{\frac{1}{(2\pi)^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_1)^2}}$, by some algebric techiques (abandon & transfer constants), we can get that the rejection region is given by $\frac{\overline{X}-\mu_0}{1/\sqrt{n}} \geq C$ (because $\mu_1 > \mu_0$). Since $P_{H_0}(\frac{\overline{X}-\mu_0}{1/\sqrt{n}} \geq C) = \alpha$, we can conclude that $C = Z_{1-\alpha}$. Hence, the test (under significance level α) is: "reject H_0 if $\frac{\overline{X}-\mu_0}{1/\sqrt{n}} \geq Z_{1-\alpha}$".

Next we move inspections to β, which is given by $\beta(\alpha) = P_{H_1}(\frac{\overline{X}-\mu_0}{1/\sqrt{n}} < Z_{1-\alpha}) = P_{H_1}(\frac{\overline{X}-\mu_1}{1/\sqrt{n}} < \frac{\mu_0-\mu_1}{1/\sqrt{n}} + Z_{1-\alpha}) = \phi\left(\frac{\mu_0-\mu_1}{1/\sqrt{n}} + Z_{1-\alpha}\right)$. β is a function of α, decreasing with α, n and the distance between assumed parameters (i.e., $|\mu_1 - \mu_0|$).

2.1.2 Simple Hypothesis on Population Proportion

Let the parameter p stands for the relative frequency of some phenomenon in the population. Suppose H_0 says that $p = p_0$ and H_1 that p_1 with assumption $p_1 > p_0$. The likelihood ratio ($X \sim Bin(n, p)$): $\frac{\binom{n}{X} p_0^X (1-p_0)^{n-X}}{\binom{n}{X} p_0^X (1-p_0)^{n-X}}$, then we can get the rejection region $\frac{X/n - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} \geq C$, which is coincident with that under CLT, $\frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} \sim N(0,1)$. So LRT under significance level α is to "reject H_0 if $\frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} \geq Z_{1-\alpha}$".

2.1.3 Simple Hypothesis on Normal Variance

Suppose $X \sim N(0, \sigma^2)$ (with some known value for the mean) and that H_0 says $\sigma^2 = \sigma_0^2$ while the alternative H_1 says that $\sigma^2 = \sigma_1^2$, where $\sigma_1^2 < \sigma_0^2$. In case of a random sample of size n, by Likelihood ratio, we can get LRTs as "reject H_0 if $\sum_{i=1}^{n}\left(\frac{X_i}{\sigma}\right)^2 \leq C$". Because $\sum_{i=1}^{n}\left(\frac{X_i}{\sigma}\right)^2 \sim \chi^2_{(n)}$, we can know $C = \chi^2_{(n,\alpha)}$.

(*Remark I.*) In the case where μ is not given but rather being estimated, replace X_i with $X_i - \overline{X}$, $1 \leq i \leq n$, and the number of degrees of freedoms is reduced to $n-1$.

(*Remark II.*) As is in most circumstances the case when the likelihood function is monotone, regardless increasing or decreasing, the resulting LRT is based on a critical value: the statistic is compared with some critical value and the rejection region is either when it is larger than or smaller than that.

2.2 *Generalized Likelihood Ratio*

In general, what we face is the situation in which under the null hypothesis the parameter belong to some set Θ_0 (above it was a single value set) while the alternative says that it belongs to some other set Θ_0^c. Let Θ denote the set of all a-priori possible value for θ (namely, the total parameter space), meaning that $\Theta_0 \cup \Theta_0^c = \Theta$. Then, the generalized likelihood ratio is defined by

$$\frac{\sup_{\theta\in\Theta_0} L(X_1,\cdots,X_n;\theta)}{\sup_{\theta\in\Theta} L(X_1,\cdots,X_n;\theta)}$$

Inspecting this ratio we can see that it is a fraction between 0 and 1. Generalized likelihood ratio test is of significance for composite hypotheses.

3 Composite Hypotheses and Test Conductions

3.1 *One-sided Alternative*

Let us go back to our leading example where $X \sim N(\mu, 1)$. Suppose now that H_1 is less specific and it says that $\mu > \mu_0$ or $\mu < \mu_0$. This is called a *one-sided alternative*. If we break that alternative to all its detailed options, our analysis will be as before. That is, $\frac{\overline{X}-\mu_0}{1/\sqrt{n}} \geq Z_{1-\alpha}$ or $\frac{\overline{X}-\mu_0}{1/\sqrt{n}} \leq Z_{1-\alpha}$. We can attain this by generalized Likelihood ratio (consider $\mu > \mu_0$ case). $\frac{L(X_1,\cdots,X_n;\mu_0)}{\sup_{\mu\geq\mu_0} L(X_1,\cdots,X_n;\mu)} = \begin{cases} e^{-\frac{n}{2}(\overline{X}-\mu_0)^2} & \mu_0 \leq \overline{X} \\ 1 & \mu_0 > \overline{X} \end{cases}$. This leads to 'automatic' acceptance if $X < \mu_0$ (as the LR equals 1). Then we can get that is: " reject H_0 if $\frac{\overline{X}-\mu_0}{1/\sqrt{n}} \geq Z_{1-\alpha}$ ".

3.2 *Two-sided Alternative*

3.2.1 For Normal Mean with Known Variance

Still being with the model dealt with above but suppose now that the alternative H_1 says that $\mu \neq \mu_0$. The test statistic of $\overline{X}$ will not lead to a monotone increasing or decreasing likelihood ratio. A sensible test will be based on the idea that the further away $\overline{X}$ is from μ_0, the more is the indication against H_0. Specifically, we look for a critical value C, where the test is: "reject H_0 if $\left|\frac{\overline{X}-\mu_0}{1/\sqrt{n}}\right| \geq Z_{1-\frac{\alpha}{2}}$". This can also be obtained from generalized LRTs, $\frac{L(X_1,\cdots,X_n;\mu_0)}{\sup_{\mu\in\mathbb{R}} L(X_1,\cdots,X_n;\mu)} = e^{-\frac{n}{2}(\overline{X}-\mu_0)^2}$, thus $(\overline{X}-\mu_0)^2 \geq \frac{\chi^2_{(1,1-\alpha)}}{n}$, thus $\left|\frac{\overline{X}-\mu_0}{1/\sqrt{n}}\right| \geq \sqrt{\chi^2_{(1,1-\alpha)}} = Z_{1-\frac{\alpha}{2}}$. In other words, the Z-test is a generalized LRT.

(*Remark.*) **Z-test**: suppose $X_1,\cdots,X_n \sim N(\mu,\sigma^2)$, and null hypothesis: $\mu = \mu_0$ with alternative $\mu \neq \mu_0$. For a test whose significance level is α, we can get Z-test: "reject H_0 if $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| \geq Z_{1-\frac{\alpha}{2}}$".

3.2.2 For Exponential Distribution

Suppose $X_i \sim exp(\lambda)$, $1 \le i \le n$, are n independent random variables. H_0 says that $\lambda = \lambda_0$ while H_1 says that $\lambda \neq \lambda_0$. The generalized LR is $\dfrac{\lambda_0^n e^{-\lambda_0 \sum\limits_{i=1}^{n} X_i}}{\widehat{\lambda}^n e^{-\widehat{\lambda} \sum\limits_{i=1}^{n} X_i}}$, up to a multiplicative constant, the rejection region is $\left(\lambda_0 \sum\limits_{i=1}^{n} X_i\right)^n e^{-\lambda_0 \sum\limits_{i=1}^{n} X_i} \le C$. This is unimodal single-peak function in $\sum\limits_{i=1}^{n} X_i$ (it is the kernel of density function of gamma distributed random variable with parameters $n+1$ and λ_0.). In summary, the test is to "reject H_0 if $\sum\limits_{i=1}^{n} X_i \le q_1$ or $\sum\limits_{i=1}^{n} X_i \ge q_2$". Since under H_0, $\sum\limits_{i=1}^{n} X_i \sim \Gamma_{(n,\lambda_0)}$, we have $F_{\Gamma_{(n,\lambda_0)}}(q_2) - F_{\Gamma_{(n,\lambda_0)}}(q_1) = 1 - \alpha$, which makes the significance level α.

3.3 *P-value and Ways to Perform Hypothesis Test*

3.3.1 P-value as Test Statistics

(*Defn.*) The ***p*-value** is the probability that the test statistic will take on a value that is *at least as extreme as* the observed value of the statistic when the null hypothesis H_0 is true.

(Namely, suppose the outcome of the sample was $T(X_1, \cdots, X_n) = t$, have in mind a fresh similar sample, $X_1', \cdots, X_n'$. The P-value of the test result is defined as $P_{H_0}\left[T(X_1', \cdots, X_n') \le t\right]$. Note that 'smaller' is taken here for convenience, we could have said that it is larger than C or belongs to some interval or region.)

To get a conclusion by p-value: given a significance level α, then if the p-value is *smaller* than α, we **reject** H_0; it the p-value is *larger* than α, there is **NOT enough evidence to reject** H_0.

As importantly, note that the P-value is the all important test statistic: It is a random variable, its distribution under H_0 is in our hand (i.e., **uniform distribution** $N(0,1)$ because $F_p(x) = P(\text{p-value} \le x) = P[P(T \le t) \le x] = P[F_T(t) \le x] = P[t \le F_T^{-1}(x)] = x$ under H_0), and it is this value which needs to be compared to the significance level in order to take the decision. All test statistics, regardless of what the parameter is, are collapsed into one: The P-value. This value needs to be compared with α. If smaller, reject.

3.3.2 The t-test

The **t-test** deals with the same set of hypotheses of Z-test above, but now the statistic is $\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$, which, under the null hypothesis, follows a t-distribution with $n-1$ degrees of freedom. Hence, we "reject if $\left|\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}\right| \ge t_{(n-1,1-\frac{\alpha}{2})}$". This is in fact a generalized LRT.

(*Remark.*) To prove that t-test is a generalized LRT, we need to consider both μ and σ^2 for making the likelihood function largest. Namely, The restricted MLE is $\left(\mu_0, \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i - \mu_0)^2\right)$, while the unrestricted MLE is $\left(\overline{X}, \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2\right)$.

3.3.3 Testing for Difference in the Means between Two Populations

Recall the point and interval estimations for the difference between means across two populations. The rest will follow in a similar fashion. Consider the paired sample case. Let H_0: $\mu_2 - \mu_1 = d$ and H_1: $\mu_2 - \mu_1 > d$, for some numerical value for d. Then a test whose significance level is $1 - \alpha$ is "reject H_0 if $\dfrac{\overline{D} - d}{S_d/\sqrt{n}} \geq t_{(n-1,1-\alpha)}$". This is the case since under H_0 the test statistic, which is $\dfrac{\overline{D} - d}{S_d/\sqrt{n}}$, follows a t-distribution with $n - 1$ degrees of freedom.

3.3.4 Testing for Equal Variances

Suppose there are two normal populations. The null hypothesis says that they have equal variances (i.e., $\dfrac{\sigma_2^2}{\sigma_1^2} = 1$), while the one-sided alternative says that $\dfrac{\sigma_2^2}{\sigma_1^2} > 1$. Denote by S_1^2 the sample variance at the first population based on a sample of size n_1, and likewise S_2^2 be the sample variance at the second based on a sample of size n_2. Then under the null hypothesis, $\dfrac{S_1^2}{S_2^2} = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(n_1-1,n_2-1)}$. Hence, a sensible test with a significance level of α is to "reject the null hypothesis if $\dfrac{S_1^2}{S_2^2} \geq F_{(n_1-1,n_2-1,1-\alpha)}$" This is clearly not a likelihood ratio test as the framework of this test does not apply here, but we claim without a proof that it is a generalized LRT where the parameter set is (σ_1^2, σ_2^2).

Additions:

Inference Difference in Means of Two Normal Distributions (unknown and unequal variances)

Welch's t-test

Two independent random samples from two Normal distributions

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2)$$
$$Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$$

$$\Rightarrow \quad \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t(r) \qquad r = \left\lfloor \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2} \right\rfloor$$

Inference on equality of two Population Proportions

$$\left.\begin{array}{l} X_1, X_2, \cdots, X_{n_1} \sim Bernoulli(p_1) \\ Y_1, Y_2, \cdots, Y_{n_2} \sim Bernoulli(p_2) \end{array}\right\} \rightarrow \quad \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

For large sample sizes (CLT)

Section V – Simple Linear Model

WANG Yuzhe*

May 6, 2022

1 The Simple Linear Model

Many problems in statistics and science involve exploring the relationships between two or more variables. **Regression analysis** is a statistical technique that is very useful for these types of problems. Let (x_i, y_i), $1 \leq i \leq n$, be a number of (different) points in the two dimensional plane. Drawing them you will get the so called ***scatter diagram***.

Note that in the previous section no probability model was assumed. Our derivation there belongs to the area of *descriptive statistics*. Our approach now will be different.

Based on the scatter diagram, it is reasonable to assume that the *mean* of the random variable Y is related to x by the following *simple linear regression model*: there exists three parameters a, b and $\sigma^2 > 0$, such that for any given x_i, there exists a random variable Y_i where $Y_i = ax_i + b + \varepsilon_i$, $i = 1, 2, \cdots, n$, where $\varepsilon_i \sim N(0, \sigma^2)$ and are i.i.d. (Y_i: **response**; x_i: **regressor/ predictor**; a: **slope**; b: **intercept**; ε_i: **random error**). Note that the *mean response* is that $E(Y|x) = ax + b$ and $V(Y|x) = \sigma^2$.

1.1 *Maximum Likelihood Estimates*

It is easy to see that the likelihood function here is $L(Y_1, \cdots, Y_n; a, b, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - ax_i - b)^2}{2\sigma^2}}$, because $Y_i \sim N(ax_i + b, \sigma^2)$, $1 \leq i \leq n$, (randomness of Y_i is fully captured by the random variable ε_i). Log-likelihood function with constant abandoned: $l(a, b, \sigma^2) = -n \ln(\sigma^2) - \sum_{i=1}^{n} (Y_i - ax_i - b)^2/\sigma^2$. Take gradient w.r.t. the three parameters respectively (estimators a, b free of σ^2), we get MLEs for a, b are

$$\widehat{a} = \frac{S_{xY}}{S_{xx}} \text{ and } \widehat{b} = \overline{Y} - \widehat{a}\overline{x}$$

which are for sure the regression line coeffcients (see in Section I).

In fact, as we show later, the two estimators are negatively (positively, respectively) correlated when $\overline{x} > 0$ ($\overline{x} < 0$, respectively). Note the MLEs of each of them in the case where the other is given are different.

Similarly, MLE for σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{a}x_i - \widehat{b})^2$$

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

We claim without a proof that $\widehat{\sigma^2}$ and $(\widehat{a}, \widehat{b})$ are independent. It is similar to what we have already said (without a proof) that in a normal population, S^2 and $\overline{X}$ are independent. We also claim without a proof that $\dfrac{\sum_{i=1}^{n}(Y_i - \widehat{a}x_i - \widehat{b})^2}{\sigma^2} \sim \chi^2_{(n-2)}$. Thus, $E(\widehat{\sigma^2}) = \dfrac{n-2}{n}\sigma^2$, indicating that $\dfrac{n\widehat{\sigma^2}}{n-2}$ is an UBE of σ^2.

1.2 *Confidence Interval for Parameters*

By observation, we can firstly get that both $\widehat{a}$ and $\widehat{b}$ are UBEs.

(*proof.*) $\widehat{a} = \dfrac{S_{xY}}{S_{xx}} = \sum_{i=1}^{n} \dfrac{x_i - \overline{x}}{S_{xx}} Y_i$. Then, $E(\widehat{a}) = \sum_{i=1}^{n} \dfrac{x_i - \overline{x}}{S_{xx}} E(Y_i) = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(ax_i + b)}{S_{xx}} = a$. And also $E(\widehat{b}) = E(\overline{Y}) - xE(\widehat{a}) = b$.

Observe that both estimators are linear functions of Y_i, $1 \le i \le n$, then $\widehat{a}$ and $\widehat{b}$ follow normal distribution. It is immediate to see that the coeffcient of Y_i in the estimator for a is $(x_i - \overline{x})/S_{xx}$ and for b is $(\overline{x}^2 - \overline{x}x_i)/S_{xx} + 1/n$, $1 \le i \le n$. Hence, $\mathrm{MSE}(\widehat{a}) = \mathrm{Var}(\widehat{a}) = \dfrac{\sigma^2}{S_{xx}}$, $\mathrm{MSE}(\widehat{b}) = \mathrm{Var}(\widehat{b}) = \left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right)\sigma^2 = \dfrac{\overline{x^2}\sigma^2}{S_{xx}}$ and $\mathrm{Cov}(\widehat{a}, \widehat{b}) = -\dfrac{\overline{x}}{S_{xx}}\sigma^2$. Then, $\widehat{a} \sim N\left(a, \dfrac{\sigma^2}{S_{xx}}\right)$ and $\widehat{b} \sim N\left(b, \dfrac{\overline{x^2}}{S_{xx}}\sigma^2\right)$.

As to $100(1-\alpha)\%$ CI for a and b, we can get it (take b as an instance) from $\dfrac{\widehat{b} - b}{\sigma\sqrt{\overline{x^2}/S_{xx}}} \sim N(0,1)$ when σ is in hand and $\dfrac{(\widehat{b} - b)}{\sqrt{\overline{x^2}/S_{xx}}\sqrt{\frac{n}{n-2}\widehat{\sigma^2}}} \sim t_{(n-2)}$. The width of confidence interval indicates the overall quality of regression line.

1.2.1 Confidence Interval on Mean Response

The **mean response**, as defined above, is $\mu_{Y|x} = E(Y|x)$. One estimator of mean response at $x = x_0$ is $\widehat{\mu}_{Y|x_0} = \widehat{b} + \widehat{a}x_0$. Then $E(\widehat{\mu}_{Y|x_0}) = b + ax_0 = \mu_{Y|x_0}$, and $\mathrm{Var}(\widehat{\mu}_{Y|x_0}) = x_0^2\mathrm{Var}(\widehat{a}) + 2x_0\mathrm{Cov}(\widehat{a}, \widehat{b}) + \mathrm{Var}(\widehat{b}) = \sigma^2\left[\dfrac{(\overline{x} - x_0)^2}{S_{xx}} + \dfrac{1}{n}\right]$. We claim without prove that $\widehat{a}x_0 + \widehat{b}$ follows a normal distribution (explanation: a linear function of Y_is), then the $(1-\alpha)$ CI is $\widehat{\mu}_{Y|x_0} - t_{(n-2,1-\frac{\alpha}{2})}\sqrt{\dfrac{n\widehat{\sigma^2}}{n-2}\left[\dfrac{(\overline{x} - x_0)^2}{S_{xx}} + \dfrac{1}{n}\right]} \le \mu_{Y|x_0} \le \widehat{\mu}_{Y|x_0} + t_{(n-2,1-\frac{\alpha}{2})}\sqrt{\dfrac{n\widehat{\sigma^2}}{n-2}\left[\dfrac{(\overline{x} - x_0)^2}{S_{xx}} + \dfrac{1}{n}\right]}$.

1.2.2 Prediction Interval

Note that the new observation x_{n+1} is independent of data used to build linear regression model. Suppose there is an interest in predicting $Y_{n+1} = ax_{n+1} + b + \varepsilon_{n+1}$, which is a new entrant to be considered after the size n sampled is conducted.

Similar to above, $\widehat{a}x_{n+1} + \widehat{b} \sim N\left(ax_{n+1} + b, \left[\dfrac{1}{n} + \dfrac{(x_{n+1} - \overline{x})^2}{S_{xx}}\right]\sigma^2\right)$, as the end random variable is a linear function of the normal errors ε_i, $1 \le i \le n$. Also, in the case where σ^2 is not

known, it can be replaced with $\frac{n}{n-2}\widehat{\sigma^2}$, but now the underlying distribution is t with $n-2$ degrees of freedom.

1.2.3 An Alternative Perspective

Sometimes, in order to simplify the model and computation, we may write $Y_i = \beta + \alpha(x_i - \overline{x}) + \varepsilon_i$. In this case, $\widehat{\alpha} = \frac{S_{xY}}{S_{xx}}$ and $\widehat{\beta} = \overline{Y}$. At this time, $\widehat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{S_{xx}}\right)$ and $\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{n}\right)$. From that we can get the corresponding $100(1-\alpha)\%$ confidence intervals.

1.3 *Hypothesis Test on Regression Parameters*

Suppose one wishes to test the null hypothesis that $a = a_0$ v.s. the alternative that $a \neq a_0$ (take a as an instance). If σ^2 is in hands, we will "reject H_0 if $\left|\frac{\widehat{a} - a_0}{\sigma/\sqrt{S_{xx}}}\right| \geq Z_{1-\frac{\alpha}{2}}$". If σ^2 is unknown, we will "reject H_0 if $\left|\frac{\widehat{a} - a_0}{\sqrt{n\widehat{\sigma^2}/(n-2)S_{xx}}}\right| \geq t_{(n-2,1-\frac{\alpha}{2})}$".

1.3.1 Testing Correlation: Independence

Let X, Y have a bivariate normal distribution (i.e., the joint pdf is that $f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right)$). Observe n pairs of data (each pair is sampled from the bivariate normal distribution and independent from other pairs): $(X_1, Y_1), \cdots, (X_n, Y_n)$. Consider the null hypothesis (about the correlation) $\rho = 0$ v.s. the alternative $\rho \neq 0$ (or $\rho > 0$ or $\rho < 0$).

Rewrite them as $Y_i = c_0 + c_1 X_i + \epsilon_i$. Then, under null hypothesis $\rho = 0 \Rightarrow c_1 \mathrm{Var}(X_i) = 0$. Define **sample correlation coefficient**: $R = \frac{S_{XY}}{S_X S_Y}$ (with sample covariance $S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$ and sample variances S_X, S_Y). Conditioned on $X_1, \cdots, X_n$ (given $X_1 = x_1, \cdots, X_n = x_n$), under H_0: $\rho = 0 \Longrightarrow c_1 = 0$, $\widehat{c_1} = \frac{S_Y}{S_x} R \sim N\left(0, \frac{\sigma_Y^2}{S_{xx}}\right)$. Because $\frac{n\widehat{\sigma_Y^2}}{\sigma_Y^2} \sim \chi^2_{(n-2)}$, and $\frac{n\widehat{\sigma_Y^2}}{\sigma_Y^2} = \frac{\sum_{i=1}^{n}\left[Y_i - \overline{Y} - \left(\frac{S_Y}{S_x}R\right)(x_i - \overline{x})\right]^2}{\sigma_Y^2} = \frac{(n-1)S_Y^2(1-R^2)}{\sigma_Y^2}$. Thus, $\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{(n-2)}$.

Then, to test H_0: $\rho = 0$, we have two alternative options: (i) compute $r = \frac{S_{xy}}{S_x S_y}$; (ii) Way I: compute $r\sqrt{\frac{n-2}{1-r^2}}$ and "reject H_0 if $r\sqrt{\frac{n-2}{1-r^2}} \geq t_{(n-2,1-\frac{\alpha}{2})}$". or (iii) Way II: "reject H_0 if $r \geq r_{(n-2,1-\frac{\alpha}{2})}$", by checking the table of CDF of R.

1.3.2 Testing Correlation: Dependence

When it comes to testing null hypothesis H_0: $\rho = \rho_0$, for some $\rho_0 \neq 0$. It turns out that we can prove that approximately $\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$. As a result, we can use $\frac{\ln\frac{1+R}{1-R} - \ln\frac{1+\rho_0}{1-\rho_0}}{2\sqrt{1/(n-3)}}$ and compare it with $Z_{1-\frac{\alpha}{2}}$.

2 Best Linear Unbiased Estimator (BLUE)

2.1 *Sample Mean as the BLUE*

The point estimator $\overline{X}$ for the mean of a given distribution was used frequently above. At most times it was the MLE. But the derivation was always model based in the sense that we assumed some parameter based family of distributions. We next claim that $\overline{X}$ is the best UBE for the mean among all linear functions of the data.

(*proof.*) Note that we do not impose any model. All needed is to assume that the observations are i.i.d. Assume X_i, $1 \leq i \leq n$, are n independent with a common mean μ and a common variance σ^2. Indeed, as shown in Section II.3.2, $\mathrm{Var}(\sum_{i=1}^{n} w_i X_i) = \sigma^2 \sum_{i=1}^{n} w_i^2$ under the constraint $\sum_{i=1}^{n} w_i = 1$, is minimize when $w_i = \frac{1}{n}$, $1 \leq i \leq n$.

2.2 *The Gauss-Markov Theorem*

Back to the simple linear model, suppose we relaxed the model assumption as now that

* The errors ε_i are with $E(\varepsilon_i) = 0, 1 \leq i \leq n$, and $E(\varepsilon_i \varepsilon_j) = 0, 1 \leq i \neq j \leq n$. In particular, the errors are uncorrelated.

* $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $1 \leq i \leq n$, for some σ^2 (denoted as homoscedastic errors).

then the pair of estimators derived above for the parameters a and b, are **BLUE**, namely they are *best (in terms of MSE) among all UBEs which are also linear functions of the sample* (in our case of $(Y_1, \cdots, Y_n)$). This known as **the Gauss-Markov theorem**.

(*proof.*) Denote the linear estimator by $\sum_{i=1}^{n} w_i Y_i$, then the mission becomes to minimize $\mathrm{Var}(\sum_{i=1}^{n} w_i Y_i) = \sigma^2 \sum_{i=1}^{n} w_i^2$ subject to $E(\sum_{i=1}^{n} w_i Y_i) = a$ or b. Take the estimator for b as an instance. Then, $E(\sum_{i=1}^{n} w_i Y_i) = b \Leftrightarrow \sum_{i=1}^{n} w_i = 1 \& \sum_{i=1}^{n} w_i x_i = 0$. By considering the Lagrangian function $L(w_1, \cdots, w_n; \lambda_1, \lambda_2) = \frac{1}{2} \sum_{i=1}^{n} w_i^2 + \lambda_1 (1 - \sum_{i=1}^{n} w_i) - \lambda_2 \sum_{i=1}^{n} w_i x_i$, we get $w_i = \lambda_1 + \lambda_2 x_i$, $1 \leq i \leq n$. And two Lagrangian multipliers $\lambda_1 = -\frac{\overline{x}}{S_{xx}}$; $\lambda_2 = \frac{\overline{x^2}}{S_{xx}}$, which gives $\widehat{b}$ as above shows.

3 Multiple Linear Regression

Sometimes we may have multiple variables $x_1, x_2, \cdots, x_k$, which all as predictors participate in deciding the response Y. For example, house price and house size are both related to property tax. To deal with these types of problems, we design multiple linear regression model, with the form $Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \epsilon$.

Estimating coefficiences by least-square, we get $\frac{\partial}{\partial \beta_i} \sum_{i=1}^{n} \varepsilon_i^2 \Big|_{\widehat{\beta}_i} = 0,\ i = 0, 1, \cdots, n.$ Thus,

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \cdots & \sum_{i=1}^n x_{ik}x_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}$$

For simplicity, we can get that $\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ (matrix approach). Estimating variance, we get that $\widehat{\sigma^2} = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \sum_{i=1}^{k}\widehat{\beta}_i x_i)^2$.

We can show in the same way above that $E(\widehat{\beta}) = \beta$, which means that β is an UBE. **Covariance matrix** of the regression coeffcient is $\mathrm{Cov}(\widehat{\beta}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$, then **estimated standard error** of $\widehat{\beta}_j$ is $se(\widehat{\beta}_j) = \sqrt{\widehat{\sigma^2}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}}$. From this we can get the test for β_j (e.g., $\beta_j = \beta_{j0}$ v.s. $\beta_j \neq \beta_{j0}$) and CIs by test statistics $T_0 = \frac{\widehat{\beta}_j - \beta_{j0}}{se(\widehat{\beta}_j)} \sim t_{(n-k-1)}$. For mean response, using $\sqrt{\widehat{\sigma^2}\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}$ to replace $se(\widehat{\beta}_j)$ above.

3.1 *Deal with Categorical Variables*

All examples exhibited above are about quantitative variables, like temperature, pressure, distance or voltage. Nonetheless, there are some categorical variables (qualitative), like location, description, type or gender. To represent these variables, we use the **dummy variables** (indicator functions), which is $x = \begin{cases} 0, & \text{type I} \\ 1, & \text{type II} \end{cases}$.

In general, a qualitative variable with r-levels can be modeled by $r-1$ indicator variable. (e.g. for $r = 3$ case, we use (x_1, x_2), $(0,0)$ to represent type I, $(1,0)$ for type II and $(0,1)$ for type III). A real example is shown as follows.

EXAMPLE 12-13 Surface Finish

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown in Table 12-15. Note that the data have been collected using two different types of cutting tools. Since the type of cutting tool likely affects the surface finish, we will fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y is the surface finish, x_1 is the lathe speed in revolutions per minute, and x_2 is an indicator variable denoting the type of cutting tool used; that is,

$$x_2 = \begin{cases} 0, \text{for tool type 302} \\ 1, \text{for tool type 416} \end{cases}$$

Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool
1	45.44	225	302
2	42.03	200	302
3	50.10	250	302
4	48.75	245	302
5	47.92	235	302
6	47.79	237	302
7	52.26	265	302
8	50.52	259	302
9	45.58	221	302
10	44.78	218	302
11	33.50	224	416
12	31.23	212	416
13	37.52	248	416
14	37.13	260	416
15	34.70	243	416
16	33.92	238	416
17	32.13	224	416
18	35.47	251	416
19	33.49	232	416
20	32.29	216	416

The fitted model is

$$\hat{y} = 14.27620 + 0.14115x_1 - 13.28020x_2$$

Section VI – Analysis of Variance (ANOVA)

WANG Yuzhe*

May 11, 2022

1 One-way ANOVA

The technique of analysis of variance (ANOVA) deals with comparison between a number of populations. Note that the special cases where the number of populations is two were dealt with in the previous sections. Equivalently, we can say that we try to explain the variability of a single population with a qualitative variable.

Suppose m populations exist and we are interested in inspecting one numerical characteristic combining them. The levels of the factor are sometimes called **treatment**s, and each treatment has some **observation**s or **replicate**s. Now we have m treatments, with the i^{th} treatment resulting in observations sampled from a normal distribution $N(\mu_i, \sigma^2)$. Suppose totally there are n_i samples from the i^{th} treatment and they are i.i.d., and $X_{ij} \sim N(\mu_i, \sigma^2)$, $j = 1, \cdots, n_i$.

These random samples are drawn in order to estimate (with point or interval estimators) some relationships between the population means or in order to test hypotheses of the type "*all means are identical*" or the "*mean in population 1 equals the average between the means of population 2 and 3*". We denote the i^{th} sample mean $\sum_{j=1}^{n_i} \frac{X_{ij}}{n_i}$, by $\overline{X_i}$. Clearly, $\overline{X_i} \sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right)$.

Let $\mathbf{a} \in \mathbb{R}^m$ be some vector and suppose there is an interest in the parameter $\mathbf{a}^T\mu = \sum_{i=1}^{m} a_i\mu_i$. A point estimation of this parameter is of course $\sum_{i=1}^{m} a_i\overline{X_i} \sim N\left(\mathbf{a}^T\mu, \sum_{i=1}^{m} a_i^2\frac{\sigma^2}{n_i}\right)$. Hence, had σ^2 been given, we have no problem to construct $1-\alpha$ confidence interval for $\mathbf{a}^T\mu$ or checking some hypothesis of the type H_0: $\mathbf{a}^T\mu = 0$. When, σ^2 is not in hand, we will use **pooled** sample variance $S_p^2 = \sum_{i=1}^{m} \frac{(n_i-1)S_i^2}{N-m}$, where $N = \sum_{i=1}^{m} n_i$ is the total size of observations. At this time $(N-m)\frac{S_p^2}{\sigma^2} \sim \chi^2_{(N-m)}$, one $1-\alpha$ CI for $\mathbf{a}^T\mu$ is $\sum_{i=1}^{m} a_i\overline{X_i} - S_p\sqrt{\sum_{i=1}^{m} \frac{a_i^2}{n_i}}t_{(N-m,1-\frac{\alpha}{2})} \le \mathbf{a}^T\mu \le \sum_{i=1}^{m} a_i\overline{X_i} + S_p\sqrt{\sum_{i=1}^{m} \frac{a_i^2}{n_i}}t_{(N-m,1-\frac{\alpha}{2})}$. In the case where $\mathbb{1}^T\mathbf{a}$ the function $\mathbf{a}^T\mu$ is called a **contrast**.

1.1 *Testing for Homogeneity*

Apparently, the most important to be tested null hypothesis is the one which says that "all μ_is, $1 \le i \le m$, are identical" against the alternative which says that this is not the case.

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

Then, the MLE for the common mean, which is denoted by $\widehat{\mu}$ under the null hypothesis, is $\widehat{\mu} = \sum_{i=1}^{m} \frac{n_i}{N} \overline{X_i}$, while MLEs for population i is $\widehat{\mu_i} = \overline{X_i}$, $1 \le i \le m$.

Using $\widehat{\mu}$ and $\widehat{\mu_i}$ to invoke μ_i, we get the generalized LRT is $\frac{\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu})^2/\sigma^2\right]}{\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu_i})^2/\sigma^2\right]}$, and then we get that the generalized LRT statistic is in fact $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu_i})^2 - \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu})^2$.

Claim: $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu})^2 = \sum_{i=1}^{m} n_i(\widehat{\mu_i}-\widehat{\mu})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu_i})^2$.

Note that it resembles variance (in Section I) as here $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu})^2$ (total sum-of-squares, aka, **SSTO**) is decomposed to two additive terms, $\sum_{i=1}^{m} n_i(\widehat{\mu_i}-\widehat{\mu})^2$ (between-treatment sum-of-squares, aka, **SST**) reflects the variability *between* groups, while the $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu_i})^2$ (error sum-of-squares, aka, **SSE**) reflects the variability within groups. Thus, the above equation becomes $\boldsymbol{SSTO} = \boldsymbol{SST} + \boldsymbol{SSE}$.

(*proof.*) $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu_i}+\widehat{\mu_i}-\widehat{\mu})^2 = \sum_{i=1}^{m} n_i(\widehat{\mu_i}-\widehat{\mu})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\widehat{\mu_i})^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n_i} n_i(\widehat{\mu_i}-\widehat{\mu})(X_{ij}-\widehat{\mu_i})$. The last term equals 0 because $\sum_{j=1}^{n_i} n_i(\widehat{\mu_i}-\widehat{\mu})(X_{ij}-\widehat{\mu_i}) = n_i(\widehat{\mu_i}-\widehat{\mu})\sum_{j=1}^{n_i}(X_{ij}-\overline{X_i}/n_i) = 0$, $\forall\ 1 \le i \le m$.

Assuming normal populations, then under the null hypothesis, $\boldsymbol{SSTO}/\sigma^2 \sim \chi^2_{(N-1)}$, and $\boldsymbol{SSE}/\sigma^2 \sim \chi^2_{(N-m)}$. Note that the there is no need to invoke the null hypothesis assumption in order to get the latter. We do not prove but claim that under normal population conditions that **SST** and **SSE** are independent (resembles the fact that $\overline{X}$ and S^2 are independent). Then, $\boldsymbol{SST}/\sigma^2 \sim \chi^2_{(m-1)}$ (we can also directly use the expression to show this result).

Hence, assume σ^2 is known, a test for homogeneity of treatments is to " reject H_0 if $\boldsymbol{SST}/\sigma^2 \ge \chi^2_{(m-1,1-\alpha)}$". Our final comment deals with the case where the population variance is not given, because **SSE** and **SST** are independent, we get $\frac{\boldsymbol{SST}/(m-1)}{\boldsymbol{SSE}/(N-m)} \sim F_{(m-1,N-m)}$. Thus, the test with significance level of α is to "reject the null-hypothesis if and only if $\frac{\boldsymbol{SST}/(m-1)}{\boldsymbol{SSE}/(N-m)} \ge F_{(m-1,N-m,1-\alpha)}$".

Claim: the test based on F distribution is also the generalized LRT. To prove it, just consider the variable set $(\mu_1, \cdots, \mu_m, \sigma^2)$ instead of only means, using $\boldsymbol{SSTO}/(N-1)$ and $\boldsymbol{SSE}/(N-m)$ to invoke σ^2 for the numerator and denominator, respectively.

We conclude with one-way ANOVA table:

One-way ANOVA Table				
Source	Sum of Squares (SS)	Degrees of Freedom	Mean Suqare (MS)	F Ratio
Treatment	**SST**	$m-1$	$\boldsymbol{MST} = \boldsymbol{SST}/(m-1)$	$\boldsymbol{MST}/\boldsymbol{MSE}$
Error	**SSE**	$N-m$	$\boldsymbol{MSE} = \boldsymbol{SSE}/(N-m)$	
Total	**SSTO**	$N-1$		

2 Two-way ANOVA

Assume now there are two factors (attributes), one of which has a levels and the other b levels. There are totally thus $n = ab$ possible combinations. Assume for the $(i,j)^{\text{th}}$ combination, data are sampled from a distribution $X_{(i,j)} \sim N(\mu_{ij}, \sigma^2)$. Note that *only one sample* from the $(i,j)^{\text{th}}$ combination of treatment. Using row variable α_i and column variable β_j to control the change of μ_{ij}, let $\mu_{ij} = \mu + \alpha_i + \beta_j$, w.l.o.g., let $\sum_{i=1}^{a} \alpha_i = 0 = \sum_{j=1}^{b} \beta_j$.

Our aim now is to test the row and column effects on μ, for instance, is there any effect on μ due to row? (i.e., H_{0A}: $\alpha_1 = \cdots = \alpha_a = 0$ and H_{1A}: $\neg H_{0A}$); is there any effect on μ due to column? (i.e., H_{0B}: $\beta_1 = \cdots = \beta_b = 0$ and H_{1B}: $\neg H_{0B}$);

Some notations: (sample means) $\overline{X_{i\cdot}} = \frac{1}{b}\sum_{j=1}^{b} X_{ij}$; $\overline{X_{\cdot j}} = \frac{1}{a}\sum_{i=1}^{a} X_{ij}$; $\overline{X_{\cdot\cdot}} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}$. (*Overall/Total Sum-of-squares*) $\boldsymbol{SSTO} = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \overline{X_{\cdot\cdot}})^2$; (*Sum-of-squares of Factor A/row factor*) $\boldsymbol{SSA} = b\sum_{i=1}^{a}(\overline{X_{i\cdot}} - \overline{X_{\cdot\cdot}})^2$. (*Sum-of-squares of Factor B/column factor*) $\boldsymbol{SSB} = a\sum_{i=1}^{b}(\overline{X_{\cdot j}} - \overline{X_{\cdot\cdot}})^2$. (*Error Sum-of-squares*) $\boldsymbol{SSE} = \sum_{i=1}^{a}\sum_{i=1}^{b}\left[X_{ij} - (\overline{X_{i\cdot}} - \overline{X_{\cdot\cdot}}) - (\overline{X_{\cdot j}} - \overline{X_{\cdot\cdot}}) - \overline{X_{\cdot\cdot}}\right]^2$.

Relationship: **SSTO** = **SSA** + **SSB** + **SSE**. Moreover, $\boldsymbol{SSTO} \sim \chi^2_{(ab-1)}$; $\boldsymbol{SSA} \sim \chi^2_{(a-1)}$; $\boldsymbol{SSB} \sim \chi^2_{(b-1)}$ and $\boldsymbol{SSE} \sim \chi^2_{(a-1)(b-1)}$. **SSA**, **SSB** and **SSE** are mutually independent. We can then generate test hypotheses. Table for two-way ANOVA is as follows.

Table 9.4-1 Two-way ANOVA table, one observation per cell

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F
Factor A (row)	SS(A)	$a-1$	$\text{MS(A)} = \frac{\text{SS(A)}}{a-1}$	$\frac{\text{MS(A)}}{\text{MS(E)}}$
Factor B (column)	SS(B)	$b-1$	$\text{MS(B)} = \frac{\text{SS(B)}}{b-1}$	$\frac{\text{MS(B)}}{\text{MS(E)}}$
Error	SS(E)	$(a-1)(b-1)$	$\text{MS(E)} = \frac{\text{SS(E)}}{(a-1)(b-1)}$	
Total	SS(TO)	$ab-1$		

2.1 *Generalize: with Over 1 Observations per Cell*

Similar to above case, but for now we have for the $(i,j)^{\text{th}}$ combination, we have $c > 1$ observations that are sampled from a distribution $X_{(i,j),k} \sim N(\mu_{ij}, \sigma^2)$, $k = 1.\cdots, c$. Currently, the model for μ_{ij} becomes $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, where the last term indicates **intraction effects** of combination (i,j). Some notations and relationships are listed as follows.

Sample mean of the (i,j)-th interaction effect $\overline{X}_{ij\cdot} = \frac{1}{c}\sum_{k=1}^{c} X_{ijk}$,

Sample mean of the i-th row factor $\overline{X}_{i\cdot\cdot} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk}$,

Sample mean of the j-th column factor $\overline{X}_{\cdot j\cdot} = \frac{1}{ac}\sum_{i=1}^{a}\sum_{k=1}^{c} X_{ijk}$,

Overall sample mean $\overline{X}_{\cdots} = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk}$

Overall sum-of-squares or total sum-of-squares

$$SS(TO) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \overline{X}...)^2$$

Sum-of-squares of Factor A (row factor)

$$SS(A) = bc\sum_{i=1}^{a}(\overline{X}_{i..} - \overline{X}...)^2$$

Sum-of-squares of Factor B (column factor)

$$SS(B) = ac\sum_{j=1}^{b}(\overline{X}_{.j.} - \overline{X}...)^2$$

Interaction sum-of-squares

$$SS(AB) = c\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}...)^2$$

Error sum-of-squares

$$SS(E) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \overline{X}_{ij.})^2$$

Relations

$$SS(TO) = SS(A) + SS(B) + SS(E) + SS(AB)$$

If H_{0A}, H_{0B} and H_{0AB} are true, then

$$H_{0AB}: \gamma_{ij} = 0, \text{ for all } i, j, \quad H_{1AB}: H_{0AB} \text{ is not true } \underbrace{\frac{SS(TO)}{\sigma^2}}_{\sim\chi^2(abc-1)} = \underbrace{\frac{SS(A)}{\sigma^2}}_{\sim\chi^2(a-1)} + \underbrace{\frac{SS(B)}{\sigma^2}}_{\sim\chi^2(b-1)} + \underbrace{\frac{SS(E)}{\sigma^2}}_{\sim\chi^2(ab(c-1))} + \underbrace{\frac{SS(AB)}{\sigma^2}}_{\sim\chi^2((a-1)(b-1))},$$

Table 9.4-4 Two-way ANOVA table, c observations per cell

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F
Factor A (row)	SS(A)	$a-1$	$\text{MS(A)} = \frac{\text{SS(A)}}{a-1}$	$\frac{\text{MS(A)}}{\text{MS(E)}}$
Factor B (column)	SS(B)	$b-1$	$\text{MS(B)} = \frac{\text{SS(B)}}{b-1}$	$\frac{\text{MS(B)}}{\text{MS(E)}}$
Factor AB (interaction)	SS(AB)	$(a-1)(b-1)$	$\text{MS(AB)} = \frac{\text{SS(AB)}}{(a-1)(b-1)}$	$\frac{\text{MS(AB)}}{\text{MS(E)}}$
Error	SS(E)	$ab(c-1)$	$\text{MS(E)} = \frac{\text{SS(E)}}{ab(c-1)}$	
Total	SS(TO)	$abc-1$		

2.2 *Three-way ANOVA*

Table for three-way ANOVA is as follows for an intial taste.

Table 9.5-1 ANOVA table

Source	SS	d.f.	MS	F
A	SS(A)	$a-1$	MS(A)	MS(A)/MS(E)
B	SS(B)	$b-1$	MS(B)	MS(B)/MS(E)
C	SS(C)	$c-1$	MS(C)	MS(C)/MS(E)
AB	SS(AB)	$(a-1)(b-1)$	MS(AB)	MS(AB)/MS(E)
AC	SS(AC)	$(a-1)(c-1)$	MS(AC)	MS(AC)/MS(E)
BC	SS(BC)	$(b-1)(c-1)$	MS(BC)	MS(BC)/MS(E)
ABC	SS(ABC)	$(a-1)(b-1)(c-1)$	MS(ABC)	MS(ABC)/MS(E)
Error	SS(E)	$abc(d-1)$	MS(E)	
Total	SS(TO)	$abcd-1$		

Three-way ANOVA Table

Section VII – Chi-square Goodness of Fit

WANG Yuzhe*

May 13, 2022

1 Goodness of Fit

The hypothesis-testing procedures that we have discussed in previous sections are mostly designed for problems in which the population or probability distribution is known and the hypotheses involve the parameters of the distribution. Nonetheless, under most circumstances in our real life, we do NOT know the underlying distribution of the population, and we wish to test the hypothesis w.r.t. that population (e.g., whether a particular distribution will be satisfactory as a population model).

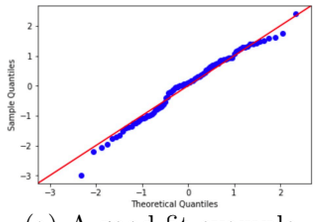
Consider a special case, namely, we want to check if our data is statistically similar to a normal distribution. **Order statistics** and **QQ-plot** (**quantiles-quantiles plot**) can be used as a direct and intuitive method.

Let $X_1, \cdots, X_n$ be i.i.d. random samples from a common distribution f. Denote $X_{(k)}$ as the k^{th} smallest one in these samples. CDF of $X_{(k)}$: $F_{X_{(k)}}(x) = P(X_{(k)} \leq x) = P(\text{at least } k \text{ samples } \leq x) = \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1-F(x)]^{(n-r)}$, thus pdf of $X_{(k)}$ is $f_{X_{(k)}} = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{(k-1)}[1-F(x)]^{(n-k)} f(x)$. (e.g., if $Y_i \sim U[0,1]$, $Y_{(k)} \sim \beta_{(k,n+1-k)}$, $E(Y_{(k)}) = \frac{k}{n+1}$).

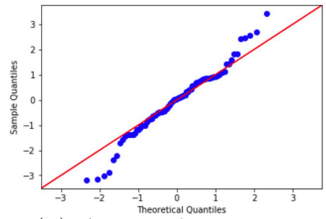
As we all know, p^{th} **theoretical quantile** of the distribution F is $\pi_p = F^{-1}(p)$ and $F(X_i) \sim U[0,1]$, $\forall 1 \leq i \leq n$. Then, $E[F(X_{(k)})] = \frac{k}{n+1}$ for arbitrary F, which means if $p = \frac{k}{n+1}$ for some $1 \leq k \leq n$, $X_{(k)}$ can be an UBE of π_p.

For fixed p, let $r = (n+1)p$, we derive a linear combination of estimators for π_p, that is, $\widehat{\pi_p} = X_{\lfloor r \rfloor} + (r - \lfloor r \rfloor)(X_{\lfloor r \rfloor + 1} - X_{\lfloor r \rfloor})$. Hence, if the data is indeed statistically similar to some given distribution, we expect that $\widehat{\pi}_{\frac{i}{n+1}} = x_{(i)} \approx \pi_{\frac{i}{n+1}}$. If we plot a scatterplot of the pairs $(\pi_{\frac{i}{n+1}}, x_{(i)})$, it should be close to the line $y = x$.

Some examples are listed as follows. (Hypothesis: the data comes from standard normal)



(a) A good fit example



(b) A poor fit example

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

As we can see above, **QQ-plot** is too intuitive to work as a persuasive enough evidence. A more formal way to test the hypothesis is to use **Chi-square goodness of fit test**.

1.1 *The Two-cell Case*

Considering the test for population proportion with H_0: $p = p_0$. In previous sections we use $Y = \dfrac{N - np_0}{\sqrt{np_0(1-p_0)}}$ and by invoking CLT, the test was "to reject H_0 if $Y \geq Z_{1-\frac{\alpha}{2}}$". An equivalent test will hence be "to reject H_0 if $Y^2 \geq \chi^2_{(1,1-\alpha)}$", because $Y^2 \sim \chi^2_{(1)}$ under the assumption of CLT. Thus, $\dfrac{(N-np_0)^2}{np_0} + \dfrac{[(n-N)-n(1-p_0)]^2}{n(1-p_0)} \sim \chi^2_{(1)}$.

1.2 *Any Number of Cells*

Suppose there are some m categories, so each of the individuals in the sample belongs to one and only one of the m cells. An individual belongs to cell i with probability $p_i \geq 0$, where $\sum_{i=1}^{m} p_i = 1$ (mutinomial distribution), The null-hypothesis says that $p_i = p_{i0} \geq 0$, $1 \leq i \leq m$, where $\sum_{i=1}^{m} p_{i0} = 1$. Let O_i be the random variable counting how many are in cell i (**observed/actual** value in cell i), $1 \leq i \leq m$, and E_i denote the **expected** value there (under H_0, $E_i = np_{i0}$). Define test statistic $\chi^2 = \sum_{i=1}^{m} \dfrac{(O_i - E_i)^2}{E_i}$, which converges in distribution to $\chi^2_{(m-1)}$ by invoking CLT (comparing with 2 cell case).

Hence, a test with significance level of α is "to reject the null hypothesis if $\chi^2 \geq \chi^2_{(m-1,1-\alpha)}$". Alternatively, one can derive the P-value once the realization of the test statistic is in one's hand. A possible use of this test for inferring with regard to a continuous random variable is to decompose its possible values into a number of intervals and consider each interval as a cell.

1.3 *Checking for Normality*

As we mentioned above, Chi-square goodness of fit test can be used for formally testing whether some data are from a target distribution. Here, we take normal distribution as an instance.

Suppose one wishes to check if a variable under consideration at a given population is normally distributed. Note that we do not specify the corresponding parameter μ and σ^2. They will be estimated, and as claimed, this fact *reduces the number of degree of freedom by two*.

Suppose a sample of n was conducted and $\overline{X}$ and S^2 turned out to equal $\overline{x}$ and s^2, respectively. Divide the real line into ten equally likely intervals. Thus, look for the eleven edges $Z_0 = 0, Z_{0.1}, Z_{0.2}, \cdots, Z_{0.9}, Z_1 = 1$. Cell i comes with all sampled observation X_j which obey $Z_{\frac{i-1}{10}} \leq \dfrac{X_j - \overline{x}}{s} \leq Z_{\frac{i}{10}}$, $1 \leq i \leq 10$. We can then conduct $\chi^2 = \sum_{i=1}^{10} \dfrac{(O_i - n/10)^2}{n/10}$. A test with significance leve α shows that we had better "reject H_0 if $\chi^2 \geq \chi^2_{(7,1-\alpha)}$".

Note that under the null, $\chi^2 \sim \chi^2_{(m-1-d)}$, where d is the number of parameters that have been estimated (estimators used).

2 Contingency Tables

2.1 *Testing for Homogeneity*

Suppose there exist two populations and each of the individuals in both population can fall in one out of a common set of k cells. Suppose the null hypothesis says that p_{1j}, $1 \le j \le k$, are the probabilities at the first population, while p_{2j}, $1 \le j \le k$, are the corresponding ones at the second. Samples of sizes n_1 (from the former population) and n_2 (from the latter) are performed. Let N_{ij} denote the random number of observations from sample i in cell j, we then get the test statistic $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(N_{ij} - n_i p_{ij})^2}{n_i p_{ij}} \sim \chi^2_{(2k-2)}$ by CLT according to above.

If now H_0 becomes $p_{1j} = p_{2j}$, $1 \le j \le k$, with no real value p in hand, we need to estimate $p_j \overset{\triangle}{=\!=} p_{1j} = p_{2j}$ by $\widehat{p_j} = \frac{N_{1j} + N_{2j}}{n_1 + n_2}$, $1 \le j \le k$. Note that it says that with respect to the partition specified by this qualitative variable, the two populations are identical.

A contingency table is the derived for the construction of test statistic $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(N_{ij} - n_i\widehat{p}_i)^2}{n_i\widehat{p}_i} \sim \chi^2_{k-1}$. Note that after $\widehat{p_1}, \cdots, \widehat{p}_{k-1}$ is in hand, $\widehat{p_k}$ is for sure known because the summation is 1, thus we only estimate $k-1$ parameters.

	A_1	A_2	A_3	⋯	A_k	Total
First experiment	$Y_{1,1}$ (Observed) $n_1\hat{p}_1$ (expected)	$Y_{1,2}$ $n_1\hat{p}_2$	$Y_{1,3}$ $n_1\hat{p}_3$	⋯	$Y_{1,k}$ $n_1\hat{p}_k$	n_1
Second experiment	$Y_{2,1}$ $n_2\hat{p}_1$	$Y_{2,2}$ $n_2\hat{p}_2$	$Y_{2,3}$ $n_3\hat{p}_3$	⋯	$Y_{2,k}$ $n_2\hat{p}_k$	n_2
Total	$Y_{1,1}+Y_{2,1}$ $\left(\hat{p}_1 = \frac{Y_{1,1}+Y_{2,1}}{n_1+n_2}\right)$	$Y_{1,2}+Y_{2,2}$ $\left(\hat{p}_2 = \frac{Y_{1,2}+Y_{2,2}}{n_1+n_2}\right)$	$Y_{1,3}+Y_{2,3}$ $\left(\hat{p}_3 = \frac{Y_{1,3}+Y_{2,3}}{n_1+n_2}\right)$		$Y_{1,k}+Y_{2,k}$ $\left(\hat{p}_k = \frac{Y_{1,k}+Y_{2,k}}{n_1+n_2}\right)$	n_1+n_2

Generalize to h populations with k cells, we can use the same method to derive test statistic and contingency table. At this time, $\chi^2 \sim \chi^2_{(h-1)(k-1)}$.

	A_1	A_2	A_3	⋯	A_k	Total
First	$Y_{1,1}$ (Observed) $n_1\hat{p}_1$ (expected)	$Y_{1,2}$ $n_1\hat{p}_2$	$Y_{1,3}$ $n_1\hat{p}_3$	⋯	$Y_{1,k}$ $n_1\hat{p}_k$	n_1
Second	$Y_{2,1}$ $n_2\hat{p}_1$	$Y_{2,2}$ $n_2\hat{p}_2$	$Y_{2,3}$ $n_3\hat{p}_3$	⋯	$Y_{2,k}$ $n_2\hat{p}_k$	n_2
…	…	…	…		…	…
h-th	$Y_{h,1}$ $n_h\hat{p}_1$	$Y_{h,2}$ $n_h\hat{p}_2$	$Y_{h,3}$ $n_h\hat{p}_3$		$Y_{h,k}$ $n_h\hat{p}_k$	n_k
Total	$Y_{1,1}+\cdots+Y_{h,1}$ $\hat{p}_1 = \frac{Y_{1,1}+\cdots+Y_{h,1}}{n_1+\cdots+n_h}$	$Y_{1,2}+\cdots+Y_{h,2}$ $\hat{p}_2 = \frac{Y_{1,2}+\cdots+Y_{h,2}}{n_1+\cdots+n_h}$	$Y_{1,3}+\cdots+Y_{h,3}$ $\hat{p}_3 = \frac{Y_{1,3}+\cdots+Y_{h,3}}{n_1+\cdots+n_h}$		$Y_{1,k}+\cdots+Y_{h,k}$ $\hat{p}_k = \frac{Y_{1,k}+\cdots+Y_{h,k}}{n_1+\cdots+n_h}$	$n_1+\cdots+n_h$

2.2 *Testing for Independence*

Suppose the population can be partitioned in accordance to two categorical variables, A and B. Category A comes with k options, while B with h. Thus, there exists kh cells and each

individual in the sample belongs to one and only one of them. Denote by p_{ij} the probability of cell (i,j). In a similar way, the statistic $\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(N_{ij}-np_{ij})^2}{np_{ij}} \sim \chi^2_{(hk-1)}$.

Now, denote $p_{i\cdot} = \sum_{j=1}^{h} p_{ij}$, $1 \le i \le k$ and $p_{\cdot j} = \sum_{i=1}^{k} p_{ij}$, $1 \le j \le h$. Suppose a sample of size n is conducted. Let N_{ij} be the random variable counting how many among the n sampled belong to cell (i,j). Then, estimators for p_{ij}, $p_{i\cdot}$ and $p_{\cdot j}$ are $\widehat{p}_{ij} = \frac{N_{ij}}{n}$, $\widehat{p}_{i\cdot} = \sum_{j=1}^{h}\frac{N_{ij}}{n}$ and $\widehat{p}_{\cdot j} = \sum_{i=1}^{k}\frac{N_{ij}}{n}$, respectively.

Suppose we wish to test the null hypothesis that A and B are independent. Under H_0 here, $p_{i\cdot}p_{\cdot j} = p_{ij}$, and we can then estimate p_{ij} by $\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}$. In summary, the test statistic is $\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(N_{ij}-n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j})^2}{n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}} \sim \chi^2_{(h-1)(k-1)}$, here $(h-1)+(k-1)$ degrees of freedom lose due to the same reason as above (no need to estimate the last one).

	B_1	B_2	⋯	B_k	Total
A_1	$Y_{1,1}$ (Observed) $n\hat{p}_{1,\cdot}\hat{p}_{\cdot,1}$ (expected)	$Y_{1,2}$ $n\hat{p}_{1,\cdot}\hat{p}_{\cdot,2}$	⋯	$Y_{1,k}$ $n\hat{p}_{1,\cdot}\hat{p}_{\cdot,k}$	$Y_{1,\cdot}$ $\hat{p}_{1,\cdot} = Y_{1,\cdot}/n$
A_2	$Y_{2,1}$ $n\hat{p}_{2,\cdot}\hat{p}_{\cdot,1}$	$Y_{2,2}$ $n\hat{p}_{2,\cdot}\hat{p}_{\cdot,2}$	⋯	$Y_{2,k}$ $n\hat{p}_{2,\cdot}\hat{p}_{\cdot,k}$	$Y_{2,\cdot}$ $\hat{p}_{1,\cdot} = Y_{2,\cdot}/n$
⋯	⋯	⋯	⋯	⋯	⋯
A_h	$Y_{h,1}$ $n\hat{p}_{h,\cdot}\hat{p}_{\cdot,1}$	$Y_{h,2}$ $n\hat{p}_{h,\cdot}\hat{p}_{\cdot,2}$	⋯	$Y_{h,k}$ $n\hat{p}_{h,\cdot}\hat{p}_{\cdot,k}$	$Y_{h,\cdot}$ $\hat{p}_{h,\cdot} = Y_{h,\cdot}/n$
Total	$Y_{\cdot,1}$ $\hat{p}_{\cdot,1} = Y_{\cdot,1}/n$	$Y_{\cdot,2}$ $\hat{p}_{\cdot,2} = Y_{\cdot,2}/n$		$Y_{\cdot,k}$ $\hat{p}_{\cdot,k} = Y_{\cdot,k}/n$	n

3 Addition: The Method of Monte Carlo

Recall that in STA2001, we learn about the **random number generator**, which means generating observations from a specified distribution or sample using $U[0,1]$. This is one method of which is called **Monte Carlo generation**s. Here comes a example.

(e.g.1 **Monte Carlo Integration**) Suppose we want to obtain the integral $\int_a^b g(x)dx$ for a continuous function g over a closed and bounded interval $[a,b]$. If the anti-derivative of g does not exist, then numerical integration is in order. A simple numerical technique is the method of Monte Carlo. We can write the integral as $\int_a^b g(x)dx = (b-a)\int_a^b \frac{g(x)}{b-a}dx = (b-a)E[g(X)]$, where $X \sim U[a,b]$. Generate samples $X_1, \cdots, X_n$ to get $\overline{Y} = \overline{(b-a)g(X)}$, which is a consistent estimate of the integral.

According to the **random number generator**, if we can obtain $F_X^{-1}(u)$ in closed form then we can easily generate observations with cdf F_X. In many cases where this is not possible, techniques have been developed to generate observations. Note that the normal distribution serves as an example of such a case and, in the next example, we show how to generate normal observations.

(e.g.2 **Generating Normal Observations**). To simulate normal variables, Box and Muller (1958) suggested the following procedure. Let Y_1, Y_2 be a random sample from the uniform

distribution over $0 < y < 1$. Define X_1 and X_2 by $\begin{cases} X_1 = (-2\log Y_1)^{\frac{1}{2}}\cos(2\pi Y_2) \\ X_2 = (-2\log Y_1)^{\frac{1}{2}}\sin(2\pi Y_2) \end{cases}$. This transformation is one-to-one and maps $\{(Y_1, Y_2)|0 < Y_1 < 1, 0 < Y_2 < 1\}$ onto $\{(X_1, X_2)|-\infty < X_1 < \infty, -\infty < X_2 < \infty\}$ except for sets involving $X_1 = 0$ and $X_2 = 0$, which have probability zero. The inverse transformation is given by $\begin{cases} y_1 = \exp\left(-\frac{x_1^2 + x_2^2}{2}\right) \\ y_2 = \frac{1}{2\pi}\tan\frac{x_2}{x_1} \end{cases}$. By calculating the Jacobian, we get the joint pdf of X_1 and X_2 is $\frac{1}{2\pi}\exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$. That is, X_1 and X_2 are independent, standard normal random variables.

Another algorithm proposed by Marsaglia and Bray (1964) is to
(i) generate $X, Y \sim U[-1, 1]$ i.i.d. and (ii) $W = U^2 + V^2$;
(iii) If $W > 1$, back to (i);
(iv) $Z = \sqrt{(-2\log W)/W}$, let $X_1 = UZ$; $X_2 = VZ$. Then, $X_1, X_2 \sim N(0, 1)$ i.i.d.

(*proof.*) Consider $J = -\frac{1}{2}\exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$, by changing variables in $f_{U,V|W\leq 1}(u, v) = \frac{1}{\pi}$.

3.1 *Accept-Reject Generation Algorithm*

In this section, we develop the **accept-reject procedure** that can often be used to simulate random variables whose inverse CDF cannot be obtained in closed form. Let X be a continuous random variable with pdf $f(x)$. For this discussion, we will call this pdf the **target** pdf. Suppose it is relatively easy to generate an observation of the random variable Y which has pdf $g(y)$ and that for some constant M we have $f(x) \leq Mg(x)$, $x \in \mathbb{R}$. We will call g(x) the **instrumental** pdf.

(*Algorithm.* **Accept-Reject** Algorithm) Let $f(x)$ be a pdf. Suppose that Y is a random variable with pdf $g(y)$, $U \sim U[0, 1]$, Y and U are independent and $f(x) \leq Mg(x)$, $x \in \mathbb{R}$. The following algorithm generates a random variable X with pdf $f(x)$.
(i) Generate Y and U;
(ii) If $U \leq \frac{f(Y)}{Mg(Y)}$ then take $X = Y$. Else return to step (i);
(iii) X has pdf $f(x)$.

(*proof.*) $P(X \leq x) = P\left(Y \leq x \middle| U \leq \frac{f(Y)}{Mg(Y)}\right) = \frac{P\left(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \leq \frac{f(Y)}{Mg(Y)}\right)} = \frac{\int_{-\infty}^{x}\frac{f(y)}{Mg(y)}g(y)dy}{\int_{-\infty}^{\infty}\frac{f(y)}{Mg(y)}g(y)dy} = \int_{\infty}^{x} f(y)dy.$

Section VIII – Non-parametric Statistics

WANG Yuzhe*

May 16, 2022

1 Distribution-free Confidence Intervals

1.1 *Distribution-free CI for Percentiles*

Recall the estimate of the population(theoretical) percentiles/quantiles $\pi_p = F^{-1}(p)$, (note that $\pi_{\frac{1}{2}}$ is called the median of the distribution F) is $X_{(k)}$, $k = (n+1)p$ and the p^{th} sample quantile $\widehat{\pi}_p$. A CI for median for $m = \widehat{\pi}_p$ can be derived from the order statistic, as $(X_{(i)}, X_{(j)})$. Define $W = \sum \mathbb{1}_{\{X_i \leq m\}}$, then $W \sim Bin\left(n, \frac{1}{2}\right)$. Then, confidence level is $P(X_i \leq m \leq X_j) = P(i \leq W \leq j-1) = \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k$.

1.2 *CLT Approximation for Large Size*

When n goes large enough, we get $W \sim N\left(\frac{n}{2}, \frac{n}{4}\right)$ approximately. Note that we need to do **half-unit continuity correction**, $P(X_i \leq m \leq X_j) = P(i \leq W \leq j-1) \overset{\text{correction}}{=\!=} P(i - \frac{1}{2} \leq W \leq j - \frac{1}{2}) \approx \phi\left(\frac{j - 1/2 - n/2}{\sqrt{n/4}}\right) - \phi\left(\frac{i - 1/2 - n/2}{\sqrt{n/4}}\right)$. Similarly, for π_p estimate, we can define $W_p = \sum \mathbb{1}_{\{X_{(i)} \leq \pi_p\}}$, by invoking CLT, $W_p \sim N\left(np, np(1-p)\right)$.

The advantages for distribution-free CI is that (i) *little assumptions* (all we assume is that the distribution is of the continuous type); (ii) *works better and more robust* for distributions deviate a lot from normal (highly skewed or heavy-tailed); (iii) can be used to get confidence intervals for various percentiles, not just the median.

2 Distribution-free Hypothesis Test

We here remove the assumption that he distribution of a variable in a population comes from a parameter-based distribution (which does not mean that we do not invoke probabilistic techniques). Thus, we will not encounter any estimation here, but as we will see throughout a few examples, how hypotheses are postulated and then tested. Testing will be based on framing some expectations based on the null-hypothesis (which, as said, do not involve with

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

parameters) and checked if some suitable test-statistic meets these expectations. The less it meets them, the more we tend to reject the null-hypothesis.

2.1 *The Sign Test for Matched Samples*

Suppose one wishes to test if husbands and wives are statistically identical with respect to some quantitative variable. Towards this end, a sample of n couples was taken. Let X_i and Y_i be the values of this variable for husband and wife, respectively, in couple i, $1 \leq i \leq n$. Clearly, for a given i, X_i and Y_i are not independent but across couples, for example X_i and Y_j, for $i \neq j$, we can assume independence.

Denote by $X_i - Y_i$ by D_i, $1 \leq i \leq n$. In particular, tests for testing a null hypothesis which says that the two populations (of husbands and wives) are identical vs. the alternative that they are not, are designed.

An option to design is the **sign-test**. Specifically, look at the signs of D_i, $1 \leq i \leq n$. Assuming no ties, namely D_i can never be equal to zero, we get that under the null-hypothesis, $N \sim Bin(n, 0.5)$, where N is the number of positive signs among the n-size sample of D_i. A two-sided symmetric test will be "reject H_0 if $N \leq k$ or $N \geq n - k$". In particular, significance level $\alpha = \left(\sum_{x=0}^{k} + \sum_{x=n-k}^{n} \right) P_{H_0}(N = x) = \frac{1}{2^{n-1}} \sum_{x=0}^{k} \binom{n}{x}$.

In the case where n is large, we can invoke the CLT (on the sample, not on the populations), Specifically, since, under the null-hypothesis, the standardization of N, i.e., $\frac{N - n/2}{\sqrt{n/4}} \sim N(0, 1)$. Thus, a two sided test will be to "reject the null-hypothesis if $\frac{|N - n/2|}{\sqrt{n/4}} \geq Z_{1-\frac{\alpha}{2}}$." Note that for large values for n the fact that N is discrete stands less on our approximation procedure.

2.2 *Wilcoxon Rank Test for Unmatched Samples*

The sign-test was based on the assumption that the samples involved were large enough and if not, that the populations were normal. When these assumptions are too hard to agree with we can do the following quite popular **Wilcoxon test** (also known as the **Mann-Whitney test**).

Suppose n_i observations are taken from population i, $i = 1, 2$. One ends up with $n_1 + n_2$ observations. Reorder them from the smallest to the largest. In particular, each observation gets a rank initiating at 1 and ending at $n_1 + n_2$. Assume that there are no ties, an assumption which is a certainty in the case of continuous distributions.

Construct a statistic by sum up the ranks of those observations which correspond to population 2. The larger is this sum, the more we will tend to reject the null-hypothesis $\mu_1 = \mu_2$ where the alternative is $\mu_1 < \mu_2$.

Considering the distribution of that statistic we are after under H_0, it is a function of n_1 and n_2 and deriving it is a (not easy) question in combinatorics. Yet, this is what is done in tables, where you can see the probabilities that this sum of ranks in greater than or equal to k for any $\frac{n_2(n_2+1)}{2} \leq k \leq \frac{n_2(n_2+1)}{2} + n_1 n_2$. In particular, any threshold value k leads to its own significance level.

As an approximation for its distribution, note that the expected value of this statistic equals

$\frac{n_2(n_1+n_2+1)}{2}$ (easily by symmetry) while its variance equals $\frac{n_1 n_2(n_1+n_2+1)}{12}$ (proving this fact calls for a longer argument). Once we have these two parameters (n_1 and n_2), we can use the normal approximation for the distribution of the sum of the corresponding ranks, $K \sim N(n_2(n_1+n_2+1)/2,\ n_1 n_2(n_1+n_2+1)/12)$.

(*Remark.*) An alternative test: compare all n_1 and all n_2 values one-by-one, respectively. Consider the test statistic as the winning times the 2^{nd} sample against the 1^{st}, totally from 0 to $n_1 n_2$. A sensible test is "to reject H_0 if winning times is larger than some threshold in $[0, n_1 n_2]$". It is equivalent to the Wilcoxon test because given winning times k, the sum of the ranks of one population is $k + \frac{n_2(n_2+1)}{2}$.

2.3 *Fisher's Test for Independence*

Consider the example, a group comes with 20 men and 30 women. A committee of 5 is selected at random from this group. Consider the number of men and women it contains.

gender \ member	members	non-members	total
men	x	$20-x$	20
women	$5-x$	$25+x$	30
total	5	45	50

Suppose both horizontal and vertical margins are given. The joint figures are random but in fact we have only one proper random variable: Given the margins (which are not random) and one of the joints, we are able to derive the other three joint values.

In the case (H_0) where the two ways of classifications (gender and membership vs. non-member) are independent, the figures we expect for x is $20 \times \frac{5}{50} = 2$.

Suppose the alternative hypothesis H_1 is that the selection process discriminates against men and prefers more women there. A possible answer to judge how supportive to the null hypothesis is some x we get here in light of this alternative is the following P-value: What is the (prior) probability of getting results such as the one we got (e.g., $x = 1$) and more extreme from the null hypothesis' point of view (e.g., $x < 1$). The answer is $\sum_{x=0}^{k} \frac{\binom{20}{x}\binom{30}{5-x}}{\binom{50}{5}}$.

(*Remark.*) Hypegeometric distribution: $\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$, with the range of m: $\min\{0, n+M-N\} \le m \le \min\{n, M\}$.

2.4 *Kolmogorov-Smirnov Test*

Initially, we define a measure for the distance between two CDFs, F and G. Of course, there is more than one way to do that but $d(F,G) := \max_x |F(x) - G(x)|$ is a sensible option (explicitly, $d(F,G) = d(G,F)$).

Secondly, for a given sample $X_1, \cdots, X_n$, define the corresponding **empirical distribution function**, $\widehat{F}_n$. Specifically, recall $X_{(1)}, \cdots, X_{(n)}$ is the smallest to largest rearrangement for samples. For simplicity, assume that there are no ties in the sample. Then, for any x, define $\widehat{F}_n(x)$ as the relative frequency of the number of observations which are smaller than or equal

to x. It is easy to see that $\forall\ x \in [X_{(i)}, X_{(i+1)})$, $\widehat{F}_n(x) = \dfrac{i}{n}$, $0 \le i \le n$,(set $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$). In particular, we get a step function with jumps at the observed values.

Not been too formal, we can say that if H_0 says that the population is distributed with a CDF of F, then the statistic $d(\widehat{F}_n, F)$ can serve as a measure quantifying the correctness of this hypothesis. In particular, the larger it is, the more is the evidence against the null-hypothesis. In passing, note that the maximization defining $d(\widehat{F}_n, F)$ is attained at one of the n points, $X_1, \cdots, X_n$. In short, $d(\widehat{F}_n, F) = \max\limits_{1 \le i \le n} |F(X_i) - \frac{i}{n}|$.

(*Theorem.*) Assuming F is strictly monotone increasing along its support, then under the null-hypothesis, the distribution of $d(\widehat{F}_n, F)$ is not a function of F.

(*Proof.*) First, a formal definition of the test statistic is $d(\widehat{F}_n, F) = \max\limits_x \left| \dfrac{1}{n} \sum\limits_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}} - F(x) \right| = \max\limits_x \left| \dfrac{1}{n} \sum\limits_{i=1}^{n} \mathbb{1}_{\{F(X_i) \le F(x)\}} - F(x) \right|$. But under the null-hypothesis, $F(X) \sim U[0,1]$. This implies that under the null-hypothesis the above statistic follows the same distribution as the statistic $\max\limits_x \left| \dfrac{1}{n} \sum\limits_{i=1}^{n} \mathbb{1}_{\{U_i \le F(x)\}} - F(x) \right| \overset{u=F(x)}{=\!=\!=} \max\limits_{u \in [0,1]} \left| \dfrac{1}{n} \sum\limits_{i=1}^{n} \mathbb{1}_{\{U_i \le u\}} - u \right|$, which is free of F.

(*Remark.*) The exact distribution we are after is of course a function of n. There is no closed-form function and hence tables were constructed. Specifically, as $\overline{X} \to \mu_0$ (convergence

For example, the critical values for the case where $n = 10$ are

α	0.1	0.05	0.02	0.01
$1 - \alpha$ percentile	0.3687	0.4093	0.4566	0.4889

in probability) under the null-hypothesis and hence, we need to inflate it by $\sqrt{n}$ in order to get a non-degenerate random variable at the limit, here for any x we get that $d(\widehat{F}_n(x), F(x))$ converges to zero when n goes to infinity because $\left| \widehat{F}_n(x), F(x) \right| \to \infty$. Thus, the test statistic will now be $\sqrt{n} d(\widehat{F}_n(x), F(x))$. Of course, this distribution varies with n and some tables for small and medium size values for n exist. For larger value of n a limit distribution was derived and it can be found in the literature.

Finally, it is better to point out that there exists a version of the test where one compares two populations, where the null-hypothesis says that the corresponding distributions are identical.

Section IX– Bayesian Statistics

WANG Yuzhe*

May 21, 2022

1 Subjective Probabilities

A **subjective probability**, as its name shows, is someone's opinion of what the probability is for an event. Although this may not seem very scientific, it is often the best you can do when you have no past experience (so you cannot use relative frequency) and no theory (so you cannot use theoretical probability).

Suppose now a person has assigned $P(C) = p$ as his/her subjective probability to some event C, then if that person is willing to bet, he/she is willing to accept either side of the bet:
(1) win 1 units if C occurs and lose p if it does not occur;
(2) win p units if C does not occur and lose 1 if it does.
If that is not the case, that person should review his/her subjective probability of event C.

Then it turns out, all rules (definitions and theorems) on probability found in STA2001 follow for subjective probabilities. The proofs come from **Dutch book** and induction methods. Here is a example.

(*Theorem.*) For **mutually exclusive** C_1 and C_2, we have $P(C_1 \cup C_2) = P(C_1) + P(C_2)$.

(*proof.*) W.o.l.g., suppose a person belive that $P(C_1 \cup C_2) = p_3 < p_1 + p_2 = P(C_1) + P(C_2)$, $P(C_i) = p_i$, respectively. Denote $d = p_1 + p_2 - p_3$, and assume there is a gambler (really good at gambling), who will purchase $C_1 \cup C_2$ by $p_3 + \frac{d}{4}$ (adding price) and sell C_i to that person for $p_i - \frac{d}{4}$ (discount), $i = 1, 2$. According to that person's subjective probabilities, that is a good deal (with that gambler). That is, the person is down $p_1 + p_2 - \frac{d}{2} - \left(p_3 + \frac{d}{4}\right) = \frac{d}{4}$ before any bets are settled. Two case after bets:
(1) C_i happens: the gambler has $C_1 \cup C_2$ and the person has C_i; so they exchange their winning units and the person is still down $\frac{d}{4}$;
(2) Neither happens, then the gambler and that person receive zero, and the person is still down $\frac{d}{4}$.
Thus we see that it is bad for that person to assign $p_3 < p_1 + p_2$, the other side follows the same proof.

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R.C. Email: yuzhewang@link.cuhk.edu.cn.

2 Bayesian Procedures

To understand the Bayesian inference, let us review **Bayes Theorem**, in a situation in which we are trying to determine something about a parameter of a distribution.

(*Theorem.*) **Bayes' Theorem**. Let B_1, B_2, $\cdots$, B_m constitute a partition of the sample space S and A is an event. The **prior probability** of the event B_i is $P(B_i) > 0$, $i = 1, \cdots, n$. Then, we can compute the **posterior probability** of B_k: $P(B_k|A) = \dfrac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$

Suppose we have a distribution (e.g., Poisson distribution) with parameter $\theta > 0$, and we believe that either $\theta = \theta_1$ or $\theta = \theta_2$. In Bayesian inference, the parameter is treated as a random variable Θ. Suppose for this example, we assign **subjective prior probabilities** of $P(\theta = \theta_1) = p$ and $P(\theta = \theta_2) = 1 - p$ to the two possible values. These subjective probabilities are based upon past experiences, and it might be *unrealistic* that can only take one of two values, instead of a continuous $\Theta > 0$.

By **Bayes Theorem**, we can compute $\widetilde{p} = P(\Theta = \theta_1 | X_1 = x_1, \cdots, X_n = x_n)$ which is conditioned on n samples $X_1, \cdots, X_n$. Suppose $\widehat{p} < p$ now. That means, with the observations $X_1 = x_1, \cdots, X_n = x_n$, the *posterior probability* of $\Theta = \theta_1$ was smaller than the *prior probability* of $\Theta = \theta_1$. Similarly, the *posterior probability* of $\Theta = \theta_2$ was greater than corresponding *prior*. That is, the observations $X_1 = x_1, \cdots, X_n = x_n$ seemed to favor $\Theta = \theta_2$ more than $\Theta = \theta_1$. Now let us address in general a more realistic situation in which we place a prior pdf $h(\theta)$ on a support which is a continuium.

2.1 *Prior and Posterior Distributions*

We shall now describe the Bayesian approach to the problem of estimation. This approach takes into account any *prior knowledge of the experiment* that the statistician has and it is one application of a principle of statistical inference that may be called **Bayesian statistics**.

Initially, let us introduce Bayes Theorem to continuous random variables. Assume $f_{X,Y}(x, y)$ is the joint density of X and Y, recall that conditional density follows $f_{X|Y=y}(x) f_Y(y) = f_{X,Y}(x, y) = f_{Y|X=x}(y) f_X(x)$. Then, we can find that $f_{X|Y=y}(x) = \dfrac{f_{Y|X=x}(y) f_X(x)}{f_Y(y)} = \dfrac{f_{Y|X=x}(y) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X=\xi}(y) f_X(\xi) d\xi} \propto f_{Y|X=x}(y) f_X(x)$, $-\infty < x < \infty$. This technique is useful for finding marginal density function by checking the kernel. We also allow hybrid models, namely where one of the variables is discrete while the there is continuous.

Hence, consider a random variable X that has a distribution of probability that depends upon the symbol θ, where θ is an element of a well-defined set Ω, and look upon θ as a possible value of the random variable Θ (summary): $X|\theta \sim f_{X|\Theta=\theta}(x)$ and $\Theta \sim f_\Theta(\theta)$. ($f_\Theta(\theta)$ is called the **prior distribution**.)

Suppose that $X_1, \cdots, X_n$ is a random sample from the conditional distribution of X given $\Theta = \theta$. Let $\mathbf{X} = (X_1, \cdots, X_n)$ and $\mathbf{x} = (x_1, \cdots, x_n)$, then the joint pdf of $\mathbf{X}$ (given $\Theta = \theta$) is $f_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) = \prod_{i=1}^{n} f_{X|\Theta=\theta}(x_i)$. Thus, the joint pdf of $\mathbf{X}$ and Θ is $f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) = f_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) f_\Theta(\theta)$. Similarly to above, we define **posterior distribution** $\Theta|\mathbf{X}$ with **posterior pdf** $f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) \propto f_{\mathbf{X}|\Theta=\theta}(x) f_\Theta(\theta)$

(*Definition.*) A class of prior pdfs for the family of distributions with pdfs $f(x|\theta)$, $\theta \in \Omega$ is

said to define a **conjugate family of distribution**s if the posterior pdf of the parameter is in the same family of distributions as the prior.

(*Example I.*) Suppose the random variable $\Lambda \sim \Gamma(\alpha, \beta)$ for some $\alpha > 0$ and $\beta > 0$. Also, assume that $X|\Lambda = \lambda \sim Pois(\lambda)$. Hence, the random sample is drawn from a Poisson distribution with mean λ and the prior distribution is a $\Gamma(\alpha, \beta)$ distribution. Suppose n samples are drawn from X conditioned on $\Lambda = \lambda$, we then follow the steps above, and get that the *posterior distribution* is $f_{\Lambda|\mathbf{X}=\mathbf{x}}(\lambda) \propto f_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x}) f_{\Lambda}(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$, $\forall\ -\infty < \lambda < \infty$, from this we know the *kernel* is $\lambda^{\sum x_i + \alpha - 1} e^{-(\beta+n)\lambda}$, which means $\Lambda|\mathbf{X} = \mathbf{x} \sim \Gamma(\alpha + \mathbb{1}^T\mathbf{x}, \beta + n)$. Notice that the *posterior pdf* reflects both prior information (α, β) and sample information $(\mathbb{1}^T\mathbf{x})$. Recall that the both *prior* and *posterior* distributions belong to the same family of distributions, namely, the Gamma family. Such family is called **conjugate**.

(*Example II.*) Now let $\Theta \sim N(\mu, \tau^2)$ for known μ and τ and $X|\Theta = \theta \sim N(\theta, \sigma^2)$ with known σ. n random sample drawn from $X|\theta$ are now $\mathbf{X} = (X_1, \cdots, X_n)$, and $Y = \overline{X}$ is a sufficient statistic. Then, *posterior distribution* is $f_{\Theta|Y=y}(\theta) \propto f_{Y|\Theta=\theta}(y) f_{\Theta}(\theta) \propto e^{-\frac{(\theta-\mu)^2}{2\tau^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2/n}}$, $\forall\ \theta \in \mathbb{R}$, from this we know the *kernel* is $exp(-\frac{(x - \frac{\mu/\tau^2 + yn/\sigma^2}{1/\tau^2 + n/\sigma^2})^2}{2(1/\tau^2 + n/\sigma^2)^{-1}})$, which means $\Theta|\overline{X} = y \sim N(\frac{\mu/\tau^2 + yn/\sigma^2}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2})$. Note that the *prior* and the *posterior* distributions are normal, so we can conclude that the normal family is a conjugate one.

2.2 *Bayesian Point Estimation*

Suppose we want a point estimator of θ. From the Bayesian viewpoint, this really amounts to selecting a decision function $\delta(\mathbf{x})$, so that $\delta(\mathbf{x})$ is a predicted value of θ (an experimental value of the random variable Θ) when both the computed value $\mathbf{x}$ and the conditional pdf $f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta)$ are known. Sometimes we use the mean $E(W)$ or the median.

It seems desirable that the choice of the decision function should depend upon a loss function $\mathcal{L}(\Theta; \delta(\mathbf{x}))$. One way in which this dependence upon the loss function can be reflected is to select the decision function δ in such a way that the *conditional expectation* of the loss is a minimum. A **Bayes' estimate** is a decision function δ, which minimizes $E(\mathcal{L}(\Theta; \delta(\mathbf{x}))|\mathbf{X} = \mathbf{x}) = \int_{-\infty}^{\infty} \mathcal{L}(\theta; \delta(\mathbf{x})) f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) d\theta$, when Θ is a continuous type random variable. The associated random variable $\delta(\mathbf{X})$ is called a **Bayes' point estimator** of θ.

Note that if the loss function is given by $\mathcal{L}(\theta; \delta(\mathbf{x})) = [\theta - \delta(\mathbf{x})]^2$, then the Bayes' estimate is $\delta(\mathbf{x}) = E(\Theta|\mathbf{x})$, the mean of the conditional distribution of Θ, given $\mathbf{X} = \mathbf{x}$. If the loss function is given by $\mathcal{L}(\theta; \delta(\mathbf{x})) = |\theta - \delta(\mathbf{x})|$, then a median of the conditional distribution of Θ, given $\mathbf{X} = \mathbf{x}$, is the Bayes' solution.

(*Example I* (cont'd).) Suppose we use $\mathcal{L}(\lambda; \delta(\mathbf{x})) = [\lambda - \delta(\mathbf{x})]^2$, then, the Bayesian point estimator of δ is the mean of this Gamma pdf which is $\delta(\mathbf{X}) = E(\Lambda|\mathbf{X}) = \frac{\alpha + \mathbb{1}^T\mathbf{X}}{\beta + n}$. Note that when n gets larger as larger, this random variable convergences in distribution to $\overline{X}$. This limit is free of the prior's parameters as it should be with a good prior: It is overwhelmed by the data and its effect disappears when the sample size is large.

(*Example II* (cont'd).) Similarly to above, we get a Bayes' point estimator for θ, which is $\frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\mu + \frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2}\overline{X}$. In words, it is a weighted average between μ and $\overline{X}$, with

weights which are inversely proportional to their variances (where in the case of $\overline{X}$ it is its conditional variance given θ). It makes sense: the less the variability is, the more you trust the observation. Here too we can see that the estimator convergences in distribution to $\overline{X}$, so again we see a prior which is overwhelmed by the collected data. Note that the Bayes' estimator changes, as it should, with different loss functions.

2.3 *Bayesian Interval Estimation*

If an interval estimate of θ is desired, we can find two functions $u(\mathbf{x})$ and $v(\mathbf{x})$ so that the conditional probability $P[u(\mathbf{x}) < \Theta < v(\mathbf{x})|\mathbf{X} = \mathbf{x}] = \int_{u(\mathbf{x})}^{v(\mathbf{x})} f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta)d\theta$ is large, for example, 0.95. Then the interval $u(\mathbf{x})$ to $v(\mathbf{x})$ is an *interval estimate* of θ in the sense that the conditional probability of Θ belonging to that interval is equal to 0.95. These intervals are often called **credible** or **probability interval**s, so as not to confuse them with confidence intervals.

(*Example II* (cont'd).) Note that the distribution of the conditional distribution $\Theta|\overline{X} = y$ is normal, we can then get the **credible interval** of probability 0.95 for θ as $\frac{\mu/\tau^2 + yn/\sigma^2}{1/\tau^2 + n/\sigma^2} \pm Z_{0.025}\sqrt{\frac{1}{1/\tau^2 + n/\sigma^2}}$.

STA2002: Probability and Statistics II
Midterm Exam

March 24, 2022

The questionnaire comes with three questions. You need to answer them all. The number of points each question carries is stated next to it. The total is 100. You have 80 minutes in order to do so. You are entitled to bring along to the exam class two A4 sheets of papers which you have prepared in advance. Finally, you are asked to present a valid Student I.D. Card for inspection during the examination session.

1. (25) Let $X_1, X_2, \ldots, X_{12}$ be a series with 12 entries. It is known that $X_3 = 4$, $X_5 = 6$, $\overline{X} = 5$ and that $\mathrm{Var}(X) = 4$. Let $Y_1, \ldots, Y_{12}$ be a similar series but with two exceptions $Y_3 = 3$ and $Y_5 = 7$.

   (a) (5) what are the standardization of X_3 and X_5?

   (b) (10) what is the value $\overline{Y}$?

   (c) (10) what is the value $\mathrm{Var}(Y)$?

   Solution:

   (a) -0.5 and $+0.5$, respectively.

   (b) the mean has not changed: $\overline{Y} = 5$

   (c) in the sum of squares leading to the variance, two entries of value of 1, were replaced by 4. The other ten have not changed. This leads to a total increase of 6 in this summation, adding a value of 6/12 to the variance, making the new variance equal to $4 + 0.5 = 4.5$.

2. (30) A sampled of some size n was conducted on three variables, X, W and Y. Both X and W are measured in kilograms, while Y is measured in meters. The following figures were computed $\overline{X} = 4.5$, $\overline{W} = 3.2$ and $\overline{Y} = 11.2$, $\mathrm{Var}(X) = 4.05$, $\mathrm{Var}(W) = 4.36$ and $\mathrm{Var}(Y) = 22.56$. Finally, $\mathrm{Cov}(X, Y) = 8.4$ and $\mathrm{Cov}(W, Y) = 6.76$.

   (a) (6) compute $\mathrm{Corr}(X, Y)$ and $\mathrm{Corr}(W, Y)$.

   (b) (6) what is the regression line of Y on X. Next to the numerical values for the slope and for the intercept state their units of measurement.

   (c) (6) repeat the previous item but now replace X with W.

   (d) (6) which of the above two regression lines do you recommend to use?

   (e) (6) for the regression line you have stated in the previous item, compute the variance of Y it explains. How much is it in percents?

   Solution:

   (a) $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)} = \frac{8.4}{\sqrt{4.05 \times 22.56}} = 0.8788$
   $\mathrm{Corr}(W, Y) = \frac{\mathrm{Cov}(W,Y)}{\mathrm{SD}(W)\mathrm{SD}(Y)} = \frac{6.76}{\sqrt{4.36 \times 22.56}} = 0.6816$

   (b) The regression line is $\frac{y-\bar{Y}}{\mathrm{SD}(Y)} = \mathrm{Corr}(X, Y)\frac{x-\bar{X}}{\mathrm{SD}(X)}$. In particular,

$$\frac{y - 11.2}{\sqrt{22.56}} = 0.8788 \times \frac{x - 4.5}{\sqrt{4.05}}$$

$$\Rightarrow y = 2.0741x + 1.8667$$

   The slope 2.0741 is measured in meters per kilogram, while in intercept 1.8667 is measured in meters.

   (c) $y = 1.5505x + 6.2385$
   The slope 1.5505 is measured in meters per kilogram, while in intercept 6.2385 is measured in meters.

   (d) The regression line of Y on X: $y = 2.0741x + 1.8667$.

   (e) $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \mathrm{Corr}^2(X, Y)n\mathrm{Var}(Y)$. In percent, it is $\mathrm{Corr}^2(X, Y) = 77.23\%$.

3. (45) Let $X \sim \chi^2_{(2m)}$ for some integer m.

   (a) (4) what is the density function of X?

   (b) (8) what are the first two moments of X?

   (c) (4) what is the variance of X?

   (d) (10) based on a random sample X_i, $1 \leq i \leq n$, all of which are independent and distributed as X, design by the method of moments, two estimators for m based on the two moments derived above. Note that they need not be integers.

   (e) (6) what is the likelihood function of the random sample?

   (f) (9) what is the MLE for m and how it is related to the sample geometric mean? Note that it needs to be an integer. Hint: Consider the ratio between the likelihood functions for two consecutive values for m.

   (g) (4) What is the MLE for 2m?

   Solution:

   (a)
   $$f_X(x) = \frac{(1/2)^m}{(m-1)!} x^{m-1} e^{-x/2}, \ x \geq 0.$$

   (b) $\mathrm{E}(X) = 2m$, $\mathrm{E}(X^2) = 4m(m+1)$

   (c) $\mathrm{Var}(X) = 4m$.

   (d) Since $\mathrm{E}(\overline{X}) = 2m$, the first estimator is derived from $\overline{X} = 2\hat{m}$. Hence,
   $$\hat{m} = \frac{\overline{X}}{2}.$$
   Since $\mathrm{E}(\overline{X^2}) = 4m(m+1)$, the second estimator is derived from $\overline{X^2} = 4\hat{m}(\hat{m}+1)$. This leads to a quadratic equation in $\hat{m}$. Hence,
   $$\hat{m} = \frac{-1 + \sqrt{1 + \overline{X^2}}}{2}.$$
   Note that the other root leads to a negative value.

(e)

$$L(X_1, \ldots, X_n; m) = \Pi_{i=1}^n \frac{(1/2)^m}{(m-1)!} X_i^{m-1} e^{-X_i/2}$$

$$\frac{(1/2)^{nm}}{((m-1)!)^n} (\Pi_{i=1}^n X_i)^{m-1} e^{-\sum_{i=1}^n X_i/2}$$

(f)

$$L(X_1, \ldots, X_n; m+1) = \frac{(1/2)^{n(m+1)}}{(m!)^n} (\Pi_{i=1}^n X_i)^m e^{-\sum_{i=1}^n X_i/2}$$

From the above we learn that

$$\frac{L(X_1, \ldots, X_n; m+1)}{L(X_1, \ldots, X_n; m)} = (\frac{1}{2m})^n \Pi_{i=1}^n X_i.$$

Since $X \geq 0$, this ratio is clearly monotone decreasing with m and we look for the first m such that

$$(\frac{1}{2m})^n \Pi_{i=1}^n X_i \leq 1$$

or

$$\frac{(\Pi_{i=1}^n X_i)^{\frac{1}{n}}}{2} \leq m$$

Now $\hat{m}$ needs to be an integer, so it is the ceiling of what is written in the left hand side, which in fact is the ceiling of half the geometric sample mean.

(g) The MLE for $2m$ is $2\hat{m}$.

Probability and Statistics II
Final Exam

Course code: STA2002
Course name: Probability and Statistics II
Date: May 15, 2021
Time: 4:00pm-7:00pm (three hours)

The questionnaire comes with five questions. You need to solve them all. You are entitled to bring along into the exam class a calculator and four double side A4 sheets of papers prepared by yourself in advance. The total number of points is 100. The number of points each item carries is stated next to it.
Good luck!

1. Let $N \sim Pois(\lambda)$ and let X be such that $X|N \sim Bin(N, p)$.

   (a) (8) Derive $\mathrm{E}(X|N)$, $\mathrm{Var}(X|N)$, $\mathrm{Var}(\mathrm{E}(X|N))$ and $\mathrm{E}(\mathrm{Var}(X|N))$ in terms of λ and p.

   (b) (6) Derive $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ in terms of λ and p.

   (c) (3) Suppose it is known that X follows a Poisson distribution. In particular, you are not asked to prove that. What is the corresponding parameter?

   (d) (3) What is the distribution of $N - X$?

   Solution:

   (a) $\mathrm{E}(X|N) = Np$, $\mathrm{Var}(X|N) = Np(1-p)$, $\mathrm{Var}(\mathrm{E}(X|N)) = \mathrm{Var}(Np) = p^2\mathrm{Var}(N) = p^2\lambda$ and $\mathrm{E}(\mathrm{Var}(X|N)) = \mathrm{E}(N(p(1-p))) = p(1-p)\mathrm{E}(N) = \lambda p(1-p)$

   (b) $\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|N)) = \mathrm{E}(Np) = p\mathrm{E}(N) = p\lambda$ and $\mathrm{Var}(X) = \mathrm{Var}(\mathrm{E}(X|N)) + \mathrm{E}(\mathrm{Var}(X|N)) = p^2\lambda + p(1-p)\lambda = \lambda p$

   (c) $\mathrm{E}(X) = \lambda p$. Hence, $X \sim Pois(\lambda p)$.

   (d) By symmetry, $(N - X)|N \sim Bin(N, 1-p)$. Hence, $N - X \sim Pois(\lambda(1-p))$

2. (a) (4) Let X be a non-negative continuous random variable with a cumulative distribution function (CDF) of $F_X(x)$. Prove that $\mathrm{E}(X) = \int_{x=0}^{\infty}(1 - F_X(x))\,dx$. Hint: Recall that $1 - F_X(x) = \int_{t=x}^{\infty} f_X(t)\,dt$, where $f_X(t)$ is the density function of X.

   (b) (4) Let $X, Y \sim \exp(\lambda)$ be two independent random variables. Denote $\frac{X}{Y}$ by R. Prove that

$$\mathrm{P}(R \geq x) = \frac{1}{1+x},\ x \geq 0.$$

   Hint: $\mathrm{P}(X \geq xY) = \int_{y=0}^{\infty} \mathrm{P}(X \geq xy) f_Y(y)\,dy$.

   (c) (4) Suppose a random sample of ratios of the type $R_i = \frac{X_i}{Y_i}$, $1 \leq i \leq n$, is conducted. How useful can these ratios be in estimating λ? Note: one can observe only the ratios and not the actual values of X_i and Y_i, $1 \leq i \leq n$.

   (d) (6) Based on a random sample R_i, $1 \leq i \leq n$, suggest a goodness-of-fit test for testing the null hypothesis that says that the X and Y variables follow an exponential distribution with a common parameter, against the alternative which says that this is not the case. Specifically, define five possible cells (intervals) for the possible values for R which, under the null-hypothesis, are equally likely.

Solution:

(a)

$$\int_{x=0}^{\infty}(1-F_X(x))\,dx = \int_{x=0}^{\infty}\int_{t=x}^{\infty} f_X(t)\,dt\,dx = \int_{t=0}^{\infty} f_X(t)\int_{x=0}^{t} 1\,dx\,dt$$

$$\int_{t=0}^{\infty} f_X(t)t\,dt = \mathrm{E}(X).$$

(b)

$$\mathrm{P}(R \geq x) = \mathrm{P}(X \geq xY) = \int_{y=0}^{\infty} \mathrm{P}(X \geq xy) f_Y(y)\,dy = \int_{y=0}^{\infty} e^{-\lambda xy}\lambda e^{-\lambda y}\,dy$$

$$= \frac{1}{1+x}\int_{y=0}^{\infty} \lambda(1+x)e^{-\lambda(1+x)y}\,dy = \frac{1}{1+x},$$

since the last integral is the integral of an exponential density with parameter $\lambda(1+x)$ it equals 1.

(c) Since the distribution of R is free of the parameter λ, a random sample of random variables which comes with its distribution is worthless from the point of view of estimating R.

(d) If the null-hypothesis is correct, we get that is CDF equals

$$F_R(x) = \mathrm{P}(R \le x) = 1 - \mathrm{P}(R \ge x) = 1 - \frac{1}{1+x} = \frac{x}{1+x}, \quad x \ge 0.$$

We look for its five $j \times 0.2$ percentiles, to be denoted by x_j, $1 \le j \le 5$. Thus,

$$\frac{x_j}{1+x_j} = \frac{j}{5}, \ 1 \le j \le 5.$$

The solutions are 0.25, 0.66, 1.5, 4 and ∞. Set $x_0 = 0$. Then, cell j corresponds to all observationsR_i, $1 \le i \le n$, with $x_{j-1} \le R_i < x_j$, $1 \le j \le 5$. Under the null-hypothesis, the expected the number of observations in cell j is $n/5$. Denote the (random) number of observed who are actually in cell j, the observed values, by O_j, $1 \le j \le 5$. Then, the goodness-of-fit statistic is

$$\sum_{j=1}^{5} \frac{(O_j - n/5)^2}{n/5},$$

which under the null-hypothesis follows a chi-square distribution with 4 degrees of freedom. Finally, a test with a significance level of α will be to reject if

$$\sum_{j=1}^{5} \frac{(O_j - n/5)^2}{n/5} \ge \chi^2_{(4,1-\alpha)}.$$

3. Suppose $X_i \sim N(0, \sigma^2)$, $1 \leq i \leq n$, are n independent random variables.

   (a) (6) Show that $\overline{X^2}$ is an MLE for σ^2 and that it is an UBE.

   (b) (4) What is the variance of this estimator?

   (c) (4) Construct a $1 - \alpha$ two-sided confidence interval on σ.

   (d) (3) Prove that

$$\mathrm{MSE}(a\overline{X^2}) = (\frac{2}{n}a^2 + (a-1)^2)\sigma^4.$$

   Note that you need to find both the square of the bias and the variance of the estimator $a\overline{X^2}$ for σ^2.

   (e) (3) What is value for a which minimizes the MSE of the previous item? Is it $a = 1$? Whatever is the case, what is the corresponding conclusion on the MLE?

Solution:

(a)

$$\mathrm{E}(\overline{X^2}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i^2) = \frac{1}{n}n\sigma^2 = \sigma^2,$$

which proves that it is an UBE for σ^2. Also,

$$L(X_1, \ldots, X_n; \sigma^2) = \frac{1}{\sqrt{2\pi}^n(\sigma^2)^{n/2}}e^{-\frac{1}{2}\sum_{i=1}^{n}X_i^2/\sigma^2}$$

and its log (ignoring additive terms which are not functions of σ^2) equals

$$-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}X_i^2.$$

Taking derivative with respect to σ^2 (and not with respect to σ), leads to

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}X_i^2.$$

Equating the derivative to zero, leads to $\overline{X^2}$ being the unique solution. Thus, it is the MLE for σ^2.

(b)
$$\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i^2) = \frac{1}{n^2} n \mathrm{Var}(X^2) = \frac{\mathrm{Var}(X^2)}{n}.$$

But $X^2/\sigma^2 \sim \chi^2_{(1)}$. Since $\mathrm{Var}(\chi^2_{(n)}) = 2n$, we get that $\mathrm{Var}(X^2/\sigma^2) = 2$ and hence $\mathrm{Var}(X^2) = 2\sigma^4$. In summary, the variance of the MLE equals $2\sigma^4/n$.

(c) Since $\Sigma_{i=1}^n X_i^2/\sigma^2 \sim \chi^2_{(n)}$, we conclude that for any value for σ^2,

$$\mathrm{P}(\chi^2_{(n,\alpha/2)} \leq \Sigma_{i=1}^n X_i^2/\sigma^2 \leq \chi^2_{(n,1-\alpha/2)}) = 1-\alpha,$$

which is equivalent to

$$\mathrm{P}(\frac{\Sigma_{i=1}^n X_i^2}{\chi^2_{(n,1-\alpha/2)}} \leq \sigma^2 \leq \frac{\Sigma_{i=1}^n X_i^2}{\chi^2_{(n,\alpha/2)}}) = 1-\alpha.$$

Hence,
$$[\frac{\Sigma_{i=1}^n X_i^2}{\chi^2_{(n,1-\alpha/2)}}, \frac{\Sigma_{i=1}^n X_i^2}{\chi^2_{(n,\alpha/2)}}]$$
is a $1-\sigma$ confidence interval for σ^2 and

$$\left[\sqrt{\frac{\Sigma_{i=1}^n X_i^2}{\chi^2_{(n,1-\alpha/2)}}}, \sqrt{\frac{\Sigma_{i=1}^n X_i^2}{\chi^2_{(n,\alpha/2)}}}\right]$$

is a $1-\alpha$ confidence level for σ.

(d) The MSE of $a\overline{X^2}$ equals

$$\mathrm{Var}(a\overline{X^2}) + (\mathrm{E}(a\overline{X^2}) - \sigma^2))^2 = \alpha^2 \mathrm{Var}(\overline{X^2}) + a^2 \mathrm{E}^2(\overline{X^2}) - 2a\mathrm{E}(\overline{X^2})\sigma^2 + \sigma^4 =$$

$$a^2 \frac{2\sigma^4}{n} + a^2\sigma^4 - 2a\sigma^4 + \sigma^4 = (2\frac{a^2}{n} + (a-1)^2)\sigma^4.$$

(e) The MSE is a quadratic function of a. The minimum is attained in $a = n/(n+2)$. Since the optimal $a \neq 1$, we conclude that the MLE for σ^2 does not come with the minimal possible MSE.

4. Consider the following no-constant simple linear regression model. Specifically, for some given constants x_i, $1 \leq i \leq n$, and a parameter a, $Y_i \sim N(ax_i, 1)$, $1 \leq i \leq n$, are n independent random variables.

   (a) (4) What is the MLE for a? Is it an UBE?

   (b) (4) What is the distribution of the above MLE?

   (c) (4) Construct a symmetric $1 - \alpha$ confidence interval for a around the MLE.

   (d) (6) There is an interest in testing the null-hypothesis that $a = 0$ versus the alternative which says that $a = 1$. In terms of only the two series x_i and Y_i, $1 \leq i \leq n$, construct the LRT with a significance level of α.

   (e) (2) What is the power of this test? Its log, up to an additive constant which is not a function of a, equals

Solution:

(a) $Y_i \sim N(ax_i, 1)$, $1 \leq i \leq n$, and they are independent. Hence,

$$-\frac{1}{2}(\sum_{i=1}^{n} Y_i^2 - 2\beta \sum_{i=1}^{n} Y_i x_i + a^2 \sum_{i=1}^{n} x_i^2),$$

Taking derivative with respect to a and equating it to zero, we get that

$$\hat{a} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}.$$

$$L(Y_1, \ldots, Y_n; \beta) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\sum_{i=1}^{n} (Y_i - a x_i)^2}{2}}.$$

(b) It is an UBE. Indeed,

$$\mathrm{E}(\frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}) = \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i \mathrm{E}(Y_i)$$

$$\frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i a x_i = a.$$

Probability and Statistics II
Final Exam

May 19, 2022

Course code: STA2002
Course name: Probability and Statistics II
Date: May 19, 2022
Time: 16:00-19:00 (three hours)

The questionnaire comes with five questions. You need to solve them all. You are entitled to bring along into the exam class a calculator and four double sided A4 sheets of paper prepared by yourself in advance. The total number of points is 100. The number of points each item carries is stated next to it. Note that in the case where a few items form a question, you are entitled to use an earlier item for proving a later one, even if you did not prove the former. Unless it is said otherwise, you need to justify any of your claims.

Good luck!

1. (15)

   (a) (6) Given that $X \sim Pois(\lambda)$ and $Y|X \sim \chi^2_{(2X+1)}$, derive $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

   (b) (4) Consider a series of random variables X_n, where $X_n \sim \chi^2_{(n)}$, $n \geq 1$. Argue that the CLT is applicable to this series of random variables and their limit. In particular, what is the series of standardized random variables which converges in distribution to a standard normal random variable? Hint: Note that a chi-square random variable can be represented as the sum of independent and identically distributed random variables.

   (c) (5) Given that $X \sim \chi^2_{(m)}$ for some integer m, $m \geq 1$,

      i. (3) Argue that the ratio $\frac{\mathrm{Var}(X)}{\mathrm{E}(X)}$ is free of the parameter m.

      ii. (2) Explain how the statistic $S^2/\overline{X}$ based on a random sample where $X_i \sim \chi^2_{(m)}$, $1 \leq i \leq n$, can be used in order to test a null-hypothesis which says that the population is distributed in accordance with a chi-square distribution (no parameter being specified) against the hypothesis that this is not the case. Note that there is neither a requirement to derive the distribution of this statistic, nor a need to compute its significance level.

2. (20) Let $X \sim \exp(\lambda)$.

   (a) (5) Show that $\mathrm{E}(\sqrt{X}) = \frac{\sqrt{\pi}}{2\sqrt{\lambda}}$. Hint: Use the facts that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ and that $\Gamma(0.5) = \sqrt{\pi}$.

   (b) (3) Let $X_i \sim \exp(\lambda)$, $i = 1, 2$, be two independent random variables. What is the value for a such that $a\sqrt{X_1 X_2}$ is an UBE for $1/\lambda$.

   (c) (4) What is the MSE of the estimator you have derived in item (b)?

   (d) (4) Show that $(X_1 + X_2)/2$ is also an UBE for $1/\lambda$.

   (e) (4) Which of the two UBEs in item(b) or item (d) comes with a lower MSE?

3. (25) For some parameter θ let x^θ, $0 \leq x \leq 1$, be the kernel a density density function of a random variable X.

   (a) (4) What is the density function of X, its CDF, its mean value and its variance?

   (b) (2) What are the possible values of θ?

   (c) (5) Based on a (single) random variable X distributed as above, design the LRT for the case where $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$ when $\theta_1 < \theta_0$.

   (d) (2) What is the actual test in the previous item when $\theta_0 = 0$ and the significance level is α?

   (e) (5) Consider now a random sample some of X_i, $1 \leq i \leq n$, all distributed as X above. What is the density function of the statistic $T = \max_{i=1}^{n} X_i$? Hint: find first the CDF of T.

   (f) (3) What is $\mathrm{E}(T)$?

   (g) (3) Based on the previous item design an estimator for θ where the only information you have from the sample is T.

4. (25) Assume $X_i \sim N(\mu, \sigma^2)$, $1 \leq i \leq n$, are n independent random variables. The following questions deal with the parameter σ^2, while the actual value of μ is not in your hands.

   (a) (6) What is the MLE for σ^2? It is an UBE? If not, can you multiply is by a constant and have in hand a UBE? If so, what is this constant? Note that the constant might be a function of n.

   (b) (3) Repeat the previous item but now for σ.

   (c) (6) Construct a $1 - \alpha$ confidence interval for σ^2.

   (d) (3) Repeat the previous item but now for σ.

   (e) (4) You want to test the null-hypothesis which says that $\sigma^2 = 9$ against the alternative which says that $\sigma^2 = 4$. Construct a test with a significance level of α.

   (f) (3) What is the power of the test suggested in the previous item?

5. (15) A researcher wants to check if having brown eyes or not is independent of being a boy or a girl. Towards this goal, she samples 80 individuals from the target population and got the following results:

	brown eyes	not brown	total
boys	10	11	21
girls	46	13	59
total	56	24	80

   (a) (4) State the chi-square test with a significance level of 0.05 (you can use the table below). What is the distribution of the test statistic under the null hypothesis?

   (b) (4) What is the value of the chi-square statistic based on the above data? (you are asked to plug in all required numerical values in the formula but there is no need to find the actual value).

   (c) (4) Suppose the value you got above turned out to equal 4.5. Based on the table given below, give the narrowest possible interval the corresponding P-value belongs too.

   (d) (3) Suppose now that Fisher's test is conducted for the same purpose but where the alternative is that there are proportionally less boys than girls with brown eyes. What is the resulting P-value? Here you are asked to express your answer using the summation sign, without the need to derive numerically the final value.

   Some percentiles for a chi-square distributions (the number of degrees of freedom is suppressed).

p	0.9	0.95	0.975	0.99	0.995
$\chi^2_{(.,p)}$	2.71	3.84	5.02	6.63	7.88