Friday, August 25, 2023 10:30 PM

- Counting methods
  - Multiplication principle

MORE EVENT OPERATIONS

(a) $A \cap A' = \emptyset$ (b) $A \cap \emptyset = \emptyset$

(c) $A \cup A' = S$ (d) $(A')' = A$

- (e) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- (f) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(g) $A \cup B = A \cup (B \cap A')$

(h) $A = (A \cap B) \cup (A \cap B')$

DE MORGAN'S LAW

For any n events $A_1, A_2, \ldots, A_n$,

(i) $(A_1 \cup A_2 \cup \ldots \cup A_n)' = A_1' \cap A_2' \cap \ldots \cap A_n'$.

- A special case: $(A \cup B)' = A' \cap B'$.

(j) $(A_1 \cap A_2 \cap \ldots \cap A_n)' = A_1' \cup A_2' \cup \ldots \cup A_n'$.

A special case: $(A \cap B)' = A' \cup B'$.

For any two events A and B with $P(A) > 0$, the ***conditional probability*** *of B given that A has occurred is defined by*

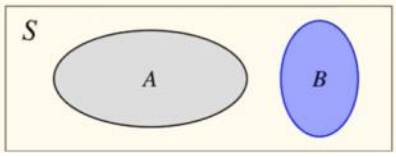
- $$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

INDEPENDENT VS MUTUALLY EXCLUSIVE

Independence and **mutually exclusivity** are totally different concepts:

$$A, B \text{ independent} \Leftrightarrow P(A \cap B) = P(A)P(B)$$
$$A, B \text{ mutually exclusive} \Leftrightarrow A \cap B = \emptyset$$

- "Mutually exclusivity" can be illustrated by a Venn diagram (like below), but we can not do that for "independence".



THEOREM 12 (BAYES' THEOREM)

Let $A_1, A_2, \ldots, A_n$ be a partition of S, then for any event B and $k = 1, 2, \ldots, n$,

- $$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_k)}.$$

- Ch 2

REMARK

So a random variable X is a function from S to $\mathbb{R}$:

- $$X : S \mapsto \mathbb{R}.$$

- Uppercase letters used to denote random variables
- Lowercase letters to denote their observed values in experiment
- The set {X=x} = {s subset S: X(s) = x} is a subset of S

(iii) The set $\{X = x\} = \{s \in S : X(s) = x\}$ is a subset of S.

(iv) If A is a subset of $\mathbb{R}$, the set $\{X \in A\} = \{s \in S : X(s) \in A\}$ is a subset of S.

- (v) With the above expressions, we define $P(X = x)$ and $P(X \in A)$ as

$$P(X = x) = P(\{s \in S : X(s) = x\});$$
$$P(X \in A) = P(\{s \in S : X(s) \in A\}).$$

- Discrete and continuous are 2 main types of random var
- Discrete random var - # of values in Rx is finite or countable - can be written Rx = {x1, x2 …}
- Continuous random variable - Rx is an interval or collection of intervals
- PROBABILITY MASS FUNCTION

DEFINITION 3 (PROBABILITY MASS FUNCTION)
For a discrete random variable X, define

$$f(x) = \begin{cases} P(X = x), & \text{for } x \in R_X; \\ 0, & \text{for } x \notin R_X. \end{cases}$$

- *Then $f(x)$ is known as the **probability function (pf)**, or **probability mass function (pmf)** of X.*

  *The collection of pairs $(x_i, f(x_i)), i = 1, 2, 3, \ldots$, is called the **probability distribution** of X.*

PROPERTIES OF THE PROBABILITY MASS FUNCTION
The probability mass function $f(x)$ of a discrete random variable **must** satisfy:

(1) $f(x_i) \geq 0$ for all $x_i \in R_X$;

(2) $f(x) = 0$ for all $x \notin R_X$;

- (3) $\sum_{i=1}^{\infty} f(x_i) = 1$, or $\sum_{x_i \in R_X} f(x_i) = 1$.

For any set $B \subset \mathbb{R}$, we have

$$P(X \in B) = \sum_{x_i \in B \cap R_X} f(x_i).$$

Continous random variable

- For a cont. random var X, Rx is an interval or a collection of intervals
- the probability function (pf) or probability density function (pdf) is defined to quantify the prob. The probability that X is in a certain range

DEFINITION 4 (PROBABILITY DENSITY FUNCTION)
*The **probability density function** of a continuous random variable X denoted by $f(x)$, is a function that satisfies:*

(1) $f(x) \geq 0$ for all $x \in R_X$; and $f(x) = 0$ for $x \notin R_X$;

- *(2) $\int_{R_X} f(x)\,dx = 1$;*

(3) For any a and b such that $a \leq b$,

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

REMARK

- Note that Condition (2) is equivalent to

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

since $f(x) = 0$ for $x \notin R_X$.

- For any specific value x_0, we have

$$P(X = x_0) = \int_{x_0}^{x_0} f(x)\,dx = 0.$$

The gives an example of "$P(A) = 0$, but A is not necessarily $\emptyset$."

- To check that a function $f(x)$ is a probability density function, it suffices to check Conditions (1) and (2). Namely,

  (1) $f(x) \geq 0$ for all $x \in R_X$; and $f(x) = 0$ for $x \notin R_X$.

  (2) $\int_{R_X} f(x)\,dx = 1$.

Tutorial 01

Thursday, August 31, 2023 10:14 AM

Tutorial 01

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE
ST2334 PROBABILITY AND STATISTICS
SEMESTER I, AY 2023/2024

Tutorial 01

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer.**
   Let A and B be two events of sample space S. Which of the following is INCORRECT:
   (a) If $A \cup B = A$, then we must have $B \subset A$. ✓
   (b) If $A \cap B = A$, then we must have $A \subset B$. ✓
   (c) If $A \cup B \subset A$, then we must have $B \cup A \subset B$. ×
   (d) All are correct.

2. **Multiple choice question: choose the unique correct answer.**
   $((A \cup B) \cap C)' = ?$
   (a) $(A' \cap C') \cap (B' \cap C')$
   (b) $(A' \cap C') \cup (B' \cap C')$
   (c) $(A' \cup C') \cap (B' \cup C')$
   (d) $(A' \cup C') \cup (B' \cup C')$

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

3. **Fill in the blank.**
   There are 5 vowels and 21 consonants among 26 alphabets. If a sample of 3 alphabets are selected without replacement, how many samples have at least 1 vowels?
   Answer: ~~7020~~ 1270

   not permutation
   $26 \cdot 25 \cdot 24$ − no vowels
   $21 \cdot 20 \cdot 19$
   ${}_{26}C_3 - {}_{21}C_3$

4. **Fill in the blank.**
   How many ways can 4 men and 3 women sit in a row if no two women are allowed to sit together?
   Answer: 1440

   $4! \times 3! \cdot {}_5C_3$
   _ M _ M _ M _ M _ 5 space for 3 W

5. **Fill in the blank.**
   A contractor wishes to build 9 houses, each of different in design in 9 plots of land. In how many ways can be placed these houses on a street if 6 lots are on South side of the street and 3 lots are on the North side? (**Note:** The 9 lots are fixed.)
   Answer: 3628800

   $(6! \cdot 3!) \times$ ways of grouping into 3 and 6
   ${}_9C_3 (6! \cdot 3!)$

Analytical Questions

1. The NUS library has five copies of a certain text on reserve. Two copies (1 and 2) are first editions, and the other three (3, 4 and 5) are second editions. A student examines these books in random order, stopping only when a second edition has been selected. One possible outcome is 5, and another is 213.

1 2 3 2 1 3 3
1 2 4 2 1 4 4
1 2 5 2 1 5 5
1 3 2 3
1 4 2 4
1 5 2 5

(i) List the outcomes in the sample space S. 15 outcomes

3 4 5

14 24
15 25

(i) List the outcomes in the sample space S. 15 outcomes

(ii) Let A denote the event that exactly one book must be examined. List the outcomes in A. 3 4 5

(iii) Let B be the event that book 5 is the one selected. List the outcomes in B. 5, 15, 25, 125, 215

(iv) Let C be the event that book 1 is not examined. List the outcomes in C. 3, 4, 5, 23, 24, 25

(v) List the outcomes in $A \cap B$, $A \cup B$, and $A \cap B \cap C$ respectively. Are A and B mutually exclusive? no

5 — 5, 15, 25, 125, 215 — 5

2. Consider the digits 0, 2, 4, 6, 8 and 9. If each digit can be used only once,

(i) how many three-digit numbers can be formed? $6 \cdot 5 \cdot 4 - 1 \cdot 5 \cdot 4 = 100$

(ii) how many of these numbers in (i) are odd numbers? $5 \cdot 4 \cdot 4 - 1 \cdot 4 \cdot 1 = 16$

(iii) how many of these odd numbers in (ii) are greater than or equal to 620? 869 809 849 829 — 9 — 689 649 629 625 — 7

3. An exam paper consists of seven questions. Candidates are asked to answer five questions. Find the number of ways to select five questions (in each of the following cases) if

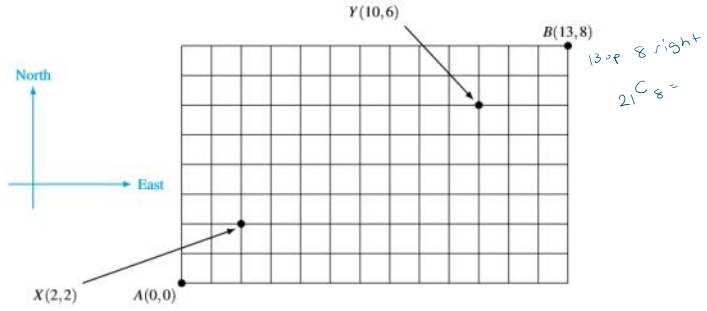
(i) there are no restrictions; ${}_7C_5 = \frac{7!}{5!2!} = 21$

(ii) the first two questions must be answered; _ _ _ _ _ ${}_5C_3 = 10$

(iii) at least one of the first two questions must be answered; and _ _ _ _ _ ${}_6C_5 = 21 - 1 = 20$

(iv) exactly two from the first three questions must be answered. ${}_3C_2 \cdot {}_4C_3 = 12$

4. Little Red Riding Hood lives at point $A : (0,0)$, and wants to visit her grandmother at point $B : (13,8)$. At each step, she can only go East (Right) or North (Up) along the grid as shown below. The Big Bad Wolf lives at $Y : (10,6)$.



13 up 8 right
${}_{21}C_8 =$

(i) How many ways can Little Red Riding Hood go to visit her grandmother regardless of whether she will pass by the Big Bad Wolf? 203490

(ii) How many ways can she go to visit her grandmother avoiding the Big Bad Wolf?

(iii) Little Red Riding Hood wants to buy a gift for her grandmother at $X : (2,2)$. How many ways can she go to visit her grandmother stopping by X but avoiding Y?

Answers for Some of the Analytical Questions

2. (i) 100; (ii) 16; (iii) 7.
3. (i) 21; (ii) 10; (iii) 20; (iv) 12.
4. (i) 203490; (ii) 123410; (iii) 44556.

ii.
# of ways from A → Y
${}_{16}C_{10}$
# of ways from Y → B
is ${}_5C_2$
# of ways to get to B and stop at Y is
${}_{16}C_{10} \cdot {}_5C_2 = 80080$
total − 80080 = 123410

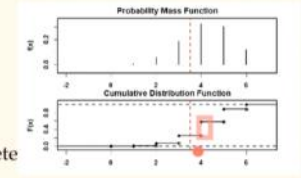
iii. ways to visit X ${}_4C_2$
ways to visit X → Y = ${}_{12}C_8$
visit y → B = ${}_5C_2$
total ways X → B = ${}_{17}C_{11}$
${}_4C_2 \cdot ({}_{17}C_{11} - {}_{42}C_8 \cdot {}_5C_2)$
= 44556

- **DEFINITION 5 (CUMULATIVE DISTRIBUTION FUNCTION)**
  *For any random variable X, we define its **cumulative distribution function (cdf)** by*
  $$F(x) = P(X \leq x).$$

- **CDF: DISCRETE RANDOM VARIABLE**

  If X is a discrete random variable, we have

  $$F(x) = \sum_{t \in R_X, t \leq x} f(t) = \sum_{t \in R_X, t \leq x} P(X = t)$$

  The cumulative distribution function of a discrete random variable is a step function.

- For any two numbers $a < b$, we have

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

  where "$a-$" represents the "largest value in R_X that is smaller than a". Mathematically,

  $$F(a-) = \lim_{x \uparrow a} F(x).$$

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

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

- Since F() has only 4 possible vals, distrib has to be discrete

- **CDF: CONTINUOUS RANDOM VARIABLE**

  If X is a continuous random variable,

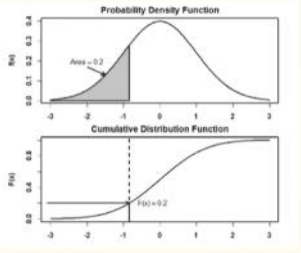
  $$F(x) = \int_{-\infty}^{x} f(t)dt,$$

  and

  $$f(x) = \frac{dF(x)}{dx}.$$

  Further

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

Expectation and Variance

DEFINITION 6 (EXPECTATION: DISCRETE RANDOM VARIABLE)
*Let X be a discrete random variable with $R_X = \{x_1, x_2, x_3, \ldots\}$ and probability function $f(x)$. The **expectation** or **mean** of X is defined by*

$$E(X) = \sum_{x_i \in R_X} x_i f(x_i).$$

By convention, we also denote $\mu_X = E(X)$.

EXAMPLE 2.10
The probability density function of weekly gravel sales X is

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

We then have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{dx} = \int_0^1 \tfrac{3x}{2}(1-x^2)\,\mathrm{dx} \\ &= \tfrac{3}{2}\int_0^1 (x-x^3)\,\mathrm{dx} = \tfrac{3}{2}\left(\tfrac{x^2}{2} - \tfrac{x^4}{4}\right)\Big|_0^1 = 3/8. \end{aligned}$$

PROPERTIES OF EXPECTATION

(1) Let X be a random variable, and let a and b be any real numbers. Then

$$E(aX+b) = aE(X)+b.$$

(2) Let X and Y be two random variables. We have

$$E(X+Y) = E(X)+E(Y).$$

X~fx Z = X+Y
Y~fy fz ?

(3) Let $g(\cdot)$ be an arbitrary function.

- If X is a **discrete** random variable with probability mass function $f(x)$ and range R_X,
$$E[g(X)] = \sum_{x \in R_X} g(x) f(x).$$
- If X is a **continuous** random variable with probability density function $f(x)$ and range R_X,
$$E[g(X)] = \int_{R_X} g(x) f(x)\,\mathrm{dx}.$$

DEFINITION 8 (VARIANCE)
*Let X be a random variable. The **variance** of X is defined as*

$$\sigma_X^2 = V(X) = E(X-\mu_X)^2.$$

- $V(X) \geq 0$ for any X. Equality holds if and only $P(X = E(X)) = 1$, that is, when X is a **constant**.
- Let a and b be any real numbers, then $V(aX + b) = a^2 V(X)$.
- The variance can also be computed by an alternative formula:

$$V(X) = E(X^2) - [E(X)]^2.$$

- The positive square root of the variance is defined as the **standard deviation** of X:

$$\sigma_X = \sqrt{V(X)}.$$

Tut 2

Wednesday, September 6, 2023 8:26 PM

Tutorial 02

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE

ST2334 PROBABILITY AND STATISTICS

SEMESTER I, AY 2023/2024

Tutorial 02

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer.**

Let A and B be events satisfying $P(A \cup B) = P(A) + P(B)$. Which of the following statement is **NOT** true?

(a) If A and B are independent, then $P(A) = 0$ or $P(B) = 0$.
(b) If $A \neq B$, then A and B are mutually exclusive.
(c) If $P(A) > P(B) > 0$, then A and B are not independent.
(d) If $A = S$, the sample space, then $P(B) = 0$.

2. **Multiple choice question: choose the unique correct answer.**

Draw 4 balls randomly without replacement from a basket containing 4 blue balls, 4 green balls, and 2 red balls. What is the probability to get 2 blue balls, 1 green ball, and 1 red ball?

(a) 7/105
(b) 8/105
(c) 9/35
(d) 8/35

3. **Multiple choice question: choose the unique correct answer.**

The probability that a Singaporean company will set up a factory in City A is 0.7. The probability that it will set up a factory in City B is 0.4, and the probability that it will set up in either City A or City B or both is 0.8. Which of the following statements is **INCORRECT**?

(a) The probability that the company will set up a factory in both cities is 0.3.
(b) The probability that the company will set up a factory in neither city is 0.2.
(c) Whether the company will set up a factory in City A will **not** affect whether it will set up a factory in City B.
(d) Whether the company will set up a factory in City B will affect whether it will set up a factory in City A.
(e) None of the given options.

4. **Fill in the blank.**

Consider 5-card poker hands dealt from a standard 52 card deck. Two important events are

$$A = \{\text{You draw a flush}\}, \quad B = \{\text{You draw a straight}\}.$$

- A flush means that you have 5 cards from the same suit.

- A straight means that the 5 cards are in numerical order, e.g., 9 of diamonds, 10 of hearts, jack of hearts, queen of spades and king of spades.
- We assume that the ace can come before 2, or after the king.
- A straight flush, i.e., 5 consecutive cards of the same suit, is not a straight.

If you are dealt a 5-card hand, find the following probabilities:

(a) $P(A) =$ 0.001381

(b) $P(B) =$ 0.003925

Analytical Questions

1. Suppose there are 500 applicants for five *equivalent* positions at a factory. The company is able to narrow the field to 30 equally qualified applicants. Seven of the finalists are minority candidates. Assume that the five who are chosen are selected at random from this final group of thirty.

   (a) In how many ways can the selection be made?

   (b) What is the probability that none of the minority candidates are hired?

   (c) What is the probability that no more than one minority candidates is hired?

2. There are two intersections with traffic lights along the route taken by a motorist driving to work. The probability that he must stop at the first light is 0.4, the probability that he must stop at the second light is 0.5, and the probability that he must stop at at least one of the two lights is 0.6. What is the probability that he must stop

   (a) at both lights? 0.3

   (b) at exactly one light? 0.6 0.3

   (c) at neither light? 0.7 0.4

   (d) at the second light given that he has stopped at the first light? 0.75

   Is the event "stopping at the first traffic light" independent of the event "stopping at the second traffic light"? No

3. A soft-drink bottling company maintains records concerning the number of unacceptable bottles of soft drink obtained from the filling and capping machines. Based on the past data, the probability that a bottle came from machine I and was nonconforming is 0.01, and the probability that a bottle came from machine II and was nonconforming is 0.025. Half the bottles are filled on machine I and the other half are filled on machine II. If a filled bottle of soft drink is selected at random, what is the probability that

   (a) it is a nonconforming bottle? 0.035

   (b) it was filled on machine II? 0.5

   (c) it was filled on machine II and is a conforming bottle? 0.0475

   (d) It was filled on machine I or is a conforming bottle? 0.975

   (e) Suppose you know that the bottle was produced on machine I. What is the probability that it is nonconforming?

   (f) Suppose you know that the bottle is nonconforming. What is the probability that it was produced on machine I?

   Explain the difference in the answers to (3e) and (3f).

Answers for Some of the Analytical Questions

1. (a) 142506; (b) 0.2361; (c) 0.6711.
2. (a) 0.3; (b) 0.3; (c) 0.4; (d) 0.75. Not independent
3. (a) 0.035; (b) 0.5; (c) 0.475; (d) 0.975; (e) 0.02; (f) 0.2857.

CDF

Thursday, September 7, 2023 10:46 AM

- Cumulative distribution functions are used to calc area under the curve to the left from a point of interest. Evaluates the accumulated probability
- For continous probability distributions, the probability = area under curve since total area = 1
- Probability density function is f(x) which describes the shape of the distributions

Tut3

Wednesday, September 13, 2023 12:10 AM

Tutorial 03 - Revised

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE
ST2334 PROBABILITY AND STATISTICS
SEMESTER I, AY 2023/2024

Tutorial 03

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer.**
   Which of the following can define probability distributions?

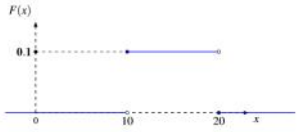
   (a) $f(x) = x/14$ for $x = 0, 1, 2, 3, 4$.
   (b) $f(x) = 3 - x^2/4$ for $x = 0, 1, 2$.
   (c) $f(x) = 1/5$ for $x = 5, 6, 7, 8, 9$.
   (d) $f(x) = 2x + 1/50$ for $x = 1, 2, 3, 4, 5$.

2. **Multiple choice question: choose the unique correct answer.**
   Which of the following figures draws the cumulative distribution function, $F(x)$, (solid blue curve) for the random variable X, whose probability function is given by

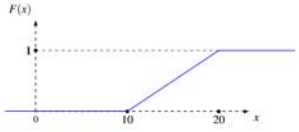
$$f(x) = \begin{cases} 0.1 & 10 \le x \le 20 \\ 0 & \text{elsewhere} \end{cases}.$$

Once pdf given cdf is fixed

   (a)

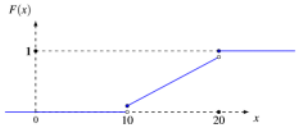


   (b)

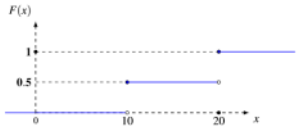


   (c)

1



(d)



3. **Multiple choice question: choose the unique correct answer**.

A worker needs to drive to work from his home daily. There is only one route available, on which there are two speeding cameras working independently. The speeding cameras at each of these locations operates 50% and 75% of the time respectively. Based on the worker's driving habit, he will speed 40% of the time; and whether he will speed at different time points are independent. What is the probability that the worker will not receive a speeding ticket for each day?

Note: whether the camera is working at any time is also independent with whether a driver is speeding when s/he drives through that camera.

(a) 0.56
(b) 0.48
(c) 0.36
(d) 0.72

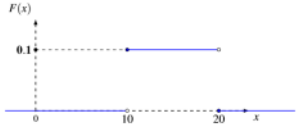
4. **Multiple response question: choose all that apply**.

Which of the following figures draws the cumulative distribution function, $F(x)$, (solid blue curve) for some random variable X?

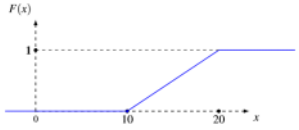
should be non decreasing

$0 \le F(x) \le 1$

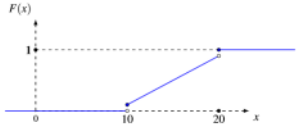
(a)



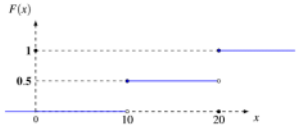
(b)



(c)



(d)



5. **True/False**.

Let X be a continuous random variable; and let Y be a discrete random variable. It is possible to find a real number a, such that $P(X = a) > P(Y = a)$.

⊙TRUE
• FALSE

Analytical Questions

1. A manufacturer of printed circuit boards exposes all finished boards to an online automated verification test. During one period, 900 boards were completed and 890 passed the test. The test is not infallible. Of 30 boards intentionally made to have noticeable defects, 25 were detected by the test.

(a) Approximate P(board passes test|board has defects).

$P(P|D) = 1 - \frac{25}{30} = \frac{5}{30}$

$\frac{890}{900} = P(\text{pass})$ $P(F|D) = \frac{25}{30}$

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

1.)

A = pass test, D = defects

$P(A|D) = \frac{25}{30}$ $P(A|D) = 1 - \frac{25}{30} = \frac{5}{30}$

$P(A) = \frac{890}{900}$

$P(A|D') = 1$ $P(A'|D') = 0$

$P(A) = P(A|D)\cdot P(D) + P(A|D')P(D') = \frac{890}{900}$

$\frac{5}{30}\cdot P(D) + 1\cdot(1-P(D)) = \text{ans}$ ←

P(defect)=

NOTE

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

2.) if $A_1 \cdots A_n$ is a partition

$P(B) = \sum_{i=1}^{n} P(A_i)\,P(B|A_i)$

3.) Bayes theorem

$P(A_K|B) = \frac{P(A_K \cap B)}{P(B)} = \frac{P(A_K \cap B)}{\sum_{i=1} P(A_i)P(B|A_i)}$

P(defect)=

(b) Give an approximate value for the probability that a manufactured board will have defects. In order to answer the question, you need information about the conditional probability that a good board will fail the test. This is important to know but was not available at the time an answer was required. To proceed, you can assume that this probability is zero. P(F | good)

(c) Approximate the probability that a board has defects given that it passed the automated test.

2. For customers purchasing a full set of tires at a particular tire store, consider the events

A = {tires purchased were made in the United States},
B = {purchaser has tires balanced immediately},
C = {purchaser requests for front-end alignment}.

Assume the following unconditional and conditional probabilities:

$P(A) = 0.75, \quad P(B|A) = 0.9, \quad P(B|A') = 0.8, \quad P(C|A \cap B) = 0.8, \quad P(C|A' \cap B) = 0.7.$

Compute the following probabilities:

(a) $P(A \cap B \cap C)$.
(b) $P(B)$.
(c) $P(A|B)$, the probability that the tires purchased were made in the United States, given that the purchaser has tires balanced immediately.
(d) $P(B \cap C)$.
(e) $P(A|B \cap C)$, the probability that the tires purchased were made in the United States, given that the purchaser has tires balanced immediately and requests for front-end alignment.

3. Total quality management (TQM) is a management philosophy and system of management techniques to improve product and service quality and worker productivity. TQM involves such techniques as teamwork, empowerment of workers, improved communication with customers, evaluation of work processes, and statistical analysis of processes and their output. One hundred Singapore companies were surveyed and it was found that 30 had implemented TQM. Among the 100 companies surveyed, 60 reported an increase in sales last year. Of those 60, 20 had implemented TQM. Suppose one of the 100 surveyed companies is to be selected randomly for additional analysis.

(a) What is the probability that a firm that implemented TQM is selected? That a firm whose sales increased is selected?
(b) Are the two events {TQM implemented} and {Sales increased} independent or dependent? Explain.
(c) Suppose that among the 60 firms reporting sales increases, there were only 18 TQM-implementers (instead of 20). Now are the events {TQM implemented} and {Sales increased} independent or dependent? Explain.

4. A company uses three different assembly lines, A_1, A_2, and A_3, to manufacture a particular component. Of those manufactured by line A_1, 5% need rework to remedy a defect, whereas 8% of A_2's components need rework, and 10% of A_3's components need rework. Suppose that 50% of all components are produced by line A_1, while 30% are produced by line A_2, and 20% come from line A_3. If a randomly selected component needs rework, what is the probability that it came

(a) from line A_1?
(b) from line A_2?
(c) from line A_3?

Answers for Some of the Analytical Questions

1. (a) 5/30; (b) 0.013; (c) 0.00225.
2. (a) 0.54; (b) 0.875; (c) 0.7714; (d) 0.68; (e) 0.7941.
3. (a) 0.3; 0.6. (b) $A \not\perp B$; (c) $A \perp B$.
4. (a) 0.3623; (b) 0.3478; (c) 0.2899.

$P(A) = P(A|D) \cdot P(D) + P(A|D')P(D') = \frac{890}{900}$

$\frac{5}{30} \cdot P(D) + 1 \cdot (1 - P(D)) = \text{ans}$ ←

$P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{P(A|D) \cdot P(D)}{P(A|D)P(D) + P(A|D')P(D')} = \text{ans}$ ←

2.) a. $P(A \cap B \cap C) = P(A|(B \cap C)) \cdot P($ (consider as one) $= P(A) \cdot P(B|A) \cdot P(C|A \cap B) = 0.54$

b. $P(B) = P(B|A) \cdot P(A) + P(B|A') \cdot P(A') = 0.875$

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

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

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

e. $P(A|B \cap C) = \frac{P(A \cap (B \cap C))}{P(B \cap C)} = \frac{0.54}{0.68} = \text{ans}$

3.) A ← TQM implemented
B ← sale ↑
$P(A) = \frac{30}{100}$ $P(B) = \frac{60}{100}$

b.

$P(A \cap B) \neq P(A) \cdot P(B)$

$P(A|B)\ P(B) = \frac{1}{3} \cdot 0.6 = 0.2$

c. $P(A|B) = \frac{18}{60} = \frac{3}{10}$

$P(A \cap B) = P(A|B)P(B) = 0.3 \cdot 0.6 = 0.18$

$P(A)\ P(B) = 0.18$

4.) B is component needs remark

$P(B) = P(A_1)P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3)P(B|A_3)$

$= 0.5 \cdot 0.05 + 0.3 \cdot 0.08 + 0.2 \cdot 0.1$

$= 0.069$

$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(B|A_1) \cdot P(A_1)}{P(B)} = \frac{0.5 \cdot 0.05}{0.069} = 0.3623$

$P(A_2|B)$ = " " " "

$P(A_3|B)$ = " " " "

Wk5

Friday, September 15, 2023 1:36 PM

Joint Distribution

- Joint distributions for multiple random var
  - When we are interested in >1 random var simultaneously
- 2 dimensional random vector
  - **Definition 1 (Two-dimensional Random Vector)**
    Let E be an experiment and S be a corresponding sample space. Suppose X and Y are two functions each assigning a real number to each $s \in S$.
    *We call (X,Y) a **two-dimensional random vector**, or a **two-dimensional random variable**.*

```
X: height
Y: weight
```

- **Definition 3 (n-dimensional Random Vector)**
  *Let $X_1, X_2, \ldots, X_n$ be n functions each assigning a real number to **every outcome** $s \in S$.*
  *We call $(X_1, X_2, \ldots, X_n)$ a **n-dimensional random vector**, or a **n-dimensional random variable**.*

- **Definition 4**
  *(X,Y) is a **discrete two-dimensional random variable** if the number of possible values of $(X(s), Y(s))$ are finite or countable. That is, the possible values of $(X(s), Y(s))$ may be represented by*

  $$(x_i, y_j), \quad i = 1,2,3,\ldots; j = 1,2,3,\ldots$$

  *(X,Y) is a **continuous two-dimensional random variable** if the possible values of $(X(s), Y(s))$ can assume any value in some region of the Euclidean space $\mathbb{R}^2$.*

- If both X and Y are discrete/continuous then (X,Y) are discrete/continuous respectively

- **Definition 5 (Discrete joint probability function)**
  *Let (X,Y) be a 2-dimensional **discrete** random variable. Its **joint probability (mass) function** is defined by*

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

  for $(x,y) \in R_{X,Y}$.

Properties of the discrete joint probability function

The joint probability mass function has the following properties:

(1) $f_{X,Y}(x,y) \geq 0$ for any $(x,y) \in R_{X,Y}$.

(2) $f_{X,Y}(x,y) = 0$ for any $(x,y) \notin R_{X,Y}$.

- (3) $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} f_{X,Y}(x_i,y_j) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = 1.$
  Equivalently, $\sum\sum_{(x,y)\in R_{X,Y}} f(x,y) = 1.$

(4) Let A be any subset of $R_{X,Y}$, then

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

DEFINITION 6 (CONTINUOUS JOINT PROBABILITY FUNCTION)
Let (X,Y) be a 2-dimensional ***continuous*** *random variable. Its* ***joint probability (density) function*** *is a function $f_{X,Y}(x,y)$ such that*

$$P((X,Y)\in D)=\iint_{(x,y)\in D} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x,$$

for any $D\subset\mathbb{R}^2$. More specifically,

$$P(a\le X\le b, c\le Y\le d)=\int_a^b\int_c^d f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x\,.$$

PROPERTIES OF THE CONTINUOUS JOINT PROBABILITY FUNCTION
The joint probability density function has the following properties:

(1) $f_{X,Y}(x,y)\ge 0$, for any $(x,y)\in R_{X,Y}$.

(2) $f_{X,Y}(x,y)=0$, for any $(x,y)\notin R_{X,Y}$.

(3) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y=1$.

Equivalently, $\iint_{(x,y)\in R_{X,Y}} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y=1$.

Marginal and Conditional Distributions

DEFINITION 7 (MARGINAL PROBABILITY DISTRIBUTION)
Let (X,Y) be a two-dimensional random variable with joint probability function $f_{X,Y}(x,y)$. We define the ***marginal distribution*** *of X as follows.*

If Y is a discrete random variable, then for any x,

$$f_X(x)=\sum_y f_{X,Y}(x,y).$$

If Y is a continuous random variable, then for any x,

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,.$$

REMARK

- $f_Y(y)$ for Y is defined in the same way as that of X.
- We can view the marginal distribution as the "projection" of the 2D function $f_{X,Y}(x,y)$ to the 1D function.
- Intuitively, it is the distribution of X by ignoring the presence of Y.

  For example, consider a person from a certain community.
  - Suppose $X=$ body weight, $Y=$ height, and (X,Y) has joint distribution $f_{X,Y}(x,y)$.
  - The marginal distribution $f_X(x)$ of X is the **distribution of body weights for all people in the community**.

- $f_X(x)$ should not involve the variable y. This can be viewed from its definition: y is either summed out or integrated over.
- $f_X(x)$ is a **probability function**; so it satisfies all the properties of the probability function.

EXAMPLE 3.4

We revisit Example 3.2. The joint probability function is given by

$$f(x,y) = \frac{1}{36}xy, \quad \text{for } x = 1,2,3 \text{ and } y = 1,2,3.$$

Note that X has three possible values: 1, 2, and 3. The marginal distribution for X is given by

- for $x = 1$, $f_X(1) = f(1,1) + f(1,2) + f(1,3) = 6/36 = 1/6$.
- for $x = 2$, $f_X(2) = f(2,1) + f(2,2) + f(2,3) = 12/36 = 1/3$.
- for $x = 3$, $f_X(3) = f(3,1) + f(3,2) + f(3,3) = 18/36 = 1/2$.

For other values of x, $f_X(x) = 0$.

Alternatively, for each $x \in \{1,2,3\}$,

$$f_X(x) = \sum_y f(x,y) = \sum_{y=1}^{3} \frac{1}{36}xy = \frac{1}{36}x\sum_{y=1}^{3} y = \frac{1}{6}x.$$

Conditional Distribution

DEFINITION 8 (CONDITIONAL DISTRIBUTION)

*Let (X,Y) be a random variable with joint probability function $f_{X,Y}(x,y)$. Let $f_X(x)$ be the marginal probability function for X. Then for any x such that $f_X(x) > 0$, the **conditional probability function of Y given $X = x$** is defined to be*

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

REMARK

- For any y such that $f_Y(y) > 0$, we can similarly define the **conditional distribution of X given $Y = y$** as

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

- $f_{Y|X}(y|x)$ is defined only for x such that $f_X(x) > 0$; likewise $f_{X|Y}(x|y)$ is defined only for y such that $f_Y(y) > 0$.
- The intuitive meaning of $f_{Y|X}(y|x)$: the distribution of Y given that the random variable X is observed to take the value x.

- Considering y as the variable (and x as a fixed value), $f_{Y|X}(y|x)$ is a probability function, so it must satisfy all the properties of a probability function.
- However, $f_{Y|X}(y|x)$ is not a probability function for x. This means that there is **NO** requirement that
  - $\int_{-\infty}^{\infty} f_{Y|X}(y|x)\,dx = 1$, for X continuous; or
  - $\sum_x f_{Y|X}(y|x) = 1$, for X discrete.
- With this definition, we immediately have
  - If $f_X(x) > 0$, $f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x)$.
  - If $f_Y(y) > 0$, $f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x|y)$.

- One immediate application of the conditional distribution is to compute, for

- One immediate application of the conditional distribution is to compute, for continuous random variable,

$$P(Y \le y|X=x) = \int_{-\infty}^{y} f_{Y|X}(y|x)\,dy;$$
$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,dy.$$

Y = weight
X = height
y = 50 kg
x = 1.7 m

Their interpretations are clear: the former is the probability that $Y \le y$, given $X = x$; the latter is the average value of Y given $X = x$.

For the discrete case, the results can be similarly established, based on the definition of $f_{Y|X}(y|x)$.

EXAMPLE 3.5

We revisit Examples 3.2 and 3.4. The joint probability function for (X,Y) is given by

$$f(x,y) = xy/36, \quad \text{for } x = 1,2,3 \text{ and } y = 1,2,3.$$

The marginal probability function for X is

$$f_X(x) = x/6, \quad \text{for } x = 1,2,3.$$

Therefore $f_{Y|X}(y|x)$ is defined for any $x = 1,2,3$:

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{(xy/36)}{(x/6)} = y/6, \quad \text{for } y = 1,2,3.$$

We can also compute

$$P(Y=2|X=1) = f_{Y|X}(2|1) = \frac{1}{6} \times 2 = 1/3;$$

$$P(Y \le 2|X=1) = P(Y=1|X=1) + P(Y=2|X=1)$$
$$= f_{Y|X}(1|1) + f_{Y|X}(2|1) = 1/6 + 1/3 = 1/2;$$

$$E(Y|X=2) = 1 \cdot f_{Y|X}(1|2) + 2 \cdot f_{Y|X}(2|2) + 3 \cdot f_{Y|X}(3|2)$$
$$= 1 \cdot (1/6) + 2 \cdot (2/6) + 3 \cdot (3/6) = 7/3.$$

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{(xy/36)}{(x/6)} = y/6, \quad \text{for } y = 1,2,3.$$

$f_X(1) \cdot 2$

$\int \frac{xy}{36}\,dx$

$\frac{x^2 y}{2 \cdot 36}\Big|_1^3 = \frac{9y}{72} - \frac{y}{72} = \frac{8y}{72}$

$\frac{y}{9}$

Tutorial 04

Wednesday, September 20, 2023 11:54 AM

Tutorial 04

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE
ST2334 PROBABILITY AND STATISTICS
SEMESTER I, AY 2023/2024

Tutorial 04

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer**.

   An insurance company offers its policyholders a number of different premium payment options. For a randomly selected policyholder, let X be the number of months between successive payments. The cumulative distribution function of X is given as follows.

$$F_X(x) = \begin{cases} 0, & x < 1; \\ 0.3, & 1 \le x < 3; \\ 0.4, & 3 \le x < 4; \\ 0.45, & 4 \le x < 6; \\ 0.6, & 6 \le x < 12; \\ 1, & 12 \le x. \end{cases}$$

   Which of the following is **INCORRECT**?

   (a) $P(3 \le X \le 6) = 0.3$
   (b) $P(X \ge 4) = 0.6$
   (c) $P(X = 3) = 0.15$
   (d) $P(2 < X < 4) = 0.1$

2. **Multiple choice question: choose the unique correct answer**.

   An insurance company offers its policyholders a number of different premium payment options. For a randomly selected policyholder, let X be the number of months between successive payments. The cumulative distribution function of X is given as follows.

$$F_X(x) = \begin{cases} 0, & x < 1; \\ 0.3, & 1 \le x < 3; \\ 0.4, & 3 \le x < 4; \\ 0.45, & 4 \le x < 6; \\ 0.6, & 6 \le x < 12; \\ 1, & 12 \le x. \end{cases}$$

   Which of the following is the probability function of X?

   (a)

x	1	3	4	6	12
$f_X(x)$	0.3	0.1	0.05	0.15	0.4

   (b)

x	1	3	4	6	12
$f_X(x)$	0.3	0.1	0.45	0.05	0.4

(c)

x	1	3	4	6	12
$f_X(x)$	0	0.3	0.4	0.6	1

(d) None of the given options.

3. **Multiple response question: choose all that apply**.

Which of the following is/are valid cumulative distribution function(s)?

(a) $F(x) = 0.5e^x$ when $x \le 0$ and $F(x) = 1 - 0.5e^{-x}$ when $x > 0$.

(b) $F(x) = x$ when $0 \le x < 1$ and $F(x) = 0$ elsewhere.

(c) $F(x) = e^x$ when $x \le 0$ and $F(x) = 1 - e^{-x}$ when $x > 0$.

(d) $F(x) = 1 - e^{-x}$ when $x \ge 0$ and $F(x) = 0$ elsewhere.

4. **Multiple response question: choose all that apply**.

Let $F(x)$ be the cumulative distribution function of a random variable X. Which of the following is/are **TRUE**?

(a) If R_X is a subset of the integers, then for any integers a and b such that $a < b$, we must have $F(b) = F(a-1) + P(a \le X \le b)$.

(b) If X is a continuous random variable and $a < b$ are two real numbers, we must have $F(b) = F(a) + P(a \le X \le b)$.

(c) Let X be a continuous random variable, then $E(X)$ must be a value in R_X.

(d) $F(x)$ is a non-increasing function x.

(e) If $a < b$, then $F(a) < F(b)$.

5. **Fill in the blank**.

Determine the value c, such that the following function can serve as a probability function of a random variable X.

$$f_X(x) = \begin{cases} c(x^2+4), & x = 0,1,2,3; \\ 0, & \text{elsewhere.} \end{cases}$$

$c =$ 1/30

$4c + 5c + 8c + 13c = 1 = 30c$

Analytical Questions

1. Consider the probability function

$$f_X(x) = \begin{cases} k\sqrt{x}, & 0 < x < 1; \\ 0, & \text{elsewhere.} \end{cases}$$

a. $\int_0^1 k x^{1/2}\,dx = 1$

$k\left[\frac{2}{3}x^{3/2}\right]_0^1 = k\left[\frac{2}{3}\right] = 1 \quad k = \frac{3}{2}$

b. $F_X(x) = \int_0^x \frac{3}{2}\sqrt{x}\,dx = x^{3/2}$ for $0 < x < 1$; 0 for $x < 0$; 1 for $x \ge 1$

(a) Find the value of the constant k. $= \frac{3}{2}$

(b) Find the cumulative distribution function $F_X(x)$, and use it to evaluate $P(0.3 < X < 0.6)$. $F_X(0.6) - F_X(0.3) = 0.6^{3/2} - 0.3^{3/2} = 0.300\,41$

2. The waiting time, in hours, between successive speeders spotted by a radar unit is a continuous random variable with cumulative distribution

$$F_X(x) = \begin{cases} 0, & x \le 0; \\ 1 - e^{-8x}, & x > 0. \end{cases}$$

$2\text{min} = \frac{12}{60} = 0.2\text{hr}$

$P(x < 12\text{min}) = F_X(0.2) = 1 - e^{-8(0.2)} = 0.7981$

(a) Find the probability of waiting less than 12 minutes between successive speeders. $= 0.7981$

(b) Find the probability density function of X.

$f_X(x) = \begin{cases} 0, & x \le 0 \\ 8e^{-8x}, & x > 0 \end{cases}$

$\frac{d}{dx}\left(1 - e^{-8x}\right) = 8e^{-8x}$

3. The random variable X, representing the number of errors per 100 lines of software code, has the following probability function:

x	2	3	4	5	6
$f_X(x)$	0.01	0.25	0.40	0.30	0.04

(a) Find $E(X)$ and $E(X^2)$.

(b) Find the variance of X using (i) the definition of variance and (ii) $V(X) = E(X^2) - [E(X)]^2$.

(c) Find the mean and variance of the discrete variable $Z = 3X - 2$.

(d) Find the probability function of the random variable Z. Hence, find the mean and variance of Z directly from its probability function).

(e) Suppose that $W = aZ + b$. Find the mean and variance of W in terms of a and b.

4. The probability function of a random variable X is given by

$$f(x) = \begin{cases} x, & \text{for } 0 < x < 1; \\ 2 - x, & \text{for } 1 \le x < 2; \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find the probability that the random variable will take on a value between 0.6 and 1.2.

(b) Find $E(X)$ and $V(X)$.

a. $E(X) = 2(0.01) + 3(0.25) + 4(0.4) + 5(0.3) + 6(0.04) = 4.11$

$E(X^2) = 4(0.01) + 9(0.25) + 16(0.4) + 25(0.3) + 36(0.04) = 17.63$

b. $V(X) = E(X - \mu_x)^2$

$= (2-4.11)^2(0.01) + (3-4.11)^2(0.25) + (4-4.11)^2(0.4) + (5-4.11)^2(0.3) + (6-4.11)^2(0.04)$

$= 0.7379$

$V(X) = E(X^2) - [E(X)]^2 = 17.63 - 4.11^2 = 0.7379$

c. $E(X) = 4.11$ $E(3X-2) = 10.33$

$V(X) = 0.7379$ $V(3X-2) = 3^2(V(X)) = 6.6411$

d.

z	4	7	10	13	16
$f_Z(z)$	0.01	0.25	0.4	0.3	0.04

$\mu_Z = 4(0.01) + 7(0.25) + 10(0.4) + 13(0.3) + 16(0.04) = 10.33$

$V(Z) = 0.01(10.33-4)^2 + 0.25(10.33-7)^2 + 0.4(10.33-10)^2 + 0.3(10.33-13)^2 + 0.04(10.33-16)^2 = 6.6411$

e. $\mu(W) = 10.33a + b$ $V(W) = 6.6411a^2$

a. 0.6 to 1 $\int_{0.6}^{1} x\,dx = \left[\frac{x^2}{2}\right]_{0.6}^{1} = \frac{1}{2} - \frac{0.36}{2} = \frac{0.64}{2}$

1 to 1.2 $\int_{1}^{1.2} (2-x)\,dx = \left[2x - \frac{x^2}{2}\right]_{1}^{1.2} = 2.4 - \frac{1.44}{2} - 2 + \frac{1}{2} = 0.18$

$P(0.6 < X < 1.2) = 0.5$

b. $E(X) = \int_0^1 x^2\,dx + \int_1^2 x(2-x)\,dx = 1$

$V(X) = E(X^2) - (E(X))^2 = \frac{7}{6} - 1 = \frac{1}{6}$

$E(X^2) = \int_0^1 x^3\,dx + \int_1^2 x^2(2-x)\,dx = \frac{7}{6}$

Answers for Some of the Analytical Questions

1. (a) 3/2; (b) $F_X(x) = \begin{cases} 0, & x \le 0 \\ x^{3/2}, & 0 < x < 1; \\ 1, & x \ge 1 \end{cases}$
$P(0.3 < X < 0.6) = 0.3004$

2. (a) 0.7981; (b) $f_X(x) = 8e^{-8x}$ for $x \ge 0$.

3. (a) 4.11, 17.63; (b) 0.7379; (c) 10.33, 6.6411; (d) $E(Z) = 10.33$, $V(Z) = 6.6411$; (e) $10.33a + b$, $6.6411a^2$.

4. (a) 0.5; (b) 1; (c) 1/6.

Wk6

Thursday, September 21, 2023 2:18 AM

Independent Random Variables

- **DEFINITION 9 (INDEPENDENT RANDOM VARIABLES)**
  *Random variables X and Y are **independent** if and only if for **any** x and y,*

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

  *Random variables $X_1, X_2, \ldots, X_n$ are **independent** if and only if for **any** $x_1, x_2, \ldots, x_n$,*

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

- If joint probability function = marginal functions multiplied

- **REMARK**
  - The above definition is applicable whether (X,Y) is continuous or discrete.
  - The "product feature" in the definition implies one necessary condition for independence: $R_{X,Y}$ needs to be a product space. In the sense that if X and Y are independent, for any $x \in R_X$ and any $y \in R_Y$, we have

    $$f_{X,Y}(x,y) = f_X(x)f_Y(y) > 0,$$

    implying $R_{X,Y} = \{(x,y)|x \in R_X; y \in R_y\} = R_X \times R_Y$.

    Conclusion:
    If $R_{X,Y}$ is not a product space, then X and Y are not independent!

- **PROPERTIES OF INDEPENDENT RANDOM VARIABLES**
  Suppose X,Y are independent random variables.

  (1) If A and B are arbitrary subsets of $\mathbb{R}$, the events $X \in A$ and $Y \in B$ are independent events in S. Thus

  $$P(X \in A; Y \in B) = P(X \in A)P(Y \in B).$$

  In particular, for any real numbers x, y,

  $$P(X \le x; Y \le y) = P(X \le x)P(Y \le y).$$

  $$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

- Joint cdf is = product of the 2 marginal cdf

(2) For arbitrary functions $g_1(\cdot)$ and $g_2(\cdot)$, $g_1(X)$ and $g_2(Y)$ are independent. For example,

- X^2 and Y are independent.
- $\sin(X)$ and $\cos(Y)$ are independent.
- e^X and $\log(Y)$ are independent.

(3) Independence is connected with conditional distribution.

- If $f_X(x) > 0$, then $f_{Y|X}(y|x) = f_Y(y)$.
- If $f_Y(y) > 0$, then $f_{X|Y}(x|y) = f_X(x)$.

- In order to check if x and y are independent, must check each value

EXAMPLE 3.6

The joint probability function of (X,Y) is given below.

x	y			$f_X(x)$
	1	3	5	
2	0.1	0.2	0.1	0.4
4	0.15	0.3	0.15	0.6
$f_Y(y)$	0.25	0.5	0.25	1

Are X and Y independent? `0.1 = 0.4 x 0.25 ?` `0.2 = 0.4 x 0.5 ?`

Expectation and Covariance

DEFINITION 10 (EXPECTATION)

Consider any two variable function $g(x,y)$.

If (X,Y) is a discrete random variable,

$$E(g(X,Y)) = \sum_x \sum_y g(x,y) f_{X,Y}(x,y).$$

If (X,Y) is a continuous random variable,

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dy\, dx.$$

REMARK

If X and Y are discrete random variables,

$$\operatorname{cov}(X,Y) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y).$$

If X and Y are continuous random variables,

$$\operatorname{cov}(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\,.$$

PROPERTIES OF THE COVARIANCE

The covariance has the following properties.

(1) $\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y)$.

This is true because

$$\begin{aligned}\operatorname{cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] = E[XY - Y\mu_X - X\mu_Y + \mu_X\mu_Y \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y.\end{aligned}$$

(2) If X and Y are independent, then $\operatorname{cov}(X,Y) = 0$. However, $\operatorname{cov}(X,Y) = 0$ does not imply that X and Y are independent.

Take note that the two statements can be summarised as:

(i) $X \perp Y \Rightarrow \operatorname{cov}(X,Y) = 0$;

(ii) $X \perp Y \not\Leftarrow \operatorname{cov}(X,Y) = 0$.

For (i), note that if X and Y are independent, then $f_{X,Y}(x,y) = f_X(x) f_Y(y)$. So

$$\begin{aligned}E(XY) &= \sum_i \sum_j x_i y_j f_{X,Y}(x_i,y_j) = \sum_i \sum_j x_i y_j f_X(x_i) f_Y(y_j) \\ &= \sum_i x_i f_X(x_i) \sum_j y_j f_Y(y_j) = E(X)E(Y).\end{aligned}$$

(3) $\text{cov}(aX+b, cY+d) = ac \cdot \text{cov}(X,Y)$.

This can be derived using the following 3 formulas:

(i) $\text{cov}(X,Y) = \text{cov}(Y,X)$;

(ii) $\text{cov}(X+b,Y) = \text{cov}(X,Y)$;

(iii) $\text{cov}(aX,Y) = a\,\text{cov}(X,Y)$.

(4) $V(aX+bY) = a^2V(X) + b^2V(Y) + 2ab \cdot \text{cov}(X,Y)$.

This can be derived using the following 2 formulas:

(i) $V(aX) = a^2V(X)$;

(ii) $V(X+Y) = V(X) + V(Y) + 2\,\text{cov}(X,Y)$.

EXAMPLE 3.7

We are given the joint distribution for (X,Y):

x	y				$f_X(x)$
	0	1	2	3	
0	1/8	1/4	1/8	0	1/2
1	0	1/8	1/4	1/8	1/2
$f_Y(y)$	1/8	3/8	3/8	1/8	1

(i) Find $E(Y-X)$.

(ii) Find $\text{cov}(X,Y)$.

Solution:

(i) Method 1:

$$\begin{aligned} E(Y-X) &= (0-0)(1/8)+(1-0)(1/4)+(2-0)(1/8) \\ &\quad +\ldots+(3-1)(1/8)=1. \end{aligned}$$

- Method 2:

$$E(Y-X)=E(Y)-E(X)=1.5-0.5=1,$$

where

$$\begin{aligned} E(Y) &= 0\cdot(1/8)+1\cdot(3/8)+2\cdot(3/8)+3\cdot(1/8)=1.5 \\ E(X) &= 0\cdot(1/2)+1\cdot(1/2)=0.5. \end{aligned}$$

(ii) We use $\text{cov}(X,Y)=E(XY)-E(X)E(Y)$ to compute. Note that we have computed $E(X)$ and $E(Y)$ in Part (i).

$$\begin{aligned} E(XY) &= (0)(0)(1/8)+(0)(1)(1/4)+(0)(2)(1/8) \\ &\quad +\ldots+(1)(3)(1/8)=1. \end{aligned}$$

- Therefore

$$\text{cov}(X,Y)=E(XY)-E(X)E(Y)=1-(0.5)(1.5)=0.25.$$

$$\begin{aligned} E(Y) &= 0\cdot(1/8)+1\cdot(3/8)+2\cdot(3/8)+3\cdot(1/8)=1.5 \\ E(X) &= 0\cdot(1/2)+1\cdot(1/2)=0.5. \end{aligned}$$

Tutorial 05

Wednesday, October 4, 2023 7:55 PM

Tutorial 05

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE

ST2334 PROBABILITY AND STATISTICS

SEMESTER 1, AY 2023/2024

Tutorial 05

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer.**

   Let $f_{X,Y}(x,y)$ be the joint probability function for the random vector (X,Y). Which of the following statement is INCORRECT?

   (a) If both X and Y are discrete random variables, then for any set of real numbers: $a_1,\ldots,a_n$, and b, we must have $\sum_{i=1}^{n} f_{X,Y}(a_i,b) \le 1$.

   ✓ (b) If both X and Y are continuous random variables, then for any interval (a,b) and a real number c, we must have $\int_a^b f_{X,Y}(x,c)dx \le 1$.

   ✓ (c) Suppose that X and Y are discrete random variables; $a_1,\ldots,a_n$, and b are real numbers. If $\sum_{i=1}^{n} f_{X,Y}(a_i,b) > 0$, then we must have $P(Y=b) > 0$

   ✓ (d) Suppose that X and Y are continuous random variables; (a,b) is an interval; c is a real number. If $\int_a^b f_{X,Y}(x,c)dx > 0$, then we must have $f_Y(c) > 0$, where $f_Y(y)$ denotes the marginal probability function of Y.

2. **Multiple choice question: choose the unique correct answer.**

   Each rear tire on an experimental airplane is supposed to be filled to a pressure of 40 pound per square inch (psi). Let X denote the actual air pressure (in 10 pound per square inch) for the right tire and Y denote the actual air pressure (in 10 pound per square inch) for the left tire. Suppose that X and Y are random variables with the joint density

   $$f_{X,Y}(x,y) = \begin{cases} k(x^2+y^2), & 3 \le x \le 5, 3 \le y \le 5; \\ 0, & \text{elsewhere.} \end{cases}$$

   Which of the following is INCORRECT?

   ✓ (a) We must have $k = 3/392$.

   ✓ (b) $P(3 \le X \le 4, 4 \le Y \le 5) = 0.25$.

   (c) The marginal probability function for X is given by $f_X(x) = \frac{1}{196}(3x^2+49)$.

   (d) $P(3.5 < X < 4) = 0.1925$.

3. **Multiple choice question: choose the unique correct answer.**

   The joint probability function of (X,Y) is given below.

$k\left(\frac{98}{3}y + \frac{2y^3}{3}\right)\Big|_3^5 = k\left[80.667\right] = 1$

$= k\left[\frac{y^3}{3} + y^2x\right]_3^5$

$k\int_3^5\int_4^5 (x^2+y^2)\,dx\,dy$

$\left[\frac{x^3}{3} + y^2x\right]_3^5 = \frac{37}{3} + y^2$

$\int_4^5 \left(\frac{37}{3} + y^2\right)dy$

$= \left[\frac{37}{3}y + \frac{y^3}{3}\right]_4^5 = \frac{98}{3}$

$\iint k(x^2+y^2)\,dx\,dy$

$k\left(\frac{x^3}{3} + y^2x\right)\Big]_3^5 = k\left(\frac{98}{3} + 2y^2\right)$

$x=3,\ f(3) = f(3,3)+f(3,4)+f(3,5)$

$= \frac{3}{392}\cdot 18 + \frac{3}{392}\cdot 25 + \frac{3}{392}\cdot 34 = k\cdot 77 = \frac{33}{56}$

$x=4 \quad = \frac{3}{392}\cdot 25 + \frac{3}{392}\cdot 36 + \frac{3}{392}\cdot 41 = k\cdot 102 \quad \frac{153}{196}$

$x=5 \quad = \frac{3}{392}\cdot 34 + \frac{3}{392}\cdot 41 + \frac{3}{392}\cdot 50 = k\cdot 125 \quad \frac{375}{392}$

x 0 0 3 3
y 0 2 0 2

y \ x	0	1
0	0.25	0.35
1	0.23	0.17

$0 \cdot 0.25 + 2 \cdot 0.23 + 3 \cdot 0.35 + 5 \cdot 0.17$

Then $E(3X+2Y) = ?$ $E(3X)+E(2Y)$

(a) 1.87
(b) 2.36
(c) 2.45
(d) 3.11
(e) None of the given options.

4. **Fill in the blank.**

Let X denote the number of times a certain numerical control machine will malfunction: 1, 2, or 3 times on any given day. Let Y denote the number of times a technician is called on an emergency call. Their joint probability distribution is given below.

$f_{X,Y}(x,y)$		x = 1	x = 2	x = 3
y	1	0.05	0.05	0.1
	2	0.05	0.10	0.35
	3	0	0.2	0.1

$f_X(2) = 0.35$

Compute $P(Y=3|X=2) =$
Answer: 0.5714

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

$\frac{0.2}{0.35}$

5. **Fill in the blank.**

Let X denote the number of times a certain numerical control machine will malfunction: 1, 2, or 3 times on any given day. Let Y denote the number of times a technician is called on an emergency call. Their joint probability distribution is given below.

$f_{X,Y}(x,y)$		x = 1	x = 2	x = 3
y	1	0.05	0.05	0.1
	2	0.05	0.10	0.35
	3	0	0.2	0.1

compute $E(X|Y=2)$.
Answer: 1.3

$= 1 \cdot f_{X|Y}(1|2) + 2 \cdot f_{X|Y}(2|2) + 3 \cdot f_{X|Y}(3|2)$
$=$

Analytical Questions

1. From a sack of fruit containing 3 oranges, 2 apples, and 3 bananas, a random sample of 4 pieces of fruit is selected. If X is the number of oranges and Y is the number of apples in the sample, find

(a) the joint probability distribution of X and Y;
(b) $P(X=1, Y=1)$;

1.)

a. $f_{X,Y} = \dfrac{\binom{3}{x}\binom{2}{y}\binom{3}{4-x-y}}{\binom{8}{4}}$

b. $= \dfrac{\binom{3}{1}\binom{2}{1}\binom{3}{2}}{\binom{8}{4}} = \dfrac{18}{70} = 0.257$

c. $P(X+Y \le 2) = \dfrac{\binom{3}{1}\binom{2}{0}\binom{3}{3} + \binom{3}{1}\binom{2}{1}\binom{3}{2} + \binom{3}{0}\binom{2}{1}\binom{3}{3} + \binom{3}{2}\binom{2}{0}\binom{3}{2} + \binom{3}{0}\binom{2}{2}\binom{3}{3}}{70}$

$X=2, Y=0$
$X=1, Y=0$
$X=1, Y=1$
$X=0, Y=1$
$X=0, Y=2$
$= \frac{35}{70}$

d. $f_X(x) = \dfrac{\binom{3}{x}\binom{5}{4-x}}{\binom{8}{4}}$

e. $f_{Y|X}(y|2) = \dfrac{\binom{2}{y}\binom{3}{2-y}}{\binom{5}{2}} = \frac{1}{10}\binom{2}{y}\binom{3}{2-y}$

$P(Y=0|X=2) = \frac{3}{10}$

(c) $P(X+Y \le 2)$;
(d) $f_X(x)$;
(e) $f_{Y|X}(y|2)$ and hence $P(Y=0|X=2)$.

2. Two random variables have the joint density

$$f(x_1,x_2) = \begin{cases} x_1 x_2, & \text{for } 0 < x_1 < 2, 0 < x_2 < 1; \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find the probability that both random variables will take on values less than 1.
(b) Find the marginal densities of the two random variables, and check whether the two random variables are independent.
(c) Find the expected value of the random variable whose values are given by $g(x_1,x_2) = x_1 + x_2$.

2) a. $\int_0^1 \int_0^1 x_1 x_2 \, dx_1 dx_2 = \int \left[\frac{x_1^2}{2} x_2\right]_0^1 dx_2$
$= \int_0^1 \frac{x_2}{2} dx_2 = \left[\frac{x_2^2}{4}\right]_0^1 = \boxed{\frac{1}{4}}$

b. $f_{X_1}(x) = \sum_{x_2=0}^{1} x_1 x_2 = x_1 \sum_{x_2=0}^{1} x_2 = x_1$ (?) $\int_0^1 f(x_1,x_2)\, dx_1 = \frac{x_2}{2}$

$f_{X_2}(x) = x_2 \sum_{x_1=0}^{2} x_1 = 2x_2$

$f(x_1,x_2) = f(x_1) \cdot f(x_2)$

$x_1 x_2 = x_1 x_2 \quad x_1 \perp x_2$

c. $E(x_1+x_2) = E(x_1) + E(x_2)$

$= \int_0^1 \int_0^2 \underbrace{g(x_1,x_2)}_{(x_1+x_2)} \underbrace{f(x_1,x_2)}_{x_1 x_2} dx_1 dx_2 = 2$

$= \int \int x_1^2 x_2 + x_1 x_2^2 \, dx_1 dx_2$

$\int \left[\frac{x_1^3}{3} x_2 + \frac{x_1^2}{2} x^2\right]_0^2 dx_2$

3. Suppose that X and Y are random variables having the joint probability function below.

$f(x,y)$		x = 2	x = 4	$f_Y(y)$
y	1	0.10	0.15	0.25
	3	0.20	0.30	0.50
	5	0.10	0.15	0.25
$f_X(x)$		0.4	0.6	

(a) Determine whether X and Y are independent.
(b) Find $E(Y|X=2)$.
(c) Find $E(X|Y=3)$.
(d) Find $E(2X-3Y)$.
(e) Find $E(XY)$.
(f) Find $V(X)$ and $V(Y)$.

d. $E(2X-3Y) = 2E(X) - 3E(Y)$
$= -2.6$

e. $E(XY) = \sum_x \sum_y g(x,y) f(x,y) \qquad g(xy) = xy$
$= 9.6$

a. $f_X(x) \cdot f_Y(y) = f(x,y)$? holds for each value in table

(e) Find $E(XY)$.

(f) Find $V(X)$ and $V(Y)$.

a. $f_X(x) \cdot f_Y(y) = f(x,y)$? holds for each value in table

$x \perp y$

b. $E(Y|X=2) =$ given $x=2$

$y = 1$ or 2 or 3

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

$= 0.1$

since $x \perp y$

$= E(Y) = E(Y|X=2)$

$= 1(0.15) + 2(0.5) + 3(0.25) = 2$

same method

c. $E(X|Y=3) = 2(0.4) + 4(0.6) = 3.2$

e. $E(XY) = \sum_x \sum_y g(x,y) f(x,y)$ $\quad g(x,y) = xy$

$= 9.6$

f. $V(X) = E(X^2) - [E(X)]^2$

$= \sum x^2 f_X(x) - \quad = 0.96$

$0.8 + 2.4$

similar method

$V(Y)$ $\quad = 2$

Answers for Some of the Analytical Questions

1. (a) $f(x,y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{4-x-y}}{\binom{8}{4}}$, $x = 0,1,2,3$, $y = 0,1,2$, $1 \le x+y \le 4$; (b) 0.2571; (c) 0.5; (d) $f_X(x) = \frac{\binom{3}{x}\binom{5}{4-x}}{\binom{8}{4}}$, $x = 0,1,2,3$. (e) $f_{Y|X}(y|2) = \frac{1}{10}\binom{2}{y}\binom{3}{2-y}$, $y = 0,1,2$; 0.3.

2. (a) 1/4; (b) $f_1(x_1) = x_1/2$, $0 \le x_1 \le 2$; $f_2(x_2) = 2x_2$, $0 \le x_2 \le 1$; (c) 2.

3. (a) Independent; (b) 3; (c) 3.2; (d) −2.6; (e) 9.6; (f) 0.96; 2.

Midterm-Sample-Paper-2

Thursday, October 5, 2023 5:25 PM

Midterm-Sample-Pap...

NATIONAL UNIVERSITY OF SINGAPORE

DEPARTMENT OF STATISTICS AND DATA SCIENCE

ST2334 PROBABILITY AND STATISTICS

MID-SEMESTER TEST SAMPLE PAPER 2

(SEMESTER I, AY 2023/2024)

TIME ALLOWED: 60 MINUTES

INSTRUCTIONS TO STUDENTS

1. Please write your student number only. **Do not write your name.**
2. This assessment contains 15 questions and comprises 18 printed pages.
3. The total marks is 25; marks are equal distributed for all questions.
4. Please answer ALL questions.
5. Calculators may be used.
6. This is an **OPEN BOOK** assessment. Only **HARD COPIES** of materials are allowed.

1 MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY

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

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

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

(c) $A \cup B' \cup C'$

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

2 FILL IN THE BLANK

How many ways are there to choose an arbitrary number of students (including the possibility of choosing 0 student) from 6 students?

(Provide your answer in numerical form.)

$\binom{6}{0} + \binom{6}{1} + \binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5} + \binom{6}{6}$

$2 + 2\left(\frac{6!}{5!}\right) + 2\left(\frac{6!}{4!\,2!}\right) + \frac{6!}{3!\,3!}$

$2 + 12 + 30 + 20 = 64$

3 **FILL IN THE BLANK**

Suppose $P(A)=1/2$ $P(B')=5/8$ $P(B|A)=1/4$

$$P(A') = 1/2, \quad P(B) = 3/8, \quad \text{and,} \quad P(B'|A) = 3/4.$$

Find $P(B \cap A)$.

(Provide your answer in decimal form and round it to three decimal places if necessary)

$P(B|A) = \frac{P(A \cap B)}{P(A)}$ $P(B \cap A) = P(B|A) \cdot P(A)$ $= \frac{1}{2} \cdot \frac{1}{4} = \boxed{\frac{1}{8}}$

4 **FILL IN THE BLANK**

A group of 8 friends A,B,C,D,E,F,G,H go to a restaurant. Due to safe-distancing measures, the group needs to split up into two groups of 4. How many ways are there to split the group such that A and B are together but away from C?

(Provide your answer in numerical form.)

total

5 MULTIPLE RESPONSE: CHOOSE ALL ANSWERS THAT APPLY

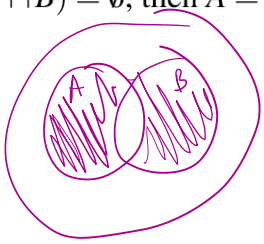
Let A and B be two events. Which of the following statements is/are true?

a) If $A \neq B$, then $P(A) \neq P(B)$.

b) If A and B are independent, then we must have $P(A \cup B) = 1 - \{1 - P(A)\}\{1 - P(B)\}$.

c) If $P(A) = 1 - P(B')$, then $P(A) = P(B)$.

d) $(A \cap B') \cup (A' \cap B) = \emptyset$, then $A = B$.



6 MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY

Consider the following statements about Peter whom you have not met before.

(A): He is not married.

(B): He is not married and smokes.

(C): He is married.

(D): He is married and does not smoke.

You are to assign probabilities to these statements. Which answer below is consistent with the laws of probability?

(a) $P(A) = 0.45, \quad P(B) = 0.5, \quad P(C) = 0.55, \quad P(D) = 0.4$

(b) $P(A) = 0.45, \quad P(B) = 0.1, \quad P(C) = 0.6, \quad P(D) = 0.3$

(c) $P(A) = 0.45, \quad P(B) = 0.2, \quad P(C) = 0.55, \quad P(D) = 0.5$

(d) $P(A) = 0.45, \quad P(B) = 0.4, \quad P(C) = 0.55, \quad P(D) = 0.6$

7 **TRUE/FALSE**

Let A and B be mutually exclusive events. If $P(A) = 0.1$, $P(B) = 0.01$, then A and B are not independent.

- TRUE
- FALSE

Mutually exclusive events cannot be independant unless P is 0

8 **TRUE/FALSE**

Cumulative distribution function can not take on values greater than 1 or smaller than 0.

- TRUE
- FALSE

9 **FILL IN THE BLANK**

Suppose that random variable X has the cumulative distribution function given by

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{9}, & 0 \le x \le 3 \\ 1, & x > 3 \end{cases}$$

$\frac{d}{dx}\left(\frac{x^2}{9}\right) = \frac{2}{9}x$

Compute $P(X = 1.5)$.

(Provide your answer in decimal form and round it to two decimal places if necessary.)

probability a continous random variable has a given value is 0

10 **FILL IN THE BLANK**

Let X be a random variable, whose cumulative distribution function is given by

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.2, & 0 \le x < 2 \\ 0.6, & 2 \le x < 3 \\ 0.7, & 3 \le x < 5 \\ 1 & x \ge 5 \end{cases}.$$

pmf

0	0.2
2	0.4
3	0.1
5	0.3

Compute $E(X)$.

(Provide your answer in decimal form and round it to two decimal places if necessary.)

$0 + 2(0.4) + 3(0.1) + 5(0.3) = 2.6$

11 FILL IN THE BLANK

Let X have probability mass function given by the following table.

x	0	2	5	6
$f(x)$	0.3	0.5	0.1	0.1

Compute $E(X)$.

(Provide your answer in decimal form and round it to two decimal places if necessary.)

$0+2(0.5)+5(0.1)+6(0.1)=2.1$

12 MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY

Let X be a random variable. Which of the following statement is **INCORRECT**?

(a) If $P(X=1)=0.1$ and $E(X)$ exists, then we must have $E(X^2)>(E(X))^2$. ✓

(b) If $V(X)>0$, then for any x, $P(X=x)<1$. ✓

(c) By the definition of the random variable, the range of X is a subset of $\mathbb{R}$; therefore, it is impossible that $P(X=x)=0$ for any $x\in\mathbb{R}$. ✗ for continous Random var

(d) There are cases under which $E(X)$ does not exist. ✓

13 **FILL IN THE BLANK**

A service station has both self-service and full-service islands. On each island, there is a single regular unleaded pump with two hoses. Let X denote the number of hoses being used on the self-service island at a particular time, and let Y denote the number of hoses on the full-service island in use at that time. The joint probability mass function of X and Y is given in the table below.

x	y		
	0	1	2
0	0.10	0.04	0.02
1	0.08	0.20	0.06
2	0.06	0.14	0.30

$P_{Y|1}$

$P_{X|1}$

$\left(\frac{1 \cdot 0.2}{0.38} + \frac{2 \cdot 0.14}{0.35}\right)$

$= 1.26$

38

Compute $E(X|Y=1)$.

(Provide your answer in decimal form and round it to two decimal places if necessary.)

14 **TRUE/FALSE**

Let $f(x,y)$ be the joint probability function of a discrete random vector (X,Y). If $f_X(1)=0$, then $f(1,y)=0$ for any y being a real number.

- TRUE
- FALSE

15 The joint probability function of (X,Y) is given by

$$f(x,y) = \begin{cases} \frac{1}{8}(x+y) & 0 \le x \le 2; 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

$P(X) = \int_0^2 \frac{1}{8}(x+y)\,dy = \frac{1}{8}\left[xy + \frac{y^2}{2}\right]_0^2$

$= \frac{1}{8}\left[2x+2\right] = \frac{x}{4} + \frac{1}{4}$

$P(X \ge 1) = \frac{1}{4}\int_1^2 (x+1)\,dx$

$= \frac{1}{4}\left(\frac{x^2}{2} + x\right)_1^2$

8-1-2

Compute $P(Y > 1|X > 1)$.

$\begin{cases} & \\ 0 & \text{elsewhere} \end{cases}$

Compute $P(Y \geq 1 | X \geq 1)$.

(Provide your answer in decimal form and round it to three decimal places if necessary.)

$P(X \geq 1; Y \geq 1) = \frac{1}{8} \int_1^2 \int_1^2 \frac{1}{8}(x+y)\, dy\, dx$

$P(Y \geq 1 | X \geq 1) = \frac{P(X \geq 1; Y \geq 1)}{P(X \geq 1)} = \frac{3/8}{5/8} = \boxed{\frac{3}{5}}$

$\frac{1}{8} \int_1^2 \left[xy + \frac{y^2}{2} \right]_1^2 dx = \frac{1}{8} \int_1^2 \left[2x + 2 - x - \frac{1}{2} \right] dx$

$= \frac{1}{8} \left[\frac{x^2}{2} + \frac{3}{2}x \right]_1^2 = \frac{1}{8} \left(2 + 3 - \frac{1}{2} - \frac{3}{2} \right)$

$= \frac{1}{8}(3)$

$\frac{1}{4} \left(\frac{x^2}{2} + x \right)_1^2 = \frac{1}{4} \left(4 - \frac{1}{2} - 1 \right) = \frac{5}{8}$

$x + \frac{3}{2}$

8-1-2

END OF PAPER

Midterm-Sample-Paper-1

Thursday, October 5, 2023 5:25 PM

Midterm-Sample-Pap...

NATIONAL UNIVERSITY OF SINGAPORE

DEPARTMENT OF STATISTICS AND DATA SCIENCE

ST2334 PROBABILITY AND STATISTICS

MID-SEMESTER TEST SAMPLE PAPER 1

(SEMESTER I, AY 2023/2024)

TIME ALLOWED: 60 MINUTES

INSTRUCTIONS TO STUDENTS

1. Please write your student number only. **Do not write your name.**
2. This assessment contains 15 questions and comprises 18 printed pages.
3. The total marks is 25; marks are equal distributed for all questions.
4. Please answer ALL questions.
5. Calculators may be used.
6. This is an **OPEN BOOK** assessment. Only **HARD COPIES** of materials are allowed.

1 MULTIPLE RESPONSE: CHOOSE ALL ANSWERS THAT APPLY

Which of the following can be used as the sample space for the problem: "choose two students from four students to complete a project"? Assume students are labeled as S_1, S_2, S_3, and S_4.

a) $\Big\{(S_1,S_2),(S_1,S_3),(S_1,S_4),(S_2,S_1),(S_2,S_3),(S_2,S_4),$
$(S_3,S_1),(S_3,S_2),(S_3,S_4),(S_4,S_1),(S_4,S_2),(S_4,S_3)\Big\}$.

b) $\Big\{\{S_1,S_2\},\{S_1,S_3\},\{S_1,S_4\},\{S_2,S_3\},\{S_2,S_4\},\{S_3,S_4\}\Big\}$.

c) $\Big\{\{S_1,S_1\},\{S_1,S_2\},\{S_1,S_3\},\{S_1,S_4\},\{S_2,S_2\},\{S_2,S_3\}\Big\}$.

d) $\Big\{S_1,S_2,S_3,S_4,S_5,S_6\Big\}$

2 FILL IN THE BLANK

In how many ways can 3 oaks, 4 pines, and 2 maples be arranged along a property line if one does not distinguish among trees of the same kind?

(Provide your answer in numerical form.)

$\frac{9!}{3!\,4!\,2!} = \frac{9\cdot 8\cdot 7\cdot 6\cdot 5\cdot 4}{} = 1260$

3 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

A new Covid test kit detects the virus 90% of the time if a patient is infected. However, it also detects the virus 5% of the time if a patient is uninfected. Given that the overall Covid infection rate is 1%, what is the probability of being infected if your test kit detects the virus?

(a) 0.114
(b) 0.215
(c) 0.154
(d) 0.322

$P(\text{pos}|\text{infect}) = 0.9$

$P(\text{pos} \mid \text{Not infect}) = 0.05$

$P(\text{neg}|\text{not infect}) = 0.95$

$P(\text{not infect}|\text{pos}) = 0.05 \cdot \frac{P(\text{pos})}{P(\text{not infect})}$

$P(\text{infect}) = 0.01$

$P(\text{not infect}) = 0.99$

$P(\text{infect}|\text{pos}) = P(\text{pos}|\text{infect}) \cdot \frac{P(\text{infect})}{P(\text{pos})}$

4 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Suppose $P(F) = P(G) = 0.4$. Which of the following statements must be true?

(a) $P(F \cup G) = 0.8$
(b) $P(F \cup G) = 0.4$
(c) $P(F \cup G) > 0.4$
(d) $P(F \cup G) \leq 0.8$

5 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

There are 10 women and 20 men in a class. Find the number of samples of three that can be formed with two women and one man.

a) $\binom{30}{3}$

b) $\binom{30}{1} \cdot \binom{10}{2}$

c) $\binom{10}{2} \cdot \binom{20}{1}$

d) $\binom{30}{2} \cdot \binom{20}{1}$

6 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Player *A* has entered a golf tournament but it is not certain whether *B* will enter. Player *A* has probability 1/6 of winning the tournament if player *B* enters, and probability 3/4 of winning if player *B* does not enter the tournament. If the probability that player *B* enters is 1/3, what is the probability that player *A* wins the tournament?

a) 5/9

b) 7/9

c) 3/7

d) 9/11

$P(A|Be) = P(A) \cdot P(Be)$

$\frac{1}{6} = \quad \cdot \frac{1}{3}$

$\frac{1}{3} \cdot \frac{1}{6}$

$P(Be) = \frac{1}{3}$

$P(A|Be) = \frac{1}{6}$

$P(A|\text{Not } Be) = \frac{3}{4}$

7 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Suppose that A and B are any two events where $P(A) = 0.4$ and $P(A \cap B) = 0.2$. Then $P(A|B) =$?

(a) 0.4

(b) 0.5

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

(c) Not enough information to determine

(d) None of the above

8 **TRUE/FALSE**

Probability density function can not take on values greater than 1.

- TRUE
- FALSE

9 **FILL IN THE BLANK**

Suppose that random variable X has the cumulative distribution function given by

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{100}, & 0 \le x \le 10 \\ 1, & x > 10 \end{cases}$$

Compute $P(X \ge 4)$.

(Provide your answer in decimal form and round it to two decimal places if necessary.)

$1 - F(4) = 1 - \frac{16}{100} = \frac{84}{100}$ 0.84

10 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let X be a random variable, whose cumulative distribution function is given by

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.2, & 0 \le x < 2 \\ 0.6, & 2 \le x < 3 \\ 0.7, & 3 \le x < 5 \\ 1 & x \ge 5 \end{cases}.$$

0.7 - 0.2

Then $P(1 \le X < 5) = ?$

a) 0.1

b) 0.4

c) 0.5

d) 0.8

11 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

The continuous random variable X has the following probability density function

$$f_X(x) = \begin{cases} \frac{1}{8}(1+3x), & 0 \le x \le 2 \\ 0, & \text{elsewhere} \end{cases}$$

The median of a continuous random variable Y, denoted by m_Y, is a real number satisfying $P(Y \le m_Y) = 0.5$. What is the median of X?

(a) 4/3
(b) 2/3
(c) 1
(d) 5/3

$\frac{1}{8}\int_0^y (1+3x)\,dx = 0.5$ 4

$x + \frac{3}{2}x^2 \Big]_0^y = 4$

$\frac{3}{2}x^2 + x - 4 = 0$

$x =$

$\frac{2}{3}$

12 **FILL IN THE BLANK**

nok down expected val formula

The probability function for random variable X is given by

$$f(x) = \begin{cases} x, & 0 \le x \le 1 \\ 0.5, & 2 \le x \le 3 \\ 0, & \text{elsewhere} \end{cases}$$

Compute $E(X)$.

(Provide your answer in decimal form and round it to two decimal places if necessary.)

$\int_0^1 x \cdot x\,dx + \int_2^3 0.5 \cdot x\,dx$

$= \frac{x^3}{3}\Big]_0^1 + \frac{x^2}{4}\Big]_2^3 = \frac{1}{3} + \frac{9}{4} - 1 = \frac{1}{3} + \frac{5}{4} = \frac{19}{12}$

1.58

13 **FILL IN THE BLANK**

A service station has both self-service and full-service islands. On each island, there is a single regular unleaded pump with two hoses. Let X denote the number of hoses being used on the self-service island at a particular time, and let Y denote the number of hoses on the full-service island in use at that time. The joint probability mass function of X and Y is given in the table below.

$P(X+Y \geq 2) =$

x	y		
	0	1	2
0	0.10	0.04	0.02
1	0.08	0.20	0.06
2	0.06	0.14	0.30

x: 0 1 2 / 2 2 2
y: 2 1 2 / 0 1 2

$0.02+0.2+0.06$
$+0.06+0.14+0.3$
$= 0.78$

Compute $P(X+Y \geq 2)$.

(Provide your answer in decimal form and round it to two decimal places if necessary.)

14 **TRUE/FALSE**

Let $f(x,y)$ be the joint probability function of a random vector (X,Y) (discrete or continuous). If $f_X(1) > 0$, then there must exist a y such that $f(1,y) > 0$.

- TRUE
- FALSE

notes

15 The joint probability function of (X,Y) is given by

$$f(x,y) = \begin{cases} \frac{1}{8}(x+y) & 0 \le x \le 2; 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

Compute $P(Y \ge 1 | X = 1)$.

(Provide your answer in decimal form and round it to three decimal places if necessary.)

$\frac{5}{8}$

$f_x(x) = \int_0^2 \frac{1}{8}(x+y)\,dy = \frac{1}{8}\left[xy + \frac{y^2}{2}\right]_0^2$

$= \frac{1}{8}\left[2x+2\right]$

$= \frac{1}{4}(x+1)$

$f_{x|y}(y \mid x=1) = \frac{f_{x,y}(1,y)}{f_x(x)} = \frac{\frac{1}{8}(1+y)}{\frac{1}{4}(1+1)} = \frac{1}{4}(1+y)$

$\frac{1}{4}\int_1^2 (1+y)\,dy = \frac{1}{4}\left[y + \frac{y^2}{2}\right]_1^2 = \frac{1}{4}\left(2 + 2 - 1 - \frac{1}{2}\right)$ $\quad 5/2$

END OF PAPER

Tutorial 06

Thursday, October 19, 2023 10:03 AM

Tutorial 06

National University of Singapore
Department of Statistics and Data Science
ST2334 Probability and Statistics
Semester I, AY 2023/2024

Tutorial 06

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple response question: choose all that apply.**
Given that $V(X) = 5$ and $V(Y) = 3$, we define $Z = -2X + 4Y - 3$. Which of the following is/are **CORRECT**?
(a) If X and Y are independent $V(Z) = 68$.
(b) If $\text{cov}(X, Y) > 0$, $V(Z) > 68$.
(c) If $\text{cov}(X, Y) < 0$, $V(Z) > 68$.
(d) No matter what is the value of $\text{cov}(X, Y)$, $V(Z) > 0$.

2. **Fill in the blank.**
The random variables X and Y have the joint probability density function given by
$$f(x,y) = \begin{cases} x+y, & 0 \le x \le 1, 0 \le y \le 1; \\ 0, & \text{elsewhere.} \end{cases}$$
Compute $E(Y|X = 0.2)$.

3. **Multiple choice question: choose the unique correct answer.**
A box contains 2 red marbles and 98 blue ones. Draws are made at random with replacement. In n draws from the box, there is a better than 50% chance for a red marble to appear at least once. What is the smallest possible value for n?
(a) 35
(b) 45
(c) 55
(d) None of the given options.

4. **Multiple choice question: choose the unique correct answer.**
If X and Y are independent with cumulative distribution functions $F_X(x)$ and $F_Y(y)$, which of the following is **INCORRECT**?
(a) $P(X^2 > 2; Y^4 > 0) = \{1 + F_X(-\sqrt{2}-) - F_X(\sqrt{2})\}\{1 + F_Y(0-) - F_Y(0)\}$.
(b) $P(X \ge x; Y \le y) = (1 - F_X(x-))F_Y(y)$.
(c) $E\left((X^2 - E(X^2))g(Y)\right) = 0$ for any function g.
(d) None of the given options.

Analytical Questions

1. A service facility operates with two service lines. On a randomly selected day, let X be the proportion of time that the first line is in use whereas Y is the proportion of time that the second line is in use. Suppose that the joint probability density function for (X, Y) is given below.
$$f(x,y) = \begin{cases} \frac{3}{7}(x^2 + y^2), & 0 \le x \le 1, 0 \le y \le 1; \\ 0, & \text{elsewhere.} \end{cases}$$
(a) Determine whether X and Y are independent.
(b) Find the mean and variance of X and Y.
(c) Find the covariance of X and Y.
(d) Find the mean and variance of $X + Y$.

2. According to Chemical Engineering Progress (Nov, 1990), approximately 30% of all pipework failures in chemical plants are caused by operator error. We assume that pipework failures occur independently of one another.
What is the probability that out of the next 20 pipework failures,
(a) at least 10 are due to operator error?
(b) no more than 4 are due to operator error?
(c) exactly 5 are due to operator error?

3. Suppose that, on average, 1 person in 1000 makes a numerical error in preparing his or her income tax return. 10,000 forms are selected at random and examined.
(a) Use a suitable approximation to find the probability that 6, 7, or 8 forms contain an error.
(b) Find the mean and variance of the number of persons among 10,000 who make an error in preparing their tax returns.

4. A couple decides they will continue to have children until they have two sons. Assume that $P(\text{son}) = 0.5$, and only one child is born each time.
(a) What is the probability that their second son is their seventh child?
(b) What is the expected number of children for the couple?

Answers for Some of the Analytical Questions

1. (a) Dependent; (b) $E(X) = E(Y) = 5/8$; $V(X) = V(Y) = 73/960$. 3. (a) 0.2657; (b) 10; 9.99.
(c) $-1/64$; (d) 5/4, 29/240.
2. (a) 0.0480; (b) 0.2375; (c) 0.1789. 4. (a) 0.0469; (b) 4.

$\text{Var}(X \pm Y) = V(X) + V(Y) \pm 2\text{Cov}(X, Y)$

$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$

$\text{Var}(aX + b) = a^2 \text{Var}(X)$

$\text{Cov}(XY) = \text{Cov}(X, Y)\sqrt{\text{Var}(X)\text{Var}(Y)}$

$V(Z) = 4V(X) + 16V(Y) + 2(-2)(4)\,\text{cov}(X,Y)$

$20 + 48 - 16\,\text{cov}(X, Y)$

$\frac{2}{100}$ $\frac{98}{100}$ $P(x \ge 1) > 0.5 \iff P(X=0) < 0.5$

$\binom{n}{0}(0.02)^0(1-0.02)^{n-0} < 0.5$

$n > 34.31$

35

1) if $X \perp Y$
a. then $f(x,y) = f(x)f(y)$
$\int_0^1 \frac{1}{2}(x^2+y^2)\,dx$
$= \frac{1}{2}\left(\frac{x^3}{3} + y^2 x\right)_0^1 = \frac{3}{2}\left(\frac{1}{3} + y^2\right) = \frac{1}{2} + \frac{3}{2}y^2 = f(x)$
$f(y) = \frac{1}{2} + \frac{3}{2}x^2$
$f(x,y) = \frac{3}{2}(x^2+y^2) = \left(\frac{1}{2} + \frac{3}{2}y^2\right)\left(\frac{1}{2} + \frac{3}{2}x^2\right)$
$\frac{3}{2}x^2 + \frac{3}{2}y^2 \ne \frac{1}{4} + \frac{3}{4}x^2 + \frac{3}{4}y^2 + \frac{9}{2}x^2y^2$
Dependent

b. $\int_0^1 x\left(\frac{1}{2} + \frac{3}{2}x^2\right) = \frac{5}{8} = E(X) = E(Y)$
$V(X) = E(X^2) - [E(X)]^2$
$\int_0^1 x^2\left(\frac{1}{2} + \frac{3}{2}x^2\right)dx$
$= \frac{7}{15} - \left(\frac{5}{8}\right)^2 = \frac{73}{960} = V(X) = V(Y)$

c. $\text{Cov}(X, Y)$

2.) a. $P(\text{operation}) = 0.3$
$P(0 \ge 10) = P(10) \cdots P(20)$
$\binom{20}{10}(0.3)^{10}(0.7)^{10} + \binom{20}{11}(0.3)^{11}(0.7)^{9} + \binom{20}{12}(0.3)^{12}(0.7)^{8} + \binom{20}{13}(0.3)^{13}(0.7)^{7}$
$+ \binom{20}{14}(0.3)^{14}(0.7)^{6} + \binom{20}{15}(0.3)^{15}(0.7)^{5} + \binom{20}{16}(0.3)^{16}(0.7)^{4} + \binom{20}{17}(0.3)^{17}(0.7)^{3}$
$+ \binom{20}{18}(0.3)^{18}(0.7)^{2} + \binom{20}{19}(0.3)^{19}(0.7)^{1} + \binom{20}{20}(0.3)^{20}(0.7)^{0}$
$\lambda = np$

Negative binomial

a. $P(X=7) = \binom{x-1}{k-1}p^k(1-p)^{x-k} = \binom{6}{1}(0.5)^2(0.5)^5 = 0.046875$
$k = 2$
$x = 7$ trials needed

b. $\frac{k}{p} = \frac{2}{0.5} = 4$

3.) a. $P(\text{error}) = 0.001$
$= P(X=6) + P(X=7) + P(X=8)$
$= \binom{10000}{6}p^6(1-p)^{9994} + \binom{10000}{7}p^7(1-p)^{9993} + \binom{10000}{8}p^8(1-p)^{9992}$
$= 0.2657$

b. $E(\text{error}) = 10000(0.001) = np$
$= 10$
$V(\text{error}) = np(1-p) = 9.99$

Tutorial 07

Wednesday, October 25, 2023 2:13 PM

Tutorial 07

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE
ST2334 PROBABILITY AND STATISTICS
SEMESTER 1, AY 2023/2024

Tutorial 07

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Fill in the blank**.
   Let X_1 and X_2 be independent random variables; $X_1 \sim \text{Poisson}(2)$; $X_2 \sim \text{Poisson}(3)$. Compute $E(X_1|X_1+X_2=10)$.
   Answer: _____.

2. **Multiple choice question: choose the unique correct answer**.
   Three people toss a fair coin and the odd one pays for coffee. If the coins all turn up the same, they toss again. What is the probability that at most x tosses are needed? Here $x \geq 1$ is a whole number.
   (a) $1-\frac{1}{2^x}$
   (b) $1-\frac{1}{2^{2x}}$
   (c) $1-\frac{1}{2^{3x}}$
   (d) None of the given options

3. **Multiple choice question: choose the unique correct answer**.
   A company rents time on a computer for periods of t hours, for which it receives \$600 an hour. The number of times the computer breaks down during t hours is a random variable having the Poisson distribution with $\lambda = 0.8t$. If the computer breaks down x times during t hours, it costs $50x^2$ dollars to fix it. What is the value of t such that the company maximizes its expected profit?
   (a) $t = 5.75$
   (b) $t = 7.75$
   (c) $t = 8.75$
   (d) $t = 10.75$

   $600t - 50(0.8t)^2 = 600t - 32t^2 = P$
   $E = 0.8t$
   $t = 9.375$

4. **True/False**.
   Let $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$. Assume that X and Y are independent, $p > 0$, and $n > m+10$. Then we must have $X - Y \sim \text{Bin}(n-m,p)$.
   - TRUE
   - FALSE

5. **Fill in the blank**.
   The time (in hours) required to repair a machine is an exponentially distributed random variable with parameter $\lambda = 1/2$. What is the probability that a repair takes at least 10 hours, given that its duration exceeds 9 hours?

1

Analytical Questions

1. Hospital administrators in large cities anguish about problems with traffic in emergency rooms in hospitals. For a particular hospital in a large city, the staff on hand cannot accommodate the patient traffic if there are more than 10 emergency cases in a given hour. It is assumed that patient arrival follows a Poisson process and historical data suggest that, on the average, 5 emergencies arrive per hour. Find the probability that
   (a) in a given hour, there is no emergency.
   (b) in a given hour, the staff can no longer accommodate the traffic?
   (c) more than 20 emergencies arrive during a 3-hour shift of personnel?

2. A notice is sent to all owners of a certain type of automobile, asking them to bring their cars to a dealer to check for the presence of a particular type of defect. Suppose that only 0.05% of the cars have the defect. Consider a random sample of 10,000 cars.
   (a) What are the expected value and variance of the number of cars in the sample that have the defect?
   (b) What is the (approximate) probability that at least 10 sampled cars have the defect?
   (c) What is the (approximate) probability that no sampled cars have the defect?

1.) a. $E(X) = 5 = \lambda$
$P(X=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-5} \approx 0.0067$

b. $P(X>10) = ppois(10, 5, lower.tail=F)$
$= 0.0137$

c. $E(Y) = 15$
$P(X>20) = ppois(20, 15, lower.tail=F) = 0.0829$

2.) a. $E(X) = \frac{0.05}{100}(10000) = 5$
$Var(X) = np(1-p) = 10000\left(\frac{0.05}{100}\right)\left(1-\frac{0.05}{100}\right) = 4.9975$

b. $P(X \geq 10) = pbinom(9, 10000, \frac{0.05}{100}, lower.tail=F)$
$= 0.03179$

c. $P(X=0) = \binom{10000}{0} p^0 (1-p)^{n-0} = 0.0067$

(a) What are the expected value and variance of the number of cars in the sample that have the defect?

(b) What is the (approximate) probability that at least 10 sampled cars have the defect?

(c) What is the (approximate) probability that no sampled cars have the defect?

3. You arrive at the bus stop at 10 a.m., knowing that the bus will arrive at some time uniformly distributed between 10 a.m. and 10:30 a.m.

(a) What is the probability that you will have to wait longer than 10 minutes?

(b) If the bus has not yet arrived at 10:15 a.m., what is the probability that you will have to wait at least an additional 10 minutes?

4. The amount of time needed to serve one customer at a cafeteria is a random variable that follows an exponential distribution with a mean of 4 minutes. Find the probability that

(a) a person is served in more than 3 minutes.

(b) a person is served in less than 3 minutes.

(c) a person is served in less than 3 minutes, on at least 4 of the next 6 days.

This question combines exponential distrib with binomial distrib
first calc person served in <3min
Then to find at least 4 of 6 days, we set up binomial distrib with variable Y = # of days of

Answers for Some of the Analytical Questions

1. (a) 0.00673; (b) 0.0137; (c) 0.0830.
2. (a) 5, 4.9975; (b) 0.0318; (c) 0.0067.
3. (a) 2/3; (b) 1/3.
4. (a) 0.4724; (b) 0.5276; (c) 0.3968.

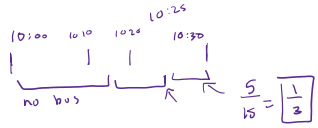
b. $P(X \geq 10) =$ pbinom(9, 10000, $\frac{0.05}{100}$, lower.tail=F)
$= 0.03179$

c. $P(X=0) = \binom{10000}{0} p^0 (1-p)^{n-0} = 0.0067$

3.) a.

2/3

b.



4.) a. $E(X) = 4 = \frac{1}{\lambda}$ $\lambda = \frac{1}{4}$

$P(X > 3) = e^{-\lambda x} = e^{-3/4} = 0.4724$

b. $P(X < 3) = 1 - e^{-3/4} = 0.5276$

c. Y = # of days person served in <3min

$P(Y \geq 4) =$ pbinom(3, 6, 0.5276, lower.tail=F)
$= 0.3968$

Tutorial 08

Thursday, November 2, 2023 12:37 AM

Tutorial 08

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE
ST2334 PROBABILITY AND STATISTICS
SEMESTER I, AY 2023/2024

Tutorial 08

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer.**
   Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed Exp(λ) distributions. Suppose $X = \min\{X_1, X_2, \ldots, X_n\}$. What is the distribution of X?

   (a) Exp($n\lambda$)
   (b) Exp(λ^n)
   (c) Poisson($n\lambda$)
   (d) Poisson(λ/n)
   (e) Insufficient information to derive

2. **True/False.**
   Let $Z \sim N(0,1)$ and $z \geq 0$, then we must have

   $$P(|Z| < z) = 2\Phi(z) - 1,$$

   where $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution.

   • TRUE
   • FALSE

3. **True/False.**
   Suppose $X \sim N(0,1)$ and $Y \sim N(0, 1.5^2)$. Then $P(X < 1) > P(Y < 1)$.

   • TRUE
   • FALSE

4. **Multiple choice question: choose the unique correct answer.**
   A fair coin is tossed 400 times. With the normal approximation, which of the following is closest to the probability of obtaining between 185 and 210 heads inclusive?

   (a) $\Phi(1.55) - \Phi(1.05)$
   (b) $\Phi(1.05) - \Phi(-1.55)$
   (c) $\Phi(1.85) - \Phi(1.15)$
   (d) $\Phi(1.15) - \Phi(-1.85)$

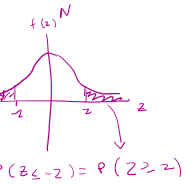
5. **Multiple choice question: choose the unique correct answer.**
   A fair coin is tossed 400 times. With the normal approximation, which of the following is closest to the probability of obtaining exactly 205 heads?

   (a) 0
   (b) $\Phi(0.75) - \Phi(0.65)$
   (c) $\Phi(0.65) - \Phi(0.55)$
   (d) $\Phi(0.55) - \Phi(0.45)$

1

1) $P(X \geq x) = P(\min(X_1 \cdots X_n) > x)$
$= P(X_1 > x, \cdots X_n > x_n)$
$P(X_1 > x) \cdot P(X_2 > x) \cdots P(X_n > x)$
$= e^{-\lambda x} \cdot e^{-\lambda x} \cdots$
$= e^{-n\lambda x}$
$X \sim Exp(n\lambda)$

2) $P(|Z| < z) = P(-z < Z < z)$
$= P(Z < z) - P(Z \leq -z)$
$=$
$P(Z < z) - P(Z \geq z)$
$P(Z < z) - (1 - P(Z < z))$
$= 2P(Z < z) - 1$



Φ cdf

$Z^* = \frac{X - np}{\sqrt{np(1-p)}}$

Analytical Questions

1. Extensive experience with fans of a certain type used in diesel engines has suggested that the exponential distribution provides a good model for time until failure. Suppose the mean time until failure is 25000 hours.

   (a) What is the probability that a randomly selected fan will last 20000 hours? At most 30000 hours? Between 20000 and 30000 hours?

   (b) What is the probability that the lifetime of a fan exceeds the mean value by more than 2 standard deviations?

2. A soft-drink machine is regulated so that it discharges an average of 200 ml per cup. Suppose that the amount of drink is normally distributed with a standard deviation equals to 15 ml.

   (a) Find the probability that a cup will contain more than 224 ml.

   (b) Find the probability that a cup contains between 191 and 209 ml.

   (c) How many cups will probably overflow if 230 ml cups are used for the next 1000 drinks?

   (d) Below what value do we get the smallest 25% of the drinks?

3. A lawyer commutes daily from his suburban home to his midtown office. The average time for a one-way trip is 24 minutes, with a standard deviation of 3.8 minutes. Assume the distribution of trip times to be normally distributed.

   (a) What is the probability that a trip will take at least half an hour?

   (b) If the office opens at 9.00 am, and he leaves his house at 8.45 am daily, what percentage of the time is he late for work?

   (c) Find the probability that 2 of the next 3 trips will take at least 1/2 hour.

4. Physicians recommend that children with type-I (insulin dependent) diabetes keep up with their insulin shots to minimize the chance of long-term complications. In addition, some diabetes researchers have observed that growth rate of weight during adolescence among diabetic patients is affected by level of compliance with insulin therapy.

   Suppose 12-year-old type-I diabetic boys who comply with their insulin shots have a weight gain over 1 year that is normally distributed, with mean 12 lb and variance 12 lb.

   Conversely, 12-year-old type-I diabetic boys who do not take their insulin shots have a weight gain over 1 year that is normally distributed with mean 8 lb and variance 12 lb.

   It is generally assumed that 75% of type-I diabetics comply with their insulin regimen. Suppose that a 12-year-old type-I diabetic boy comes to clinic and shows a 5-lb weight gain over 1 year (actually, because of measurement error, assume this is an actual weight gain from 4.5 to 5.5 lb). The boy claims to be taking his insulin medication. What is the probability that he is telling the truth?

5. The random variable X representing the number of cherries in a cherry puff, has the following probability distribution:

x	4	5	6	7
$f(x)$	0.2	0.4	0.3	0.1

   (a) Find the mean μ and the variance σ^2 of X.

   (b) Find the mean $\mu_{\overline{X}}$, and the variance $\sigma^2_{\overline{X}}$ of the mean $\overline{X}$ for random samples of 36 cherry puffs from the above probability distribution.

   (c) Find the probability that the average number of cherries in 36 cherry puffs will be less than 5.5.

2) a. $X \sim N(200, 15^2)$ $Z = \frac{X-200}{15} \sim N(0,1)$

$P(X>224) = P\left(Z > \frac{224-200}{15}\right) = 1 - \Phi\left(\frac{224-200}{15}\right) = 0.548$

b. $P(\quad < X < \quad) = P\left(\frac{\text{min}}{15} < Z < \frac{209-200}{\sigma}\right) = \Phi(0.1) - \Phi(-\quad = 0.4514$

c. $P(X>230) = P\left(Z > \frac{230-200}{15}\right) = P(Z>2) = 1 - \Phi(2) = 0.02275$

$y \to$ # of cups that will overflow

$y \to \text{bin}(1000, 0.02275) \Rightarrow E(Y) = n \cdot p = 1000(0.02275)$

$= 22.75$

$\approx \boxed{23}$

d. $z \leftarrow P(Z<z) = 25\%$ $P(Z>z)$ $75\% = \alpha$

$z = z_{75\%} = -0.6745$

$\frac{X-200}{15} = Z < z_{0.75} \Rightarrow X < 15 \cdot z_{0.75} + 200 = \boxed{189.88}$

3.) $X \sim N(24, 3.8)$ $Z = \frac{X-24}{3.8} \sim N(0,1)$

a. 0.0571

b.

c. $n=3$ $p = P(X>30)$

$y = \#$ $\Rightarrow y \sim \text{Bin}(3, p)$

$P(Y=2) = \binom{3}{2} p^2 (1-p)^1 = 0.00922$

4.) $C \leftarrow$ compliant

$D \leftarrow$ weight gain $(4.5, 5.5)$ in lbs

$P(C) = 0.75$ $P(C') = 0.25$

$P(D \mid C) = P(4.5 < X < 5.5) = \Phi\left(\frac{5.5-12}{\sqrt{12}}\right) - \Phi\left(\frac{4.5-12}{\sqrt{12}}\right) = 0.015$

$\hookrightarrow X \sim N(12, 12)$

$P(D \mid C') = P(4.5 < Y < 5.5) = \Phi\left(\frac{5.5-8}{\sqrt{12}}\right) - \Phi\left(\frac{4.5-8}{\sqrt{12}}\right) = 0.079$

$\hookrightarrow Y \sim N(8, 12)$

Bayes $P(C \mid D) = \frac{P(D \mid C) P(C)}{P(D \mid C) P(C) + P(D \mid C') P(C')} = \frac{0.015 \cdot 0.75}{0.015(0.75) + (0.079)(0.25)} = \boxed{0.364}$

5.) a. $\mu = E(x) = 5.3$ $\}$ X

$\sigma^2 = 0.81$ $\}$

b. $n = 36$

$\mu_{\bar{x}} = \mu = 5.3$

$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{0.81}{36} = 0.0225$

CLT c. $P(\bar{X} < 5.5)$

CLT: $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) = N(5.3, 0.0225)$

$P(\bar{X} < 5.5) = \Phi\left(\frac{5.5-5.3}{0.0225}\right) = 0.9082$

Answers for Some of the Analytical Questions

1. (a) 0.4493, 0.6988, 0.1481; (b) 0.0498.
2. (a) 0.0548; (b) 0.4514; (c) 23; (d) 189.88.
3. (a) 0.0571; (b) 99.11%; (c) 0.00922.
4. 0.364.
5. (a) 5.3; 0.81; (b) 0.0225; (c) 0.9082.

Wk 10 ch 5

Thursday, November 2, 2023 9:41 AM

DEFINITION 1 (POPULATION & SAMPLE)
The totality of all possible outcomes or observations of a survey or experiment is called a ***population****.*

A ***sample*** *is any subset of a population.*

Population can be finite or infinite

DEFINITION 2 (SIMPLE RANDOM SAMPLE)
A set of n members taken from a given population is called a ***sample*** *of size n.*

A ***simple random sample (SRS)*** *of n members is a sample that is chosen such that every subset of n observations of the population has the same probability of being selected.*

REMARK
More generally, let N denote the population size. The population has $\binom{N}{n}$ possible samples of size n.

For large values of N and n, one can use software easily to select the sample from a list of the population members using a random number generator.

Finite pop ^^

DEFINITION 3 (SIMPLE RANDOM SAMPLE: INFINITE POPULATION)
Let X be a random variable with certain probability distribution $f_X(x)$.

Let $X_1, X_2, \ldots, X_n$ be n independent random variables each having the same distribution as X. Then $(X_1, X_2, \ldots, X_n)$ is called a ***random sample of size n*** *from a population with distribution $f_X(x)$.*

The joint probability function of $(X_1, X_2, \ldots, X_n)$ is given by

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

where $f_X(x)$ is the probability function of the population.

DEFINITION 4 (STATISTIC)
Suppose a random sample of n observations $(X_1, \ldots, X_n)$ has been taken. A function of $(X_1, \ldots, X_n)$ is called a ***statistic****.*

EXAMPLE 5.2 (SAMPLE MEAN)
The **sample mean**, defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

is a statistic.

If the values in a random sample are observed and they are $(x_1, \ldots, x_n)$, then the realization of the statistic $\overline{X}$ is given by

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Lower case x are real numbers

EXAMPLE 5.3 (SAMPLE VARIANCE)
The **sample variance**, defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2,$$

is a statistic.

Similarly, if the values in a random sample are observed and they are $(x_1, \ldots, x_n)$, then the realization of the statistic S^2 is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2.$$

STATISTICS ARE RANDOM VARIABLES

- Note that X_1 is a random variable and so are $X_2, \ldots, X_n$.
- Thus $\overline{X}$ and S^2 are random variables as well.
- As many random samples are possible from the same population, we expect the statistic to vary somewhat from sample to sample.
- Hence a statistic is a random variable. It is meaningful to consider the probability distribution of a statistic.

Two Results

We next present two key results about the sampling distribution of the sample mean.

- Theorem 6 provides formulas for the center and the spread of the sampling distribution.
- Theorem 9 describes the shape of the sampling distribution, showing that it is often approximately normal.

THEOREM 6 (MEAN AND VARIANCE OF $\overline{X}$)
For random samples of size n taken from an infinite population with mean μ_X and variance σ_X^2, the sampling distribution of the sample mean $\overline{X}$ has mean μ_X and variance $\frac{\sigma_X^2}{n}$. That is,

$$\mu_{\overline{X}} = E(\overline{X}) = \mu_X \quad \text{and} \quad \sigma_{\overline{X}}^2 = \operatorname{var}(\overline{X}) = \frac{\sigma_X^2}{n}.$$

Validity of $\overline{X}$ as an estimator for μ_X

- The expectation of $\overline{X}$ is equal to the population mean μ_X.
- In "the long run", $\overline{X}$ does not introduce any systematic bias as an estimator of μ_X. So $\overline{X}$ can serve as a valid estimator of μ_X.
- For an infinite population, when n gets larger and larger, $\sigma_{\overline{X}}^2/n$, the variance of $\overline{X}$, becomes smaller and smaller, that is, the accuracy of $\overline{X}$ as an estimator of μ_X keeps improving.

Remark

The standard error of $\overline{X}$ describes how much $\overline{x}$ tends to vary from sample to sample of size n.

The symbol $\sigma_{\overline{X}}$ (instead of σ) and the terminology standard error (instead of standard deviation) distinguishes this measure from the standard deviation σ of the population.

Theorem 8 (Law of Large Numbers (LLN))

If $X_1, \ldots, X_n$ are independent random variables with the same mean μ and variance σ^2, then for any $\varepsilon \in \mathbb{R}$,

$$P(|\overline{X} - \mu| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

When sample size increases, probability that sample mean differs from population mean goes to 0
Increasingly likely that X bar is close to mu x as n get larger

Theorem 9 (Central Limit Theorem (CLT))

If $\overline{X}$ is the mean of a random sample of size n taken from a population having mean μ and finite variance σ^2, then, as $n \to \infty$,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to Z \sim N(0,1).$$

Equivalently, this means

$$\overline{X} \to N\left(\mu, \frac{\sigma^2}{n}\right).$$

Rule of Thumb

The Central Limit Theorem says that, if you take the mean of a large number of independent samples, then the distribution of that mean will be approximately normal.

- If the population you are sampling from is symmetric with no outliers, a good approximation to normality appears after as few as 15-20 samples.
- If the population is moderately skewed, such as exponential or χ^2, then it can take between 30-50 samples before getting a good approximation.
- Data with extreme skewness, such as some financial data where most entries are 0, a few are small, and even fewer are extremely large, may not be appropriate for the Central Limit Theorem even with 1000 samples.

Example 5.4 (Bowling League)

In a bowling league season, bowlers bowl 50 games and the average score is ranked at the end of the season. Historically, John averages 175 a game with a standard deviation of 30. What is the probability that John will average more than 180 this season?

Solution:

We do not know the distribution of X, but we know that $\mu = 175$, $\sigma = 30$ and $n = 50$. Let $\overline{X}$ be the sample mean.

By CLT, we can approximate $\overline{X}$ by $N(\mu, \sigma^2/n)$. The question asks for the probability

$$\begin{aligned} P(\overline{X} > 180) &= P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{180 - \mu}{\sigma/\sqrt{n}}\right) \\ &\approx P(Z > 1.18) = 0.119. \end{aligned}$$

c^2, t, F distrib

Definition 10 (The χ^2 Distribution)

*Let Z be a standard normal random variable. A random variable with the same distribution as Z^2 is called a χ^2 **random variable with one degree of freedom**.*

*Let $Z_1, \ldots, Z_n$ be n independent and identically distributed standard normal random variables. A random variable with the same distribution as $Z_1^2 + \cdots + Z_n^2$ is called a χ^2 **random variable with n degrees of freedom**.*

Remark

We denote a χ^2 random variable with n degrees of freedom as $\chi^2(n)$.

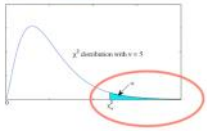
Properties of χ^2 Distributions

1. If $Y \sim \chi^2(n)$, then $E(Y) = n$ and $\text{var}(Y) = 2n$.
2. For large n, $\chi^2(n)$ is approximately $N(n, 2n)$.
3. If Y_1 and Y_2 are independent χ^2 random variables with m and n degrees of freedom respectively, then $Y_1 + Y_2$ is a χ^2 random variable with $m+n$ degrees of freedom.
4. The χ^2 distribution is a family of curves, each determined by the degrees of freedom n. All the density functions have a long right tail.

Definition 11

Define $\chi^2(n;\alpha)$ such that for $Y \sim \chi^2(n)$,

$$P(Y > \chi^2(n;\alpha)) = \alpha.$$



The sampling distribution of $(n-1)S^2/\sigma^2$

Recall that for $X_1,\ldots,X_n$ independent and identically distributed with $E(X)=\mu$ and $\text{var}(X)=\sigma^2$, the sample variance is defined as

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Though it can be shown that $E(S^2)=\sigma^2$, the sampling distribution of the random variable S^2 has little practical application in statistics.

THEOREM 12

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the random variable

$$\frac{(n-1)S^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2}$$

has a χ^2 distribution with $n-1$ degrees of freedom.

DEFINITION 13 (THE t-DISTRIBUTION)

Suppose $Z\sim N(0,1)$ and $U\sim\chi^2(n)$. If Z and U are independent, then

$$T=\frac{Z}{\sqrt{U/n}}$$

*follows the **t-distribution with n degrees of freedom**.*

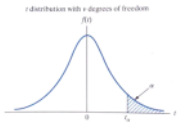
PROPERTIES OF THE t-DISTRIBUTION

- The t-distribution with n degrees of freedom, also called the Student's t-distribution, is denoted by $t(n)$.
- The t-distribution approaches $N(0,1)$ as the parameter $n\to\infty$. When $n\geq 30$, we can replace it by $N(0,1)$.
- If $T\sim t(n)$, then $E(T)=0$ and $\text{var}(T)=n/(n-2)$ for $n>2$.
- The graph of the t-distribution is symmetric about the vertical axis and resembles the graph of the standard normal distribution.

DEFINITION 14

Define $t_{n;\alpha}$ such that for $T\sim t(n)$,

$$P(T>t_{n;\alpha})=\alpha.$$



THEOREM 15

If $X_1,\ldots,X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then

$$\frac{X-\mu}{S/\sqrt{n}}$$

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

EXAMPLE 5.5 (MIDTERM SCORE)

The lecturer of a class announced that the mean score of the midterm is 16 out of 30. A student doubts it, so he randomly chose 5 classmates and asked them for their scores: 20, 19, 24, 22, 25.

Should the student believe that the mean score is 16? Assume the scores are approximately normally distributed.

Solution:

The student has $n=5$ sampled data

$$x_1=20, x_2=19, x_3=24, x_4=22, x_5=25.$$

If $\mu=16$,

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}=\frac{\overline{X}-16}{S/\sqrt{5}}$$

should follow a t-distribution with $5-1=4$ degrees of freedom.

With the observed data $\overline{x}=22$ and $s=2.55$ so

$$t=\frac{22-16}{2.55/\sqrt{5}}=5.26.$$

Using software, $P(t(4)>5.26)=0.003$. This says that there is only a 0.003 chance that T is 5.26 (or larger), provided the lecturer is telling the truth that $\mu=16$.

So should the student believe him based on his findings?

DEFINITION 16 (THE F-DISTRIBUTION)

Suppose $U\sim\chi^2(m)$ and $V\sim\chi^2(n)$ are independent. Then the distribution of the random variable

$$F=\frac{U/m}{V/n}$$

*is called a **F-distribution with (m,n) degrees of freedom**.*

PROPERTIES OF THE F-DISTRIBUTION

- The F-distribution with (m,n) degrees of freedom is denoted by $F(m,n)$.
- If $X\sim F(m,n)$, then

$$E(X)=\frac{n}{n-2},\quad \text{for } n>2$$

and

$$\text{var}(X)=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)},\quad \text{for } n>4.$$

EXAMPLE 5.6
For example,

$$F(4,5;0.05) = 5.19$$

means that $P(F > 5.19) = 0.05$, where $F \sim F(4,5)$.

- If $F \sim F(n,m)$, then $1/F \sim F(m,n)$. This follows immediately from the definition of the F-distribution.
- Values of the F-distribution can be found in the statistical tables or software. The values of interests are $F(m,n;\alpha)$ such that

  $$P(F > F(m,n;\alpha)) = \alpha,$$

  where $F \sim F(m,n)$.
- It can be shown that

  $$F(m,n;1-\alpha) = 1/F(n,m;\alpha).$$

Wk 11 ch6

Wednesday, November 8, 2023 3:15 PM

Statistical inference methods

Two types of Estimations

Point estimation

Based on sample data, a single number is calculated to estimate the population parameter. The rule or formula that describes this calculation is called the point estimator. The resulting number is called a point estimate.

Interval estimation

Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie.

Definition 1 (Estimator)

*An **estimator** is a rule, usually expressed as a formula, that tells us how to calculate an **estimate** based on information in the sample.*

Example 6.2 (Point Estimator)

We want to estimate the average waiting time for a bus (μ) for students attending ST2334. The lecturer asked 4 students their waiting times $X_1, \ldots, X_4$ for a bus. The (observed) results are

$$x_1 = 6, x_2 = 1, x_3 = 4, x_4 = 9.$$

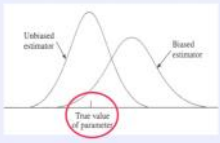
We can use $\overline{X} = \frac{1}{4}(X_1 + \cdots + X_4)$ to estimate μ. In this case, $\overline{X}$ is the estimator (for μ), and the computed value $\bar{x} = 5$ is the estimate.

Definition 2 (Unbiased Estimator)

*Let $\hat{\Theta}$ be an estimator of θ. Then $\hat{\Theta}$ is a random variable based on the sample. If $E(\hat{\Theta}) = \theta$, we call $\hat{\Theta}$ an **unbiased estimator** of θ.*

Remark

An unbiased estimator has mean value equals to the true value of the parameter.



Example 6.3 (Unbiased Estimator)

Let $X_1, X_2, \ldots, X_n$ be a random sample from the same population with mean μ and variance σ^2. Then

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator of σ^2 since $E(S^2) = \sigma^2$.

Maximum Error of Estimate

Typically $\overline{X} \neq \mu$, so $\overline{X} - \mu$ measures the difference between the estimator and the true value of the parameter.

Recall that if the population is normal or if n is large, $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ follows a standard normal or an approximately standard normal distribution.

Definition 4 (Maximum Error Of Estimate)

The quantity

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

*is called the **maximum error of estimate**.*

Example 6.4 (TV Time for Internet Users)

An investigator is interested in the amount of time internet users spend watching television per week.

Based on historical experience, he assumes that the standard deviation is $\sigma = 3.5$ hours.

He proposes to select a random sample of $n = 50$ internet users, poll them, and take the sample mean to estimate the population mean μ.

What can he assert with probability 0.99 about the maximum error of estimate?

Solution:

As $n = 50 \geq 30$ is large, $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is approximately normal.

So we can use the previous result, with $\sigma = 3.5$, $\alpha = 0.01$ and $z_{\alpha/2} = z_{0.005} = 2.576$.

With probability 0.99, the error is at most

$$E = 2.576 \times \frac{3.5}{\sqrt{50}} \approx 1.27.$$

Example 6.1

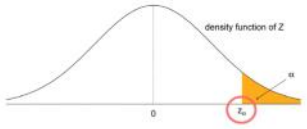
One survey asked, "Do you believe in hell?"

From sample data, the point estimate for the proportion of adult (in the population) who would respond "yes" is 0.73. The adjective "point" refers to using a single number as the parameter estimate.

An interval estimate predicts that the proportion of adult (in the population) who believe in hell falls between 0.71 and 0.75.

Definition 3 (z_α)

Define z_α to be the number with an upper-tail probability of α for the standard normal distribution Z. That is, $P(Z > z_\alpha) = \alpha$.



From the above definition, we then have

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

In other words,

$$P\left(\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(|\overline{X} - \mu| \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

This means that, with probability $1 - \alpha$, the error $|\overline{X} - \mu|$ is less than

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

REMARK

$z_{0.005}$ is the same as the 0.995 quantile of the standard normal. The value of 2.576 can be obtained from tables or software.

Use the command `qnorm(0.995)` or `qnorm(0.005, lower.tail=F)` to obtain the value via `https://rdrr.io/snippets/`.

Alternatively, you may use Radiant to get the same value as well.

Determination of Sample Size

We often want to know what the minimum sample size should be, so that with probability $1-\alpha$, the error is at most E_0.

To answer this, consider the fact that we want

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq E_0.$$

Solving for n, we have

$$n \geq \left(\frac{z_{\alpha/2}\cdot\sigma}{E_0}\right)^2.$$

DIFFERENT CASES

	Population	σ	n	Statistic	E	n for desired E_0 and α
I	Normal	known	any	$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\cdot\sigma}{E_0}\right)^2$
II	any	known	large	$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\cdot\sigma}{E_0}\right)^2$
III	Normal	unknown	small	$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2}\cdot\frac{s}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2}\cdot s}{E_0}\right)^2$
IV	any	unknown	large	$Z=\frac{\overline{X}-\mu}{S/\sqrt{n}}$	$z_{\alpha/2}\cdot\frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\cdot s}{E_0}\right)^2$

- A point estimate is almost never right
- So we use confidnece intervals

DEFINITION 5 (CONFIDENCE INTERVAL)

*An **interval estimator** is a rule for calculating, from the sample, an interval (a,b) in which you are fairly certain the parameter of interest lies in.*

*This "fairly certain" can be quantified by the **degree of confidence** also known as **confidence level** $(1-\alpha)$, in the sense that*

$$P(a<\mu<b)=1-\alpha.$$

*(a,b) is called the $(1-\alpha)$ **confidence interval**.*

Case I: σ known, data normal

Consider the case where σ is known, and data comes from a normal population.

We learnt previously that

$$P\left(-z_{\alpha/2}\leq\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\leq z_{\alpha/2}\right)=1-\alpha.$$

Rearranging, we have

$$P\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\leq\mu\leq\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha.$$

So

$$\overline{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}=\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is a $(1-\alpha)$ confidence interval.

EXAMPLE 6.5

In order to set inventory levels, a computer company samples demand during lead time over 25 time periods:

```
235 374 309 499 253 421 361 514 462 369 394 439
348 344 330 261 374 302 466 535 386 316 296 332 334
```

It is known that the (population) standard deviation of demand over lead time is 75 computers. Given that $\overline{x}=370.16$, estimate the mean demand over lead time with 95% confidence. Assume a normal distribution for the population.

Solution:

Note that $z_{\alpha/2}=z_{0.025}=1.96$. The 95% confidence interval is

$$\overline{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}=370.16\pm 1.96\frac{75}{\sqrt{25}}=370.16\pm 29.4$$

or $(340.76, 399.56)$.

REMARK

Notice that our $(1-\alpha)$ confidence interval can be written as $\overline{X}\pm E$.

This is not a coincidence: recall that there is $(1-\alpha)$ confidence that the error $|\overline{X}-\mu|$ is within E.

For the other cases, based on our understanding of the sampling distribution of $\overline{X}$, we can construct our confidence intervals for the different cases $\overline{X}\pm E$, based on the conditions given.

CONFIDENCE INTERVALS FOR THE MEAN

The table below gives the $(1-\alpha)$ confidence interval (formulas) for the population mean.

Case	Population	σ	n	Confidence Interval
I	Normal	known	any	$\overline{x}\pm z_{\alpha/2}\cdot\sigma/\sqrt{n}$
II	any	known	large	$\overline{x}\pm z_{\alpha/2}\cdot\sigma/\sqrt{n}$
III	Normal	unknown	small	$\overline{x}\pm t_{n-1;\alpha/2}\cdot s/\sqrt{n}$
IV	any	unknown	large	$\overline{x}\pm z_{\alpha/2}\cdot s/\sqrt{n}$

Note that n is considered large when $n\geq 30$.

EXAMPLE 6.6 (WHICH CASE?)

Solution:

Note that σ is unknown, and n is large. So we are in Case IV.

Solution:

Note that σ is unknown, and n is large. So we are in Case IV.

Example 6.6 (Which Case?)
The following data set collects $n = 41$ randomly sampled waiting times of students from ST2334 to receive reply for their email from a survey in the day time.

```
 2.50  23.28  19.34   4.74   7.03  21.85   2.72
17.73  21.55   9.71  30.24   0.37  31.26  35.24
 7.81  16.69  66.54   1.88  14.14  46.59  28.17
 0.06   9.32   0.03  10.75   6.97  56.86   2.89
 7.67  30.16   0.33   0.44   3.77  25.07   7.05
 0.08  10.64  13.10   7.92 112.77  11.93
```

Given that $\bar{x} = 17.736$ and $s = 21.7$, construct a 98% confidence interval for the mean waiting time of *all ST2334 students*.

Solution:
Note that σ is unknown, and n is large. So we are in Case IV.

Note that $z_{\alpha/2} = z_{0.01} = 2.326$. So our 98% confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 17.736 \pm 2.326 \times \frac{21.7}{\sqrt{41}} = (9.85, 25.62).$$

Example 6.7 (Which Case Again?)
The contents of 7 similar containers of sulphuric acid (in litres) are

```
9.8  10.2  10.4  9.8  10.0  10.2  9.6
```

It can shown that $\bar{x} = 10$ and $s^2 = 0.08$. Find a 95% confidence interval for the mean content of all such containers, assuming an approximate normal distribution for container contents.

Confidence Intervals for the Mean
The table below gives the $(1-\alpha)$ confidence interval (formulas) for the population mean.

Case	Population	σ	n	Confidence Interval
I	Normal	known	any	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
II	any	known	large	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
III	Normal	unknown	small	$\bar{x} \pm t_{n-1;\alpha/2} \cdot s/\sqrt{n}$
IV	any	unknown	large	$\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

Note that n is considered large when $n \geq 30$.

Solution:
We are in Case III.

Using software, we obtain $t_{6;0.025} = 2.447$.

Thus a 95% confidence interval for the mean content of all such containers is given as

$$\bar{x} \pm t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}} = 10 \pm 2.447 \cdot \frac{\sqrt{0.08}}{\sqrt{7}} = (9.738, 10.262).$$

Interpreting Confidence Intervals I
- We saw that $\bar{X} \pm E$ has probability $(1-\alpha)$ of containing μ.

  This is a probability statement about the procedure by which we compute the interval — the interval estimator.
- Each time we take a sample, and go through this construction, we get a different confidence interval.
- Sometimes we get a confidence interval that contains μ, and sometimes we get one that does not contain μ.
- Once an interval is computed, μ is either in it or not. There is no more randomness.

Interpreting Confidence Intervals II
- Since μ is typically not known, there is no way to determine whether a particular confidence interval succeeded in capturing the population mean.
- However, if we repeat this procedure of taking a sample and computing a confidence interval many times, about $(1-\alpha)$ of the many confidence intervals that we get will contain the true parameter.

  This is what "confidence" means — a confidence in the method used.
- The following R Shiny app allows us to explore this fact:
  https://istats.shinyapps.io/ExploreCoverage/

3 Comparing Two Populations

In real applications, it is quite common to compare the means of two populations.

Imagine that we have two populations

- Population 1 has mean μ_1, variance σ_1^2.
- Population 2 has mean μ_2, variance σ_2^2.

Two Basic Designs for Comparing Two Treatments
- Independent samples — complete randomization.
- Matched pairs samples — randomization between matched pairs.

Example 6.8 (Independent Samples)
In order to compare the examination scores of male and female students attending ST2334,

- 10 scores of female students are randomly sampled — Sample I,
- 8 scores of male students are randomly sampled — Sample II.

Note that all observations are independent —

- Sample I and Sample II are independent;
- Individuals within Sample I are independent;
- Individuals within Sample II are independent.

Example 6.9 (Matched Pairs Samples)
In order to study whether there exists income difference between male and female, 100 married couples are sampled, and their monthly incomes are collected.

In this example, the treatment groups are the female group and male group.

Note that observations are dependent in a special way —

- Within the pair, the observations are dependent (since they are married to one another);
- Between pairs, observations are independent.

Independent Samples (Known and Unequal Variances)
1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are known and not the same: $\sigma_1^2 \neq \sigma_2^2$
5. Either one of the following conditions holds:
   - The two populations are normal; OR
   - Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

Consider $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, random samples from the two populations of interest. Let

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \text{ and } \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

be the means of random samples. Then,

$$E(\bar{X}) = \mu_1, \quad \text{var}(\bar{X}) = \frac{\sigma_1^2}{n_1}, \quad E(\bar{Y}) = \mu_2, \quad \text{var}(\bar{Y}) = \frac{\sigma_2^2}{n_2}.$$

Thus

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 = \delta,$$

and, using the independence assumption,

$$\text{var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

When

- the two populations are normal, OR
- both samples are large,

we have

$$Z = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \approx N(0,1).$$

Confidence Intervals for $\mu_1-\mu_2$

We are interested in the difference

$$\delta = \mu_1 - \mu_2,$$

with confidence $100(1-\alpha)\%$ for any $0<\alpha<1$.

If σ_1^2 and σ_2^2 are **known**, by the distributions above, we have

$$P\left(\left|\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right| < z_{\alpha/2}\right) = 1-\alpha$$

or

$$P\left((\overline{X}-\overline{Y})-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1-\mu_2 < (\overline{X}-\overline{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) = 1-\alpha.$$

Thus the $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is

$$\left((\overline{X}-\overline{Y})-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ (\overline{X}-\overline{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

or

$$(\overline{X}-\overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

CONFIDENCE INTERVALS: KNOWN AND UNEQUAL VARIANCES

Suppose we have independent populations with known and unequal variances, and that either one of the following conditions holds:

- The two populations are normal; OR
- Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

The $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$, is then given as

$$(\overline{x}-\overline{y}) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}.$$

EXAMPLE 6.10

A study was conducted to compare two types of engines, A and B.

Gas mileage, in miles per gallon, was measured. 50 experiments were conducted using engine A. 75 experiments were done for engine type B. The gasoline used and other conditions were held constant.

- The average gas mileage for 50 experiments using engine A was 36 miles per gallon and
- The average gas mileage for the 75 experiments using machine B was 42 miles per gallon.

Find a 96% confidence interval on $\mu_B-\mu_A$, where μ_A and μ_B are the population mean gas mileage for machine types A and B, respectively.

Assume that the population standard deviations are 6 and 8 for machine types A and B, respectively.

Solution:

For a 96% confidence interval, $\alpha = 0.04$ and $z_{0.02} = 2.05$. We are also given that

$$n_1 = 50, \overline{x}_A = 36,\ \sigma_1^2 = 6^2$$
$$n_2 = 75, \overline{x}_B = 42,\ \sigma_2^2 = 8^2$$

The sample sizes are large, so a 96% confidence interval for $\mu_B-\mu_A$ is

$$\begin{aligned}&(\overline{x}_B-\overline{x}_A) \pm z_{\alpha/2}\sqrt{\sigma_2^2/n_2+\sigma_1^2/n_1}\\ &= (42-36) \pm 2.05 \cdot \sqrt{8^2/75+6^2/50}\\ &= (3.428, 8.571).\end{aligned}$$

INDEPENDENT SAMPLES (LARGE, WITH UNKNOWN VARIANCES)

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are unknown and not the same: $\sigma_1^2 \neq \sigma_2^2$
5. Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

Since σ_1 and σ_2 are unknown, let

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2, \text{ and } S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2$$

and use

$$Z = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \approx N(0,1).$$

CONFIDENCE INTERVALS: LARGE, WITH UNKNOWN VARIANCES

Suppose we have independent populations with unknown and unequal variances, and that both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

The $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$, is then given as

$$(\overline{x}-\overline{y}) \pm z_{\alpha/2}\sqrt{s_1^2/n_1+s_2^2/n_2}.$$

If σ_1^2 and σ_2^2 are **unknown**, the $100(1-\alpha)\%$ CI for $\mu_1-\mu_2$ is

$$\left((\overline{X}-\overline{Y})-z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\ (\overline{X}-\overline{Y})+z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$$

or

$$(\overline{X}-\overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}.$$

INDEPENDENT SAMPLES: SMALL, WITH EQUAL VARIANCES

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are unknown and the same: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
5. Both samples are small: $n_1 < 30$, $n_2 < 30$
6. Both populations are normally distributed.

Based upon the normal distribution and equal variance assumptions

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1).$$

Since σ is unknown, we shall estimate it.

Note that S_1^2 and S_2^2 are both unbiased estimators of σ^2 under the equal variance assumption.

We can use the pooled estimator to estimate σ^2 better.

DEFINITION 6 (THE POOLED ESTIMATOR: S_p^2)

*σ^2 can be estimated by the **pooled sample variance***

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

with S_1^2 and S_2^2 being the sample variances of the first and second samples respectively.

When we estimate σ^2 using S_p^2, the resulting statistic

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows a t-distribution with degrees of freedom $n_1 + n_2 - 2$.

We then have

$$P\left(-t_{n_1+n_2-2;\alpha/2} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{n_1+n_2-2;\alpha/2}\right) = 1 - \alpha.$$

CONFIDENCE INTERVALS: SMALL, WITH EQUAL VARIANCES

Suppose we have independent, normal populations with unknown and equal variances, and that both samples are small: $n_1 < 30$, $n_2 < 30$.

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given as

$$(\overline{X} - \overline{Y}) \pm t_{n_1+n_2-2;\alpha/2}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

EXAMPLE 6.11

A course in mathematics is taught to 12 students by the conventional classroom procedure. A second group of 10 students was given the same course by means of programmed materials.

At the end of the semester the same examination was given to each group.

- The 12 students meeting in the classroom made an average grade of 85 with standard deviation of 4.
- The 10 students using programmed materials made an average of 81 with a standard deviation of 5.

Find a 90% confidence interval for the difference between the population means, assuming the populations are approximately normally distributed with equal variances.

Solution:

Let μ_1 and μ_2 represent the average grades of all students who might take this course by the classroom and programmed presentations respectively.

So $\overline{x} - \overline{y} = 85 - 81 = 4$ is the point estimate for $\mu_1 - \mu_2$.

As we assume equal population variance, we estimate it by the pooled variance

$$s_p^2 = \frac{(12-1)\times 4^2 + (10-1)\times 5^2}{12 + 10 - 2} = 20.05.$$

In this case, $t_{n_1+n_2-2;\alpha/2} = t_{20;0.05} = 1.7247$. Thus a 90% confidence interval for $\mu_1 - \mu_2$ is given as

$$\begin{aligned}
&(\overline{x} - \overline{y}) \pm t_{n_1+n_2-2;\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= (85 - 81) \pm 1.7247 \times \sqrt{20.05} \times \sqrt{\frac{1}{12} + \frac{1}{10}} \\
&= (0.693, 7.307).
\end{aligned}$$

CONFIDENCE INTERVALS: LARGE, WITH EQUAL VARIANCES

Suppose we have independent populations with unknown and equal variances, and that both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given as

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\, S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

PAIRED DATA

1. $(X_1, Y_1), \ldots, (X_n, Y_n)$ are matched pairs, where $X_1, \ldots, X_n$ is a random sample from population 1, $Y_1, \ldots, Y_n$ is a random sample from population 2.
2. X_i and Y_i are dependent.
3. (X_i, Y_i) and (X_j, Y_j) are independent for any $i \neq j$.
4. For matched pairs, define $D_i = X_i - Y_i$. $\mu_D = \mu_1 - \mu_2$.
5. Now we can treat $D_1, D_2, \ldots, D_n$ as a random sample from a single population with mean μ_D and variance σ_D^2.

All techniques derived for a single population can now be employed.

- We consider the statistic

$$T = \frac{\overline{D}-\mu_D}{S_D/\sqrt{n}}, \quad \text{where} \quad \overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}, \quad S_D^2 = \frac{\sum_{i=1}^{n}(D_i-\overline{D})^2}{n-1}.$$

- If $n < 30$ and the population is normally distributed then

$$T \sim t_{n-1}.$$

- If $n \geq 30$, then

$$T \sim N(0,1).$$

CONFIDENCE INTERVALS: PAIRED DATA

For paired data, if n is small ($n < 30$) and the population is normally distributed, a $(1-\alpha)100\%$ confidence interval for μ_D is

$$\overline{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

If n is large ($n \geq 30$), a $(1-\alpha)100\%$ confidence interval for μ_D is

$$\overline{d} \pm z_{\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

EXAMPLE 6.12

Twenty students were divided into 10 pairs, each member of the pair having approximately the same IQ.

One of each pair was selected at random and assigned to a mathematics section using programmed materials only. The other member of each pair was assigned to a section in which the professor lectured.

At the end of the semester each group was given the same examination and the following results were recorded.

Pair	1	2	3	4	5	6	7	8	9	10
P.M.	76	60	85	58	91	75	82	64	79	88
Lecture	81	52	87	70	86	77	90	63	85	83
d	−5	8	−2	−12	5	−2	−8	1	−6	5

Given that $\overline{d} = -1.6$ and $s_D^2 = 40.71$, compute a 98% confidence interval for the true difference in the two learning procedures.

Solution:

Since $\alpha = 0.02$, we have $t_{n-1;\alpha/2} = t_{9,0.01} = 2.821$. Thus a 98% confidence interval for the true difference μ_D is given as

$$\overline{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}} = -1.6 \pm 2.821 \times \sqrt{\frac{40.71}{10}} = (-7.292, 4.092).$$

Tutorial 09

Thursday, November 9, 2023 1:02 AM

Tutorial 09

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE
ST2334 PROBABILITY AND STATISTICS
SEMESTER I, AY 2023/2024

Tutorial 09

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer.**

   Let $X \sim N(0,64)$. Then $E(X^{2k+1})$, for any $k = 0,1,2,\ldots$, must be

   (a) greater than 0.
   (b) smaller than 0.
   (c) equal to 0.
   (d) depending on the value of k.

2. **Multiple choice question: choose the unique correct answer.**

   Let $X \sim \text{Bin}(n,p)$, where n is a given constant, but p is an unknown parameter. Which of the following statements is **INCORRECT**?

   (a) Let $U = X/n$. Then U is an unbiased estimator of p.
   (b) Let $V = \dfrac{X+n/2}{3n/2}$. Then V is an unbiased estimator of p.
   (c) Let $W = \dfrac{X(X-1)}{n(n-1)}$. Then W is an unbiased estimator of p^2.
   (d) None of the given options.

3. **Multiple choice question: choose the unique correct answer.**

   A random sample of size $n_1 = 25$ taken from a normal population with a standard deviation $\sigma_1 = 5$ has a mean $\bar{x}_1 = 80$. A second random sample of size $n_2 = 36$ taken from a different normal population with a standard deviation $\sigma_2 = 3$ has a mean $\bar{x}_2 = 75$. A 94% confidence interval for $\mu_1 - \mu_2$ is approximately given by:

   Note: $\Phi(1.88) \approx 0.97$; $\Phi(1.55) \approx 0.94$.

   (a) (2.898, 7.102)
   (b) (1.982, 7.991)
   (c) (1.489, 6.511)
   (d) None of the given options

4. **Fill in the blank.**

   Let $\{X_1, X_2, \ldots, X_{10}\}$ be randomly sampled **with replacement** from the population of $\{0,1,3\}$. Find the variance of the sample mean.

   Answer: _____.

5. **True/False.**

   Let $X_1, X_2, \ldots X_n$ be a random sample from a certain population. Then the larger n, the smaller the variance of the sample mean.

   • TRUE

- FALSE

Analytical Questions

1. If S_1^2 and S_2^2 represent the variances of independent random samples of size $n_1 = 8$ and $n_2 = 12$, taken from normal populations with equal variances, find $P\left(S_1^2/S_2^2 < 4.89\right)$.

2. Assume that the helium porosity Y (in percentage) of coal samples taken from any seam is normally distributed.

   (a) If the true standard deviation of Y is 0.75, compute a 95% confidence interval for the average porosity of a certain seam. We are given that the average porosity for 20 specimens from the seam was 4.85.

   (b) How large a sample size is necessary if the length of the 95% interval is to be 0.40?

   (c) If the variance of Y is unknown, and the sample standard deviation for the sample in (a) is 0.75, compute a 95% confidence interval for the average porosity of a certain seam.

3. A random sample of 12 shearing pins is taken in a study of the Rockwell hardness of the head on the pin. Measurements on the Rockwell hardness were made for each of the 12, yielding an average value of 48.50 with a sample standard deviation of 1.5. Assuming the measurements to be normally distributed, construct a 90% confidence interval for the mean Rockwell hardness.

4. A study was conducted to determine if a certain metal treatment has any effect on the amount of metal removed in a pickling operation. A random sample of 100 pieces was immersed in a bath for 24 hours without the treatment, yielding an average of 12.2 millimeters of metal removed and a sample standard deviation of 1.1 millimeters. A second sample of 200 pieces was exposed to the treatment, followed by the 24-hour immersion in the bath, resulting in an average removal of 9.1 millimeters of metal with a sample standard deviation of 0.9 millimeters. Compute a 98% confidence interval estimate for the difference between the population means. Does the treatment appear to reduce the mean amount of metal removed?

$n_1 = 100 \quad \bar{X}_1 = 12.2 \quad S_1 = 1.1$ — σ is unknown

$n_2 = 200 \quad \bar{X}_2 = 9.1 \quad S_2 = 0.9$ — not equal

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm Z_{\frac{0.02}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Answers for Some of the Analytical Questions

1. 0.99.
2. (a) (4.5213, 5.7187); (b) 54; (c) (4.499, 5.201).
3. (47.722, 49.278).
4. (2.804, 3.396).

Ex 5.14 ch 5

CASE 1

CASE 3

CASE 3 t-distb

2 SAMPLE

CASE 2

$P\left(\frac{S_1^2}{S_2^2} < 4.89\right)$

1.)

one sample - Case 1

2.) a. $n = 20 \quad \bar{y} = 4.85 \quad \sigma_y = 0.75$

small distrib

$\bar{y} \pm Z \cdot \frac{\sigma}{\sqrt{n}} =$

(4.899, 5.201)

b. length of CI $= 2 \cdot Z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} = 0.4$ (1.96, 0.75)

$n = 54$

c. $4.85 \pm t_{19;0.025} \cdot \frac{0.75}{\sqrt{20}} =$ (2.093)

3.) $n = 12 \quad \bar{h} = 48.5 \quad s = 1.5$

small population, σ known

$\bar{x} \pm t_{11,\frac{0.1}{2}} \cdot \frac{S}{\sqrt{n}}$ (1.796)

Tutorial 10

Wednesday, November 15, 2023 6:28 PM

Tutorial 10

NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF STATISTICS AND DATA SCIENCE
ST2334 PROBABILITY AND STATISTICS
SEMESTER I, AY 2023/2024

Tutorial 10

Please work on the questions before attending the tutorial.

Exam Format Questions

1. **Multiple choice question: choose the unique correct answer**.

   In order to halve the width of a 95% confidence interval for a mean of normal distribution, by what factor should the sample size be increased?

   $\pm z^* \frac{\sigma}{\sqrt{n}}$

   (a) $\sqrt{2}$ (c) 3
   (b) 2 (d) 4

2. **Multiple response question: choose all that apply**.

   Which of the following statements is/are **CORRECT**?

   (a) Type I error can occur if we reject the null hypothesis.
   (b) Type II error can occur if we do not reject the null hypothesis.
   (c) Type I and Type II errors may occur simultaneously in a single test.
   (d) In a test, the greater the type II error probability, the smaller the power of the test.

3. **Fill in the blank**.

   In 64 randomly selected hours of production, the mean and the standard deviation of the number of acceptable pieces produced by an automatic stamping machine are $\bar{x} = 1{,}038$ and $s = 146$. The p-value for testing the hypotheses $H_0 : \mu = 1{,}000$ versus $H_1 : \mu > 1{,}000$ is given by $\Phi(c)$. Find the value for c.

   $P(X>1000)$

   $\frac{1038-1000}{146/8} = 2.082$

   Answer: $c =$ -2.08.

   (Round your answer to the second decimal place.)

   Note: $\Phi(\cdot)$ denotes the cumulative distribution of $N(0,1)$.

   $P(Z>2.082) = P(Z<-2.082)$
   $= \Phi(-2.08$

4. **Multiple choice question: choose the unique correct answer**.

   Suppose we test the hypotheses

   $$H_0 : \mu = 2 \quad \text{vs} \quad H_1 : \mu \neq 2$$

   $\frac{\bar{x}-2}{s/\sqrt{n}} = 2$

   P

   and found a two-sided p-value of 0.03.

   Separately, we constructed the 95% confidence interval for μ. Which of the following is possibly the constructed confidence interval?

   (a) (1.5, 4.0) (c) (1.9, 3.3)
   (b) (1.2, 1.9) (d) None of the given options

1

Analytical Questions

1. A manufacturer claims that the average tar content of a certain kind of cigarette is $\mu = 14.0$. In an attempt to show that it differs from this value, five measurements are made:

   ```
   14.5 14.2 14.4 14.3 14.6
   ```

   Show that the difference between the mean of this sample, $\bar{x} = 14.4$, and the average tar claimed by the manufacturer, $\mu = 14.0$, is significant at $\alpha = 0.05$. Assume normality.

2. A study based on a sample size of 36 reported a mean of 87 with a margin of error of 10 for 95% confidence.

   (a) Give the 95% confidence interval for the population mean μ.

   (b) You desire a margin of error of 2.5 with the same confidence level. What is the sample size that will give you that kind of accuracy? Assume that we know the population variance.

   (c) You are asked to test the hypothesis that $\mu = 80$ against a two sided alternative at $\alpha = 0.05$. What is your conclusion?

3. In a study, the mean CAP (cumulative average point) of a random sample of 49 final year students is calculated to be 4.5. The standard deviation for this sample is given as 0.75.

   (a) Find a 95% confidence interval for the mean CAP of the entire final year class.

   (b) The university administration claims that the mean CAP for the entire final year class is 4.3. Does our study offer evidence against this claim? Explain.

4. The dynamic modulus of concrete is obtained for two different samples of concrete mixes. For the first mix, $n_1 = 33$, $\bar{x} = 115.1$, and $s_1 = 0.47$ psi. For the second mix, $n_2 = 31$, $\bar{y} = 114.6$, and $s_2 = 0.38$ psi. Test, with $\alpha = 0.05$, the null hypothesis of equality of mean dynamic modulus versus the two-sided alternative. **Assume that the population variances are not the same.**

5. Obtain a 95% confidence interval for the difference in mean dynamic modulus in Question 4.

6. Two procedures for etching integrated circuits are to be compared. Given 10 units, five are prepared using etching procedure *A* and five are prepared using etching procedure *B*.

   (a) The response is the percent of area on the integrated circuit where the etching was inadequate. Suppose the results are

Prodedure A	5	2	9	6	3
Prodedure B	1	3	4	0	2

   Find a 95% confidence interval for the difference in means.
   Summary statistics are given as $\bar{x} = 5$, $s_1 = 2.73$ and $\bar{y} = 2$, $s_2 = 1.58$.

   (b) What assumptions did you make for your answer to part (b)?

7. A manager is considering instituting an additional 15-minute coffee break if it can be shown to decrease the number of errors that employees commit. The manager divided a sample of 20 employees into two groups of 10 each. Members of one group followed the same work schedule as before, but the members of the other group were given a 15-minute coffee break in the middle of the day. The following data give the total number of errors committed by each of the 20 workers over the next 20 working days.


```
Coffee break group:     8, 7,  5,  8, 10, 9, 7, 8,  4, 5
No-break group:         7, 6, 14, 12, 13, 8, 9, 6, 10, 9
```

Summary statistics are given as $\bar{x} = 7.1$, $s_1 = 1.91$, $\bar{y} = 9.4$ and $s_2 = 2.84$.

(a) Using a suitable test, at the 5 percent level of significance, test the hypothesis that instituting a coffee break reduces the mean number of errors. What is your conclusion?

(b) What assumptions have you made?

(c) What is the *p*-value of your test?

8. Two methods are used to determine the percentage of lidocaine in a pharmaceutical formulation that is supposed to contain 27 percent lidocaine. The contents of 10 different tubes of lidocaine solution were each analyzed by both methods with results shown here:

Method 1	31.36	24.57	27.59	29.18	24.34	25.04	26.40	26.83	28.35	31.12
Method 2	31.41	24.78	27.62	29.27	24.59	25.04	26.57	26.96	28.55	31.00
Difference	−0.05	−0.21	−0.03	−0.09	−0.25	0.00	−0.17	−0.13	−0.20	0.12

Some summary statistics are given:

$\bar{x}_{\text{Method 1}} = 27.48$ $\quad \bar{x}_{\text{Method 2}} = 27.58$ $\quad \bar{x}_{\text{Difference}} = -0.101$

$s_{\text{Method 1}} = 2.54$ $\quad s_{\text{Method 2}} = 2.47$ $\quad s_{\text{Difference}} = 0.11367$

We are interested to find out if the two methods give mean results that are significantly different. Use a suitable test at the 5% significance level to determine that. Assume normality.

matched pair

$$T = \frac{-0.101 - 0}{0.11367/\sqrt{10}} = -2.8098$$

$$P\left(-t_{9;0.025} < -2.8098 < t_{9;0.025}\right)$$

-2.262 $\qquad$ 2.262

reject H_0

$\frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{14.4-14}{\sqrt{0.9}/\sqrt{5}} = 0.9428$

$P(Z > 0.9428)$

$\frac{1}{n-1}\sum(\bar{x}-x)^2 = V$

$s = \sqrt{V} = 0.158$

$10 = z \cdot \sqrt{\frac{s^2}{36}}$ (1.96)

$s = 9.0175$

b.36*4*4 = 576

$4.5 \pm z_{0.025} \cdot \frac{s}{\sqrt{n}}$

$1.96 \cdot \frac{0.75}{7}$

$H_0 = \mu = 4.3$ $\quad H_A = \mu \neq 4.3$ $\quad z_{0.025}$

$z^* = \frac{4.5-4.3}{0.75/7} = 1.8666 < 1.96$

4.) $H_0: \mu_1 - \mu_2 = 0$ $\quad H_A: \mu_1 - \mu_2 \neq 0$

$z = \frac{\bar{x}-\bar{y}-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{115.1-114.6}{\sqrt{\frac{0.47^2}{33}+\frac{0.38^2}{31}}} = 4.69$

$z^* > z_{\alpha/2}$

$z_{\alpha/2} = 1.96$

$z^* > 1.96$ or $z^* < -1.96$ (reject H_0)

5.) $(\bar{x}-\bar{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ =

6.) a. $\bar{A} = 5$ $\quad \bar{B} = 2$

7.)

$\alpha = 0.05$ $\quad H_0: \mu_X - \mu_Y = 0$ $\quad H_A: \mu_X - \mu_Y < 0$

$P(\bar{X}-\bar{Y} < 0)$

$\frac{(\bar{x}-\bar{y})-0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = -2.125$

$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{9(1.91)^2 + 9(2.84)^2}{18} = 5.85685$

$n = 10$ $\quad$ t test $\quad t_{9;0.05}$

$\sim t_{18}$ $\quad -t_{18;0.05} = -1.734$

Since $-2.13 < -1.734$

reject H_0

reject H_0

paired t-distrib.

$P(t_{18} < -2.13) = 0.023$

-2.262 2.262

reject H_0

Answers for Some of the Analytical Questions

1. Reject H_0.
2. (a) (77,97); (b) 576; (c) Do not reject H_0.
3. (a) (4.29, 4.70); (b) No.
4. Reject H_0.
5. (0.29, 0.71).
6. (a) $(-0.26, 6.26)$; (b) Normality and independence.
7. (a) Reject H_0; (b) Normality and independence; (c) 0.023.
8. Reject H_0.

3

Wk12 ch7

Thursday, November 16, 2023 11:58 AM

Chapter 7 - Print

Seven

Hypothesis Tests

1 Hypothesis Tests

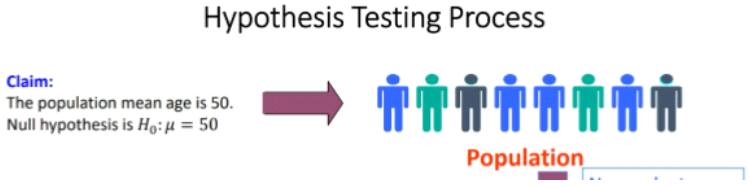
One of the most fundamental technique of statistical inference is the hypothesis test. There are many types of hypothesis tests but all follow the same logical structure, so we begin with hypothesis testing of a population mean.

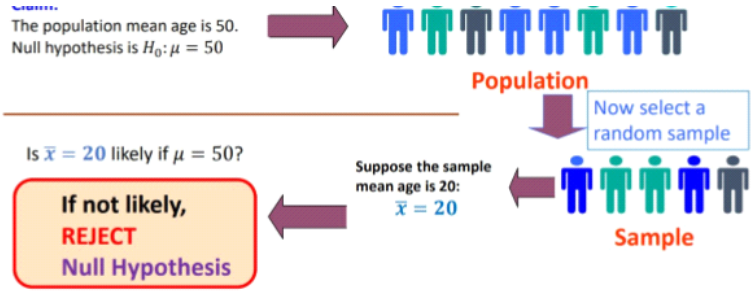
Hypothesis testing begins with a null hypothesis and an alternative hypothesis. Both the null and the alternative hypotheses are statements about a population. In this chapter, that statement will be a statement about the mean(s) of the population(s).

We will illustrate using an example.

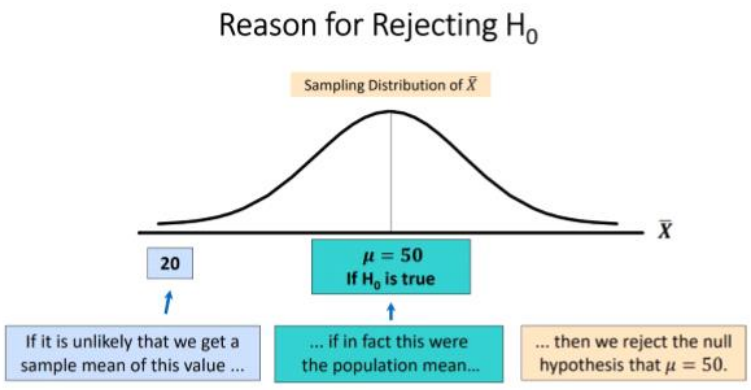
Example 7.1 (Mean Age)
We are interested to check if the mean age of a population is $\mu = 50$.

Suppose we have no access to population data. So we take a sample from the population and obtained a sample mean age of $\bar{x} = 20$. Does this gives evidence for or against the hypothesis that $\mu = 50$?





1



EXAMPLE 7.2 (NUS STUDENTS' IQ)
Consider the statement

"NUS students have higher IQ than the general population (100)."

It is difficult/expensive to ask every NUS student to take an IQ test. So we take a sample.

Suppose the sample average is 110.

- Does that mean we're right?
- What if the sample average is 101? What about 100.1?
- Does the sample size matter?

HOW TO DO A HYPOTHESIS TEST
There are five main steps to hypothesis testing.

Step 1: Set your competing hypotheses: null and alternative

HOW TO DO A HYPOTHESIS TEST

There are five main steps to hypothesis testing.

Step 1: Set your competing hypotheses: null and alternative.

Step 2: Set the level of significance.

Step 3: Identify the test statistic, its distribution and the rejection criteria.

Step 4: Compute the observed test statistic value, based on your data.

Step 5: Conclusion.

Let us have a closer look at each step.

Step 1: Null Hypothesis vs Alternative Hypothesis

Our goal is to decide between two competing hypotheses.

NULL VS ALTERNATIVE

In general, we adopt the position of the null hypothesis unless there is overwhelming evidence against it.

The null hypothesis is typically the default assumption, or the conventional wisdom about a population. Often it is exactly the thing that a researcher is trying to show is false.

We usually let the hypothesis that we want to prove be the alternative hypothesis. The alternative hypothesis states that the null hypothesis is false, often in a particular way.

The outcome of hypothesis testing is to either reject or fail to reject the null hypothesis.

A researcher would collect data relating to the population being studied and use a hypothesis test to determine whether the evidence against the null hypothesis (if any) is strong enough to reject the null hypothesis in favor of the

A researcher would collect data relating to the population being studied and use a hypothesis test to determine whether the evidence against the null hypothesis (if any) is strong enough to reject the null hypothesis in favor of the alternative hypothesis.

We usually phrase the hypotheses in terms of population parameters.

EXAMPLE 7.3 (ONE-SIDED TEST)
Let μ be the average IQ of NUS students. Consider

$$H_0 : \mu = 100 \quad \text{vs} \quad H_1 : \mu > 100.$$

This is an example of a one-sided hypothesis test.

For this alternative hypothesis, we do not care if $\mu < 100$: the goal here is just to show NUS students have IQ higher than 100.

EXAMPLE 7.4 (TWO-SIDED TEST)
Sometimes it is more natural to do a two-sided hypothesis test.

For example, let p be the probability of heads for a particular coin. You want to test if the coin is fair (that is, $p = 0.5$), as it is equally problematic if p was larger or smaller.

Hence you set your hypotheses to be

$$H_0 : p = 0.5 \quad \text{vs} \quad H_1 : p \neq 0.5.$$

Step 2: Level of Significance

For any test of hypothesis, there are two possible conclusions:

- Reject H_0 and therefore conclude H_1;
- Do not reject H_0 and therefore conclude H_0.

Whatever decision is made, there is a possibility of making an error.

	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	**Type I error**
H_0 is false	**Type II error**	Correct Decision

	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	**Type I error**
H_0 is false	**Type II error**	Correct Decision

DEFINITION 1 (TYPE I VS TYPE II ERROR)
*The rejection of H_0 when H_0 is true is called a **Type I error**.*

*Not rejecting H_0 when H_0 is false is called a **Type II error**.*

DEFINITION 2 (SIGNIFICANCE LEVEL VS POWER)
*The probability of making a Type I error is called the **level of significance**, denoted by α. That is,*

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ is true}).$$

Let

$$\beta = P(\text{Type II error}) = P(\text{Do not reject } H_0 \mid H_0 \text{ is false}).$$

*We define $1-\beta = P(\text{Reject } H_0 \mid H_0 \text{ is false})$ to be the **power of the test**.*

REMARK
The Type I error is considered a serious error, so we want to control the probability of making such an error.

Thus prior to conducting a hypothesis test, we set the significance level α to be small, typically at $\alpha = 0.05$ or 0.01.

Step 3: Test Statistic, Distribution and Rejection Region

To test the hypothesis, we first select a suitable test statistic for the parameter under the hypothesis.

The test statistic serves to quantify just how unlikely it is to observe the sample, assuming the null hypothesis is true.

The test statistic serves to quantify just how unlikely it is to observe the sample, assuming the null hypothesis is true.

As the significance level α is given, a decision rule can be found such that it divides the set of all possible values of the test statistic into two regions, one being the rejection region (or critical region) and the other, the acceptance region.

Step 4 & 5: Calculation and Conclusion

Once a sample is taken, the value of the test statistic is obtained.

We check if it is within our rejection region.

- If it is, our sample was too improbable assuming H_0 is true, hence we reject H_0.
- If it is not, we did not accomplish anything. We failed to reject H_0 and hence fall back to our original assumption of H_0.

Note that in the latter case, we did not "prove" that H_0 is true. Hence, it is prudent to use the term "fail to reject H_0" instead of "accept H_0."

L-EXAMPLE 7.1

A certain type of cold vaccine is known to be only 25% effective after a period of 2 years.

In order to determine if a new and somewhat more expensive vaccine is superior in providing protection against the same virus for a longer period of time, 20 people are chosen at random and inoculated with the new vaccine.

If more than 8 of those receiving the new vaccine surpass the 2-year period without contracting the virus, the new vaccine will be considered superior to the one presently in use.

This is equivalent to testing the hypothesis that the binomial parameter for the probability of a success on a given trial is $p = \frac{1}{4}$ against the alternative that $p > \frac{1}{4}$.

In other words, we are testing

$$H_0 : p = \frac{1}{4} \quad \text{vs} \quad H_1 : p > \frac{1}{4}.$$

Let X be the number of individuals who remain free of the virus for at least 2 years.

For the conditions given, we think of the "acceptance region" and "rejection region" as

$$\overbrace{0,1,2,\ldots,7,8,}^{\text{acceptance region}}\ \overbrace{9,10,11,\ldots,19,20}^{\text{rejection region}}$$

The above decision rule has the level of significance given by

$$\begin{aligned}\alpha &= P(\text{Type I error})\\ &= P(\text{Reject } H_0 \mid H_0 \text{ is true})\\ &= P(X > 8 \mid p = \frac{1}{4})\\ &= \sum_{i=9}^{20} \binom{20}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{20-i} = 0.0409.\end{aligned}$$

The probability of committing a Type II error is impossible to compute unless we have a specific alternative hypothesis. So let's consider

$$H_0 : p = \frac{1}{4} \quad \text{vs} \quad H_1 : p = \frac{1}{2}.$$

Note that $p = \frac{1}{2}$ satisfies $p > \frac{1}{4}$.

With this,

$$\begin{aligned}\beta &= P(\text{Type II error})\\ &= P(\text{Do not reject } H_0 \mid H_1 \text{ is true})\\ &= P(X \leq 8 \mid p = \frac{1}{2})\\ &= \sum_{i=0}^{8} \binom{20}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{20-i} = 0.2517.\end{aligned}$$

2 Hypotheses concerning the Mean

Let's apply our hypothesis steps to testing a population mean.

Case: Known variance

Let us consider the case where

- the population variance σ^2 is known; AND
- where
  - the underlying distribution is normal; OR
  - n is sufficiently large (say, $n \geq 30$).

Step 1: We set the null and alternatives hypotheses as

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

Note that in this case we are considering a two-sided alternative hypothesis.

Step 2: Set level of significance: α is typically set to be 0.05.

Step 3: Statistic & its distribution:
With σ^2 known and population normal (or $n \geq 30$),

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

When H_0 is true, $\mu = \mu_0$, the above becomes

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

and will serve as our test statistic.

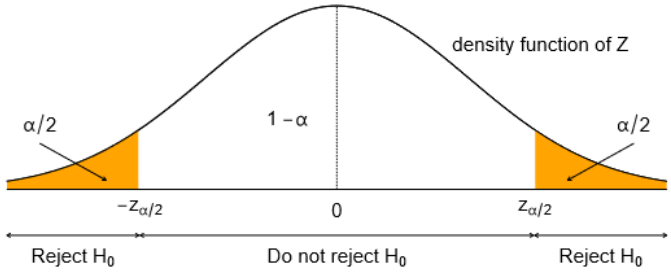
Rejection region:
Intuitively, we should reject H_0 when $\overline{X}$ is too large or too small compared with μ_0.

This is the same as when Z is too large or too small. In theory,

$$P(|Z| > z_{\alpha/2}) = \alpha.$$

Let the observed value of Z be z. Then the rejection region is defined by $|z| > z_{\alpha/2}$, which is

$$z < -z_{\alpha/2} \quad \text{or} \quad z > z_{\alpha/2}.$$



Step 4: Computations: z should be computed from the statistic above based upon the observed sample.

Step 5: Conclusion: check whether z is located within rejection region. If so, reject H_0, otherwise do not reject H_0.

WHERE DID THE VALUE 0.05 COME FROM?
In 1931, in a famous book called The Design of Experiments, *Sir Ronald Fisher discussed the amount of evidence needed to reject a null hypothesis.*

He said that it was situation dependent, but remarked, somewhat casually, that for many scientific applications, 1 out of 20 might be a reasonable value.

Since then, some people — indeed some entire disciplines — have treated the number 0.05 as sacrosanct.

Sir Ronald Fisher (1890 – 1962) was one of the founders of modern Statistics. For a biography of Fisher, browse to
`http://www-history.mcs.st-andrews.ac.uk/Biographies/Fisher.html`

EXAMPLE 7.5
The director of a factory wants to determine if a new machine A is producing cloths with a breaking strength of 35 kg with a standard deviation of 1.5 kg.

A random sample of 49 pieces of cloths is tested and found to have a mean breaking strength of 34.5 kg. Is there evidence that the machine is not meeting the specifications for mean breaking strength?
Use $\alpha = 0.05$.

Solution:
Note that $n > 30$ and $\sigma = 1.5$.

Let μ be the mean breaking strength of cloths manufactured by the new machine.

Step 1: We test

$$H_0 : \mu = 35 \quad \text{vs} \quad H_1 : \mu \neq 35.$$

Step 2: Set $\alpha = 0.05$.

$$H_0 : \mu = 35 \quad \text{vs} \quad H_1 : \mu \neq 35.$$

Step 2: Set $\alpha = 0.05$.

Step 3: As σ^2 is known and $n \geq 30$,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

will serve as our test statistic.

Since $z_{\alpha/2} = z_{0.025} = 1.96$, the critical/rejection region is

$$z < -1.96 \quad \text{or} \quad z > 1.96.$$

Step 4: z is computed to be

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{34.5 - 35}{1.5/\sqrt{49}} = -2.3333 < -1.96.$$

Step 5: The observed z value, $z = -2.3333$, falls inside the critical region. Hence the null hypothesis $H_0 : \mu = 35$ is rejected at the 5% level of significance.

One-sided alternatives

Now the above procedures are establish under

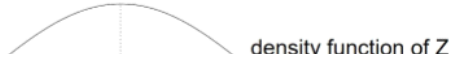
$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

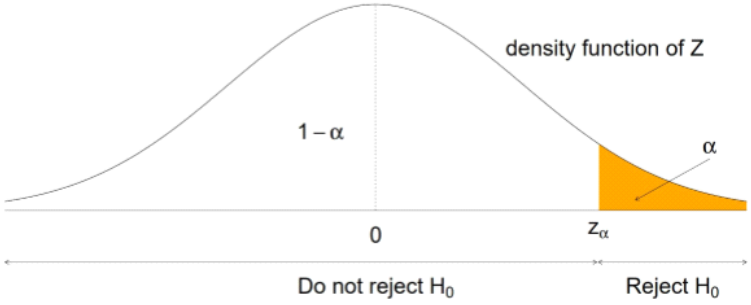
Suppose instead we are considering

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0.$$

Similar steps can be used to address this problem, we only need to do the following changes:

- Step 1: H_1 is replaced with $H_1 : \mu > \mu_0$.
- Step 3: The test statistic and its distribution are kept the same. The rejection region should be replaced with $z > z_\alpha$, since now, we should reject only when $\overline{x}$ (and therefore z) is large.





The case for

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0$$

should be self-evident.

HYPOTHESIS TEST FOR THE MEAN: KNOWN VARIANCE

Consider the case where

- the population variance σ^2 is known; AND
- where
  - the underlying distribution is normal; OR
  - n is sufficiently large (say, $n \geq 30$).

For the null hypothesis $H_0 : \mu = \mu_0$, the test statistics is given by

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

Let z be the observed Z value. For the alternative hypothesis

- $H_1 : \mu \neq \mu_0$, the rejection region is

$$z < -z_{\alpha/2} \quad \text{or} \quad z > z_{\alpha/2}.$$

- $H_1 : \mu < \mu_0$, the rejection region is

$$z < -z_\alpha.$$

- $H_1 : \mu > \mu_0$, the rejection region is

$$z < -z_\alpha.$$

- $H_1 : \mu > \mu_0$, the rejection region is

$$z > z_\alpha.$$

p-value approach to testing

The above technique introduced by Fisher is based on a pre-declared significance level α.

Today, there is little reason to stick to the arbitrary 1% or 5% levels that Fisher suggested. We can instead use the idea of the p-value.

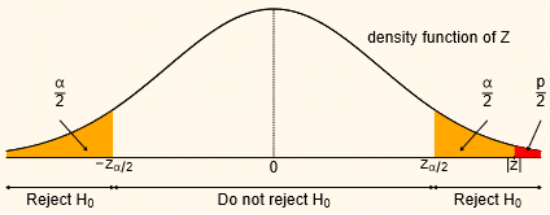
> **DEFINITION 3 (p-VALUE)**
> *The **p-value** is the probability of obtaining a test statistic at least as extreme ($\leq$ or $\geq$) than the observed sample value, given H_0 is true.*
>
> *It is also called the observed level of significance.*

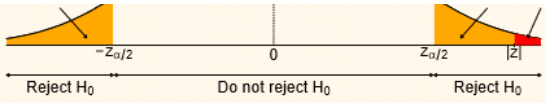
p-VALUE FOR HYPOTHESIS TESTS

Suppose our computed test statistic was z. For a two sided test, a "worse" result would be if $Z > |z|$ or $Z < -|z|$, in other words, $|Z| > |z|$.

So the p-value is given by

$$p\text{-value} = P(|Z| > |z|) = 2P(Z > |z|) = 2P(Z < -|z|)$$

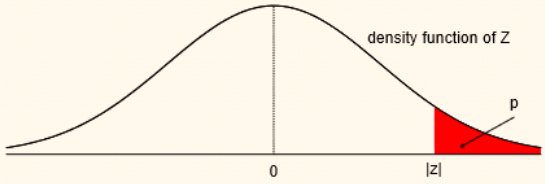




z falls in the rejection region if and only if $p < \alpha$

For the alternative hypothesis $H_1 : \mu < \mu_0$, the p-value is $P(Z < -|z|)$.
That is, only the area in the left tail is used.

For the alternative hypothesis $H_1 : \mu > \mu_0$, the p-value is $P(Z > |z|)$.
That is, only the area in the right tail is used.



One−sided test: p−value is represented by the shaded area

REJECTION CRITERIA USING p-VALUE

We see that the p-value is smaller than the significance level *if and only if* our test statistic is in the rejection region.

Thus our rejection criteria would be

- If p-value $< \alpha$, reject H_0; else
- If p-value $\geq \alpha$, do not reject H_0.

REMARK

In practice, it is better to report the p-value than to indicate whether H_0 is rejected.

- The p-values of 0.049 and 0.001 both result in rejecting H_0 when $\alpha = 0.05$, but the second case provides much stronger evidence.
- p-values of 0.049 and 0.051 provide, in practical terms, the same amount of

- The p-values of 0.049 and 0.001 both result in rejecting H_0 when $\alpha = 0.05$, but the second case provides much stronger evidence.
- p-values of 0.049 and 0.051 provide, in practical terms, the same amount of evidence about H_0.

Most research articles report the p-value rather than a decision about H_0. From the p-value, readers can view the strength of evidence against H_0 and make their own decision, if they want to.

EXAMPLE 7.6 (MIDTERM EXAM SCORE)
Recall the midterm exam scores example in an earlier chapter. The data obtained are

$$20, 19, 24, 22, 25.$$

We were told that the exam scores are approximately normal.

The lecturer announced that the variance of the exam score over the class is 5 (just believe that this is the truth). Test at $\alpha = 0.01$ significance level whether the average midterm score is different from 16.

Solution:
Let μ be the average midterm score for the whole class.

Step 1: $H_0 : \mu = 16 \quad \text{vs} \quad H_1 : \mu \neq 16.$

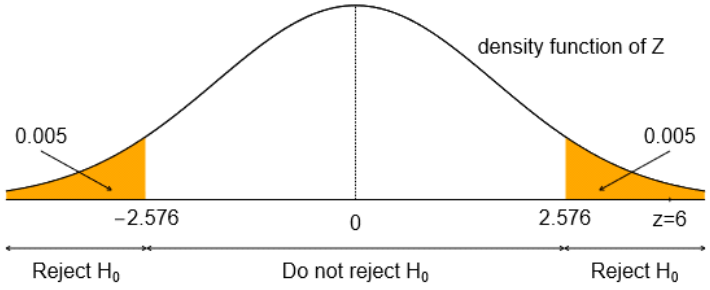
Step 2: Choose $\alpha = 0.01$.

Step 3: In this example $\sigma = \sqrt{5}$ is known, data are normal, and $n = 5$. Therefore the test statistic and its distribution is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

Now $z_{\alpha/2} = z_{0.005} = 2.576$. Thus the rejection region is

$$z < -2.576 \quad \text{or} \quad z > 2.576.$$



Step 4: $z = (22-16)/(\sqrt{5}/\sqrt{5}) = 6 > 2.576$.

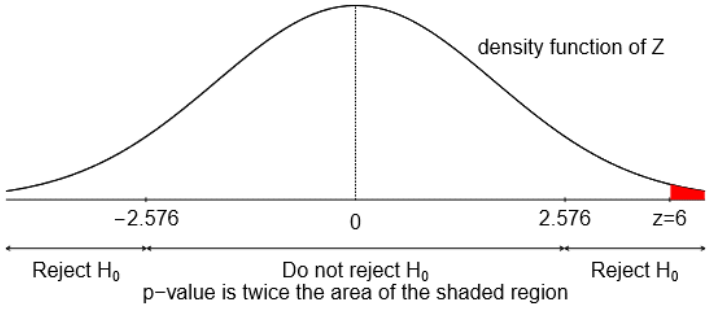
Step 5: As $z = 6$ falls in rejection region, H_0 is rejected.

Alternatively, we can use the *p*-value approach.

Note that the *p*-value is given, using a computer, as

$$2P(Z > 6) = 1.973175 \times 10^{-9},$$

which is smaller than $\alpha = 0.01$. So we reject H_0.



We can use our knowledge of the sampling distribution to determine the test statistic for other situations.

HYPOTHESIS TEST FOR THE MEAN: UNKNOWN VARIANCE
Consider the case where

- the population variance σ^2 is unknown; AND
- the underlying distribution is normal.

For the null hypothesis $H_0 : \mu = \mu_0$, the test statistics is given by

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

Let t be the observed T value. For the alternative hypothesis

- $H_1 : \mu \neq \mu_0$, the rejection region is

$$t < -t_{n-1,\alpha/2} \quad \text{or} \quad t > t_{n-1,\alpha/2}.$$

- $H_1 : \mu < \mu_0$, the rejection region is

$$t < -t_{n-1,\alpha}.$$

- $H_1 : \mu > \mu_0$, the rejection region is

$$t > t_{n-1,\alpha}.$$

REMARK
When $n \geq 30$, we can replace t_{n-1} by Z, the standard normal distribution.

L-EXAMPLE 7.2 (MIDTERM EXAM SCORE II)
Continuing from the previous example. Let's say the lecturer didn't announce the variance, that is, σ is unknown.

The data given has $\bar{x} = 22$ and $s = 2.55$.

Test again at $\alpha = 0.01$ significance level whether the average midterm score is different from 16.

Solution:
We are now in Case III since σ is unknown.

Again, we let μ be the average midterm score for the whole class.

Step 1: $H_0 : \mu = 16$ vs $H_1 : \mu \neq 16$.

Step 2: Choose $\alpha = 0.01$.

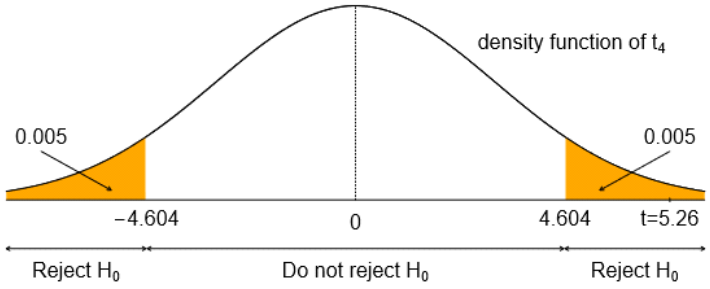
Step 2: Choose $\alpha = 0.01$.

Step 3: Since σ is unknown, data are normal, and n=5, the test statistics is

$$T = \frac{\overline{X} - 16}{S/\sqrt{n}} \sim t_{n-1} = t_4.$$

Now $t_{n-1,\alpha/2} = t_{4,0.005} = 4.604$. So the rejection region is

$$t < -4.604 \quad \text{or} \quad t > 4.604.$$



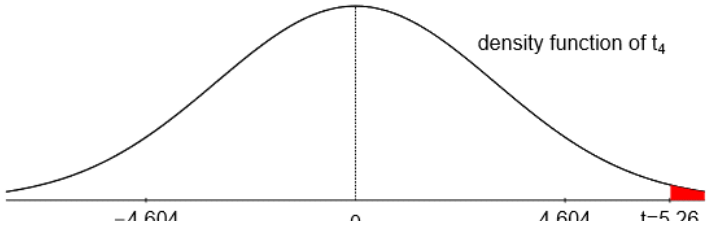
Step 4: $t = (22 - 16)/(2.55/\sqrt{5}) = 5.26$.

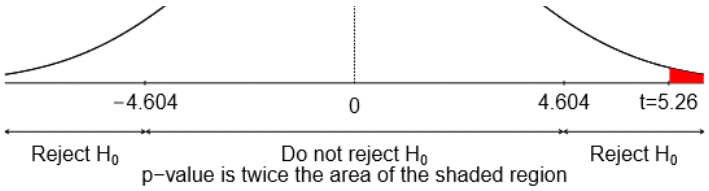
Step 5: Since $t = 5.26$ falls within the rejection region, we reject H_0.

Alternatively, the p-value can be found to be

$$2P(t_4 > 5.26) = 0.0063,$$

which is smaller than $\alpha = 0.01$, so we reject H_0.





L-EXAMPLE 7.3 (DEPARTMENT STORE)
A department store manager determines that a new billing system will be cost-effective only if the mean monthly account is more than \$170. It is known that the accounts has standard deviation \$65.

A random sample of 400 monthly accounts is drawn, for which the sample mean is \$178. Can we conclude that the new system will be cost-effective at 5% level of significance?

Solution:
Let μ be the mean monthly account.

Step 1: We test

$$H_0 : \mu = 170 \quad \text{vs} \quad H_1 : \mu > 170.$$

Step 2: Choose $\alpha = 0.05$.

Step 3: We are in Case II. So we use the following test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Under H_0, by the Central Limit Theorem, we have, $Z \sim N(0,1)$.

At a 5% significance level ($\alpha = 0.05$), we get

$$z_\alpha = z_{0.05} = 1.645.$$

Step 4: We are given that

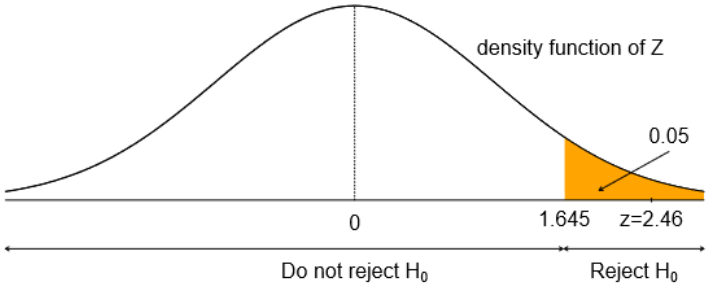
$$n = 400, \quad \bar{x} = 178, \quad \sigma = 65, \quad \alpha = 0.05$$

and so

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{178 - 170}{65/\sqrt{400}} = 2.46 > z_\alpha = 1.645.$$

Step 5: Therefore, we reject the null hypothesis and conclude that the mean monthly account is more than \$170.

Step 5: Therefore, we reject the null hypothesis and conclude that the mean monthly account is more than $170.



3 Two-sided Tests and Confidence Intervals

In this section, we establish that the two-sided hypothesis test procedure is equivalent to finding a $100(1-\alpha)\%$ confidence interval for μ.

We illustrate using Case III: normal population, small n, unknown σ.

Once again, consider

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

The $100(1-\alpha)\%$ confidence interval for μ in this case is given by

$$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right).$$

If the $100(1-\alpha)\%$ confidence interval contains μ_0, we will have

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Rearranging the above inequality, we obtain

$$-t_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq t_{\alpha/2}.$$

This means that the computed test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ satisfies

$$-t_{\alpha/2} \leq \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \leq t_{\alpha/2}.$$

This means that the computed test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ satisfies

$$-t_{\alpha/2} \leq t \leq t_{\alpha/2}.$$

Note that the rejection region for this case is

$$t < -t_{\alpha/2} \quad \text{or} \quad t > t_{\alpha/2}.$$

This means that when the confidence interval contains μ_0, H_0 will not be rejected at level α.

Similarly, when the confidence interval does not contain μ_0, then

$$t > t_{\alpha/2} \quad \text{or} \quad t < -t_{\alpha/2}.$$

Thus t falls within the rejection region and so H_0 will be rejected.

Therefore confidence intervals can be used to perform two-sided tests.

EXAMPLE 7.7 (MIDTERM EXAM SCORE III)
Back to Example 7.6, regarding midterm exam scores. Assume that the lecturer did not announce the variance, i.e., σ is unknown.

The student constructed a 99% ($\alpha = 0.01$) confidence interval for the average score of students for the midterm:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 22 \pm 4.604 \times \frac{2.55}{\sqrt{5}} = (16.75, 27.25).$$

The interval does not contain 16, so the following test of hypothesis should be rejected at $\alpha = 0.01$:

$$H_0 : \mu = 16 \quad \text{vs} \quad H_1 : \mu \neq 16.$$

What about

$$H_0 : \mu = 16 \quad \text{vs} \quad H_1 : \mu \neq 16.$$

What about

$$H_0 : \mu = 17 \quad \text{vs} \quad H_1 : \mu \neq 17?$$

L-EXAMPLE 7.4
A study based on a sample size of 36 reported a mean of 87 with a margin of error of 10 for 95% confidence.

Give the 95% confidence interval for the population mean μ.

You are then asked to test the hypothesis that $\mu = 80$ against a two sided alternative at $\alpha = 0.05$. What is your conclusion?

Solution:
The 95% confidence interval for μ is given as

$$\bar{x} \pm E = 87 \pm 10 = (77, 97).$$

The 95% confidence interval contains the the value 80 so there is no evidence to reject the null at $\alpha = 0.05$.

4 TESTS COMPARING MEANS: INDEPENDENT SAMPLES

Suppose two independent samples are drawn from two populations with means μ_1 and μ_2. We are interested in testing

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

against a suitable alternative hypothesis.

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

against a suitable alternative hypothesis.

COMPARING MEANS: INDEPENDENT SAMPLES I

(A) Consider the case where

- the population variances σ_1^2 and σ_2^2 are known; AND
- where
  - the underlying distributions are normal; OR
  - n_1, n_2 are sufficiently large (say, $n_1 \geq 30, n_2 \geq 30$).

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, the test statistics is given by

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

(B) Consider the case where

- the population variances σ_1^2 and σ_2^2 are unknown; AND
- n_1, n_2 are sufficiently large (say, $n_1 \geq 30, n_2 \geq 30$).

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, the test statistics is given by

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1).$$

The rejection regions or p-values can be established similarly as before.

REJECTION REGIONS AND p-VALUES

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, and specified alternative H_1, the rejection regions and p-values are given below.

H_1	Rejection Region	p-value
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$P(Z > \lvert z \rvert)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$P(Z < -\lvert z \rvert)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \lvert z \rvert)$

EXAMPLE 7.8
Analysis of a random sample consisting of $n_1 = 20$ specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of $\bar{x} = 29.8$ ksi.

A second random sample of $n_2 = 25$ two-side galvanized steel specimens gave a sample average strength of $\bar{y} = 34.7$ ksi.

Assuming that the two yield strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$, does the data indicate that the corresponding true average yield strengths μ_1 and μ_2 are different?
Use $\alpha = 0.01$.

Solution:
Let μ_1 and μ_2 be the mean strength of cold-rolled steel and two-side galvanized steel respectively.

Step 1: Note that $\delta_0 = 0$ in this example. So the hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

Step 2: Set $\alpha = 0.01$.

Step 3: Test statistic and its distribution is given below:

$$Z = \frac{(\overline{X} - \overline{Y}) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1).$$

Note that $z_{\alpha/2} = z_{0.005} = 2.5782$. Thus the rejection region is

$$z > 2.5782 \quad \text{or} \quad z < -2.5782.$$

Step 4: Plug in the data,

$$z = \frac{(29.8 - 34.7) - 0}{\sqrt{\frac{16}{20} + \frac{25}{25}}} = -3.652 < -2.5782 = -z_{\alpha/2}.$$

Step 5: Since $z = -3.652$ falls inside the critical region, hence $H_0 : \mu_1 = \mu_2$ is rejected at the 1% level of significance. We conclude that the sample data strongly suggest that the true average yield strength for cold-rolled steel differs from that for galvanized steel.

Alternatively, we can compute the p-value to be

$$2 \times P(Z < -3.652) = 0.00026 < 0.01 = \alpha.$$

Thus we reject the null hypothesis at $\alpha = 0.01$ level.

L-EXAMPLE 7.5 (ELECTRICAL USAGE II)
As a baseline for a study on the effects of changing electrical pricing for electricity during peak hours, July usage during peak hours was obtained for $n_1 = 45$ homes with air-conditioning and $n_2 = 55$ homes without. The summarized results are provided below

	Samples		
population	Size	Mean	Variance
With	45	204.4	13,825.3
Without	55	130.0	8,632.0

Perform a hypothesis test at $\alpha = 0.05$ that the mean on-peak usage for homes with air-conditioning is higher than that for homes without.

Solution:
Let μ_1 and μ_2 be the mean on-peak usage for homes with and without air-conditioning respectively.

Step 1: Again we have $\delta_0 = 0$. So we test

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 > 0.$$

Step 2: Set $\alpha = 0.05$.

Step 3: Test statistic and its distribution is given below:

$$Z = \frac{(\overline{X} - \overline{Y}) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1).$$

The rejection region is $z > z_{0.05} = 1.645$.

Step 4: Plug in the data,

$$z = \frac{204.4 - 130.0 - 0}{\sqrt{\frac{13,825.3}{45} + \frac{8632.0}{55}}} = 3.45.$$

Step 5: We reject H_0 since $z = 3.45 > z_{0.05} = 1.645$.

COMPARING MEANS: INDEPENDENT SAMPLES II
Consider the case where

- the population variances σ_1^2 and σ_2^2 are unknown but equal;
- the underlying distributions are normal;
- n_1, n_2 are small (say, $n_1 < 30, n_2 < 30$).

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, the test statistics is given by

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

L-EXAMPLE 7.6 (COAL SPECIMENS)
The following are measurements of heat-producing capacity (millions of calories per ton) of sample specimens of coal from two mines:

```
Mine 1: 8260 8130 8350 8070 8340
Mine 2: 7950 7890 7900 8140 7920 7840
```

The sample summary statistics are

$$\bar{x} = 8230, \quad s_1 = 125.5, \quad \bar{y} = 7940, \quad s_2 = 104.5.$$

Assume that both populations are normal with equal variance. Test at $\alpha = 0.01$ level if the means between these two mines are different.

Solution:
Let μ_1 and μ_2 be the mean heat-producing capacity for the two mines.

Step 1: $H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0.$

Step 2: $\alpha = 0.01.$

Step 3: We are given that $s_1 = 125.5, s_2 = 104.5$ and that the equal variance assumption holds. The test statistics is

$$T = \frac{(\overline{X} - \overline{Y}) - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Since $t_{n_1+n_2-2,\alpha/2} = t_{9,0.005} = 3.250$, the rejection region is

$$t < -3.250 \quad \text{or} \quad t > 3.250.$$

Step 4: Plug in everything, we get

$$\begin{aligned} s_p^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{(5-1)\times 125.5^2 + (6-1)\times 104.5^2}{5+7-2} = 13066.92, \end{aligned}$$

and

$$t = \frac{(8230-7940)-0}{\sqrt{13066.92} \times \sqrt{\frac{1}{5}+\frac{1}{6}}} = 4.18963.$$

Step 5: Since $t = 4.18963$ is in the rejection region, we reject H_0.

$\sqrt{13066.92 \times \sqrt{\frac{1}{5} + \frac{1}{6}}}$

Step 5: Since $t = 4.18963$ is in the rejection region, we reject H_0.

5 Tests Comparing Means: Paired Data

Comparing means with matched-pairs data is easy. We merely use methods we have already learned for single samples.

Comparing Means: Paired Data

For paired data, define $D_i = X_i - Y_i$.

For the null hypothesis $H_0 : \mu_D = \mu_{D_0}$, the test statistics is given by

$$T = \frac{\overline{D} - \mu_{D_0}}{S_D/\sqrt{n}}.$$

- If $n < 30$ and the population is normally distributed then
$$T \sim t_{n-1}.$$
- If $n \geq 30$, then
$$T \sim N(0,1).$$

Example 7.9 (Treating catalyst surfaces)

Prof X developed a new procedure for treating catalyst surfaces which he claims will result in a significant enhancement in the number of active sites.

The number of active sites can be determined by absorption of H_2 gas.

Prof X tested each sample before and after the treatment and obtained the following H_2 uptake in terms of mmol/g.

Sample No.	Before treatment (X)	After treatment (Y)	Difference (D)
1	165	172	7
2	146	189	43

Sample No.	Before treatment (X)	After treatment (Y)	Difference (D)
1	165	172	7
2	146	189	43
3	174	168	−6
4	186	176	−10
5	147	198	51
6	153	184	31
7	132	188	56
8	175	197	22

The summary statistics for the variable D are $\overline{d} = 24.25$ and $s_D = 25.34$.

Has the treatment resulted in an increase in the number of active sites on the catalyst surfaces? Assume normality, and test at $\alpha = 0.05$ level.

Solution:
Note that in such a setup the two samples are not independent, and so the two sample t-test does not apply.

Define $D_i = Y_i - X_i$, where X_i and Y_i are the "before treatment" and "after treatment" readings.

The question is now reduced to:

Do the data give any evidence that $\mu_D > 0$?

Step 1: We set the null and alternative to be

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D > 0.$$

Step 2: Set $\alpha = 0.05$.

Step 3: We use the paired t-test with the test statistics

$$T = \frac{\overline{D} - 0}{S_D/\sqrt{n}}.$$

The rejection region is $t > t_{7,0.05} = 1.895$.

Step 4: The observed t value is

$$t = \frac{\overline{d} - 0}{s_D/\sqrt{n}} = \frac{24.25 - 0}{25.34/\sqrt{8}} = 2.70 > 1.895.$$

Step 5: Since $t = 2.70 > t_{7,0.05} = 1.895$, we reject H_0 and conclude that there is evidence that treatment of catalysts increases the number of active sites.

Step 5: Since $t = 2.70 > t_{7,0.05} = 1.895$, we reject H_0 and conclude that there is evidence that treatment of catalysts increases the number of active sites.

As an aside, the p-value is

$$P(t_7 > t) = P(t_7 > 2.70) = 0.0153,$$

which is smaller than 0.05.

L-EXAMPLE 7.7 (WATER TREATMENT)
A state law requires municipal waste water treatment plants to monitor their discharges into rivers and streams. A treatment plant could choose to send its samples to a commercial laboratory of its choosing.

Concern over this self-monitoring led a civil engineer to design a matched pairs experiment. Exactly the same bottle of effluent cannot be sent to two different laboratories. To match "identical" as closely as possible, she would take a sample of effluent in a large sample bottle and pour it back and forth over two open specimen bottles.

When they were filled and capped, a coin was flipped to see if the one on the right was sent to commercial laboratory or the state laboratory.

This process was repeated 11 times. The results, for the response suspended solids (SS) are

Sample	1	2	3	4	5	6	7	8	9	10	11
Commercial lab	27	23	64	44	30	75	26	124	54	30	14
State lab	15	13	22	29	31	64	30	64	56	20	21
Difference $X_i - Y_i$	12	10	42	15	−1	11	−4	60	-2	10	−7

The summary statistics for $D = X_i - Y_i$ are

$$\overline{d} = 13.27, s_D^2 = 418.61.$$

Conduct a hypothesis test to check if the SS from the commercial lab is higher than those from state lab at significance level 0.05. Assume a normal distribution for the population.

Solution:
We shall test

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D > 0.$$

We shall test

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D > 0.$$

The test statistics is

$$T = \frac{\overline{D} - 0}{S_D/\sqrt{n}},$$

and the rejection region is $t > t_{10,0.05} = 1.812$.

Computations gives the observed test statistics as

$$t = \frac{\overline{d} - 0}{\sqrt{418.61/11}} = 2.15 > 1.812.$$

Since $t = 2.15 > t_{10,0.05} = 1.812$, we reject H_0 and conclude that the response from commercial lab is higher than those from the state lab.

Final-Sample-Paper-1

Sunday, November 19, 2023 11:39 PM

Final-Sample-Paper-1

NATIONAL UNIVERSITY OF SINGAPORE

DEPARTMENT OF STATISTICS AND DATA SCIENCE

ST2334 PROBABILITY AND STATISTICS

FINAL EXAMINATION SAMPLE PAPER 1

(SEMESTER I, AY 2023/2024)

TIME ALLOWED: 120 MINUTES

INSTRUCTIONS TO STUDENTS

1. Please write your student number only. **Do not write your name.**
2. This assessment contains 30 questions and comprises 33 printed pages.
3. The total marks is 60; marks are equal distributed for all questions.
4. Please answer ALL questions.
5. Calculators may be used.
6. This is an **OPEN BOOK** assessment. Only **HARD COPIES** of materials are allowed.

1 MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY

A random variable X has the following probability function.

$$f_X(x) = \frac{x}{4}, \quad \text{for } x = 0.4, 0.6, 0.9, 2.1;$$

and $f_X(x) = 0$ elsewhere.

What type of random variable is X?

(a) Continuous

(b) Discrete

(c) Unable to determine

(d) None of the given options

2 **TRUE/FALSE**

Suppose $X \sim N(1, \sigma_1^2)$ and $Y \sim N(0, \sigma_2^2)$. $P(X < 1)$ is larger than $P(Y < 1)$ when $\sigma_1^2 > \sigma_2^2 > 0$.

- TRUE
- FALSE

The probability of $N(Y<1) = 0.5$

$N($

3 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

A professor receives, on average, 21.7 emails from students the day before the midterm exam. To compute the probability of receiving at least 10 emails on such day, what type of probability distribution will he use?

(a) Binomial distribution.
(b) Poisson distribution.
(c) Normal distribution.
(d) Negative Binomial distribution.

4 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Which of the following would be an appropriate null hypothesis?

(a) The mean of a population is equal to 55. ✓
(b) The mean of a sample is equal to 55.
(c) The mean of a population is greater than 55.
(d) None of the given options

5 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let A, B be events in sample space S. Which of the following may **NOT** be true?

(a) $A \cap A' = \emptyset$ T
(b) $A \cup A' = S$. T
(c) $(A \cup B)' = A' \cup B'$ X
(d) $A \cup B = A \cup (B \cap A')$

$(A \cup B) \cap (A \cup A')$

S

6 **FILL IN THE BLANK**

Let X and Y be independent random variables such that $E(X) = 1$, $E(Y) = 2$, $V(X) = 3$, $V(Y) = 4$. Compute $V(2X - Y)$.

Answer: 16 $\quad V(X)^2 + V(Y) = 2^2 \cdot 3 + 4 = 12 + 4 = 16$

(Provide your answer in decimal form and round it to two decimal places if necessary.)

7 **FILL IN THE BLANK**

We toss a fair die until the outcome "6" appears twice. Find the probability that it takes 5 tosses.

Answer: 0.064

(Provide your answer in decimal form and round it to three decimal places if necessary.)

negative binomial

$$P(X=x) = \binom{x-1}{k-1} p^k q^{x-k}$$

$$P(X=5) \quad \binom{4}{1}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.064$$

8 **TRUE/FALSE**

Under the usual random sampling setup, we can halve the standard deviation of the sample mean by doubling the sample size.

- TRUE
- FALSE

9 **TRUE/FALSE**

Let (X, Y) be a random vector, then for any real numbers x and y, we must have

$$P(X \leq x, Y \leq y) = 1 - P(X > x, Y > y).$$

- TRUE
- FALSE

10 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Which of the following is a valid cumulative distribution function?

a) $F(x) = \begin{cases} 0 & x \le -1 \\ 0.3 & -1 < x \le 1 \\ 0.7 & 1 < x \le 10 \\ 1 & \text{elsewhere} \end{cases}$

b) $F(x) = \begin{cases} 0 & x < -1 \\ 0.5 & -1 < x < 1 \\ 0.7 & 1 < x < 10 \\ 1 & \text{elsewhere} \end{cases}$

c) $F(x) = \begin{cases} 0 & x < -1 \\ 0.6 & 1 < x \le 10 \\ 0.7 & -1 < x \le 1 \\ 1 & \text{elsewhere} \end{cases}$

d) $F(x) = \begin{cases} 0 & x < -1 \\ 0.6 & -1 \le x < 1 \\ 0.7 & 1 \le x < 10 \\ 1 & \text{elsewhere} \end{cases}$

11 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

The life time (in years) of a certain brand of light bulb follows an exponential distribution with the probability density function: $\frac{1}{2}\exp(-x/2)$. What is the probability that the bulb will last for more than 5 years, given that it has been working for 3 years?

a) $\frac{1}{2}\exp(-5/2)$

b) $\exp(-5/2)$

c) $\frac{1}{2}\exp(-1)$

d) $\exp(-1)$

$P(X>5) = e^{-\alpha \cdot 5} = e^{-5/2}$

$P(X>3) = e^{-3/2}$

$P(X>5 \mid X>3) = \frac{P(X>5)\cdot P(X>3 \mid X>5)}{P(X>3)}$

$\frac{e^{-5/2}}{e^{-3/2}} \quad -\frac{5}{2}+\frac{3}{2} = e^{-2/2} = e^{-1}$

12 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

The Central Limit Theorem is important in statistics because

a) for a large sample size, n, it says the population is approximately normal. ✗

b) for any population, it says the sampling distribution of the sample mean is approximately normal, regardless of the sample size. ✗

c) for a large sample size, n, it says the sampling distribution of the sample mean is approximately normal, regardless of the shape of the population. ✓

d) for any sized sample, it says the sampling distribution of the sample mean is approximately normal. ✗

13 **FILL IN THE BLANK**

John rolls a fair die 6 times independently. What is the probability that he will get numbers more than 2 at least twice?

$P(X>2)$ — 2 times — 4 times $<$

3 — 3

4 — 2

5 — 1

6 — 0

Answer: 0.9822

(Provide your answer in decimal form and round it to three decimal places if necessary.)

X = # of times >2

$X \sim Bin(6, 2/3)$

$P(X \ge 2) = 1 - P(X=0) - P(X=1)$

$= 0.9822$

$1 - P(X<2)$ — 0 times, 1 times $\rightarrow P(X<2)$

$P(X>2)$ 0 times $\left(\frac{2}{6}\right)^6 =$

1 times $\left(\frac{2}{6}\right)^5\left(\frac{4}{6}\right)\binom{6}{1} =$

$= 0.0178$

14 **FILL IN THE BLANK**

Jill's bowling scores are approximately normally distributed with mean 170 and standard deviation 20, while Jack's bowling scores are approximately normally distributed with mean 160 and standard deviation 15. If Jack and Jill each bowl one game, then assuming that their scores are independent random variables, the probability that the sum of the scores is higher than 340 is approximately equal to $\Phi(c)$. Find the value of c.

Answer: $c =$ −0.4

(Provide your answer in decimal form and round it to three decimal places if necessary.)

Note: $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0,1)$.

Jill = x
$\mu_x = 170$
$\sigma_x = 20$

Jack = y
$\mu_y = 160$
$\sigma_y = 15$

$P(X+Y) > 340$

$W = X+Y \sim N(330, 20^2+15^2)$

$P(W > 340) = P\left(\frac{W-330}{25} > \frac{340-330}{25}\right) = P\left(\frac{x-330}{25} > 0.4\right)$

$= P(Z < -0.4)$

$X \perp Y$

$X \sim N(170, 20^2)$

$Y \sim N(160, 15^2)$

15 **FILL IN THE BLANK**

An experiment was carried out to test whether mean weight gain for pigs fed ration A is higher than those fed ration B. Eight pairs of pigs were used. The rations were assigned at random to the two animals within each pair. The gain (in kilograms) after 45 days, assuming normally distributed, are given as follows.

Pairs	1	2	3	4	5	6	7	8	mean	sd
Ration A	30	17	18	21	22	30	24	27	23.625	5.0409
Ration B	26	18	15	20	21	25	27	23	21.875	4.1555
Difference, A - B	4	-1	3	1	1	5	-3	4	1.75	2.7646

matched pairs
t-distribution

$T^* = \frac{\bar{D} - \mu_0}{s_D/\sqrt{n}} = \frac{1.75 - 0}{2.7646/\sqrt{8}}$

Reject H_0

$\mu_0 = A - B$

Suppose that the pigs within each pair were littermates. What is the observed value of the test statistic in testing the alternative hypothesis that ration A is better, in terms of mean weight gain, than ration B at a 5% significance level?

Answer: _____ $t = 1.7904$

$H_0: \mu_0 = 0$

$H_A: \mu_0 > 0$

(Provide your answer in decimal form and round it to two decimal places if necessary.)

16 **FILL IN THE BLANK**

The mean lifetime of 100 randomly selected pumps made by a particular factory was 200 days. Assuming it is known that the population standard deviation $\sigma = 40$, find a 95% confidence interval for the mean lifetime of pumps made by the factory.

Answer: (192.16, 207.84).

Note: $z_{0.025} = 1.96$; $z_{0.05} = 1.64$.

(Provide your answers in decimal form and round them to two decimal places if necessary.)

$200 \pm z_{\frac{0.05}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

$n = 100$

$\bar{X} = 200$

$\sigma = 40$

17 **FILL IN THE BLANK**

We roll a fair die 3 times. Find the probability that the sum is equal to 5.

Answer: 0.028

(Provide your answer in decimal form and round it to three decimal places if necessary.)

1 2 2 $\binom{3}{1} = 3$

1 3 1 $\binom{3}{1} = 3$

Total ways $= 6^3 = 216$ ways

$\frac{6}{216} = 0.028$

18 **TRUE/FALSE**

Consider the Z-test for $H_0 : \mu = 0$ based on $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$. It turns out that $\bar{X} = 2.3$. The p-value for the one-sided test ($H_1 : \mu > 0$) is half of that for the two-sided test ($H_1 : \mu \neq 0$).

- TRUE
- FALSE

$p = 2P(Z > z)$ ← 2 sided
$= P(Z > z)$

19 **FILL IN THE BLANK**

A new COVID rapid test is able to correctly diagnose that you do not have the virus 90% of the time. However, if you do have the virus, it fails to detect it 25% of the time. Given that the overall COVID infection rate at a particular worker dorm is 20%, what is the probability of a worker being infected if his rapid test does not detect the virus?

Answer: 0.0649

(Provide your answer in decimal form and round it to three decimal places if necessary.)

T^+
T^-

$P(V) = 0.2$ $P(T^1 | V^1) = 0.9$

$P(V | T^1) = \frac{P(T^1|V) \cdot P(V)}{P(T^1)}$ $P(T^1 | V) = 0.25 = \frac{P(V|T^1) \cdot P(T^1)}{P(V) = 0.2}$

$P(T^1)P(V|T^1) = P(T^1|V) \cdot 0.2$

$P(V|T^1)P(T^1) = 0.05$

$\frac{1-P(V|T^1)}{\checkmark} = \frac{P(V^1|T^1) \cdot P(T^1)}{P(V^1) \Leftarrow 1-P(V)}$ $\frac{(1-P(V|T^1))P(T^1)}{(1-P(V))} = \frac{P(T^1) - P(V|T^1)P(T^1)}{1-P(V) = 0.8}$

$P(V|T^1) = 1 - \frac{0.72}{P(T^1)} = \frac{0.05}{P(T^1)}$

$1 = \frac{0.77}{P(T^1)}$

$P(T^1) = 0.77$

$P(V|T^1) = \frac{0.05}{0.77}$

20 **MULTIPLE RESPONSE: CHOOSE ALL ANSWERS THAT APPLY**

Which of the following can happen if the hypothesis is rejected.

X a) p-value> α;
✓ b) test statistic falls in the rejection region; ✓
✓ c) type I error occurs; - false rejection of H_0
d) type II error occurs. false acceptance of H_0

21 **FILL IN THE BLANK**

Let $X_1, X_2, \ldots, X_{100}$ be independent and identically distributed continuous random variables with $E(X_i) = 5$ and $V(X_i) = 4$. Compute approximately $P\left(\sum_{i=1}^{100} X_i > 510\right)$. $\sim N(500, 400)$

Answer: _____

(Provide your answer in decimal form and round it to three decimal places if necessary.)

Note: $\Phi(0.5) = 0.6915$, $\Phi(1) = 0.8413$, $\Phi(1.5) = 0.9332$; where $\Phi(\cdot)$ denotes the cumulative distribution function for $N(0,1)$.

$P(\sum_{i=1}^{100} X_i > 510)$ $P(\sum X_i > 510)$ $n \sim bin(50.5)$

$= P\left(Z > \frac{510 - 500}{\sqrt{400}}\right) = 1 - 0.6915 = \boxed{0.3085}$

22 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Which of the following statements about probability is **INCORRECT**?

a) If A and B are two events and $P(A \cap B) = P(A)$, then $P(A \cap B') = 0$.

b) Let S be the sample space and let A be an event. If there exists an $x \in S$ but $x \notin A$, then $P(A) < 1$.

c) Let A and B be two events; then $P(A \cup B) \leq P(A) + P(B)$.

d) Let A and B be independent events. If $P(A) > 0$ and $P(B) > 0$, then A and B are not mutually exclusive.

$P(A \cap B) = P(A)P(B) \neq 0$ thus but $P(A \cap B) = 0$ if mutually ex

23 **FILL IN THE BLANK**

Consider the following game:

- First round: the gamer flips a fair coin. If he gets a head, he loses; otherwise he wins the round.
- Second round: the gamer flips two fair coins independently. If he gets two heads, he loses; otherwise he wins the round.
- Third round: the gamer flips three fair coins independently. If he gets three heads, he loses; otherwise he wins the round.
- And so on.

$P(L_1) = \frac{1}{2}$, $P(L_2) = (\frac{1}{2})^2$, $P(L_3) = (\frac{1}{2})^3$, $P(W) = 1 - (\frac{1}{2})^4$

What is the probability that the gamer will make his first win in the 4th round?

Note: you can assume that from rounds to rounds, the flips are independently conducted.

Answer: 0.0146

(Provide your answer in decimal form and round it to four decimal places if necessary.)

24 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let $\{X_1, X_2, \ldots, X_{25}\}$ be a random sample from the $N(\mu, 2^2)$ distribution. Consider the hypotheses: $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$, and the test statistic $Z = \dfrac{\bar{X}}{2/\sqrt{25}}$. Suppose that we reject H_0 if $|z_{obs}| > 2$; and do not reject H_0 otherwise. What is the probability that we will not reject H_0 given that the true value of μ is equal to 2?

Note: $\Phi(z)$ denotes the c.d.f. of the standard normal distribution: $\Phi(z) = Pr(\tilde{Z} \leq z)$ with $\tilde{Z} \sim N(0,1)$.

a) $\Phi(-3) - \Phi(-7)$

b) $\Phi(2) - \Phi(-2)$

c) $\Phi(-2) - \Phi(-6)$

d) $\Phi(-3) - \Phi(-8)$

Type II error

$\left|\frac{\bar{X}}{2/\sqrt{25}}\right| > 2$

$|\bar{X}| > \frac{4}{5}$

$P(\bar{X} > \frac{4}{5}) = P(Z > \frac{\frac{4}{5} - 2}{2/5}) = P(Z > 0.4)$

$1 -$

do not reject H_0, even tho H_A true

$P(|\bar{X}| \leq 2 \mid \mu = 2) = P(-2 \leq \frac{\bar{X}}{2/5} \leq 2 \mid \mu = 2)$

$= P(-2 - \frac{2}{2/5} \leq \frac{\bar{X} - 2}{2/5} \leq 2 - \frac{2}{2/5} \mid \mu = 2)$

$= P(-7 \leq \tilde{Z} \leq -3) = \Phi(-3) - \Phi(-7)$

25 **TRUE/FALSE**

For any $\theta \in \mathbb{R}$, the function

$$f(x) = \begin{cases} \theta - x & \text{if } \theta - 1 \le x < \theta \\ x - \theta & \text{if } \theta \le x \le \theta + 1 \\ 0 & \text{elsewhere} \end{cases}$$

can serve as a probability density function of some distribution, whose population mean is equal to θ.

- TRUE
- FALSE

i. $f(x) \ge 0$ for all $x \in \mathbb{R}$

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

$\int \theta - x\,dx = \theta x - \frac{x^2}{2}\Big|_{\theta-1}^{\theta}$

$= \theta^2 - \frac{\theta^2}{2} - \theta(\theta-1) + \frac{(\theta-1)^2}{2}$

$\frac{\theta^2 - \theta^2 - 2\theta + 1}{2} - \theta^2 + \theta$

$-\frac{2\theta+1}{2} - \theta^2 + \theta = \frac{1}{2} - \theta^2$

$\int x - \theta\,dx = \frac{x^2}{2} - \theta x \Big|_{\theta}^{\theta+1}$

$\frac{\theta^2+2\theta+1}{2} - \theta(\theta+1) - \frac{\theta^2}{2} + \theta^2 = \frac{1}{2} + \frac{\theta^2}{2}$

$-\theta^2 - \theta$

$\frac{1}{2} - \theta^2 + \frac{1}{2} + \frac{\theta^2}{2} = 1 - \frac{\theta^2}{2}$

26 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

In a course, students are graded based on a "normal curve". For example, students within 0.5 standard deviation from the mean receive a C; between 0.5 and 1.0 standard deviation above the mean receive a $C+$; between 1.0 and 1.5 standard deviation above the mean receive a B; between 1.5 and 2.0 standard deviation receive a $B+$, etc. The class average in an exam was 60 with a standard deviation of 10. What are the bounds for a B grade and the percentage of students who will receive a B grade?

a) (65, 75), 24.17%
b) (65, 75), 12.08%
c) (70, 75), 18.38%
d) (70, 75), 9.19%

Note: $\Phi(1) = 0.8413$, $\Phi(1.5) = 0.9332$, $\Phi(2) = 0.9772$, where $\Phi(\cdot)$ denotes the cumulative distribution of $N(0,1)$.

$N(60, 10^2)$

X

B

$60+10$ $60+15$

70 75

$P(X<75) - P(X<70)$

$P\left(Z < \frac{75-60}{10}\right) - P\left(Z < \frac{70-60}{10}\right)$

$0.9332 - 0.8413$

27 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Flip an unfair coin. If a head shows, roll a fair die and report the number; otherwise, roll a fair die twice and report the summation minus 1. Then $P(6 \text{ is reported}) = ?$

a) 1/6
b) 1/4
c) 1/2
d) can not tell

H die result $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$

T — — sum=7 $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$

1 6
2 5
3 4
4 3
5 2
6 1

$\frac{6}{36}$

28 **FILL IN THE BLANK**

An urn contains 3 red balls and 5 white balls. 4 balls are drawn uniformly at random without replacement from the urn. Let X be the random variable for the number of red balls drawn. If $X \le 1$, you win \$$X$. If $X > 1$, you flip a fair coin. If the coin comes up heads, you double your winnings and win a total of \$$2X$. If the coin comes up tails, you still win \$$X$. Let W be the random variable for your winnings. Find the probability that W is odd.

Answer: _____

(Provide your answer in decimal form and round it to four decimal places if necessary.)

W.O.R

3R 5W

ABBB

$X \le 1$

$X \to X$

$X = 0: \left(\frac{5}{8}\cdot\frac{4}{7}\cdot\frac{3}{6}\cdot\frac{2}{5}\right)$

$X = \frac{\binom{3}{1}\binom{5}{3}}{8\cdot7\cdot6\cdot5} \cdot 1$

$X > 1$ #R > 1

coin — H → 2X, T → X

2R
3R — P(R=3) and P(T)
4R

$= \frac{\binom{3}{3}\binom{5}{1}}{8\cdot7\cdot6\cdot5} \cdot \frac{1}{2}$

29 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let X be a random variable with density function

$$f(x) = \begin{cases} k\sqrt{x} & \text{, for } 0 \le x \le 1 \\ ke^{\frac{1-x}{2}} & \text{, for } x > 1 \\ 0 & \text{, otherwise} \end{cases}$$

where k is a constant. What is the value of k?

a) 1/2
b) 2/5
c) 3/8
d) 4/9

$\int k\sqrt{x} = k\frac{2}{3}x^{3/2}\Big]_0^1 = \frac{2k}{3}$

$-2k\int_1^\infty e^{\frac{1}{2}-\frac{x}{2}} = -2k\left[e^{\frac{1}{2}-\frac{x}{2}}\right]_1^\infty$

$= -2k\left[e^{-\infty} - e^{0}\right) = 0 + 2k$

$\frac{2k}{3} + 2k = 1$

$2k\left(\frac{1}{3}+1\right) = 1$

$2k = \frac{3}{4}$

$k = \frac{3}{8}$

pooled sample variance

30 **MULTIPLE RESPONSE: CHOOSE ALL ANSWERS THAT APPLY**

Let $X_1, X_2, \ldots, X_{n_1}$ be independent and identically distributed (i.i.d.) random variables with population mean μ_1; let $Y_1, Y_2, \ldots, Y_{n_2}$ be i.i.d. random variables with population mean μ_2; let $U_1, U_2, \ldots, U_{n_3}$ be i.i.d. random variables with population mean 4. All these random variables have the unknown but common variance σ^2. Which of the following is/are unbiased estimator(s) for σ^2?

a) $\dfrac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2+\sum_{k=1}^{n_3}(U_k-4)^2}{n_1+n_2+n_3-2}$.

b) $\dfrac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2+\sum_{k=1}^{n_3}(U_k-4)^2}{n_1+n_2+n_3-3}$.

c) $\dfrac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2+\sum_{k=1}^{n_3}(U_k-\bar{U})^2}{n_1+n_2+n_3-2}$.

d) $\dfrac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2+\sum_{k=1}^{n_3}(U_k-\bar{U})^2}{n_1+n_2+n_3-3}$.

END OF PAPER

Final-Sample-Paper-2

Sunday, November 19, 2023 11:39 PM

Final-Sampl
e-Paper-2

NATIONAL UNIVERSITY OF SINGAPORE

DEPARTMENT OF STATISTICS AND DATA SCIENCE

ST2334 PROBABILITY AND STATISTICS

FINAL EXAM SAMPLE PAPER 2

(SEMESTER I, AY 2023/2024)

TIME ALLOWED: 120 MINUTES

INSTRUCTIONS TO STUDENTS

1. Please write your student number only. **Do not write your name.**
2. This assessment contains 30 questions and comprises 33 printed pages.
3. The total marks is 60; marks are equal distributed for all questions.
4. Please answer ALL questions.
5. Calculators may be used.
6. This is an **OPEN BOOK** assessment. Only **HARD COPIES** of materials are allowed.

1 **TRUE/FALSE**
Let X be a discrete random variable; then $E(X)$ always exists.

- TRUE
- FALSE

2 **TRUE/FALSE**
Let $\{X_1, X_2, \ldots, X_{1000}\}$ be a random sample from a population with expectation μ. Then both the sample mean $\bar{X}$ and X_1 are unbiased estimators for μ.

- TRUE
- FALSE

3 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Which of the following is a random sample:

a) In order to estimate the probability of getting heads for a biased coin, flip the coin 100 times and collect the results for all flips.

b) In order to study the scores of undergraduate students of Singapore, 1000 registered undergraduates in NUS were randomly sampled.

c) In order to study the average studying hours of students in NUS every week. A random survey with size 1000 was conducted at all libraries of NUS.

d) In order to study the average life time of a brand of bulbs, all bulbs' life times in LT32 over a whole semester were recorded.

4 **FILL IN THE BLANK**

Let $X_1, X_2, \ldots, X_{10}$ be independent and identically distributed random variables having the exponential distribution Exp(1). Let $T = \min\{X_1, X_2, \ldots, X_{10}\}$. Find $E(T)$.

Answer: 1/10

(Provide your answer in decimal form and round it to two decimal places if necessary.)

5 **FILL IN THE BLANK**

Suppose that X_1 is Poisson with expectation 1, X_2 is Poisson with expectation 1, and X_3 is Poisson with expectation 2 and assume that the three random variables are independent. Let $Y_1 = X_1 + X_2$ and let $Y_2 = X_2 + X_3$. The conditional probability that $Y_1 = 1$, given that $Y_2 = 2$, is equal to _____.

Answer: _____

(Provide your answer in decimal form and round it to three decimal places if necessary.)

$X_1 = \text{pois}(1)$

$X_2 = \text{pois}(1)$

$X_3 = \text{pois}(2)$

$Y_1 = X_1 + X_2 \sim \text{pois}(2)$

$Y_2 = \text{pois}(3)$ $X_2 + X_3$

$P(Y_2 = 2) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \frac{e^{-3} \cdot 3^2}{2!}$

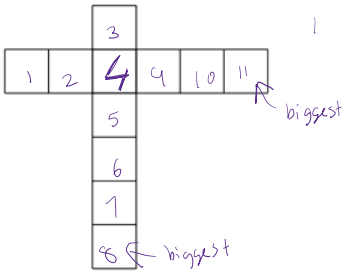
$P(Y_1 = 1) = \frac{e^{-2} \cdot 2^1}{1}$

$P(Y_1 = 1 \mid Y_2 = 2) = \frac{P(Y_1 = 1 \cap Y_2 = 2)}{P(Y_2 = 2)} = \frac{2e^{-2} + \frac{e^{-3} \cdot 9}{2} - (2e^{-2})\left(\frac{e^{3} \cdot 9}{2}\right)}{2e^{-2}}$

$P(y_2=2) = \frac{e^{-\lambda}\cdot\lambda^{x}}{x!} = \frac{e^{-3}\cdot 3^{2}}{2!}$ $\qquad P(y_1=1) = \frac{e^{-2}\cdot 2^{1}}{1}$

6 **FILL IN THE BLANK**

In how many different ways, we can put 11 different real numbers $x_1, x_2, \ldots, x_{11}$ in the boxes blow (one box each; and each number will be used once and only once), such that for any two neighbors (boxes sharing a common side), the left is always smaller than the right, and the above is always smaller than the blow?



1 2 3 4 | 5 6 7 8 9 10 11

$\binom{7}{3}\cdot 3 = 105$

Answer: 105 (Provide your answer in numerical form.)

7 **FILL IN THE BLANK**

Let A and B be independent events in the sample space S. If $P(A) = 0.4$, $P(A \cup B) = 0.7$, then $P(B) = ?$

Answer: 0.5

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

(Provide your answer in decimal form and round it to two decimal places if necessary.)

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

$0.7 = 0.4 + P(B) - P(A \cap B)$

$0.4 + P(B) - 0.4\,P(B) = 0.7$

$P(B) = 0.5$

8 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let A and B be events in the sample space, where $P(A) = 0$. Which of the following must be **TRUE**?

a) A and B are independent.

b) A and B are mutually exclusive.

c) A must be a subset of B.

d) None of the given options.

9 **TRUE/FALSE**

Let $f(x)$ be the probability function of random variable X. If $f(x) = 0$ for $x \in (0, 10)$, then $P(X \leq 0 \text{ or } X \geq 10) = 1$.

- TRUE
- FALSE

10 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let X be a random variable, such that $V(X)$ exists. Which of the following may **NOT** hold?

a) $E(3X+2X^2+e^X) = 3E(X)+2E(X^2)+E(e^X)$.

b) If $E(X)=0$, then $V(X)=E(X^2)$. $= E(X^2) = (E(X))^2$

c) $E(e^X/X) = e^{E(X)}/E(X)$.

d) $V(X+E(X)+E(e^X))+(E(X))^2 = E(X^2)$.

$= V(X) + (E(X))^2 = E(X^2) - [E(X)]^2 + [E(X)]^2 = E(X^2)$

11 **MULTIPLE RESPONSE: CHOOSE ALL ANSWERS THAT APPLY**

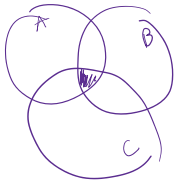
Suppose A, B, C are events in the sample space S. Which of the following may **NOT** hold?

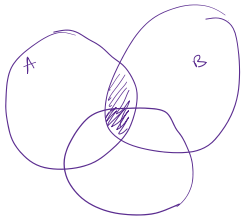
a) $(A \cup B) \cup C = A \cup (B \cup C)$

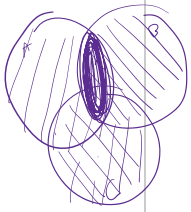
b) $(A \cap C) \cap B = (A \cap C) \cap (B \cap C)$

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

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







12 **FILL IN THE BLANK**

The number of power outrages at a power plant has a Poisson distribution with a mean of 0.06 outrages per day. What is the expected number of power outrages at this power plant in a year? Suppose that a year has 365 days.

Answer: 21.9

(Provide your answer in decimal form and round it to two decimal places if necessary.)

$\lambda = 0.06$ / day

$0.06 \cdot 365$

13 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let X be a random variable with cumulative distribution function given by

$$F(x) = \begin{cases} 0 & x < 0 \\ x/4 & 0 \le x < 1 \\ (3+x)/12 & 1 \le x < 3 \\ x/6 & 3 \le x < 6 \\ 1 & x \ge 6 \end{cases}$$

cdf

Then $P(1 \le X \le 2) = ?$

a) 1/12

b) 1/6

c) 1/4

d) None of the given options

$P(X \le 2) - P(X < 1) = \frac{5}{12} - \frac{1}{4} = \frac{2}{12} = \frac{1}{6}$

14 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Assume random variables $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. Which of the following statements must be **CORRECT**?

a) If $\sigma_X = \sigma_Y$, then $Cov(X+Y, X-Y) = 0$.

b) If $\mu_X = \mu_Y$, then $Cov(X+Y, X-Y) = 0$.

c) If X and Y are independent, then $Cov(X+Y, X-Y) = 0$.

d) If X and Y are independent, then $X+Y$ and $X-Y$ are independent.

$Cov(X+Y, X-Y) = E[(X+Y)(X-Y)] - E(X+Y)E(X-Y)$

$E(X^2) - E(Y)^2 - [E(X)]^2 + [E(Y)^2]$

$V(X) - V(Y)$

15 **FILL IN THE BLANK**

For a multiple response question with four possible choices (1), (2), (3) and (4), the number of correct answers could be 1 or 2 or 3 or 4. Only when a student answers the question exactly correct, s/he can get the mark. During the exam, suppose one particular student has no time to work on the question, and s/he decides to answer the question by independently flipping a fair coin four times: if a head shows on the ith flip, the choice (i) is included in her/his answer, otherwise, the choice is excluded; if the student gets all tails, s/he leaves the question un-answered. What is the probability that this student gets the mark?

Answer: 0.0625

(Provide your answer in decimal form and round it to four decimal places if necessary.)

1 correct $4\left(\frac{1}{2}\right)^4$ 2 correct $6\cdot\left(\frac{1}{2}\right)^4$ 3 correct $4\cdot\left(\frac{1}{2}\right)^4$ 4 correct $1\cdot\left(\frac{1}{2}\right)^4$

$1 + \binom{4}{3} + \binom{4}{2} + \binom{4}{1} = 1 + 4 + 12 + 4 =$

possible answers $= 2^4 = 16$

$1/16$

16 **FILL IN THE BLANK**

The probability function for random variable X is given by

$$f(x) = x^2/10, \quad \text{for } x = -2,-1,0,1,2,$$

and $f(x) = 0$ elsewhere. Compute the variance of X.

Answer: 3.4

(Provide your answer in decimal form and round it to two decimal places if necessary.)

−2	4/10
−1	1/10
0	0
1	1/10
2	4/10

$E(X) = 0$

$\text{Var} = E(X^2) - [E(X)]^2$ (0)

$= 3.4$

17 **FILL IN THE BLANK**

Let (X,Y) be a random vector, whose joint probability function is given by

$$f(x,y) = \begin{cases} \pi e^{-\pi(x^2+y)} & \text{for } -\infty < x < \infty; y \geq 0 \\ 0 & \text{elsewhere} \end{cases}.$$

Compute $Cov(X,Y)$.

$\pi e^{-\pi x^2 - \pi y} = \frac{\pi}{e^{\pi x^2} e^{\pi y}}$

Answer: 0

(Provide your answer in decimal form and round it to two decimal places if necessary.)

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

$f(x,y) = \pi e^{-\pi(x^2+y)}$

$f_x = \pi \int_{-\infty}^{\infty} e^{-\pi x^2 - \pi y}\, dy = -e^{-\pi(x^2+y)}$

$f_y = \pi \int_{0}^{\infty} e^{-\pi x^2 - \pi y}\, dx = \frac{-e^{-\pi(x^2-\pi y)}}{2}$

18 **FILL IN THE BLANK**

Let (X,Y) be a discrete random vector, whose joint probability function is given by:

Table of the joint probability function for (X,Y)

y \ x	0	1	2	3	4	5
0	0	0.01	0.02	0.05	0.06	0.08
1	0.01	0.03	0.04	0.05	0.05	0.07
2	0.02	0.03	0.05	0.06	0.06	0.07
3	0.02	0.04	0.03	0.04	0.06	0.05

X	Y
1	3
2	2
2	3
3	1
3	2
3	3
4	0
4	1
4	2
4	3

X	Y
5	0
5	1
5	2
6	3

Compute $P(X+Y>3)$.

Answer: 0.77

(Provide your answer in decimal form and round it to two decimal places if necessary.)

Answer: _____

(Provide your answer in decimal form and round it to two decimal places if necessary.)

19 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let X be the random variable following a normal distribution, where $E(X) = 10$, $E(X^2) = 125$. Suppose $P(-2.5 \leq X < 22.5) = 2\Phi(c) - 1$, where $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution. What is the value for c?

a) 2.5

b) 3

c) 3.5

d) Unable to tell

$N(10, 25)$

10

$E(X^2) - [E[X]]^2 = V = 125 - 100 = 25$

$P\left(\frac{-2.5-10}{5} \leq \frac{X-10}{5} \leq \frac{22.5-10}{5}\right)$

$P(-2.5 \leq Z \leq 2.5)$

$P(Z < 2.5) - P(Z < -2.5)$

$P(Z < 2.5) - P(Z > 2.5)$

$\Phi(2.5) - (1 - P(Z < 2.5))$

20 **TRUE/FALSE**

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with $E(X_1) = 10$ and $\text{var}(X_1) = 1$; denote by $\overline{X}$ the sample mean of these n random variables. Then when n is sufficiently large (approaching infinity say), we can always have

$$P\left(\left|\overline{X} - 10\right| > 1/\sqrt{n}\right) < 0.0000000000000001.$$

- TRUE
- FALSE

21 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Roll a fair die until the number 6 appears 6 times. What is the expected number of rolls needed?

(a) 36

(b) 24

(c) 6

(d) None of the given options

$\frac{6}{1/6}$

negative binomial

22 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let $A = \{1,2,4,6,8\}$, $B = \{3,4,6,8,10\}$, and $C = \{2,4,8,9\}$ be events in the sample space $S = \{1,2,3,4,6,8,9,10\}$, which of the following is **WRONG**?

(a) $A \cap B \cap C = \{4,8\}$ ✓

(b) $A \cup B \cup C = \{1,2,3,4,6,8,9,10\}$

(c) $A' \cap C' = \{3,10\}$ ✓

(d) $B' \cap C' = \{1,5,7\}$

23 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Let (X,Y) be a random vector. If $\text{var}(X+Y)=10$, $\text{var}(X)=4$, $\text{var}(Y)=8$, then $\text{var}(X-Y)=$?

a) 12

b) 13

c) 14

d) not enough information to compute

$$\text{Var}(X+Y) = E\left((X+Y)^2\right) - \left[E(X+Y)\right]^2$$

$$\text{var}(X+Y) = V(X) + V(Y) + 2\text{cov}(X,Y)$$

$$10 = 4 + 8 + 2(-1)$$

$$\text{Var}(X-Y) = 4 + 8 - 2(-1)$$

24 **TRUE/FALSE**

Let $f(x,y)$ be the probability function for the random vector (X,Y); let $f_X(x)$ be the marginal probability function for X. Then, for any real numbers x and y, we must have $f(x,y) \leq f_X(x)$.

- TRUE
- FALSE

25 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

A random sample from $N(\mu, \sigma^2)$ is given below:

$$x_1 = 7, \quad x_2 = 7, \quad x_3 = 3, \quad x_4 = 7.$$

What is the value of b, such that $(2.82, b)$ is a 95% confidence interval for μ?

Note: $t(3, 0.05) = 2.35$; $t(3, 0.025) = 3.18$; $t(4, 0.05) = 2.13$; $t(4, 0.025) = 2.78$.

a) 7.85

b) 8.11

c) 8.76

d) 9.18

6

$\bar{x} = 6$

$3.18 = \frac{6 - 2.82}{b - \sqrt{4}}$

26 **TRUE/FALSE**

Type I error may occur only when we reject the null hypothesis.

- TRUE
- FALSE

27 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Suppose that A and B are two events such that $P(A) = 0.5$, $P(A \cap B) = 0.3$. Then $P(A|B) = ?$

(a) 0.4

(b) 0.3

(c) Not sufficient information to compute

(d) None of the given options

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

$=$

$\frac{P(A \cap B)}{P(A)} = P(B|A) = \frac{0.3}{0.5} = 0.6$

$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = 0.6 \cdot 0.5$

28 **FILL IN THE BLANK**

The following are the average weekly losses of worker-hours due to accidents in three industrial plants before and after a certain safety program was put into operation.

X = avg hours

$H_0 \rightarrow X - Y = 0$

$H_A \rightarrow X - Y > 0$

	Industrial Plant		
	1	2	3
before	45	124	33
after	36	119	29

X $\bar{X} = 67.33$

Y $\bar{Y} = 61.\overline{33}$

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$= \frac{1}{2}[\quad)^2 + \quad^2] = 61$

$S^2 = 7$

Assume that the data are normally distributed.

What is the **absolute value of the computed test statistic** in testing the alternative hypothesis that the safety program reduces the average weekly losses of worker-hours due to accidents at a 5% significance level?

$t^* = \frac{6-0}{\sqrt{7}/\sqrt{3}} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$

Answer: 3.93

(Provide your answer in decimal form and round it to two decimal places if necessary.)

29 **MULTIPLE RESPONSE: CHOOSE ALL ANSWERS THAT APPLY**

Suppose X and Y are independent random variables. Which of the following must be **CORRECT**?

a) X^2 and Y^{1000} are independent. ✓

b) $X+Y$ and $X-Y$ are independent. ✗

c) The events $\{X \leq 10\}$ and $\{Y > 10\}$ are independent. ✓

d) The events $\{X \leq 10\}$ and $\{Y > 10\}$ are mutually exclusive. ✗

30 **MULTIPLE CHOICE: CHOOSE ONE ANSWER ONLY**

Assume that the survival time (in years) of a patient who has a certain cancer follows the exponential distribution with average survival time equal to 4 (years). If such a patient has survived for 4 years, what is the probability that this patient can survive for another 4 years?

(a) e^{-1}

(b) $1-e^{-1}$

(c) e^{-2}

(d) $1-e^{-2}$

$\alpha = 1/4$

$E(X) = 4 = \frac{1}{\alpha}$

$P(X>8 \mid X>4) = \frac{P(X>4 \mid X>8)\,P(X>8)}{P(X>4)} = \frac{1/e^2}{1/e} = \frac{1}{e}$

$P(X>8) = e^{-1/4 \cdot 8} = \frac{1}{e^2}$

$P(X>4) = e^{-1} = \frac{1}{e}$

END OF PAPER