

# Statistics for the Sciences - Part 1

Lecture notes by Demetrio Labate

April 5, 2023

# What is statistics?

**Statistics** is a discipline focused on the collection and analysis of data.

Its objective: to solve **quantitative** problems in the presence of uncertainty due to the variability of data.

Its method: the main tool for the study of statistics is the **mathematical theory of probability**.

## Historical note

Historically, probability originated with calculations related to games of chance.

G. Cardano, about 1560: *Liber de ludo aleae* = Book about games of dice.

The modern **mathematical theory of probability** started with A. N. Kolmogorov, about 1933. He laid out an axiomatic approach which is the basis for the current theory.

One of the cornerstone of his approach is the notion of *idealized thought experiment*. For example, one can think of a *coin tossing experiment* as an idealized experiment that can be repeated indefinitely and that admits two equally likely outcomes.

Applications of probability are very pervasive: opinion polls, stock market models, weather forecast, Mendel's theory of heredity, queues in communications, robot navigation, analysis of noise in signal processing, models of radioactive decay, spread of infectious diseases, reliability theory, neural networks, ...

# Review of set theory

**Definition:** A **set** is a collection of objects.

It can be defined by **enumeration**

$$A = \{1, 7, 10\}$$

$$B = \{\text{Adam, Jim, Paula}\}$$

or can be defined by **description**

$$E = \{\text{The students enrolled in MA3339, Sec.2}\}$$

Each objects in a set is an **element of the set**.

$$a \in A = \text{"a belongs to A"} = \text{"a is an element of A"}$$

Two sets  $A$  and  $B$  are **equal** if and only if they contain the same elements.



# Review of set theory

## Definitions:

$\emptyset$  is the empty set, that is, the set that contains no elements.

$\mathbb{R}$  is the set of real numbers

$\mathbb{N}$  is the set of natural numbers

**Definition:** A **subset**  $B$  of a set  $A$ , denoted as  $B \subset A$ , is a set such that any element of  $B$  is also an element of  $A$ .

Note: if  $B \subset A$  and  $A \subset B$ , it follows that  $A = B$ .

**Definition:** The **union** of two sets  $A$  and  $B$  is the set

$$A \cup B = \{x: x \in A \text{ or } x \in B\}$$

**Definition:** The **intersection** of two sets  $A$  and  $B$  is the set

$$A \cap B = \{x: x \in A \text{ and } x \in B\}$$

## Review of set theory

**Definition:** Let  $S$  be a set and  $A \subset S$ . The **complement of  $A$**  in  $S$ , denoted as  $A^c$  (or  $\overline{A}$ ) is the set of all elements in  $S$  that are not in  $A$ , that is

$$A^c = \{x: x \in S \text{ and } x \notin A\}$$

**Definition:** The difference of two sets  $A \setminus B$  is the set of all elements in  $A$  that are not in  $B$ , that is

$$A \setminus B = A \cap B^c$$

# Review of set theory

**Properties:** Let  $S$  be a set and  $A \subset S$ .

①  $(A^c)^c = A$

②  $A \cup A^c = S, \quad A \cap A^c = \emptyset$

③  $A \cup \emptyset = A, \quad A \cap \emptyset = \emptyset$

④  $A \cup S = S, \quad A \cap S = A$

# Review of set theory

**Properties:** Let  $A$  and  $B$  be sets.

- ①  $A \cup B = B \cup A$  (commutative property)
- ②  $A \cap B = B \cap A$  (commutative property)
- ③  $A \cup B = (A^c \cap B^c)^c$  (De Morgan's law)
- ④  $A \cap B = (A^c \cup B^c)^c$  (De Morgan's law)

**Definition:** Two sets  $A$  and  $B$  are said to **disjoint** if  $A \cap B = \emptyset$ .

# 1. Probability Theory

# Axiomatic theory of probability

We will be concerned with (thought) experiments whose outcome is subject to uncertainty.

E.g., tossing a coin, rolling a die, picking a card from deck of cards.

**Definition:** The **sample space**  $S$  is the set of all possible outcomes of an experiment.

An **event** is any collection of outcomes/events in  $S$ .

Events  $A$  and  $B$  in  $S$  such that  $A \cap B = \emptyset$  are called **mutually exclusive** or **disjoint**.

# Axiomatic theory of probability

**Definition:** Given a random experiment and a sample space  $S$ , we assign to any event  $A$  a number  $P(A)$  called **probability of  $A$**  according to the following axioms:

- ① For any  $A \in S$ ,  $P(A) \geq 0$ ;
- ②  $P(S) = 1$ ;
- ③ if  $A_1, \dots, A_n$  are disjoint events, then
$$P(A_1 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i).$$

It follows that  $P(\emptyset) = 0$ . (Exercise: Why?)

It follows that  $P(A) \leq 1$  for any event  $A$ . (Exercise: Why?)

# Axiomatic theory of probability

**Proposition.** Given a random experiment and a sample space  $S$ , let  $A$  and  $B$  be two events.

- ①  $P(A^c) = 1 - P(A)$
- ②  $P(A) \leq 1$  for any event  $A$ .
- ③  $P(A \setminus B) = P(A \cup B) - P(B)$ .
- ④  $P(A \setminus B) = P(A) - P(A \cap B)$ .
- ⑤  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

**Proof.**

3.  $(A \setminus B) \cup B = A \cup B$  where left-hand-side (LHS) is disjoint.

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

4.  $(A \setminus B) \cup (A \cap B) = A$  where LHS is disjoint.

Hence  $P(A \setminus B) + P(A \cap B) = P(A)$

5.  $(A \setminus B) \cup B = A \cup B$  where LHS is disjoint.

Hence  $P(A \setminus B) + P(B) = P(A \cup B)$ .

By 4.,  $P(A \setminus B) + P(B) = P(A) - P(A \cap B) + P(B) = P(A \cup B)$



## Axiomatic theory of probability - Exercise

**Exercise:** In Springfield, MO, 60% of households receive internet access through ACME company, 80% of households receive TV access through the same company and 50% of households receive both TV and internet through ACME company.

- (a) If a household is randomly selected, what is the probability it receives at least one of the two services through ACME company?
- (b) If a household is randomly selected, what is the probability it receives exactly one of the two services through ACME company?

**Solution:** We start by *reformulating the problem in terms of probability calculations*.

$A$  = household receives internet access through ACME company;

$B$  = household receives TV access through ACME company;

Hence, we have that  $P(A) = 0.6$ ,  $P(B) = 0.8$  and  $P(A \cap B) = 0.5$ .

Question (a) asks to find  $P(A \cup B)$ .

Question (b) asks to find  $P((A \setminus B) \cup (B \setminus A))$ .

## Axiomatic theory of probability - Equally likely events

Let  $S$  be a sample space containing a finite number of simple events  $A_1, \dots, A_n$ .

We assume that events  $A_1, \dots, A_n$  are **mutually exclusive** and **exhaustive**, that is

$$\bigcup_{i=1}^n A_i = S \quad \text{and} \quad A_i \cap A_j = \emptyset, \quad \forall i \neq j.$$

This implies that

$$1 = P(S) = \sum_{i=1}^n P(A_i) = P(A_1) + \dots + P(A_n)$$

We also assume that events  $A_1, \dots, A_n$  are **equally likely**, that is

$$P(A_i) = p \quad \forall i = 1, \dots, n.$$

This implies that

$$1 = \sum_{i=1}^n P(A_i) = np \quad \Rightarrow \quad p = \frac{1}{n}$$

# Axiomatic theory of probability - Equally likely events

In conclusion, if the events  $A_1, \dots, A_n$  are **mutually exclusive**, **exhaustive** and equally likely, then

$$P(A_i) = \frac{1}{n}, \quad \forall i = 1, \dots, n.$$

If the event  $E = \{A_1 \cup A_2 \cup \dots \cup A_k\}$  (the union of  $k$  elementary events  $A_i$ ), then

$$P(E) = P(A_1 \cup A_2 \cup \dots \cup A_k) = \sum_{i=1}^k P(A_i) = \frac{k}{n}.$$

## Equally likely events - Examples

Several classical games of chance, e.g., coin tossing, rolling dice, card games, can be modeled using equally likely events as described above.

**Example:** Consider the experiment of rolling a (fair) die. We want to compute the probability of the following events:

$E_1 = A_2 =$  outcome is 2

$E_2 = (A_2 \cup A_4 \cup A_6) =$  outcome is an even number

**Solution:**

$$P(E_1) = P(A_2) = \frac{1}{6}$$

$$P(E_2) = P(A_2 \cup A_4 \cup A_6) = \frac{3}{6} = \frac{1}{2}$$

## Equally likely events - Card example

Consider a deck of 52 playing cards (we assume cards in the deck are shuffled).

Note: A standard deck of playing cards consists of 52 cards. All cards are divided into 4 suits. There are two black suits - spades (♠) and clubs (♣) and two red suits — hearts (♥) and diamonds (♦).

In each suit there are 13 cards including a 2, 3, 4, 5, 6, 7, 8, 9, 10, a jack, a queen, a king and an ace A.

Our experiment consists in drawing a card from the deck.

## Equally likely events - Card example

We want to compute the probability of the following events:

$E_1$  = card is a king;  $E_2$  = card is a spade (♠)

$$P(E_1) = \frac{4}{52} \quad (\text{there are 4 kings in the deck})$$

$$P(E_2) = \frac{13}{52} \quad (\text{there are 13 spades in the deck})$$

$$P(E_1 \cap E_2) = \frac{1}{52} \quad (\text{there is 1 king of spade in the deck})$$

$$P(\text{not a spade}) = P(E_2^c) = 1 - \frac{13}{52} = \frac{39}{52}$$

$$P(\text{a king or a spade}) = P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{16}{52}$$

# Counting rules

If all  $n$  outcomes of a random experiment are equally likely, as we examined above, the probability of an event  $E$  is given by

$$P(E) = \frac{|E|}{n}$$

where  $|E|$  denotes the number of elements in  $E$ .

To apply this rule, we need to be able to count the number of elements in events. We shall look at:

- multiplication rules
- permutations
- combinations

# Counting rules - multiplication

Let us consider an ordered pair  $(a, b)$ .

If we can choose  $a$  in  $n_1$  ways and  $b$  in  $n_2$  ways then *the number of possible pairs is  $n_1 n_2$* .

**Example.** A survey consists of two multiple choice questions. The first question has 3 possible answers and the second has 4 possible answers. What is the total number of different ways in which this survey could be completed? Solution:  $3 \times 4 = 12$

We proceed similarly with an ordered  $m$ -tuple  $(a_1, \dots, a_m)$  where we can choose  $a_1$  in  $n_1$  ways,  $\dots$ ,  $a_m$  in  $n_m$  ways.

**Example.** A circuit board contains 4 relays where each of the first 2 can be set to any of 4 positions and each of the last 2 can be set to any of 3 positions. What is the total number of distinct configurations for the 4 relays? Solution:  $4 \times 4 \times 3 \times 3 = 144$



# Counting rules - permutations and combinations

Permutations and combinations describe the various ways in which objects from a set may be selected, generally without replacement, to form subsets. This selection of subsets is called a **permutation** when the order of selection is a factor, a **combination** when order is not a factor.

# Counting rules - permutations

**How many different arrangements (permutations) of  $n$  distinct objects are possible?**

The first object can be chosen in  $n$  ways;

the second object can then be chosen in  $n - 1$  ways and so on.

The number of ways of arranging (in order)  $n$  distinct objects is

$$n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1 = n! \quad (\text{n factorial})$$

**Example.** 6 horses run a race. The total number of possible results of this race (assuming no ties) is  $6! = (6)(5)(4)(3)(2)(1) = 720$ .

**Example.** The total number of different ways in which the letters of the word “sprint” can be arranged is  $6! = (6)(5)(4)(3)(2)(1) = 720$ . (Note: it is important here that all letters are different.)

## Counting rules - permutations

**How many different arrangements (permutations) of  $n$  distinct objects, taken  $k$  at a time, are possible?**

The first object can be chosen in  $n$  ways;

the second object can then be chosen in  $n - 1$  ways and so on.

When choosing the  $k$ -th object, we have already chosen  $k - 1$  objects, so there are still  $n - (k - 1) = n - k + 1$  possible choices.

Hence the number of size- $k$  arrangements ( $k$ -permutations) of a set of  $n$  distinct objects is

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

**Example.** 8 horses run a race. How many different possibilities are there for who finishes first, second and third? The number of 3-permutations is a set of size 8 is  $\frac{8!}{(8-3)!} = (8)(7)(6) = 336$

## Counting rules - permutations

### What if not all the objects are distinct?

What is the total number of different arrangements of the letters in the word *pill*?

Suppose the two *l* can be distinguished  $pi l_1 l_2$ . Then we would have  $4!$  arrangements.

Each arrangement of the original word *pill* would generate  $2!$  arrangements of  $pi l_1 l_2$

So the number of arrangements of the word *pill* is  $\frac{4!}{2!} = 12$ .

In general if we have  $n$  items  $k$  of which are identical, the total number of distinct permutations is  $\frac{n!}{k!}$

**Example.** How many different ways can we rearrange the letters of MISSISSIPPI?

There are 11 letters of which 4 are 'I', 4 are 'S' and 2 are 'P'. In this situation, the total number of different rearrangements is  $\frac{11!}{4!4!2!}$

# Counting rules - combinations

**How many different ways can we select a set of size  $k$  from a larger set of  $n$  distinct objects?** Here the order of selection does not matter.

We know that there are  $P_k^n$  (ordered) arrangements of  $n$  distinct objects of size  $k$ . Since each combination of  $k$  objects can be permuted in  $k!$  ways then the number of **combinations** of  $n$  objects taken  $k$  at a time is

$$C_k^n = \frac{P_k^n}{k!} = \frac{n!}{(n-k)!k!} := \binom{n}{k} \quad (\text{binomial coefficient})$$

Note:  $\binom{n}{k} = \binom{n}{n-k}$

## Counting rules - combinations

**Example.** In how many ways can a subcommittee of 5 be chosen from a panel of 20 individuals?

$$\binom{20}{5} = \frac{(20)(19)(18)(17)(16)}{(5)(4)(3)(2)(1)} = 15504$$

**Example.** In the powerball game, first a player selects 5 out of the first 55 positive integers and next a second number - the powerball - out of the first 42 integers.

What is the probability of hitting the powerball, that is, of guessing correctly all 6 numbers? What is the probability of matching the first 5 numbers but not the last one?

The size of the samples space is  $\binom{55}{5} \cdot \binom{42}{1}$ .

$$P(\text{powerball}) = \frac{1}{\binom{55}{5} \cdot \binom{42}{1}} = 6.884 \cdot 10^{-9}$$

$$P(\text{first 5 numbers}) = \frac{41}{\binom{55}{5} \cdot \binom{42}{1}} = 2.806 \cdot 10^{-7}$$

# Counting rules - combinations

## Example.

A hand of 5 cards is dealt from a well-shuffled 52-card deck. What is the probability that the hand contains:

- 1 no aces?
- 2 5 clubs?
- 3 at least 1 club
- 4 at least 1 ace?

**Solution.** Size of sample space is  $\binom{52}{5} = \frac{52!}{47!5!} = 2,598,960$ .

$$P(\text{no aces}) = \frac{\binom{48}{5}}{\binom{52}{5}}$$

$$P(5 \text{ clubs}) = \frac{\binom{13}{5}}{\binom{52}{5}}$$

$$P(\text{at least 1 club}) = 1 - P(\text{no clubs}) = 1 - \frac{\binom{39}{5}}{\binom{52}{5}}$$

$$P(\text{at least 1 ace}) = 1 - P(\text{no aces}) = 1 - \frac{\binom{48}{5}}{\binom{52}{5}}$$

# Counting rules

## Example.

A hand of 5 cards is dealt from a well-shuffled 52-card deck. What is the probability that the hand contains:

- 1 3 clubs and 2 hearts?
- 2 2 kings, 2 queens and 1 jack?

## Solution.

$$P(3 \text{ clubs and } 2 \text{ hearts}) = \frac{\binom{13}{3}\binom{13}{2}}{\binom{52}{5}}$$

$$P(2 \text{ kings, } 2 \text{ queens and } 1 \text{ jack}) = \frac{\binom{4}{2}\binom{4}{2}\binom{4}{1}}{\binom{52}{5}}$$



# Conditional probability

**Definition:** Given a random experiment and a sample space  $S$ , Let  $A$  and  $B$  be two events. The **conditional probability** of  $A$  given  $B$  is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0$$

It follows that

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

and

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

*the probability of  $A$  and  $B$  is the probability of  $A$  multiplied by the probability of  $B$  given  $A$ .*

# Conditional probability - Example

## Example.

A bowl contains 10 chips: 5 red, 3 white, 2 blue.  
We randomly draw a chip from the bowl.

Let us consider the following events:

$E_1$  = chip is red or blue

$E_2$  = chip is red or white.

$$P(E_1) = \frac{7}{10}, \quad P(E_2) = \frac{8}{10}$$

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{5/10}{8/10} = \frac{5}{8}.$$

Interpretation: the condition  $E_2$  has the effect of reducing the size of the sample space. If we know that the chip is red or white, then the sample space only includes 8 chips, not 10.

## Conditional probability - Example

### Example.

A bowl contains 10 chips: 6 white, 4 blue.

We randomly draw 2 chips from the bowl in succession (and without replacement).

What is the probability that both chips are white?

**Solution 1.** We count the number of ordered arrangements of 2 chips (where we think of every chip as a distinct object).

There are  $C_2^{10}$  possible ways to draw 2 chips.

There are  $C_2^6$  possible ways to draw 2 white chips.

Hence

$$P(2 \text{ white chips}) = \frac{\binom{6}{2}}{\binom{10}{2}} = \frac{30}{90} = \frac{1}{3}$$

## Conditional probability - Example

**Solution 2.** We consider the events

$E_1$  = first chip is white.

$E_2$  = second chip is white.

Using conditional probability, we write

$$P(E_1 \cap E_2) = P(E_1)P(E_2|E_1) = \frac{6}{10} \frac{5}{9} = \frac{30}{90} = \frac{1}{3}.$$

$P(E_2|E_1) = \frac{5}{9}$  since only 5 white chips and 9 chips in total are left after the first draw.

# Conditional probability

The definition of conditional probability is consistent with the multiplication rule.

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$$

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1 \cap A_2) P(A_3|A_1 \cap A_2) \\ &= P(A_1) P(A_2|A_1) P(A_3|A_1 \cap A_2) \end{aligned}$$

(Same idea can be extended to more events)

## Conditional probability - Example

**Example.** 3 cards are dealt from a 52-deck of playing cards successively and without replacement.

What is the probability that all cards are spades?

**Solution.** We consider the events:

$E_1$  = first card is spade.

$E_2$  = second card is spade.

$E_3$  = third card is spade.

Using conditional probability, we write

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) P(E_2|E_1) P(E_3|E_1 \cap E_2)$$

Hence

$$P(E_1 \cap E_2 \cap E_3) = \frac{13}{52} \frac{12}{51} \frac{11}{50}.$$

**Alternative solution.** Using combinations we have

$$P(E_1 \cap E_2 \cap E_3) = \frac{\binom{13}{3}}{\binom{52}{3}}.$$

## Conditional probability - Example

**Example.** 3 cards are dealt from a 52-deck of playing cards successively and without replacement.

What is the probability that the first card is a spade, the second is a spade and the third is not a spade?

**Solution.** We consider the events:

$E_1$  = first card is spade.

$E_2$  = second card is spade.

$E_3$  = third card is not a spade.

Using conditional probability, we write

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) P(E_2|E_1) P(E_3|E_1 \cap E_2)$$

Hence

$$P(E_1 \cap E_2 \cap E_3) = \frac{13}{52} \frac{12}{51} \frac{39}{50}.$$

## Conditional probability - Example

**Example.** 5 cards are dealt from a 52-deck of playing cards successively and without replacement.

What is the probability that the 3rd spade occurs in the 5th draw?

**Solution.** We consider the events:

$E_1$  = first 4 cards include 2 spades.

$E_2$  = fifth card is a spade.

Using conditional probability, we write

$$P(E_1 \cap E_2) = P(E_1)P(E_2|E_1) = \frac{\binom{13}{2}\binom{39}{2}}{\binom{52}{4}} \frac{11}{48}.$$

$P(E_2|E_1) = \frac{11}{48}$  since at the fifth draw there are only 11 spades and 48 cards in total left in the deck.



# Conditional probability - Independence

**Definition:** Two events  $A$  and  $B$  are **independent** if

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

Otherwise  $A$  and  $B$  are **dependent**.

It follows that, if  $A$  and  $B$  are independent, then

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

and similarly

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

# Independence - Example

**Example.** We toss a quarter, a nickel and a dime.

What is the probability that we get 3 heads?

**Solution.** We consider the events:

$E_1$  = head on the quarter;

$E_2$  = head on the nickel;

$E_3$  = head on the dime;

Note that  $P(E_1) = P(E_2) = P(E_3) = \frac{1}{2}$

Since the three events are independent,

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) P(E_2) P(E_3) = \left(\frac{1}{2}\right)^3$$

# Conditional probability - Bayes' theorem

## Theorem

Suppose that a random experiment results in  $k$  mutually exclusive and exhaustive events  $A_1, \dots, A_k$  with probabilities  $P(A_1), \dots, P(A_k)$ .

Suppose there is another event  $B$  for which the conditional probabilities  $P(B|A_1), \dots, P(B|A_k)$  are given. Then, for any  $i = 1, \dots, k$ , we have

$$P(A_i|B) = \frac{P(B|A_i) P(A_i)}{\sum_{j=1}^k P(B|A_j) P(A_j)}$$

In the language of Bayes' theorem, the initial probabilities  $P(A_1), \dots, P(A_k)$  are called *the prior probabilities* and the computed probabilities  $P(A_1|B), \dots, P(A_k|B)$  are called *the posterior probabilities*.

## Bayes' theorem - Example

**Example.** Three separate plants,  $M_1$ ,  $M_2$  and  $M_3$ , produce 45%, 30% and 25%, respectively, of the total parts produced in a factory. The percentages of defective production of these machines are 3%, 4% and 5%, respectively.

- a) If we choose a part randomly, what is the probability that it is defective?
- b) Suppose now that we choose a part randomly and it is defective. What is the probability that it was produced by  $M_2$ .

**Solution.** We denote as  $E_i$  the event that a part is produced by plant  $M_i$ ,  $i = 1, \dots, 3$  and by  $D$  the event that a part is defective. The word problem gives:

$$P(E_1) = 0.45, P(E_2) = 0.30, P(E_3) = 0.25$$

and

$$P(D|E_1) = 0.03, P(D|E_2) = 0.04, P(D|E_3) = 0.05$$

## Bayes' theorem - Example

a) If we choose a part randomly, what is the probability that it is defective?

$$\begin{aligned}P(D) &= P(D \cap E_1) + P(D \cap E_2) + P(D \cap E_3) \\&= P(D|E_1)P(E_1) + P(D|E_2)P(E_2) + P(D|E_3)P(E_3) \\&= (0.03)(0.45) + (0.04)(0.30) + (0.05)(0.25) = 0.038.\end{aligned}$$

b) If a randomly chosen part is defective, what is the probability that it was produced by  $M_2$ ?

By Bayes' theorem

$$\begin{aligned}P(E_2|D) &= \frac{P(D|E_2) P(E_2)}{\sum_{j=1}^3 P(D|E_j) P(E_j)} \\&= \frac{P(D|E_2) P(E_2)}{P(D)} = \frac{(0.04)(0.30)}{0.038} = 0.316\end{aligned}$$

## Bayes' theorem - Case $k = 2$

When the Bayes' theorem is applied to the situation where there are only  $k = 2$  mutually exclusive and exhaustive events  $A, A^c$ , then the formula of the theorem can be simplified.

Given an event  $B$  for which the conditional probabilities  $P(B|A), P(B|A^c)$  are given, then we have

$$P(A|B) = \frac{P(B|A) P(A)}{P(B|A) P(A) + P(B|A^c) P(A^c)}$$

$$P(A^c|B) = \frac{P(B|A^c) P(A^c)}{P(B|A) P(A) + P(B|A^c) P(A^c)}$$

## Bayes' theorem - Example ( $n = 2$ )

**Example.** Two similar bat species,  $A_1$  and  $A_2$ , occupy both highland ( $B$ ) and lowland ( $B^c$ ) areas. Species  $A_1$  makes up 90% of the population; species  $A_2$ , 10%. We know that 80% of species  $A_1$  live in the lowlands while 60% of species  $A_2$  live in the highlands. What is the probability that a randomly caught bat belongs to each species, if it is caught in the highlands?

**Solution.** Based on the word problem,  $P(A_1) = 0.9$ ,  $P(A_2) = 0.1$ . We have that  $P(B^c|A_1) = 0.8$  and  $P(B|A_2) = 0.6$ .

For the solution, we need  $P(B|A_1) = 1 - P(B^c|A_1) = 0.2$ .

By Bayes' theorem, the posterior probabilities are

$$P(A_1|B) = \frac{P(B|A_1) P(A_1)}{P(B|A_1) P(A_1) + P(B|A_2) P(A_2)} = \frac{(0.2)(0.9)}{(0.2)(0.9) + (0.6)(0.1)} = 0.75$$

$$P(A_2|B) = \frac{P(B|A_2) P(A_2)}{P(B|A_1) P(A_1) + P(B|A_2) P(A_2)} = \frac{(0.1)(0.6)}{(0.2)(0.9) + (0.6)(0.1)} = 0.25$$

# Bayes' theorem - Screening Tests and Disease Diagnosis

Clinical tests are frequently used in medicine and epidemiology to diagnose or screen for the presence ( $T^+$ ) or absence ( $T^-$ ) of a particular condition, such as pregnancy or disease. Different measures of the test's merit can then be estimated via various conditional probabilities.

**Sensitivity** or **True Positive rate**  $= P(T^+ | D^+)$

**Specificity** or **True Negative rate**  $= P(T^- | D^-)$

**False Positive rate** or **Fall-out**  $= P(T^+ | D^-)$

**False Negative rate** or **Miss rate**  $= P(T^- | D^+)$

Definitive disease status (either  $D^+$  or  $D^-$ ) is often subsequently determined by means of a *gold standard*, typically resulting from follow-up, invasive surgical or radiographic procedures or from autopsy.



# Bayes' theorem - Screening Tests and Disease Diagnosis

In order to apply such a screening test to the general population, we need accurate estimates of its **predictive values** of a positive and negative test,  $PV^+ = P(D^+ | T^+)$  and  $PV^- = P(D^- | T^-)$ , respectively, which are calculated via Bayes's theorem as

$$PV^+ = P(D^+ | T^+) = \frac{P(T^+ | D^+) P(D^+)}{P(T^+ | D^+) P(D^+) + P(T^+ | D^-) P(D^-)}$$

$$PV^- = P(D^- | T^-) = \frac{P(T^- | D^-) P(D^-)}{P(T^- | D^-) P(D^-) + P(T^- | D^+) P(D^+)}$$

To compute such formulas, we need the prior probabilities  $P(D^+)$  and  $P(D^-) = 1 - P(D^+)$ .

$P(D^+)$  is called the **prevalence** of the disease in the population and is estimate (usually grossly overestimated) by the corresponding sample-based value.

## Bayes' theorem - Example

**Example.** A patient exhibits symptoms that make the physician concerned that he/she may have a particular disease. The disease has a **prevalence** of 2%, (i.e., the disease affects 2% of the population). The screening test has a reported **sensitivity** of 85%, that is, the probability of screening positive, given the presence of disease is 85%, and a **specificity** of 95%, that is, the probability of screening negative, given the absence of disease is 95%. What is the probability that the patient is sick if the test returns positive? What is the probability that the patient is healthy if the test returns negative?

## Bayes' theorem - Example

**Solution.** From the word problem we have  $P(D^+) = 0.02$ , hence  $P(D^-) = 1 - P(D^+) = 0.98$ .

We also have  $P(T^+|D^+) = 0.85$ ,  $P(T^-|D^-) = 0.95$ , hence we derive  $P(T^+|D^-) = 1 - P(T^-|D^-) = 0.05$  and  $P(T^-|D^+) = 1 - P(T^+|D^+) = 0.15$

Hence

$$\begin{aligned} P(D^+|T^+) &= \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} \\ &= \frac{(0.85)(0.02)}{(0.85)(0.02) + (0.05)(0.98)} = 0.258 \end{aligned}$$

$$\begin{aligned} P(D^-|T^-) &= \frac{P(T^-|D^-)P(D^-)}{P(T^-|D^-)P(D^-) + P(T^-|D^+)P(D^+)} \\ &= \frac{(0.95)(0.98)}{(0.95)(0.98) + (0.15)(0.02)} = 0.997 \end{aligned}$$

# Bayes' theorem - Example

## Discussion of solution.

A positive test result increases the probability of having this disease from 2% (prior probability) to 25.8% (posterior probability); a negative test result increases the probability of not having the disease from 98% (prior probability) to 99.7% (posterior probability).

Hence, this test is extremely specific for the disease (i.e., low false positive rate,  $P(T^+|D^-) = 0.05$ ) but it is not very sensitive to its presence (i.e., high false negative rate,  $P(T^-|D^+) = 0.15$ ).

A physician may wish to use a screening test with higher sensitivity (i.e., low false negative rate). However, such tests also sometimes have low specificity (i.e., high false positive rate), e.g., MRI screening for breast cancer. An ideal test generally has both high sensitivity and high specificity but a test satisfying both conditions is often expensive.

## 2. Random Variables

# Random variables

**Definition:** Any measurements that are outcomes of a random experiment are **random variables**.

The name indicates that the outcomes of the random experiment cannot be deterministically predicted.

A random variable is a **discrete random variable** if it has a countable number of possible values.

Example: counting the number of eggs  $N$  that a hen lays in a given day.

A random variable is a **continuous random variable** if it has an uncountable number of possible values.

Example: measuring the time  $T$  for a task to be completed.

# Random variables

Next, I will present the general theory of **discrete random variables**.

Following that, I will present two important classes of discrete random variables:

- 1 Binomial random variables
- 2 Poisson random variables

## 2.1 Discrete random variables



# Random variables

**Example.** We consider the random experiment of tossing an (unbiased) coin three times. Let  $X$  be the number of times we obtain an head (H).

Sample space  $S = \{0, 1, 2, 3\}$ .

Probabilities:

$$P(X = 0) = P(\text{'TTT'}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^3$$

$$P(X = 1) = P(\text{'HTT'} \text{ or } \text{'THT'} \text{ or } \text{'TTH'}) = 3 \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 3 \left(\frac{1}{2}\right)^3$$

$$P(X = 2) = P(\text{'HHT'} \text{ or } \text{'THH'} \text{ or } \text{'HTH'}) = 3 \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 3 \left(\frac{1}{2}\right)^3$$

$$P(X = 3) = P(\text{'HHH'}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^3$$

In general,

$$p(x) = P(X = x) = \binom{3}{x} \left(\frac{1}{2}\right)^3$$

describes the values of the probability of  $X$  for any element in its range  $\{0, 1, 2, 3\}$ .

$p(x)$  is the **probability mass function (pmf)** of  $X$ .

# Random variables - pmf

**Definition:** Let  $X$  be a discrete random variables with values in  $R$  (the range of  $X$ ).  $p(x)$  is the **probability mass function (pmf)** of  $X$  if

①  $p(x) \geq 0 \quad \forall x \in R$

②  $\sum_{x \in R} p(x) = 1$

In this case  $p(x) = P(X = x)$  is the probability that  $X = x$

## Random variables - pmf

**Example.** We consider the random experiment of tossing an (unbiased) coin until head (H) appears. Let  $X$  be the number of trials needed to obtain a head (H).

Sample space  $S = \{1, 2, 3, \dots\} = \mathbb{N}$ .

Probabilities:

$$P(X = 1) = P(\text{'H'}) = \frac{1}{2}$$

$$P(X = 2) = P(\text{'TH'}) = \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^2$$

$$P(X = x) = P(\text{'T...TH'}) = \left(\frac{1}{2}\right)^{x-1} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^x$$

Hence,

$$p(x) = P(X = x) = \left(\frac{1}{2}\right)^x$$

Note that  $\sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1$ .

## Random variables - pmf and cdf

**Definition:** Let  $X$  be a discrete random variables with pmf  $p(x)$ . The **cumulative distribution function (cdf)** of  $X$  is

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

It follows that  $F(x)$  is a piece-wise constant increasing function with values in  $[0, 1]$ .

For  $x_i, x_j \in R$  and  $x_i < x_j$ , we have that

$$P(x_i \leq X \leq x_j) = F(x_j) - F(x_{i-1})$$

In fact

$$P(x_i \leq X \leq x_j) = \sum_{x=x_i}^{x_j} p(x) = \sum_{x \leq x_j} p(x) - \sum_{x \leq x_{i-1}} p(x)$$

## Random variables - pmf and cdf

**Example.** We consider the random experiment of tossing an (unbiased) coin until head (H) appears. Let  $X$  be the number of trials needed to obtain a head (H).

Sample space  $S = \{1, 2, 3, \dots\} = \mathbb{N}$ .

We found that in this case  $p(x) = P(X = x) = \left(\frac{1}{2}\right)^x$

The cdf in this case is

$$F(x) = P(X \leq x) = \sum_{i=1}^x p(i) = \sum_{i=1}^x \left(\frac{1}{2}\right)^i = \frac{\frac{1}{2} - \left(\frac{1}{2}\right)^{x+1}}{\frac{1}{2}} = 1 - \left(\frac{1}{2}\right)^x$$

**Example (continue).** We can use the cdf to compute probabilities over intervals:

$$P(3 \leq X \leq 5) = F(5) - F(2) = 1 - \left(\frac{1}{2}\right)^5 - 1 + \left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^5.$$

$$P(X \geq 2) = 1 - P(X < 2) = 1 - F(1) = 1 - \frac{1}{2} = \frac{1}{2}.$$

# Random variables - Expectation

**Definition:** Let  $X$  be a discrete random variables with pmf  $p(x)$  and range  $R$ . The **expected value** - or **mean** - of  $X$  is

$$\mu_X = E[X] = \sum_{x \in R} xp(x)$$

If  $u$  is a function on  $R$ , the **expected value** of  $u(X)$  is

$$E[u(X)] = \sum_{x \in R} u(x)p(x)$$

The expected value of  $X$  is a measure of the central tendency of the r.v.  $X$ .

## Random variables - Expectation

**Example.** Let  $X$  be the number of times we obtain an head (H) in 3 independent tosses of an (unbiased) coin.

- (a) Find the expected value of  $X$
- (b) Suppose we set up a game awarding  $X^2$  dollars to each person flipping a coin 3 times. If the game will be played "many" times, how much should each player chip in to play the game (so that there will be enough time to pay the awards)?

### Solution

Recall that  $p(x) = P(X = x) = \binom{3}{x} \left(\frac{1}{2}\right)^3$ , with  $X = 0, 1, 2, 3$ .

$$E[X] = \sum_{x=0}^3 x \binom{3}{x} \left(\frac{1}{2}\right)^3 = 0 + 1 \cdot 3 \left(\frac{1}{2}\right)^3 + 2 \cdot 3 \left(\frac{1}{2}\right)^3 + 3 \cdot 1 \left(\frac{1}{2}\right)^3 = \frac{3}{2}$$

$$E[X^2] = \sum_{x=0}^3 x^2 \binom{3}{x} \left(\frac{1}{2}\right)^3 = 0 + 1 \cdot 3 \left(\frac{1}{2}\right)^3 + 4 \cdot 3 \left(\frac{1}{2}\right)^3 + 9 \cdot 1 \left(\frac{1}{2}\right)^3 = 3$$

# Random variables - Expectation

**Proposition** Let  $X$  be a discrete random variables with pmf  $p(x)$  and range  $R$ .

- 1 If  $c$  is a constant,  $E[c] = c$ .
- 2 If  $c_1, \dots, c_n$  are constants, then
$$E\left[\sum_{i=1}^n c_i u_i(X)\right] = \sum_{i=1}^n c_i E[u_i(X)].$$

**Example.** Let  $X$  be the number of times we obtain an head (H) in 3 independent tosses of an (unbiased) coin. Suppose we set up a game awarding  $3X + 2X^2$  dollars to each person flipping a coin 3 times. If the game will be played "many" times, how much should each player chip in to play the game?

**Solution.** Using the computation from example above,

$$E[3X + 2X^2] = 3 E[X] + 2 E[X^2] = 3 \cdot \frac{3}{2} + 2 \cdot 3 = \frac{21}{2}$$



# Random variables - Expectation

**Definition:** Let  $X$  be a discrete random variables with pmf  $p(x)$  and range  $R$ . The **variance** of  $X$  is

$$\sigma_X^2 = \text{var}(X) = E[(X - \mu_X)^2] = \sum_{x \in R} (x - \mu_X)^2 p(x)$$

The **standard deviation** of  $X$  is

$$\sigma_X = \sqrt{\text{var}(X)} = \sqrt{E[(X - \mu_X)^2]} = \sqrt{\sum_{x \in R} (x - \mu_X)^2 p(x)}$$

The variance of  $X$  is a measure of the variability of the r.v.  $X$  around its mean.

# Random variables - Expectation

**Proposition** Let  $X$  be a discrete random variables.

- ①  $\sigma_x^2 \geq 0, \sigma_x \geq 0$
- ②  $\sigma_X^2 = E[X^2] - \mu_X^2$

**Proposition** Let  $X$  be a discrete random variables and  $Y = a_0 + a_1 X$  where  $a_1, a_2$  are constants.

- ①  $\mu_Y = a_0 + a_1 \mu_X$
- ②  $\sigma_Y^2 = a_1^2 \sigma_X^2$

**Example** Let  $Y = -2 - X$ , where  $\mu_X = -1, \sigma_X^2 = 0.5$ .  
Then  $\mu_Y = -2 - (-1) = -1$  and  $\sigma_Y^2 = (-1)^2(0.5) = 0.5$ .

## Random variables - Expectation

**Example.** Let  $X$  be a discrete random variable with pmf given by

$x$	1	2	3	4
$p(x)$	0.4	0.2	0.3	0.1

Compute mean and variance of  $X$

**Solution.**

$$\mu_X = E[X] = \sum_{x=1}^4 x p(x) = 1(0.4) + 2(0.2) + 3(0.3) + 4(0.1) = 2.1$$

$$E[X^2] = \sum_{x=1}^4 x^2 p(x) = 1(0.4) + 4(0.2) + 9(0.3) + 16(0.1) = 5.5$$

$$\sigma_X^2 = E[X^2] - \mu_X^2 = 5.5 - (2.1)^2 = 1.09$$

# A note about R and RStudio

# R software

R is a free software environment for statistical computing and graphics.

<https://www.r-project.org/>

It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

R provides a wide variety of statistical techniques such as linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering as well as graphical techniques, and is very extensible.

# R Studio

To run R, I recommend to use RStudio

https:

[//rstudio.com/products/rstudio/download/#download](https://rstudio.com/products/rstudio/download/#download)

RStudio is a set of integrated tools designed to help you write and execute scripts in R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

You can run RStudio on a wide variety of UNIX platforms, Windows and MacOS.

## R examples

Once you have R environment setup, you can start your R command prompt by just typing a command.

This will launch R interpreter and R will execute the command.

Example: Create and print a vector.

```
apple <- c('red','green',"yellow")  
print(apple)
```

Example: Create and print a matrix.

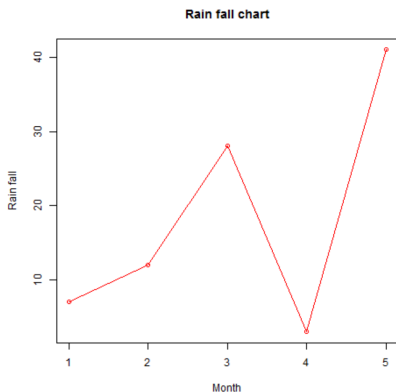
```
M = matrix( c('a','a','b','c','b','a'), nrow = 2,  
ncol = 3, byrow = TRUE)  
print(M)
```

## R examples

Example: Create a plot from a vector

```
v <- c(7,12,28,3,41)
```

```
plot(v,type = "o", col = "red", xlab = "Month", ylab =  
= "Rain fall", main = "Rain fall chart")
```





## R examples

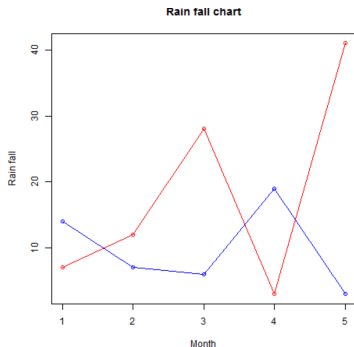
More than one line can be drawn on the same chart by using the `lines()` function. After the first line is plotted, the `lines()` function can be used to draw a second line.

```
v <- c(7,12,28,3,41)
```

```
t <- c(14,7,6,19,3)
```

```
plot(v,type = "o",col = "red", xlab = "Month", ylab =  
"Rain fall", main = "Rain fall chart")
```

```
lines(t, type = "o", col = "blue")
```



## R examples

You can use R to compute the mean, standard deviation and other statistical functions.

- `mean()` computes the mean
- `median()` computes the median
- `var()` computes the sample variance  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- `sd()` computes the sample standard deviation  $= \sqrt{s^2}$

There is no direct command to compute the variance.

To compute the variance of a vector `y` you can do as follows:

```
n=length(y); var(y)*(n-1)/n
```

## R examples

Many **tutorials on R** are freely available online such as

<https://www.tutorialspoint.com/r/index.htm>

In the first assignment, you will be asked to plot the graphs of the pmf and cdf for some given distributions. You can handle those with the command `plot`.

Here are some tutorials specifically focused on **producing graphs in R**:

<https://sites.harding.edu/fmccown/r/>

<http://www.sthda.com/english/wiki/creating-and-saving-graphs-r-base-graphs>

## 2.2 The binomial distribution

## Bernoulli trials

**Definition:** Any (discrete) random variable whose only possible values are 0 and 1 is a **Bernoulli random variable**.

**Bernoulli trial.** A Bernoulli trial is a random experiment with exactly two possible outcomes, 'success' and 'failure', in which (i) the outcome of the trials are mutually independent and (ii) the probability of success  $p$  is the same every time the experiment is conducted.

Let  $X$  be a Bernoulli random variable. We associate the random variable  $X$  to a Bernoulli trial where

$$X = 1 = \text{success}, \quad X = 0 = \text{failure}$$

$$P(X = 1) = p, \quad P(X = 0) = 1 - p = q$$

Hence, the pmf of  $X$  is

$$f(x) = P(X = x) = p^x(1 - p)^{1-x}, \quad x = 0, 1$$

# Bernoulli trials

For a Bernoulli r.v.  $X$  with  $f(x) = p^x(1-p)^{1-x}$ ,  $x = 0, 1$ ,

$$\mu_X = E[X] = \sum_{x=0}^1 xf(x) = f(1) = p$$

$$\sigma_X^2 = E[X^2] - \mu_X^2 = \sum_{x=0}^1 x^2 f(x) - p^2 = p - p^2 = p(1-p)$$

Since the outcomes of  $n$  Bernoulli trials are mutually independent, given  $n$  Bernoulli random variables  $X_1, \dots, X_n$ , then

$$P(X_1 = x_1, \dots, X_n = x_n) = f(x_1) \cdots f(x_n) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}$$

# Bernoulli trials

In practical applications, the probability  $p$  of success in a Bernoulli trial need to be estimated.

Since the mean of a Bernoulli random variable is  $\mu_X = p$ , we use the **estimator**

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

where  $X_i$  is a Bernoulli trial.

Since  $E[X_i] = p$ , then

$$E[\hat{p}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = p$$

This shows that  $\hat{p}$  is an **unbiased estimator** of  $p$ .

# The binomial distribution

**Definition:** A **binomial random variable** is a random variable of the form  $Y = \sum_{i=1}^n X_i$  where any  $X_i$  is a Bernoulli random variable with probability of success  $p$ .

The binomial random variable  $Y$  counts the number of successes over  $n$  Bernoulli trials.

The range of  $Y$  is  $R = \{0, 1, \dots, n\}$ .

To compute  $P(Y = y)$ , we observe that the event  $Y = y$  happens when, over  $n$  trials, we have  $y$  successes and  $n - y$  failures. Any such event has probability  $p^y(1 - p)^{n-y}$ . Observing that there are  $\binom{n}{y}$  possible sequences with  $y$  successes and  $n - y$  failures, then

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}$$



# The binomial distribution

Therefore, the pmf of  $Y$ , called the **binomial pmf** with parameters  $n, p$ , is the function

$$f(y) = P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y = 0, 1, \dots, n$$

To say that  $Y$  is a *binomial random variable with parameters  $n, p$* , we use the notation  $Y \sim b(n, p)$

By the classical *binomial theorem* from linear algebra:

$$\sum_{y=0}^n f(y) = \sum_{y=0}^n \binom{n}{y} p^y (1 - p)^{n-y} = 1$$

# The binomial distribution

**Mean** and **variance** of the binomial pmf are as follows:

$$\mu_Y = \sum_{y=0}^n y f(y) = \sum_{y=0}^n y \binom{n}{y} p^y (1-p)^{n-y} = np$$

$$\sigma_Y^2 = \sum_{y=0}^n y^2 \binom{n}{y} p^y (1-p)^{n-y} - \mu_Y^2 = np(1-p)$$

The **cdf of the binomial distribution** is:

$$F(y) = P(Y \leq y) = \sum_{z \leq y} f(z) = \sum_{z \leq y} \binom{n}{z} p^z (1-p)^{n-z}$$

# The binomial distribution

According to the observations made above, since the mean of a binomial random variable  $Y$  is  $\mu_Y = np$ , we define the **estimator** of  $p$

$$\hat{p} = \frac{Y}{n}$$

and we have that  $\hat{p}$  is an **unbiased estimator** of  $p$  since

$$E[\hat{p}] = E\left[\frac{Y}{n}\right] = \frac{1}{n}E[T] = p$$

We also have that

$$\begin{aligned} \text{var}(\hat{p}) &= E[(\hat{p} - p)^2] = E\left[\left(\frac{Y}{n} - p\right)^2\right] \\ &= \frac{1}{n^2}E[(Y - np)^2] \\ &= \frac{1}{n^2}np(1-p) \\ &= \frac{p(1-p)}{n} \end{aligned}$$

# The binomial distribution

## Example

Suppose that a certain medical procedure is known to have a 70% successful recovery rate (assuming independence). In a random sample of  $n = 5$  patients, (a) what is the probability that three or fewer patients will recover? (b) what is the probability that more than 3 patients will recover? (c) what is the mean number of patients expected to recover?

**Solution** Set  $Y$  to be the number of patients that recover.  $Y$  is a binomial random variable  $Y \sim \text{bin}(5, 0.7)$

- (a)  $P(Y \leq 3) = F(3)$   
 $= P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) = 0.4718$
- (b)  $P(Y > 3) = 1 - P(Y \leq 3) = 1 - F(3) = 0.5282$
- (c)  $\mu_Y = np = (5)(0.7) = 3.5$

# The binomial distribution

## Example: R solution

$Y$  is a binomial random variable  $Y \sim \text{bin}(5, 0.7)$

(a)  $P(Y \leq 3)$

```
> pbinom(3,5,0.7)
```

```
[1] 0.47178
```

(b)  $P(Y > 3) = 1 - P(Y \leq 3)$

```
> 1-pbinom(3,5,0.7)
```

```
[1] 0.52822
```

# The negative binomial distribution

**Definition:** A **negative binomial random variable** is a random variable  $Z$  that counts the number of independent Bernoulli trials needed to observe  $k$  successes (each with probability  $p$ ).

The range of  $Z$  is  $R = \{k, k + 1, \dots\}$ .

$P(Z = z)$  can be expressed as the probability of getting  $k - 1$  successes in  $z - 1$  trials times the probability  $p$  of getting a success at the  $z$ th trial. Since the probability of getting  $k - 1$  successes in  $z - 1$  trials is  $\binom{z-1}{k-1} p^{k-1} (1-p)^{z-k}$ , then

$$h(z) = P(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k}$$

$h(z)$  is the **negative binomial distribution** with parameters  $k, p$ .

When  $k = 1$ ,  $h(z)$  is called the **geometric distribution**.

# The negative binomial distribution

**Mean** and **variance** of the negative binomial pmf are as follows:

$$\mu_Z = \sum_{z=k}^{\infty} z \binom{z-1}{k-1} p^k (1-p)^{z-k} = \frac{k}{p}$$

$$\sigma_Z^2 = \sum_{z=k}^{\infty} z^2 \binom{z-1}{k-1} p^k (1-p)^{z-k} - \mu_Z^2 = \frac{k(1-p)}{p^2}$$

**Remark.** The negative binomial distribution is sometimes defined in terms of the random variable  $Y$  = number of failures before  $k$ -th success. This formulation is statistically equivalent to the one given above in terms of  $Z$  = trial at which the  $k$ -th success occurs, since  $Y = Z - k$ . The alternative form of the negative binomial distribution is

$$P(Y = y) = \binom{k+y-1}{y} p^k (1-p)^y, \quad y = 0, 1, \dots$$

# The negative binomial distribution

## Example

Suppose that a certain production process is repeated until the first defective part is produced. Assuming that the probability of producing a defective part is  $p = 0.05$ , (a) what is the probability that the first defective part occurs at the 5-th trial? (b) what is the probability that the first defective part occurs at or before the 5-th trial? (c) when do you expect to see the first defective part?

**Solution** Set  $Z$  to be the number of trials needed for the production line to produce the first defective part.  $Z$  is a negative binomial random variable with  $k = 1$ ,  $Z \sim \text{nb}(1, 0.05)$

$$(a) \quad P(Z = 5) = \binom{4}{0} (0.05)(0.95)^4 = 0.0407$$

$$(b) \quad P(Z \leq 5) = \sum_{z=1}^5 \binom{z-1}{0} (0.05)(0.95)^{z-1} = 0.2262$$

$$(c) \quad \mu_Z = \frac{1}{p} = 20$$



# The negative binomial distribution

**Example:** R solution

$$Z \sim \text{nb}(1, 0.05)$$

$Y = \text{number of failures before } k\text{-th success} = Z - 1$

(a)  $P(Z = 5) = P(Y = 4)$

```
> dnbinom(4, 1, 0.05)
```

```
[1] 0.04072531
```

(b)  $P(Z \leq 5) = P(Y \leq 4)$

```
> pnbinom(4, 1, 0.05)
```

```
[1] 0.2262191
```

# The hypergeometric distribution

**Definition:** A **hypergeometric experiment** is a statistical experiment that has the following properties:

- 1 A sample of size  $n$  is randomly selected without replacement from a population of  $N$  items.
- 2 In the population,  $N_1$  items can be classified as successes and  $N_2 = N - N_1$  items can be classified as failures.

A **hypergeometric random variable** is the number of successes  $W$  that result from a hypergeometric experiment.

The range of  $W$  is  $R = \{1, 2, \dots, N_1\}$ .

The hypergeometric pmf and the corresponding mean and variance are:

$$f(w) = P(W = w) = \frac{\binom{N_1}{w} \binom{N-N_1}{n-w}}{\binom{N}{n}}$$

$$E[W] = n \frac{N_1}{N}, \quad \text{var}(W) = \frac{n \frac{N_1}{N} (1 - \frac{N_1}{N}) (N - n)}{N - 1}$$

# The hypergeometric distribution

**Example** An animal population in certain region consists of 25 individuals in total. To monitor such populations, 5 animals are caught, tagged and the released. After a certain time, 10 such animals are caught; of those animal,  $X$  are found to be tagged. (a) What is the probability that  $X = 2$ ? (b) What is the probability that  $X \leq 2$ ? (c) What is the expected number of tagged animals?

**Solution**  $X$  is a hypergeometric random variable with  $n = 10$ ,  $N_1 = 5$ ,  $N = 25$ .

$$(a) \quad P(X = 2) = \frac{\binom{5}{2} \binom{20}{8}}{\binom{25}{10}} = 0.385$$

$$(b) \quad P(X \leq 2) = \sum_{x=0}^2 \frac{\binom{5}{x} \binom{20}{10-x}}{\binom{25}{10}} = 0.699$$

$$(c) \quad \mu_X = n \frac{N_1}{N} = 10 \frac{5}{25} = 2$$

# The hypergeometric distribution

## Example: R solution

$X$  is a hypergeometric random variable with  $n = 10$ ,  $N_1 = 5$ ,  $N = 25$ . Thus  $N_2 = N - N_1 = 20$ .  
 $X \sim \text{hypergeom}(5, 20, 10)$ .

(a)  $P(X = 2)$

```
> dhyper(2,5,20,10,log=FALSE)
[1] 0.3853755
```

(b)  $P(X \leq 2)$

```
> phyper(2,5,20,10,log=FALSE)
[1] 0.6988142
```

## Hypergeometric vs binomial distribution

As compared to the binomial distribution where the probability of success  $p$  is constant, in a hypergeometric experiment the probability of success  $\frac{N_1}{N}$  changes at every trial since the values  $N$  and possibly  $N_1$  are being decreased at each trial.

However, if  $N, N_1 \gg n$  (in practice,  $N, N_1$  about an order of magnitude larger than  $n$ ), then the probability of success  $\frac{N_1}{N}$  will remain approximately constant during each trial of the hypergeometric experiment.

In this case, the hypergeometric distribution with parameters  $N, N_1, n$  is well approximated by a binomial distribution with parameters  $p = \frac{N_1}{N}$  and  $n$ .

Note that, in this case,

$$E[W] = n \frac{N_1}{N} \approx np$$

$$\text{var}(W) = \frac{n \frac{N_1}{N} (1 - \frac{N_1}{N}) (N - n)}{N - 1} = np(1 - p) \frac{N - n}{N - 1} \approx np(1 - p)$$

# The hypergeometric distribution

**Example** An animal population in certain region consists of 1000 individuals. To monitor such populations, 100 animals are caught, tagged and the released. After a certain time, 20 such animals are caught; of those animal,  $X$  are found to be tagged. What is the probability that  $X \leq 2$ ?

**Exact solution.**  $X$  is a hypergeometric random variable with  $n = 20$ ,  $N_1 = 100$ ,  $N = 1000$ .

$$P(X \leq 2) = \sum_{x=0}^2 \frac{\binom{100}{x} \binom{900}{20-x}}{\binom{1000}{20}} = 0.6772$$

Using R: `phyper(2,100,900,20,log=FALSE)` = 0.677224

**Approximate solution.** We approximate  $X$  as a binomial random variable with  $n = 20$ ,  $p = \frac{100}{1000} = 0.1$ .

$$P(X \leq 2) \approx \sum_{x=0}^2 \binom{20}{x} (0.1)^x (0.9)^{20-x} = 0.6769$$

# A note about R and RStudio

# Computing probabilities in R

The following commands in R can be used to compute probabilities associated with various distributions.

- `dbinom(x, n, p)`:  $P(X = x)$  for  $X \sim \text{bin}(n, p)$
- `pbinom(q, n, p)`:  $P(X \leq q)$  for  $X \sim \text{bin}(n, p)$
- `dnbinom(x, n, p)`:  $P(X = x)$  for  $X \sim \text{Nbin}(n, p)$
- `pnbinom(q, n, p)`:  $P(X \leq q)$  for  $X \sim \text{Nbin}(n, p)$
- `dhyper(x, N1, N2, n, log = FALSE)`:  $P(X = x)$  for  $X \sim \text{hyper}(N_1, N_2, n)$
- `phyper(q, N1, N2, n, log = FALSE)`:  $P(X \leq q)$  for  $X \sim \text{hyper}(N_1, N_2, n)$
- `dpois(x, lambda)`:  $P(X = x)$  for  $X \sim \text{Poisson}(\lambda)$
- `ppois(q, lambda)`:  $P(X \leq q)$  for  $X \sim \text{Poisson}(\lambda)$

Note that the negative binomial rv in R counts the number of failures that occur before getting the desired success.



## 2.3 The Poisson distribution

# The Poisson distribution

The Poisson distribution is associated with counting the number of occurrences of a (rare) event in a given interval of time or space.

**Definition:** A **Poisson process** is a processes generating a certain number  $X$  of occurrences of an event  $E$  over a fixed interval in time or space of size  $T$  that satisfies the following properties:

- ① all the occurrences of  $E$  are independent in the interval;
- ② the expected number of occurrences of  $E$  in the interval is proportional to  $T$ , i.e.,  $\mu = \alpha T$ . This constant of proportionality is the **rate of the Poisson process**.

In this case, the probability of obtaining any specified number  $x$  of occurrences of the event  $E$  is given by the **Poisson distribution**

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

where the range of  $X$  is  $R = \{0, 1, \dots\}$ .

We say that  $X$  is Poisson random variable with parameter  $\mu$ .

Notation:  $X \sim \text{Poisson}(\mu)$

# The Poisson distribution

We have that

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} \frac{e^{-\mu} \mu^x}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!} = e^{-\mu} e^{\mu} = 1$$

Using the above observation,

$$\mu_X = E[X] = \sum_{x=0}^{\infty} x \frac{e^{-\mu} \mu^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\mu} \mu^x}{(x-1)!} = \mu \sum_{z=0}^{\infty} \frac{e^{-\mu} \mu^z}{z!} = \mu$$

With a similar computation, we show  $E[X^2] = \mu^2 + \mu$ , hence

$$\text{var}(X) = E[X^2] - \mu_X^2 = \mu$$

# The Poisson distribution

**Example** The number of fatal traffic accidents reported per week in a certain county was estimated to be equal to 7. What is the probability that the number of accidents in a given week is larger or equal than 10?

**Solution.** We can model the number  $X$  of fatal accidents per week as a Poisson distribution,  $X \sim \text{Poisson}(7)$ . Hence

$$P(X \geq 10) = 1 - P(X \leq 9) = 1 - \sum_{x=0}^9 \frac{e^{-7} 7^x}{x!} = 1 - 0.830 = 0.170$$

R solution:

```
> 1-ppois(9,7)
[1] 0.1695041
```

## Poisson approximation to the Binomial Distribution

When  $n$  is large and  $p$  is small, the binomial distribution is well approximated by a Poisson distribution with  $\mu = np$ .

**Example** A certain medical condition  $E$  affects 1% of the population. Let  $X$  = number of affected individuals in a random sample of size  $n = 300$ . What is the probability that 3 individuals are affected by the disease?

**Exact solution** We model  $X$  as  $X \sim \text{bin}(n = 300, p = 0.01)$ . Hence

$$P(X = 3) = \binom{300}{3} (0.01)^3 (0.99)^{300-3} = 0.22517$$

**Approximate solution** We approximate  $X$  using a Poisson distribution where  $\mu = pn = (0.01)(300) = 3$ . Note that 3 is the mean number of expected occurrences of  $E$  in the sample. Hence

$$P(X = 3) = \frac{e^{-\mu} \mu^3}{3!} = \frac{e^{-3} 3^3}{3!} = 0.22404$$

## Discrete random variables - Review problems

**Problem.** A die is thrown until the number 6 occurs for the first time. (a) What is the probability that the number 6 occurs for the first time in the 3rd throw? (b) What is the probability that it takes at most 3 throws for the number 6 to occur for the first time? (c) What is the expected number of throws for the number 6 to occur for the first time?

## Discrete random variables - Review problems

**Problem.** A die is thrown until the number 6 occurs for the first time. (a) What is the probability that the number 6 occurs for the first time in the 3rd throw? (b) What is the probability that it takes at most 3 throws for the number 6 to occur for the first time? (c) What is the expected number of throws for the number 6 to occur for the first time?

**Solution.**

$X$ , counting the number of throws until the number 6 occurs for the first time, is a geometric random variable with  $p = 1/6$

In R,  $W = X - 1$  counts the number of failures before the number 6 occurs for the first time.

$$(a) P(X = 3) = P(W = 2) = \text{dnbinom}(2, 1, 1/6) = 0.1157407$$

$$(b) P(X = 3) = P(W \leq 2) = \text{pnbinom}(2, 1, 1/6) = 0.4212963$$

$$(c) E(X) = 1/p = 0.1666667$$

# Discrete random variables - Review problems

**Problem.** In a certain town, it is known that 6% of the population is color-blind. If a random sample of 50 people is drawn from this population, what is the probability that (a) at least 3 people are color-blind? (b) at most 3 people are color blind?



## Discrete random variables - Review problems

**Problem.** In a certain town, it is known that 6% of the population is color-blind. If a random sample of 50 people is drawn from this population, what is the probability that (a) at least 3 people are color-blind? (b) at most 3 people are color blind?

**Solution.**

$X$ , counting the number of color-blind people in the population, is a binomial random variable with  $p = 0.06$  and  $n = 50$

(a)

$$P(X \geq 3) = 1 - P(x \leq 2) = 1 - \text{pbinom}(2, 50, 0.06) = 0.5837535.$$

$$(b) P(X \leq 3) = \text{pbinom}(3, 50, 0.06) = 0.6473034.$$

## Discrete random variables - Review problems

**Problem.** In a study of the effectiveness of an insecticide, a large area was sprayed. Later the area was examined by randomly selecting squares of the same size and counting the number of live insects per square. Past experience has shown that the average number of live insects per square after spraying to be 1.2. What is the probability that a selected square will contain (a) no live insects? (b) at most 2 live insects?

## Discrete random variables - Review problems

**Problem.** In a study of the effectiveness of an insecticide, a large area was sprayed. Later the area was examined by randomly selecting squares of the same size and counting the number of live insects per square. Past experience has shown that the average number of live insects per square after spraying to be 1.2. What is the probability that a selected square will contain (a) no live insects? (b) at most 2 live insects?

**Solution.**

$X$ , counting the number of live insects per square, is a Poisson random variable with  $\mu = 1.2$

$$(a) P(X = 0) = \text{dpois}(0, 1.2) = 0.3011942.$$

$$(b) P(X \leq 2) = \text{ppois}(2, 1.2) = 0.8794871.$$

## 2.4 Multivariate distributions

# Multivariate distributions

In many situations, we are interested in more than one aspect of a random experiment.

In this case, we are interested in the probability of a combination of events, that is, several random variable are involved.

Examples:

price of crude oil (per barrel) and price per gallon of unleaded gasoline at your local station (per gallon);

grades of college students across multiples disciplines;

level of multiples chemical contaminants in soil samples

# Multivariate distributions

**Definition.** Let  $X$  and  $Y$  be two discrete random variables and  $R$  be the range of  $X \times Y$ . The probability that  $X = x$  and  $Y = y$ , denoted by

$$p(x, y) = P(X = x, Y = y), \quad (x, y) \in R$$

is the **joint probability mass function** (joint pmf) of  $X$  and  $Y$  and satisfies the following properties

- ①  $p(x, y) \geq 0$ , for any  $(x, y) \in R$
- ②  $\sum_{(x, y) \in R} p(x, y) = 1$

The **cumulative density function** (cdf) of  $X$  and  $Y$  is

$$F(x, y) = P(X \leq x, Y \leq y)$$

## Multivariate distributions

**Example** We roll a pair of unbiased dice. The consider as outcome of the experiment the random variables (ordered) pair  $(X, Y)$  where  $X$  is the smaller and  $Y$  is the larger of the two outcomes. For example, if we get the number 3 and 2, then the outcome of the experiment is  $(X = 2, Y = 3)$ . Note that we get the same  $(X, Y)$  if we either the dice from the rolling experiment are 2 and 3 or 3 and 2. Clearly, if get 2 and 2, then  $(X = 2, Y = 2)$ . Since 36 ordered pairs can result from rolling a pair of dice:

$$P(X = 2, Y = 3) = P(\text{dice } 2,3) + P(\text{dice } 3,2) = \frac{1}{36} + \frac{1}{36} = \frac{2}{36}$$

$$P(X = 2, Y = 2) = P(\text{dice } 2,2) = \frac{1}{36}$$

In general

$$P(X = x, Y = y) = \begin{cases} \frac{1}{36} & \text{if } 1 \leq x = y \leq 6 \\ \frac{2}{36} & \text{if } 1 \leq x < y \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

# Multivariate distributions

**Example (continue)** Using the joint pmf of  $X$  and  $Y$ , we can solve any probability problem on  $X$  and  $Y$ .

$$P(X = x, Y = y) = \begin{cases} \frac{1}{36} & \text{if } 1 \leq x = y \leq 6 \\ \frac{2}{36} & \text{if } 1 \leq x < y \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} P(X+Y = 4) &= P(X = 1, Y = 3) + P(X = 2, Y = 2) = \frac{2}{36} + \frac{1}{36} = \frac{3}{36} \\ P(X = 4) &= P(X = 4, Y = 4) + P(X = 4, Y = 5) + P(X = 4, Y = 6) \\ &= \frac{1}{36} + \frac{2}{36} + \frac{2}{36} = \frac{5}{36} \end{aligned}$$



# Multivariate distributions - Marginals

**Definition.** Let  $X$  and  $Y$  be two discrete random variables with joint pmf  $f(x, y)$  with  $(x, y) \in R$ . The **marginal pmf** of  $X$  is

$$f_1(x) = \sum_y f(x, y), \quad x \in R_1$$

Similarly, the **marginal pmf** of  $Y$  is

$$f_2(y) = \sum_x f(x, y), \quad y \in R_2$$

the random variables  $X$  and  $Y$  are **independent** is

$$f(x, y) = f_1(x) f_2(y)$$

# Multivariate distributions

## Example (continue)

x	1	2	3	4	5	6
y						
1	1/36					
2	2/36	1/36				
3	2/36	2/36	1/36			
4	2/36	2/36	2/36	1/36		
5	2/36	2/36	2/36	2/36	1/36	
6	2/36	2/36	2/36	2/36	2/36	1/36

The marginal pmfs are computed on the margins of the table

$$f_1(x) = \sum_{y=1}^6 f(x, y), \quad f_2(y) = \sum_{x=1}^6 f(x, y)$$

# Multivariate distributions

## Example

Let the joint pmf of  $X$  and  $Y$  be given by

$$f(x, y) = \frac{x + y}{21}, \quad x = 1, 2, 3; y = 1, 2$$

Note that  $\sum_{y=1}^2 \sum_{x=1}^3 f(x, y) = 1$ . Hence  $f(x, y)$  is a pmf.  
Are  $X$  and  $Y$  independent?

**Solution.** By direct computation,

$$f_1(x) = \sum_{y=1}^2 f(x, y) = \sum_{y=1}^2 \frac{x+y}{21} = \frac{2x+3}{21}$$

$$f_2(y) = \sum_{x=1}^3 f(x, y) = \sum_{x=1}^3 \frac{x+y}{21} = \frac{6+3y}{21}$$

Since  $f(x, y) \neq f_1(x) f_2(y)$ , then  $X$  and  $Y$  are NOT independent.

# Multivariate distributions

## Example

Let the joint pmf of  $X$  and  $Y$  be given by

$$f(x, y) = \frac{xy^2}{30}, \quad x = 1, 2, 3; y = 1, 2$$

Note that  $\sum_{y=1}^2 \sum_{x=1}^3 f(x, y) = 1$ . Hence  $f(x, y)$  is a pmf.  
Are  $X$  and  $Y$  independent?

**Solution.** By direct computation,

$$f_1(x) = \sum_{y=1}^2 f(x, y) = \sum_{y=1}^2 \frac{xy^2}{30} = \frac{x}{6}$$
$$f_2(y) = \sum_{x=1}^3 f(x, y) = \sum_{x=1}^3 \frac{xy^2}{30} = \frac{y^2}{5}$$

Since  $f(x, y) = f_1(x) f_2(y)$ , then  $X$  and  $Y$  are independent.

# Multivariate distributions - Expectation

**Definition.** Let  $X$  and  $Y$  be two discrete random variables with joint pmf  $f(x, y)$  with  $(x, y) \in R$ .

The **expectations** of  $X$  and  $Y$  are

$$\mu_X = E[X] = \sum_{(x,y) \in R} x f(x, y) = \sum_x x f_1(x)$$

$$\mu_Y = E[Y] = \sum_{(x,y) \in R} y f(x, y) = \sum_y y f_2(y)$$

Similarly, the **variances** of  $X$  and  $Y$  are

$$\sigma_X^2 = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f_1(x) = \sum_x x^2 f_1(x) - \mu_X^2$$

$$\sigma_Y^2 = E[(Y - \mu_Y)^2] = \sum_y (y - \mu_Y)^2 f_2(y) = \sum_y y^2 f_2(y) - \mu_Y^2$$

## Multivariate distributions - Expectation

In addition to the variance of  $X$  (or, similarly,  $Y$ ) that measure the variability around its mean, we can also quantify the variability of  $X$  and  $Y$  with respect to each other.

**Definition.** Let  $X$  and  $Y$  be two discrete random variables with joint pmf  $f(x, y)$  with  $(x, y) \in R$ .

The **covariance** of  $X$  and  $Y$  is

$$\begin{aligned}\text{cov}(X, Y) &= \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \\ &= \sum_{(x,y) \in R} (x - \mu_X)(y - \mu_Y) f(x, y) \\ &= \sum_{(x,y) \in R} xy f(x, y) - \mu_X \mu_Y\end{aligned}$$

**Remark.** If  $X$  and  $Y$  are independent, one can show that  $\sum_{(x,y) \in R} xy f(x, y) = \mu_X \mu_Y$ , hence  $\sigma_{XY} = 0$ . However, the converse is not true in general, that is,  $\sigma_{XY} = 0$  does not imply that  $X$  and  $Y$  are independent.

## Multivariate distributions - Expectation

The correlation coefficient is a normalized version of the covariance.

**Definition.** Let  $X$  and  $Y$  be two discrete random variables with joint pmf  $f(x, y)$  with  $(x, y) \in R$ .

The **correlation coefficient** of  $X$  and  $Y$  is

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

**Remark.**  $\rho$  ranges in the interval  $[-1, 1]$ . If  $\rho > 0$ , then  $X$  and  $Y$  are **positively correlated**; if  $\rho < 0$ , then  $X$  and  $Y$  are **negatively correlated**; if  $\rho = 0$ , then  $X$  and  $Y$  are **uncorrelated**;

As for the covariance, if  $X$  and  $Y$  are independent, then  $\rho = 0$ ; that is, *independence implies no correlation*. However, the converse is not true in general, that is,  $\rho = 0$  does not imply that  $X$  and  $Y$  are independent.

# Multivariate distributions - Covariance

## Properties of the covariance

- ①  $\text{cov}(X, Y) = \text{cov}(Y, X)$
- ②  $\text{cov}(X, X) = \text{var}(X) = \sigma_X^2$
- ③ If  $X$  and  $Y$  are independent  $\text{cov}(X, Y) = 0$
- ④ If  $X$ ,  $Y$  and  $Z$  are jointly distributed, and  $a, b$  are constants, then  $\text{cov}(X, aY + bZ) = a \text{cov}(X, Y) + b \text{cov}(X, Z)$ .



## Multivariate distributions

**Example.** Let the joint pmf of  $X$  and  $Y$  be given by the table below

$x$	1	2	3
$y$			
1	0.1	0.3	0.1
2	0.2	0.1	0.2

Are  $X$  and  $Y$  positively/negatively correlated? uncorrelated?

**Solution.**

$$\mu_X = \sum_{x=1}^3 x f_1(x) = (1)(0.3) + (2)(0.4) + (3)(0.3) = 2$$

$$\mu_Y = \sum_{y=1}^2 y f_2(y) = (1)(0.5) + (2)(0.5) = 1.5$$

$$E[XY] = \sum_{y=1}^2 \sum_{x=1}^3 xy f(x, y) = (1)(1)(0.1) + (2)(1)(0.3) + (3)(1)(0.1) + (1)(2)(0.2) + (2)(2)(0.1) + (3)(2)(0.2) = 3$$

$$\sigma_{XY} = E[XY] - \mu_X \mu_Y = 3 - (2)(1.5) = 0$$

Hence:  $X$  and  $Y$  are uncorrelated (however one can verify they are NOT independent).

## Multivariate distributions

**Example.** Let the joint pmf of  $X$  and  $Y$  be given by

$x \backslash y$	1	2
	0.4	0.1
1	0.4	0.1
2	0.1	0.4

Are  $X$ ,  $Y$  positively/negatively correlated? uncorrelated? Find  $\rho$ .

**Solution.**

$$\mu_X = \sum_{x=1}^2 x f_1(x) = (1)(0.5) + (2)(0.5) = 1.5$$

$$\mu_Y = \sum_{y=1}^2 y f_2(y) = (1)(0.5) + (2)(0.5) = 1.5$$

$$E[XY] = \sum_{y=1}^2 \sum_{x=1}^2 xy f(x, y) = (1)(1)(0.4) + (2)(1)(0.1) + (1)(2)(0.1) + (2)(2)(0.4) = 2.4$$

$$\sigma_{XY} = E[XY] - \mu_X \mu_Y = 2.4 - (1.5)(1.5) = 0.15$$

Hence:  $X$  and  $Y$  are positively correlated.

$$E[X^2] = \sum_{x=1}^2 x^2 f_1(x) = (1)(0.5) + (4)(0.5) = 2.5 = E[Y^2]$$

$$\text{Hence } \sigma_X^2 = \sigma_Y^2 = 2.5 - (1.5)^2 = 0.25 \text{ and}$$

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0.15}{\sqrt{(0.25)(0.25)}} = 0.6$$

# Multivariate distributions

**Interpretation.** Consider the two joint pmf

	x	1	2
y	1	0.4	0.1
	2	0.1	0.4

and

	x	1	2
y	1	0.1	0.4
	2	0.4	0.1

In the first case,  $\rho = 0.6$ , in the second case,  $\rho = -0.6$ .

Positive correlation: it is more likely that smaller values of  $X$  occur with smaller values of  $Y$  and larger values of  $X$  occur with larger values of  $Y$ .

Negative correlation: it is more likely that smaller values of  $X$  occur with larger values of  $Y$  and larger values of  $X$  occur with smaller values of  $Y$ .

# Multivariate distributions

Let  $W = a_0 + a_1X + a_2Y$  where  $X, Y$  are random variables with joint pmf  $f(x, y)$  and  $a_0, a_1, a_2$  are constants.

Then  $W$  is also a random variable where

$$\mu_W = a_0 + a_1 \mu_X + a_2 \mu_Y$$

and

$$\sigma_W^2 = a_1^2 \sigma_X^2 + a_2^2 \sigma_Y^2 + 2a_1 a_2 \sigma_{XY}$$

NOTE: If  $X$  and  $Y$  are independent, then  $\sigma_{XY} = 0$  and

$$\sigma_W^2 = a_1^2 \sigma_X^2 + a_2^2 \sigma_Y^2$$

# Multivariate distributions

**Example** Let  $X$  be a random variable with mean  $\mu_X = -1$  and variance  $\sigma_X^2 = 2$ , and  $Y$  be another random variable with mean  $\mu_Y = 2$  and variance  $\sigma_Y^2 = 1$ . Let  $W = 1 - 2X + Y$ . Assuming that  $X, Y$  are independent, find the mean and variance of  $W$ .

By the formulas derived above, we have

$$\mu_W = 1 - 2\mu_X + \mu_Y = 1 + (-2)(-1) + 2 = 5$$

$$\sigma_W^2 = 4\sigma_X^2 + \sigma_Y^2 = (4)(2) + 1 = 9$$

## 2.5 Continuous distributions

# Continuous random variables

A continuous random variable is a random variable whose range  $R$  is uncountable.

The main difference with respect to discrete random variables is that probabilities involving continuous random variables are defined over an interval of values and represented as an integral.

Given an interval  $[a, b] \subset R \subset \mathbb{R}$ ,

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

where  $f(x)$  is the **probability density function** (pdf) of  $X$

The probability of observing any single value is equal to 0

$$P(X = c) = P(c \leq X \leq c) = \int_c^c f(x) dx = 0$$

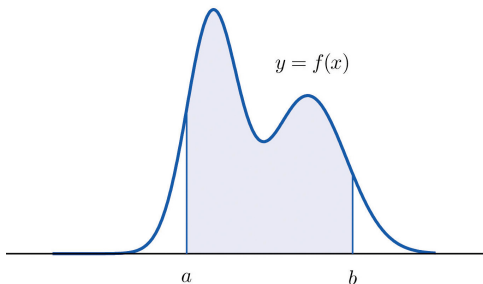
# Continuous random variables

Properties of a **probability density function**:

- ①  $f(x) \geq 0, \quad x \in R$
- ②  $\int_R f(x) dx = 1$

The probability is computed by evaluating an area under a curve:

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

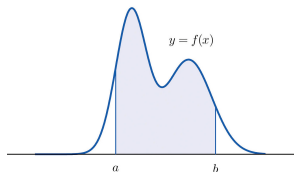




# Continuous random variables

The probability is computed by evaluating an area under a curve:

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



The **cumulative distribution function** of  $X$  is the indefinite integral of  $f$ :

$$F(x) = \int_{-\infty}^x f(w) dw$$

Hence (as in the Fundamental Theorem of Calculus)

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

# Continuous random variables: the uniform distribution

A random variable  $X$  is **uniform** on the interval  $[a, b]$  if  $X$  is equally likely to take any value in the range  $R = [a, b]$ . In this case

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases}$$

Notation:  $X \sim \text{unif}(a, b)$

**Example.** Transit time of the subway between Downtown Station and Midtown Station is uniformly distributed between 10.0 and 20.0 minutes. What is the the probability that the transit time is less than 12 minutes?

We have that uniform pdf:  $f(x) = \frac{1}{10}$  for  $10 \leq x \leq 20$

$$P(X < 12) = \int_{10}^{12} \frac{1}{10} dx = \left. \frac{x}{10} \right|_{x=10}^{x=12} = \frac{2}{10} = \frac{1}{5}.$$

# Continuous random variables: the uniform distribution

Solution using R:

Let  $X \sim \text{unif}(10, 20)$

(a) Compute  $P(X \leq 12)$

```
> punif(12,min=10,max=20)  
[1] 0.2
```

(b) Compute  $P(X > 12)$

```
> 1-punif(12,min=10,max=20)  
[1] 0.8
```

## Continuous random variables: expectation

Let  $X$  be a continuous random variable with pdf  $f(x)$ ,  $x \in R$ .

The **mean** or **expectation** of  $X$  is

$$\mu_X = E[X] = \int_R x f(x) dx$$

For a function  $u$  defined on  $R$ ,

$$E[u(X)] = \int_R u(x) f(x) dx$$

The **variance** of  $X$  is

$$\sigma_X^2 = E[(X - \mu_X)^2] = \int_R (x - \mu_X)^2 f(x) dx$$

A direct computation shows that

$$\sigma_X^2 = E[X^2] - \mu_X^2$$

# Continuous random variables: expectation

## Example: uniform pdf

Let  $X$  be a uniformly distributed continuous random variable defined in the interval  $[a, b]$ .

$$\mu_X = \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

$$\begin{aligned}\sigma_X^2 &= \int_a^b \frac{x^2}{b-a} dx - \mu_X^2 = \frac{x^3}{3(b-a)} \Big|_a^b - \frac{(b+a)^2}{4} \\ &= \frac{b^3 - a^3}{3(b-a)} - \frac{(b+a)^2}{4} \\ &= (\dots) \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

# The normal distribution

The **normal distribution** is the most important probability distribution in statistics because it fits many natural phenomena. In addition, many statistical hypothesis test require that data follow a normal distribution.

It is defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}$$

where the factor  $\frac{1}{\sigma\sqrt{2\pi}}$  ensures that

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

The normal distribution is a Gaussian function described by the parameters  $\mu$  and  $\sigma$ .

Notation of **normal random variable**:  $X \sim N(\mu, \sigma)$

# The normal distribution

Normal distribution:  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

**Mean:**

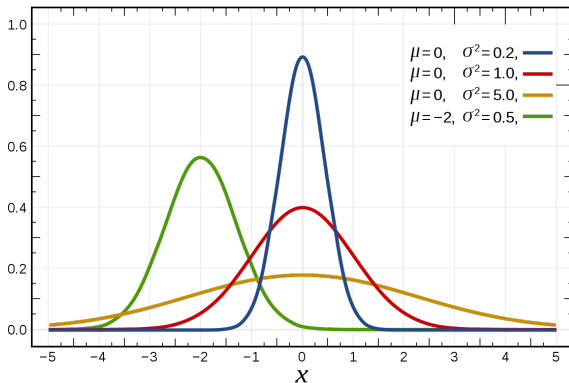
$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \mu$$

**Variance:**

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

# The normal distribution

The plot of the normal distribution is a bell-shaped curve, symmetrical with center about  $\mu$  and spread determined by  $\sigma$ .

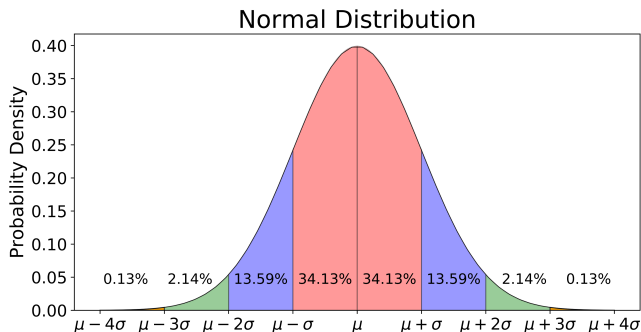


Larger  $\sigma \Rightarrow$  larger spread about  $\mu$ . Total area = 1.



# The normal distribution

The fraction of the area under the curve  $y = f(x)$  depends on  $\sigma$ .



Most of the area, 99.7%, is contained between  $\mu - 3\sigma$  and  $\mu + 3\sigma$

# The normal distribution

Probability computations when  $X \sim N(\mu, \sigma)$  require to solve

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

As in the general case, we can write

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

where  $F(x)$  is the **cumulative distribution function (cdf)**

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2} dw$$

Note: there is no analytic closed-form solution of the integral.

# The normal distribution - Using R

**Example.** The test scores of a college entrance exam fit a normal distribution where the mean test score is 72 and the standard deviation is 15.2. (a) What is the percentage of students scoring less than 84 in the exam? (b) What is the percentage of students scoring 84 or more in the exam?

(a) We compute  $P(X < 84)$  where  $X \sim N(\mu = 72, \sigma = 15.2)$

Using R

```
> pnorm(84, mean=72, sd=15.2)  
[1] 0.78508
```

(b) We compute  $P(X \geq 84)$  where  $X \sim N(\mu = 72, \sigma = 15.2)$

Using R

```
> 1-pnorm(84, mean=72, sd=15.2)  
[1] 0.21492
```

# The standard normal distribution

To solve problems involving the normal distribution, we introduce the **standard normal distribution** obtained for  $\mu = 0$ ,  $\sigma = 1$ .

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in \mathbb{R}$$

Standard normal random variable:  $z \sim N(0, 1)$

The values of

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

are available on statistical tables.

Using the tables:

$$P(z_1 \leq Z \leq z_2) = P(Z \leq z_2) - P(Z < z_1) = \Phi(z_2) - \Phi(z_1)$$

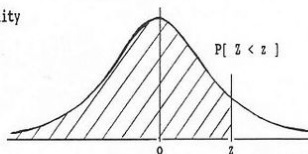
# The standard normal distribution

## STANDARD STATISTICAL TABLES

### 1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value  $z$  i.e.

$$P(Z < z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$$



$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

# The standard normal distribution

**Examples.** Let  $Z \sim N(0, 1)$ . Compute (a)  $P(Z \leq 0.92)$ ; (b)  $P(Z > 0.92)$ ; (c)  $P(0.45 < Z \leq 1.17)$

(a)  $P(Z \leq 0.92) = \Phi(0.92) = 0.8212$

```
> pnorm(0.92)
[1] 0.8212136
```

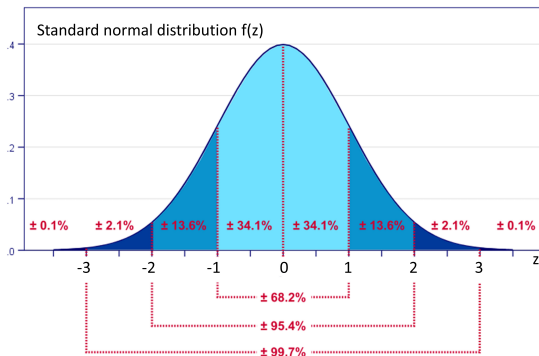
(b)  $P(Z > 0.92) = 1 - \Phi(0.92) = 1 - 0.8212 = 0.1788$

```
> 1-pnorm(0.92)
[1] 0.1787864
```

(c)  $P(0.45 < Z \leq 1.17) = \Phi(1.17) - \Phi(0.45) = 0.8790 - 0.6736 = 0.2054$

```
> pnorm(1.17)-pnorm(0.45)
[1] 0.2053547
```

# The standard normal distribution



From the plot:

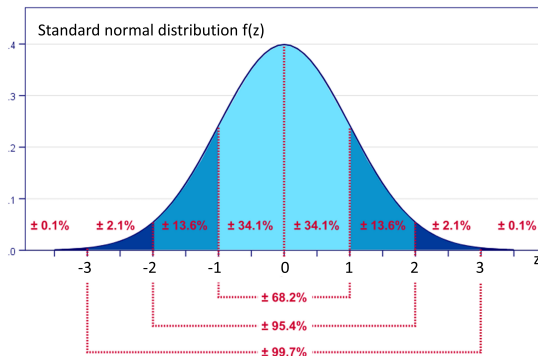
$$P(Z \leq 0) = \Phi(0) = 0.5$$

$$P(Z \leq 1) = \Phi(1) = 0.841$$

$$P(0 \leq Z \leq 1) = \Phi(1) - \Phi(0) = 0.341$$

$$P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) = 0.682$$

# The standard normal distribution



Useful properties:

$$\Phi(-1) = 1 - \Phi(1)$$

$$\Phi(-z) = 1 - \Phi(z)$$

You can derive all the value of  $\Phi(z)$  using the table for  $z \geq 0$ .



# The normal distribution

To solve probability problems for  $X \sim N(\mu, \sigma)$  we can apply a change of variable to standardize the random variable.

Setting  $Z = \frac{X - \mu}{\sigma}$ , then  $Z \sim N(0, 1)$

Hence

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

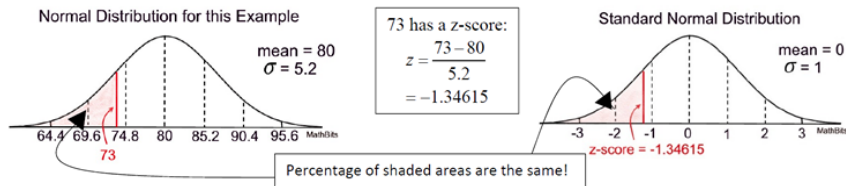
where  $\Phi$  is the cdf of the standard normal distribution.

# The normal distribution

**Example.** Let  $X \sim N(\mu = 80, \sigma = 5.2)$ .

$$P(X < 73) = P\left(\frac{X - 80}{5.2} < \frac{73 - 80}{5.2}\right) = \Phi\left(\frac{73 - 80}{5.2}\right) = \Phi(-1.346)$$

From the table  $\Phi(-1.346) = 1 - \Phi(1.346) = 0.0893$



# The normal distribution

**Example.** The body mass index (BMI) in the Canada population of 60 year old males is normally distributed and has a mean value  $= 29$  and a standard deviation  $= 6$ . What is the probability that a 60 year old male has BMI less than 35? What is the probability that a 60 year old male has BMI larger than 35?

The population is modeled as a normal random variable  $X \sim N(\mu = 29, \sigma = 6)$ .

$$P(X < 35) = P\left(\frac{X - 29}{6} < \frac{35 - 29}{6}\right) = \Phi\left(\frac{35 - 29}{6}\right) = \Phi(1) = 0.841$$

$$P(X > 35) = 1 - P(X \leq 35) = 1 - \Phi(1) = 0.159$$

Using R:

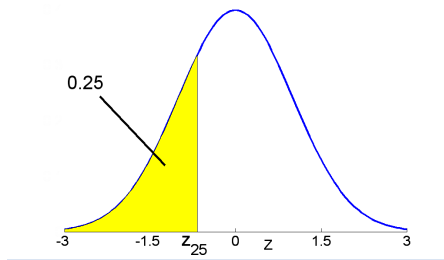
```
> pnorm(1)
[1] 0.8413447
```

# The normal distribution - Percentiles

**Definition.** A percentile is a value in the distribution that holds a specified percentage of the population below it.

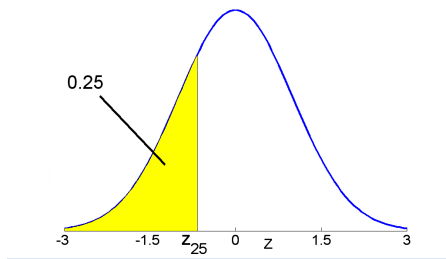
For the standard normal distribution, the **100p-th percentile** is the value  $z_p$  such that  $P(Z \leq z_p) = p$ .

Example: the 25-th percentile is the number  $z_{0.25}$  such that  $P(Z \leq z_{0.25}) = 0.25$



$$z_{0.25} = -0.675$$

# The normal distribution - Percentiles



Commonly used percentiles:

- 50-th percentile  $z_{0.50} = 0$
- 75-th percentile  $z_{0.75} = 0.675$
- 90-th percentile  $z_{0.90} = 1.282$
- 95-th percentile  $z_{0.95} = 1.645$
- 99-th percentile  $z_{0.99} = 2.326$

By the symmetry of  $f(z)$ , we have:  $z_{1-p} = -z_p$

## The normal distribution - Percentiles

For a normal distribution, the **100p-th percentile** is the value  $x_p$  such that  $P(X \leq x_p) = p$ .

**Example.** The body mass index (BMI) in the Canada population of 60 year old males is normally distributed and has a mean value  $= 29$  and a standard deviation  $= 6$ . What is the 90th percentile of BMI for 60 year old males?

The 90th percentile is the BMI that holds 90% of the BMIs below it (and 10% above). We need to find  $x_{0.90}$  such that  $P(X \leq x_{0.90}) = 0.90$

By converting into the standard normal distribution, we have

$$P\left(\frac{X - 29}{6} \leq \frac{x_{0.90} - 29}{6}\right) = P\left(Z \leq \frac{x_{0.90} - 29}{6}\right) = 0.90$$

Using the percentile values of the standard normal distribution,  $\frac{x_{0.90} - 29}{6} = z_{0.90} = 1.282$ .

Hence  $x_{0.90} = (6)(1.282) + 29 = 36.69$

# The normal distribution - Percentiles

Solution of the percentile problem in R

The function `qnorm()`, which comes standard with R, finds the boundary value  $z_p$  that determines this probability area

$$P(z < z_p) = p.$$

For example, the 90-th percentile is

```
> qnorm(0.9)
[1] 1.281552
```

If you want to find that 90-th percentile of a normal distribution whose mean is 29 and whose standard deviation is 6, then

```
> qnorm(0.90, mean=29, sd=6)
[1] 36.68931
```

# The normal distribution - Percentiles

**Example.** Suppose that SAT scores are normally distributed, and that the mean SAT score is 1000 and the standard deviation of all SAT scores is 100. How high must you score so that only 10% of the population scores higher than you?

Solution.

If 10% score higher than you, then 90% score lower. That is, you want to find the 90-th percentile  $x_{0.90}$  of the SAT scores given by

$$P(X \leq x_{0.90}) = 0.90.$$

Using R we find

```
> qnorm(0.90,mean=1000,sd=100)
[1] 1128.155
```



### 3. Statistical Inference

# Statistical inference

In scientific applications or social sciences, where we deal with data affected by randomness, we often want to extract useful information and draw conclusions from the data.

Examples: medical diagnostics, wireless communication, election polls.

**Statistical inference** is a collection of methods that deal with drawing conclusions from real data that are associated with uncertainty.

# Statistical inference

In a typical problem of statistical inference, after data collection, we want to draw some conclusion about a random variable  $X$ . We may have the following two situations:

- 1 The form of the pdf  $f(x, \theta)$  or the pmf  $p(x, \theta)$  describing the random variable  $X$  is known, but the parameter  $\theta$  is unknown. Note that  $\theta$  may be a vector.
- 2 The pdf  $f(x, \theta)$  or the pmf  $p(x, \theta)$  describing the random variable  $X$  is unknown.

In case 1, where the distribution model is known but the parameter  $\theta$  is not, the problem is addressed using classical **parametric** methods. For example, we may have reason to think that a certain collection of data can be modeled using a normal distribution  $N(\mu, \sigma)$ . Hence, we need to **estimate**  $\mu$  and  $\sigma$ .

In case 2, where the distribution model is unknown, we need to use a **nonparametric** approach.

# Statistical inference - parametric case

**Estimation** is the procedure by which we infer the value about the unknown parameter(s) of a distribution based on collected data:

- 1 There is an unknown parameter  $\theta$  that we would like to estimate;
- 2 we collect a random sample of the data;
- 3 we use the data sample to **estimate** the desired quantity.

There are two major approaches to this problem:

- **Frequentist or classical inference.** The unknown quantity  $\theta$  is assumed to be a fixed quantity. That is,  $\theta$  is a deterministic (non-random) quantity to be estimated using collected data.
- **Bayesian approach.** The unknown parameter  $\theta$  is assumed to be a random variable, and we assume that we have some initial guess about the distribution of  $\theta$ . After observing the data, we update the distribution of  $\theta$  using Bayes' Theorem.

In MATH 3339 and MATH 4310, we only consider the frequentist or classical approach.

# Statistical inference - parametric case

Estimation includes two main categories: **point estimation** and **interval estimation**.

For example, in the classical polling problem, we are interested in the percentage  $p$  of people who will vote for Candidate A. After polling  $n$  randomly chosen voters, we define the estimator

$$\hat{p} = \frac{Y}{n}$$

where  $Y$  is the number of people - among the randomly chosen voters - who say they will vote for Candidate A.

Note that, although  $p$  is a parameter, the estimator  $\hat{p}$  of  $p$  is a random variable as it depends on the random sample.

Rather than a point estimate, we can compute an interval in which the value of the unknown parameter  $p$  is highly likely to lie.

# 3.1 Point Estimation

## Random sampling

Suppose that our goal is to investigate the height distribution of people in a well defined population (i.e., adults of age 25-50 in a certain country). To do this, we define random variables  $X_1, X_2, X_3, \dots, X_n$  as follows: we choose a **random sample** of size  $n$  from the population and let  $X_i$  be the height of the  $i$ -th chosen person.

**Definition.** A collection of random variables  $X_1, X_2, \dots, X_n$  is said to be a *random sample of size  $n$*  if they are independent and identically distributed (i.i.d.), i.e.,

- 1  $X_1, X_2, \dots, X_n$  are independent random variables;
- 2 they have the same distribution.

To estimate the average height in the population, we may define an estimator as

$$\hat{\theta} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Note that the estimator is a function of the random sample and, in particular, it is a random variable.

# Point estimation

**Estimation** is a process for learning and determining the population parameter based on the model fitted to the data.

**Point estimation**, **interval estimation** and **hypothesis testing** are among the main ways of learning about the population parameter from a random sample.

An **estimator** is particular example of a statistic and becomes an **estimate** when the formula is replaced with actual observed sample values.

**Point estimation** = a single value that estimates the parameter. Point estimates are single values calculated from a random sample.



## Point estimation

Let us assume that  $\theta$  is an unknown parameter to be estimated. For example, it might be the expected value of a random variable  $X$ :

$$\theta = \mu_X = E[X]$$

To estimate  $\theta$ , we collect a random samples  $X_1, X_2, X_3, \dots, X_n$  from the unknown distribution  $p(x)$  of  $X$  and then we define a point estimator  $\hat{\theta}$  that is a function of the samples, that is

$$\hat{\theta} = h(X_1, X_2, X_3, \dots, X_n)$$

There are many possible estimators for  $\hat{\theta}$  of  $\theta$ :

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \quad \text{(sample mean)}$$

$$\tilde{X} = X_1$$

$$\tilde{\tilde{X}} = \frac{\min X_i + \max X_i}{2}$$

# Point estimation

How do we compare different possible estimators?

Roughly speaking, a good estimator  $\hat{\theta}$  should be “close” to the real value of  $\theta$ .

We make this notion more precise by defining 3 desirable properties for point estimators: **bias, mean square error and consistency**.

**Definition.** Let  $\hat{\theta} = h(X_1, X_2, X_3, \dots, X_n)$  be a point estimator of  $\theta$ . The **bias** of the estimator is

$$b(\hat{\theta}) = E[\hat{\theta}] - \theta$$

We say that  $\hat{\theta}$  is an **unbiased estimator** of  $\theta$  if  $b(\hat{\theta}) = 0$

Interpretation: The bias of an estimator  $\hat{\theta}$  tells us **on average** how far  $\hat{\theta}$  is from the real value of  $\theta$ .

## Point estimation - The sample mean

**Proposition.** Let  $X_1, X_2, \dots, X_n$  be a random samples from a distribution with  $E[X] = \mu_X$  and consider the sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n$ . Then  $\bar{X}$  is an unbiased estimator of the mean.

Proof: Since the samples  $X_1, X_2, \dots, X_n$  are i.i.d, then  $E[X_i] = \mu_X$  for any  $i$ . Hence

$$b(\bar{X}) = E[\bar{X}] - \mu_X = \frac{1}{n} \sum_{i=1}^n E[X_i] - \mu_X = \frac{1}{n} n\mu_X - \mu_X = 0 \quad \square$$

Also the estimator  $\tilde{X} = X_1$  is an unbiased estimator since

$$b(\tilde{X}) = E[\tilde{X}] - \mu_X = E[X_1] - \mu_X = \mu_X - \mu_X = 0$$

The second example shows that an unbiased estimator is not necessarily a good estimator.

## Point estimation - The sample mean

**Definition.** The **mean squared error** (MSE) of a point estimator  $\hat{\theta}$ , denoted as  $MSE(\hat{\theta})$ , is defined as

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Interpretation: the MSE is a measure of the distance between  $\hat{\theta}$  and  $\theta$ , and a smaller MSE is generally indicative of a better estimator.

**Proposition.** Let  $X_1, X_2, \dots, X_n$  be a random samples from a distribution with  $E[X] = \mu$  and  $var(X) = \sigma^2$ . We consider the following two estimators for  $\mu_X$ .

- 1  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ;
- 2  $\tilde{X} = X_1$ .

Then  $MSE(\bar{X}) < MSE(\tilde{X})$ , if  $n > 1$ . That is, the sample mean estimator has a lower MSE.

## Point estimation - The sample mean

**Proof:**

$$MSE(\tilde{X}) = E[(X_1 - \mu)^2] = \text{var}(X_1) = \sigma^2$$

$$MSE(\bar{X}) = E[(\bar{X} - \mu)^2] = \text{var}(\bar{X} - \mu) + (E[\bar{X} - \mu])^2,$$

using the observation that  $\text{var}(Y) = E[Y^2] - (E[Y])^2$  with  $Y = \bar{X} - \mu$ .

Using the observation that  $\text{var}(\bar{X} - \mu) = \text{var}(\bar{X})$  since  $\mu$  is constant and that  $E[\bar{X} - \mu] = 0$ , we conclude that

$$MSE(\bar{X}) = \text{var}(\bar{X}) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

# Point estimation - The sample mean

**Definition.** Let  $\hat{\theta}_1, \hat{\theta}_2, \dots$  be a sequence of point estimators of  $\theta$ . We say that  $(\hat{\theta}_n)$  is a **consistent estimator** of  $\theta$ , if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0$$

Interpretation: an estimator is consistent if as the sample size  $n$  gets larger, the estimator converges to the real value of  $\theta$ .

One can show that the sample mean  $\bar{X}$  is a consistent estimator of the mean.

## Point estimation - The sample variance

We examine now how to estimate the variance of a distribution.

Let  $X_1, X_2, \dots, X_n$  be a random sample with mean  $E[X_i] = \mu$ , and variance  $\text{Var}(X_i) = \sigma^2$  for all  $i$ . Suppose that we use the following estimators of  $\sigma^2$

$$\textcircled{1} \quad \bar{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\textcircled{2} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (\text{sample variance})$$

**Proposition.**  $S^2$  is an unbiased estimator of  $\sigma^2$  while  $\bar{S}^2$  is not.

The **sample standard deviation** is defined as  $S = \sqrt{S^2}$  and is commonly used as an estimator for  $\sigma$ . Nevertheless,  $S$  is a biased estimator of  $\sigma$ .

## Point estimation - The sample variance

**Proof.** Since  $\text{Var}(Y) = E[Y^2] - (E[Y])^2$ , then

$$E[\bar{X}^2] = (E[\bar{X}])^2 + \text{Var}(\bar{X}) = \mu^2 + \frac{\sigma^2}{n}$$

$$E[X_i^2] = (E[X_i])^2 + \text{Var}(X_i) = \mu^2 + \sigma^2.$$

$$\begin{aligned}\text{Thus:} \quad E[\bar{S}^2] &= \frac{1}{n} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^n X_i^2 - n\bar{X}^2\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i^2] - nE[\bar{X}^2] \\ &= \frac{1}{n} \left( n(\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right) \right) \\ &= \frac{n-1}{n} \sigma^2\end{aligned}$$

Similarly,  $E[S^2] = \sigma^2$ .



# Sampling distribution

Let  $X_1, X_2, \dots, X_n$  be a random samples from a distribution with mean  $\mu$  and variance  $\sigma^2$ .

As noted, the sample mean

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

is also a random variable and, in particular, it has its own pdf.

We have:

$$\mu_{\bar{X}} = E[\bar{X}] = \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] = \frac{1}{n} n \mu = \mu$$

$$\sigma_{\bar{X}}^2 = \text{var}(\bar{X}) = \frac{1}{n^2} \text{var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

## Sampling distribution

We can standardize  $\bar{X}$  by defining

$$Z_n = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma}$$

Note:  $E[Z_n] = \frac{1}{\sigma_{\bar{X}}} E[\bar{X} - \mu] = 0$  and  $\text{var}(Z_n) = (\frac{\sqrt{n}}{\sigma})^2 \text{var}(\bar{X}) = 1$

### Theorem

Let  $X_1, X_2, \dots, X_n$  be a random samples from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Then  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$  and, as a consequence,  $Z_n \sim N(0, 1)$

Than is, the sample mean  $\bar{X}$  is also normally distributed with mean  $\mu$  and variance  $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ .

Not surprisingly, if samples are taken from a normal distribution, the sample mean is also normally distributed. In general, though, samples are taken from a distribution which is not normal or that may even be unknown.

# Sampling distribution - Central Limit Theorem

## Central Limit Theorem.

Let  $X_1, X_2, \dots, X_n$  be random samples from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Then the random variable  $Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  converges in distribution to a standard normal random variable as  $n$  tends to infinity, that is

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x), \quad \text{for all } x,$$

where  $\Phi$  is the standard normal cdf.

Significance: Even if we take random samples from an unknown distribution, if the sample size is sufficiently large ( $n > 30$ ) then the distribution of the sample mean is approximately normal.

# Sampling distribution - Central Limit Theorem

**Exercise.** The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

**Solution.** The sample mean  $\bar{X}$  has mean  $\mu = 2.61$  and standard deviation  $\sigma_{\bar{X}} = \sigma/\sqrt{100} = 0.5/10 = 0.05$ . By the CLT, we can approximate the sampling distribution as

$X \sim N(\mu = 2.61, \sigma_{\bar{X}} = 0.05)$ .

Thus we can compute  $P(2.51 < \bar{X} < 2.71)$  using R as

```
pnorm(2.71,mean=2.61,sd=0.05)-pnorm(2.51,mean=2.61,sd=0.05)  
[1] 0.9544997
```

## Sampling distribution - Central Limit Theorem

**Exercise.** An automobile battery manufacturer claims that its midgrade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this brand is normal. (a) Find the probability that a randomly selected battery of this type will last less than 48 months. (b) Find the probability that the mean battery life of a random sample of 36 such batteries will be less than 48 months.

**Solution.** (a) Since the population of battery lives is known to have a normal distribution  $N(\mu = 50, \sigma = 6)$ , we compute  $P(X < 48)$  as

```
> pnorm(48,mean=50,sd=6)
```

```
[1] 0.3694413
```

(b) The sample mean is normally distributed with  $\mu = 50$  and standard deviation  $\sigma_{\bar{X}} = 6/\sqrt{36} = 1$ . Thus we compute  $P(\bar{X} < 48)$  as

```
> pnorm(48,mean=50,sd=1)
```

```
[1] 0.02275013
```

# Sampling distribution - Central Limit Theorem

The central limit theorem implies we can **approximate the binomial pmf** with parameters  $p, n$  using the normal distribution.

Explanation: Recall that, given  $X_1, \dots, X_n$  Bernoulli trials with probability of success  $p$ , then  $Y = \sum_{i=1}^n X_i \sim \text{binom}(n, p)$ , with  $\mu_Y = np$  and  $\sigma_Y^2 = np(1 - p)$ .

We can think of the  $n$  Bernoulli trials as a random sample. Hence, if  $n$  is sufficiently large, by CLT we can approximate

$$\frac{Y}{n} = \sum_{i=1}^n X_i \sim N\left(\mu = p, \sigma^2 = \frac{p(1-p)}{n}\right)$$

$\frac{Y}{n}$  can be interpreted as the **proportion** of successes, i.e., the number of successes over the number of trials.

Note that  $Y \sim N(\mu = np, \sigma^2 = np(1 - p))$

# Sampling distribution - Central Limit Theorem

**Exercise.** You flip a fair coin 100 times. (a) What is the expected number of heads for this experiment? what is the variance? (b) What is the probability that we observe between 45 and 60 heads?

**Solution.** (a) Because the coin is fair, the expected number of heads is  $np = (100)(0.5) = 50$ . Variance is  $np(1 - p) = (100)(0.5)(0.5) = 25$

(b) Define random variable  $Y =$  number of heads observed. BY the CLT,  $Y \sim N(\mu = 50, \sigma = 5)$ . Hence,  $P(45 \leq Y \leq 60)$  is computed using R as

```
> pnorm(60,mean=50,sd=5)-pnorm(45,mean=50,sd=5)
[1] 0.8185946
```

# Sampling distribution - Central Limit Theorem

**Exercise.** Lab results indicate that a certain drug is effective 75% of the time (success) and ineffective 25% of the time (failure). As a trial, the drug is administered to a sample of 1000 patients. a) What is the expected proportion of successes for this experiment? (b) What is the probability that between 71% and 77% of the patients are helped by the drug?

**Solution.** (a) The expected proportion of successes is  $p = 0.75$ .

(b) Define random variable  $W$  = proportion of successes. By the CLT,  $W$  is well approximated by a normal random variable with mean 0.75 and standard deviation

$$\sqrt{p(1-p)/n} = \sqrt{(0.75)(0.25)/1000} = 0.01369306$$

We compute  $P(0.71 \leq W \leq 0.77)$  using R as

```
pnorm(0.77,mean=0.75,sd=0.0137)-pnorm(0.71,mean=0.75,sd=0.0137)
[1] 0.9260831
```



# Sampling distribution - Central Limit Theorem

We will use R to illustrate the Central Limit Theorem

The following commands in R computes 5000 simulations of sample means of size 12 from a normal distribution with mean  $\mu = 100$  and standard deviation  $\sigma = 14$ .

```
require(fastR2)
samplesum <- do(5000) *
c(sample.mean=mean(rnorm(12,100,14)))
```

We next compute the approximate mean and standard deviation of the sample mean

```
mean(~sample.mean, data=samplesum)
[1] 100.0191
sd(~sample.mean, data=samplesum)
[1] 4.138746
```

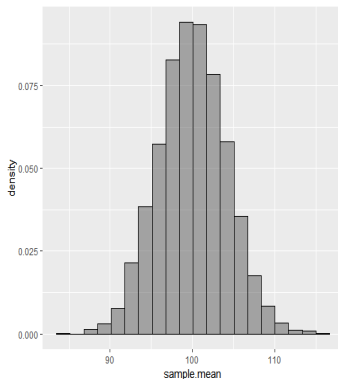
Compare with the theoretical values:

$$\mu_{\bar{X}} = \mu = 100 \text{ and } \sigma_{\bar{X}} = \sigma / \sqrt{12} = 4.041452$$

# Sampling distribution - Central Limit Theorem

The following command plots the histogram giving the approximate distribution of the sample mean.

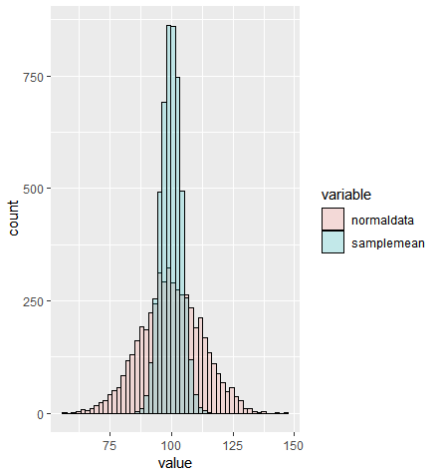
```
gf_dhistogram(~sample.mean,data=samplesum,bins=20,color="black")
```



The histogram approximates a normal distribution with the same mean as the data but a different standard deviation

# Sampling distribution - Central Limit Theorem

For clarity, the following plot compares the histogram of the sample mean and the histogram of the data.



# Sampling distribution - Central Limit Theorem

R script used to generate comparison plot:

```
library(ggplot2)
normaldata <- rnorm(5000,mean=100,sd=12)
nsamplesum <- do(5000) *
c(sample.mean=mean(rnorm(12,100,14)))
samplemean=nsamplesum$sample.mean

df <- data.frame(variable = c(rep("normaldata",
length(normaldata)),
rep("samplemean",length(samplemean))),
value=c(normaldata,samplemean))
ggplot(df, aes(x=value, fill=variable))+
geom_histogram(position = "identity",alpha =
.2,bins=50,color="black")
```

# Sampling distribution - Central Limit Theorem

We repeat the same simulation as above using now samples from a uniform distribution in the interval  $[-2, 4]$ .

Also in this case, we run a numerical test over 5000 simulations:

```
require(fastR2)
nsamplesum <- do(5000) *
c(sample.mean=mean(runif(12,-2,4)))
```

We next compute mean and standard deviation of the sample mean, and compare it to the theoretical result.

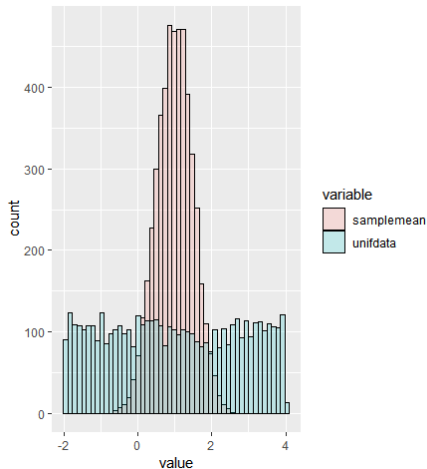
```
mean(~sample.mean, data=nsamplesum)
[1] 0.9819321
sd(~sample.mean, data=nsamplesum)
[1] 0.4991983
```

Compare with the theoretical values:

$$\mu_{\bar{X}} = \mu = \frac{4-2}{2} = 1 \text{ and } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{12}} = \frac{4+2}{\sqrt{12} \cdot \sqrt{12}} = 0.5$$

# Sampling distribution - Central Limit Theorem

Again, we compare the histogram of the sample mean and the histogram of the data.



# Sampling distribution - Central Limit Theorem

R script used to generate comparison plot:

```
library(ggplot2)
unifdata <- runif(5000,-2,4)
nsamplesum <- do(5000) *
c(sample.mean=mean(runif(12,-2,4)))
samplemean=nsamplesum$sample.mean

df <- data.frame(variable = c(rep("unifdata",
length(unifdata)),
rep("samplemean",length(samplemean))),
value=c(unifdata,samplemean))
ggplot(df, aes(x=value, fill=variable))+
geom_histogram(position = "identity",alpha =
.2,bins=50,color="black")
```

## 3.2 Confidence Intervals



# Confidence Intervals or Interval Estimation

The point estimate alone does not give much information about a parameter  $\theta$  of a distribution. Without additional information, we do not know how close the estimate  $\hat{\theta}$  is to the real  $\theta$ .

Here, we introduce the concept of interval estimation where, rather than giving just one value  $\hat{\theta}$  as the estimate for  $\theta$ , we produce an interval that is likely to include the true value of  $\theta$ .

In interval estimation, there are two important concepts:

- 1 The **length** of the reported interval which is likely to contain the true value of  $\theta$ . The length of the interval shows the precision of our estimate.
- 2 The **confidence level** that shows how confident we are about the interval. The confidence level is the probability that the interval includes the real value of  $\theta$ .

# Confidence Intervals or Interval Estimation

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with a parameter  $\theta$  that is to be estimated. A **interval estimator** with confidence level  $1 - \alpha$  consists of two estimators  $\hat{\theta}_L$  and  $\hat{\theta}_H$  such that

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_H) < 1 - \alpha, \quad \text{for all } \theta.$$

Equivalently, we say that  $[\hat{\theta}_L, \hat{\theta}_H]$  is a  $100(1 - \alpha)$  percent **confidence interval** of  $\theta$ .

We remark that  $\hat{\theta}_L$  and  $\hat{\theta}_H$  are random variables because they are functions of the observed random variables  $X_1, X_2, \dots, X_n$ . By contrast  $\theta$  is not a random variable.

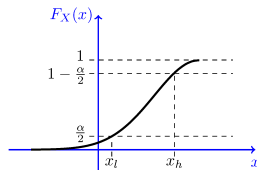
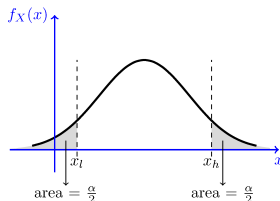
# How to find a Confidence Interval

Let  $X$  be a normal random variable with CDF  $F(x) = P(X \leq x)$ . Suppose that we are interested in finding two values  $x_h$  and  $x_l$  such that

$$P(x_l \leq X \leq x_h) = 1 - \alpha$$

We can choose  $x_l$  and  $x_h$  such that

$$P(X \leq x_l) = F(x_l) = \alpha/2, \quad P(X \geq x_h) = 1 - F(x_h) = \alpha/2.$$



In this case,  $[x_l, x_h]$  is the  $100(1 - \alpha)$  percent confidence interval of  $X$ .

# How to find a Confidence Interval

**Example** Let  $z \sim N(0, 1)$ .

We want to find  $z_l, z_h$  such that  $P(z_l \leq Z \leq z_h) = 0.95$

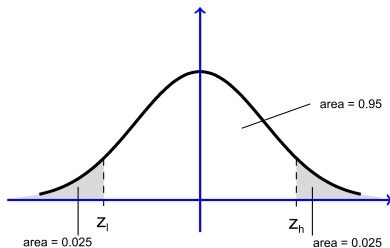
Because  $Z$  is the standard normal random variable, we can express the solution using the cdf  $\Phi$  of  $Z$ . In fact,  $z_l, z_h$  are determined by

$$\Phi(z_l) = 0.025, \quad \Phi(z_h) = 1 - 0.025 = 0.975$$

$z_h$  is the 97.5th percentile and, by symmetry,  $z_l = -z_h$ . That is

$$z_h = \Phi^{-1}(0.975) = 1.96 \quad z_l = \Phi^{-1}(0.025) = -1.96,$$

Using R: `qnorm(0.975) = 1.959964`, `qnorm(0.025) = -1.959964`



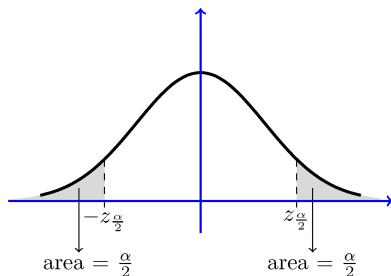
# How to find a Confidence Interval

In general, we denote  $z_{\frac{\alpha}{2}}$  and  $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$  such that

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

Hence  $\Phi(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$  and

$$z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2}) = \text{qnorm}(1 - \frac{\alpha}{2}) \quad (\text{R formula})$$



$1 - \alpha$	$\alpha$	$z_{\frac{\alpha}{2}}$
0.90	0.10	1.645
0.95	0.05	1.960
0.99	0.01	2.576

## Confidence Interval of the mean

Let  $X_1, X_2, \dots, X_n$  be a **random sample from a normal distribution**  $N(\mu, \sigma)$ , where  $\sigma$  is known.

We want to find the 95% confidence interval for  $\mu$ .

To solve the problem, we start from the estimator of the mean

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_i$$

We have that  $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ , hence  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

Because  $\mu_Z = E[Z] = 0$ , to solve the problem we need to find  $z_l, z_h$  such that

$$P(z_l \leq \left(Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \leq z_h) = 0.95$$

By our observations above,  $z_h = 1.96$ ,  $z_l = -1.96$

# Confidence Interval of the mean

Hence we can write:

$$P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

which is equivalent to

$$P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

That is, the 95% confidence interval for  $\mu$  is

$$[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}]$$

# Confidence Interval of the mean [normal pdf, $\sigma$ known ]

In general:

## Theorem

Let  $X_1, X_2, \dots, X_n$  be a **random sample from a normal distribution**  $N(\mu, \sigma)$ , where  $\mu$  is unknown and  $\sigma$  is known.

Let  $\bar{X}$  be the sample mean  $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_i$ . Then

$$[\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \quad \text{or} \quad \bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is a  $(1 - \alpha)100\%$  **confidence interval for  $\mu$** .

Note that  $z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2}) = \text{qnorm}(1 - \frac{\alpha}{2})$



## Confidence Interval of the mean [ $\sigma$ known, $n > 30$ ]

What if the distribution is not normal?

By the central limit theorem,  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  is approximately normal.  
Hence we can use the same argument above to get:

### Theorem

Let  $X_1, X_2, \dots, X_n$  be a **random sample an unknown distribution** where  $\mu = E[X_i]$  is unknown,  $Var(X_i) = \sigma^2$  is known and  $n > 30$ . Let  $\bar{X}$  be the sample mean  $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_i$ . Then

$$[\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \quad \text{or} \quad \bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is an approximate  $(1 - \alpha)100\%$  **confidence interval for  $\mu$** .

## Confidence Interval of the mean

### Example.

A scientist measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius)

102.5, 101.7, 103.1, 100.9, 100.5, 102.2

on 6 different samples of the liquid. If he knows that the standard deviation for this procedure is  $\sigma = 1.2$  degrees, what is the 95% confidence interval for the population mean?

If the measurements follow a normal distribution, then the sample mean will have the distribution  $N(\mu, \sigma_{\bar{X}} = \sigma / \sqrt{n})$ . We compute  $\bar{x} = 101.82$ . Hence a 95% confidence interval is:

$$\left[ 101.82 - 1.96 * \frac{1.2}{\sqrt{6}}, 101.82 + 1.96 * \frac{1.2}{\sqrt{6}} \right] = [100.86, 102.78]$$

As the level of confidence decreases, the size of the corresponding interval will decrease. For instance, the 90% confidence interval is:

$$\left[ 101.82 - 1.645 * \frac{1.2}{\sqrt{6}}, 101.82 + 1.645 * \frac{1.2}{\sqrt{6}} \right] = [101.01, 102.63]$$

## Confidence Interval of the mean

### Example.

A scientist measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius). If he knows that the standard deviation for this procedure is  $\sigma = 1.2$  degrees, under the assumption that the measurements follow a normal distribution, how large the sample size  $n$  should be so that the width of the 95% confidence interval is 1 degree Celsius?

We want to find  $n$  such that the interval

$$[\bar{x} - 0.5, \bar{x} + 0.5]$$

is a 95% confidence interval for the mean. Hence we need to find  $n$  such that  $1.96 * \frac{1.2}{\sqrt{n}} = 0.5$ . We solve as

$$n = \left(1.96 * \frac{1.2}{0.5}\right)^2 = 22.13$$

This shows that the scientist has to collect  $n = 23$  measurements or more.

# Confidence Interval of the mean - Sample size

For normally distributed data with known standard deviation, the  $(1 - \alpha)100\%$  confidence interval for  $\mu$

$$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is an interval of *width*  $w = 2 z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$  and *half-width*  $h = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ .

To find the sample size  $n$  such that we are  $(1 - \alpha)100\%$  confident the mean is contained in a confidence interval of width  $w$ , we solve the equation above for  $n$ , finding

$$n \geq \frac{(2 z_{\frac{\alpha}{2}} \sigma)^2}{w^2}$$

or equivalently

$$n \geq \frac{(z_{\frac{\alpha}{2}} \sigma)^2}{h^2}$$

## Confidence Interval of the mean [ $\sigma$ unknown, $n > 30$ ]

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution where  $\mu = E[X_i]$  is unknown,  $\text{Var}(X_i) = \sigma^2$  is unknown and  $n > 30$ . We want to find the  $(1 - \alpha)100\%$  confidence interval for  $\mu$ .

As above, we have that

$$P(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha.$$

However we cannot explicitly write the confidence interval because  $\sigma$  is unknown.

There are two general approaches: we can either find an **upper bound** for  $\sigma$ , or we can **estimate**  $\sigma$ .

# Confidence Interval of the mean [ $\sigma$ unknown, $n > 30$ ]

## Variance estimation.

We have already discussed a point estimator for  $\sigma^2$  and we called it the sample variance:

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

If  $n$  is large,  $S^2$  is likely to be close to the real value of  $\sigma^2$ .

Hence, given a random sample  $X_1, X_2, \dots, X_n$  from a distribution where  $\mu = E[X_i]$  is unknown,  $\text{Var}(X_i) = \sigma^2$  is unknown and  $n > 30$ , the (approximate)  $(1 - \alpha)100\%$  confidence interval for  $\mu$  is

$$\left[ \bar{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right]$$

or

$$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

## Confidence Interval of the mean [ $\sigma$ unknown, $n > 30$ ]

### Example.

We have collected a random sample  $X_1, X_2, \dots, X_{100}$  from an unknown distribution. The sample mean and the sample variance for this random sample are given by  $\bar{X} = 15.6$ ,  $S^2 = 8.4$ . Construct an approximate 99% confidence interval for  $\mu = E[X_i]$ .

Since  $1 - \alpha = 0.99$ , then  $\alpha = 0.01$  and  $z_{0.005} = 2.576$ . Hence, the approximate 99% confidence interval for  $\mu$  is

$$\left[ 15.6 - 2.576 \frac{\sqrt{8.4}}{\sqrt{100}}, 15.6 + 2.576 \frac{\sqrt{8.4}}{\sqrt{100}} \right] = [14.85, 16.34]$$

# Confidence Interval of the mean [ $\sigma$ unknown, $n > 30$ ]

## Variance upper bound.

If the variance of a distribution is unknown, we may be able to still compute the confidence interval of the mean by using an upper bound for  $\sigma^2$

Suppose we can find a bound

$$\sigma \leq \sigma_{max} < \infty.$$

Then the following interval

$$\left[ \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_{max}}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_{max}}{\sqrt{n}} \right]$$

is also a valid  $(1 - \alpha)100\%$  confidence interval for  $\mu$ .



# Confidence Interval of the proportion

We want to estimate the **proportion** of people who plan to vote for Candidate A in an upcoming election. It is assumed that the number of voters is large, and  $p$  is the portion of voters who plan to vote for Candidate A.

We define the random variable  $X$  as follows. A voter is chosen uniformly at random among all voters and we ask her/him: “Do you plan to vote for Candidate A?” If she/he says “yes,” then  $X = 1$ , otherwise  $X = 0$ . Then  $X$  is a **Bernoulli random variable** with probability of success  $p$ .

We randomly select  $n$  voters (with replacement) and we ask each of them if they plan to vote for Candidate A. That is, we collect a random sample  $X_1, X_2, \dots, X_n$  be a random sample from the Bernoulli distribution with probability of success  $p$ .

We want to find a  $(1 - \alpha)100\%$  confidence interval for  $p$ .

# Confidence Interval of the proportion

We have that  $E[X_i] = p$  and  $Var(X_i) = \sigma^2 = p(1 - p)$ .

Thus, to find  $\sigma^2$  we need to know  $p$ , which is the quantity we want to estimate.

We will find an upper bound for  $\sigma^2$  by observing that the function  $f(p) = p(1 - p)$  satisfies

$$f(p) = p(1 - p) \leq f\left(\frac{1}{2}\right) = \frac{1}{4}, \quad p \in [0, 1]$$

Hence,  $\sigma^2 \leq \frac{1}{4}$  and  $\sigma \leq \sigma_{max} = \frac{1}{2}$ .

We conclude that

$$\left[ \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_{max}}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_{max}}{\sqrt{n}} \right] = \left[ \bar{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \bar{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}} \right]$$

is also a  $(1 - \alpha)100\%$  confidence interval for  $p$ .

# Confidence Interval of the proportion

## Example.

We randomly selected 100 voters (with replacement) and we ask each of them if they plan to vote for Candidate A. We found that 53% of respondents plan to vote for candidate A.

We want to find a 99% confidence interval for the proportion  $p$  of voters who plan to vote for candidate A

Since  $1 - \alpha = 0.99$ , then  $\alpha = 0.01$  and  $z_{0.005} = 2.576$ . Hence, the approximate 99% confidence interval for  $p$  is

$$\left[0.53 - \frac{2.576}{2\sqrt{100}}, 0.53 + \frac{2.576}{2\sqrt{100}}\right] = [0.40, 0.66]$$

If we choose a 95% confidence level, we get

$$\left[0.53 - \frac{1.96}{2\sqrt{100}}, 0.53 + \frac{1.96}{2\sqrt{100}}\right] = [0.43, 0.63]$$

## Confidence Interval of the proportion

**Example.** As above, we want to determine what proportion  $p$  of voters plan to vote for candidate A. We will choose a random sample (with replacement) of  $n$  voters and ask them if they plan to vote for Candidate A. Our goal is to estimate  $p$  so that the margin of error is 3 percentage points. That is, we would like to choose  $n$  such that

$$P(\bar{x} - 0.03, \bar{X} + 0.03) \geq 0.95$$

where  $\bar{x}$  is the proportion of people in our random sample that say they plan to vote for Candidate A. Assume a 95% confidence level.

Based on the above analysis, we need to choose  $n$  such that

$$\frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}} = 0.03$$

$$\text{Hence } n = \left( \frac{z_{\frac{\alpha}{2}}}{(2)(0.03)} \right)^2 = \left( \frac{1.96}{(2)(0.03)} \right)^2 = 1067.11$$

$$\text{If the confidence level is 99\%, then } n = \left( \frac{2.576}{(2)(0.03)} \right)^2 = 1843.27$$

## Confidence Interval of the mean [ $\sigma$ unknown, $n$ small]

In the above discussion, we assumed  $n$  to be large so that we could use the CLT.

We found that the confidence interval does not depend on the details of the distribution from which we obtained the random sample but only depended on statistics such as  $\bar{X}$  and  $S^2$ .

What if  $n$  is not large?

In this case, we cannot use the CLT, so we need to use the probability distribution from which the random sample is obtained. A very important case is when we have a sample  $X_1, X_2, \dots, X_n$  from a **normal distribution**.

Next, we discuss how to find interval estimators for the mean and the variance of a normal distribution when  $\sigma^2$  is unknown and  $n$  is not large.

# Chi-Squared Distribution

If  $Z_1, Z_2, \dots, Z_n$  are independent standard normal random variables, then the random variable

$$X = Z_1 + Z_2 + \dots + Z_n$$

is also normal. Specifically,  $X \sim N(0, n)$

If we define a random variable  $Y$  as

$$Y = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

then  $Y$  is said to have a **chi-squared distribution with  $n$  degrees of freedom**, which we denote as  $Y \sim \chi^2(n)$

# Chi-Squared Distribution

Properties of the chi-squared distribution:

- 1 It is a special case of the gamma distribution

$$f_Y(y) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad y > 0$$

- 2  $E[Y] = n$ ,  $Var(Y) = 2n$ .

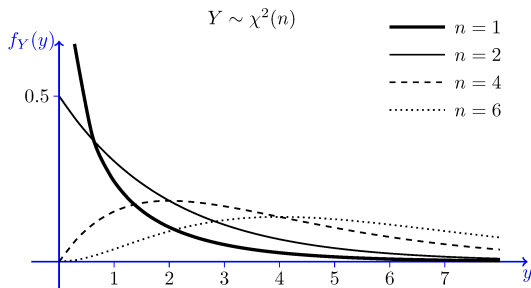


Figure: Plot of the pdf of the  $\chi^2(n)$  distribution for selected values of  $n$

# Chi-Squared Distribution

The chi-squared distribution arises in connection with the sample variance of the normal distribution.

**Theorem.** Let  $X_1, X_2, \dots, X_n$  be i.i.d.  $N(\mu, \sigma^2)$  random variables and let  $S^2$  be the sample variance for this random sample. Then, the random variable  $Y$  defined as

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

has a chi-squared distribution with  $n - 1$  degrees of freedom, i.e.,  $Y \sim \chi^2(n-1)$

Moreover,  $\bar{X}$  and  $S^2$  are independent random variables.



# Student's t-distribution

Let  $Z \sim N(0, 1)$  and  $Y \sim \chi^2(n)$ , where  $n \in \mathbb{N}$ .

Also assume that  $Z$  and  $Y$  are independent.

The random variable  $T$  defined as

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

has a **Student's t-distribution** (or simply t-distribution) with  $n$  degrees of freedom, which we denote by  $T \sim T(n)$ .

# Student's t-distribution

Properties of the Student's t-distribution:

- 1 The t-distribution has a bell-shaped curved centered at 0

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

As  $n \rightarrow \infty$ ,  $T(n) \rightarrow N(0, 1)$ .

- 2  $E[T] = 0$ , for  $n \geq 2$ .  $E[T]$  is undefined for  $n = 1$ .
- 3  $Var(T) = n - 2$ , for  $n > 2$ .  $Var(T)$  is undefined for  $n = 1, 2$ .

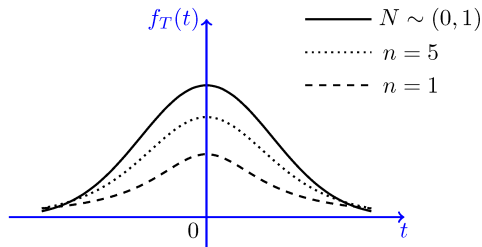


Figure: Plot of the pdf of t-distribution for some values of  $n$  compared with the standard normal pdf.

# Student's t-distribution

The t-distribution arises in connection with the sample mean of the normal distribution.

## Theorem

Let  $X_1, X_2, \dots, X_n$  be i.i.d.  $N(\mu, \sigma^2)$  random variables and let  $\bar{X}$ ,  $S^2$  be the sample mean and sample variance for this random sample, resp. Then, the random variable  $T$  defined as


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

has a t-distribution with  $n - 1$  degrees of freedom,  $T \sim T(n - 1)$ .

**Proof.** Define  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ . Then  $Z \sim N(0, 1)$ . Also define  $Y = \frac{(n-1)S^2}{\sigma^2}$ .

By our observations above,  $Y \sim \chi^2(n - 1)$ . It follows that

$$T = \frac{Z}{\sqrt{Y/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

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

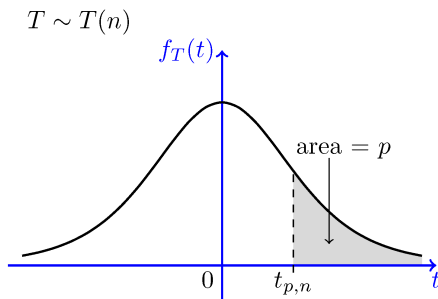
# Student's t-distribution

For any  $p \in [0, 1]$  and  $n \in \mathbb{N}$  we define  $t_{p,n}$  as the real value for which

$$P(T \geq t_{p,n}) = p$$

Since the t-distribution has a symmetric pdf, we have that

$$t_{1-p,n} = -t_{p,n}$$



# Confidence Interval for the mean of a normal distribution

Let  $X_1, X_2, \dots, X_n$  be i.i.d.  $N(\mu, \sigma^2)$  random variables.

Our goal is to find an interval estimator for  $\mu$ . We make no assumptions on  $n$ , that is,  $n$  can be any natural number.

There are two possible scenarios depending on whether  $\sigma^2$  is known or not.

If the value of  $\sigma^2$  is known, then the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

has  $N(0, 1)$  distribution.

It follows from our discussed above, we have that

$$\left[ \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

is a  $(1 - \alpha)100\%$  confidence interval for  $\mu$ .

## Confidence Interval for the mean of a normal distribution

If the value of  $\sigma^2$  is not known, then the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

where  $S^2$  is the sample variance for the random sample, has  $T(n-1)$  distribution.

To find a  $(1 - \alpha)100\%$  confidence interval for  $\mu$ , we solve

$$P(-t_{\frac{\alpha}{2}, n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha$$

which is equivalent to

$$P(\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}) = 1 - \alpha$$

That is, a  $(1 - \alpha)100\%$  confidence interval for  $\mu$  is

$$[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}]$$

# Confidence Interval for the mean of a normal distribution [ $\sigma$ unknowns]

## Theorem

Let  $X_1, X_2, \dots, X_n$  be a **random sample from a normal distribution**  $N(\mu, \sigma)$ , where  $\mu$  is unknown and  $\sigma$  is unknown. Let  $\bar{X}$  be the sample mean  $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_i$ . Then

$$[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}] \quad \text{or} \quad \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$$

is a  $(1 - \alpha)100\%$  **confidence interval for  $\mu$** .

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  is the sample variance of  $X$  and  $S = \sqrt{S^2}$  is the sample standard deviation of  $X$ .

$$t_{\frac{\alpha}{2}, n-1} = \text{qt}(1 - \frac{\alpha}{2}, \text{df} = n - 1)$$

## Confidence Interval for the mean - Example

**Example.** A farmer weights 10 randomly chosen watermelons from his farm and he obtains the following values (in lbs):

7.72, 9.58, 12.38, 7.77, 11.27, 8.80, 11.10, 7.80, 10.17, 6.00

Assuming that the weight is normally distributed, find a 95% confidence interval for the mean.

**Solution.** Using R, from the data we obtain

```
x <-c(7.72,9.58,12.38,7.77,11.27,8.8,11.10,7.8,10.17,6.0)
mean(x) = 9.259
```

```
var(x) = 3.961454
```

Hence, we set  $\bar{X} = 9.26$ ,  $S^2 = 3.96$  and, since the variance is not known, the statistics is  $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T(n-1)$ , with  $n = 10$ .

Using R, we find  $t_{0.025,9} = \text{qt}(0.975, 9) = 2.262$ . Thus, we obtain the 95% confidence interval for the mean

$$\left[ 9.26 - 2.26 \frac{\sqrt{3.96}}{\sqrt{10}}, 9.26 + 2.26 \frac{\sqrt{3.96}}{\sqrt{10}} \right] = [7.84, 10.68]$$



# Confidence Interval for the mean - Example

## Numerical solution using R. Case 1: $\sigma$ known.

Assume that we have collected a normal sample of size  $n = 20$  and found that the sample mean is 5. Assume we know the standard deviation is 2. We compute the 95% confidence interval using R.

Note:  $1 - \alpha = 0.95$ ,  $\alpha = 0.05$ ,  $\frac{\alpha}{2} = 0.025$ ,  $1 - \frac{\alpha}{2} = 0.975$

```
> xbar <- 5
> sigma <- 2
> n <- 20
> error <- qnorm(0.975)*sigma/sqrt(n)
> left <- xbar-error
> right <- xbar+error
> left
[1] 4.123477
> right
[1] 5.876523
```

# Confidence Interval for the mean - Example

## Numerical solution using R. Case 2: $\sigma$ unknown.

Assume that we have collected a normal sample of size  $n = 20$  and found that the sample mean is 5 and the sample variance is 4. We compute the 95% confidence interval using R.

Note:  $1 - \alpha = 0.95$ ,  $\alpha = 0.05$ ,  $\frac{\alpha}{2} = 0.025$ ,  $1 - \frac{\alpha}{2} = 0.975$

```
> xbar <- 5
> s <- 2
> n <- 20
> error <- qt(0.975,df=n-1)*s/sqrt(n)
> left <- xbar-error
> right <- xbar+error
> left
[1] 4.063971
> right
[1] 5.936029
```

## Confidence Interval for the mean - Example

**Numerical solution using R.** Case 2:  $\sigma$  unknown.

R has a command called `t.test` that computes the confidence interval for the mean after the significance test (that we will discuss later).

```
> watermelon =c(7.72, 9.58, 12.38, 7.77, 11.27, 8.80,  
11.10, 7.80, 10.17, 6.00)  
> t.test(watermelon,conf.level=0.95)
```

One Sample t-test

```
data: watermelon  
t = 14.711, df = 9, p-value = 1.336e-073  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
7.835196 10.682804  
sample estimates:  
mean of watermelon  
9.259
```

## Confidence Interval for the mean - Example

**Numerical solution using R.** Case 1:  $\sigma$  known.

R has a command `z.test` available with package `TeachingDemos`

Here we assume  $\sigma^2 = 3.96$

```
> watermelon =c(7.72, 9.58, 12.38, 7.77, 11.27, 8.80,  
11.10, 7.80, 10.17, 6.00)  
>stdev=sqrt(3.96)  
>library(TeachingDemos)  
>z.test(watermelon,mu=9,stdev,conf.level=0.95)
```

One Sample z-test

```
data:  watermelon  
z = 0.41158, n = 10.00000, Std. Dev. = 1.98997, Std.  
Dev. of the sample mean = 0.62929, p-value = 0.6806  
alternative hypothesis: true mean is not equal to 9  
95 percent confidence interval:  
8.025623 10.492377  
sample estimates:  
mean of watermelon = 9.259
```

## Confidence Interval for a proportion - Example

### **Numerical solution using R for a proportion.**

The R command `prop.test` can be used to construct confidence intervals for the normal approximation to the binomial.

Let us consider again the election poll example where 53 out of 100 respondents expressed intention to vote for candidate A.

```
> prop.test(53, 100, conf.level=0.95)
```

1-sample proportions test with continuity correction

data: 53 out of 100, null probability 0.5

X-squared = 0.25, df = 1, p-value = 0.6171

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

0.4280225 0.6296465

sample estimates:

p

0.53

# Confidence Interval for the difference of two means

The same method can be applied to compute confidence intervals for the difference of two means

## Theorem

Let  $X_1, X_2, \dots, X_n$  and  $Y_1, Y_2, \dots, Y_m$  be two independent random samples from two normal distribution  $N(\mu_X, \sigma_X)$  and  $N(\mu_Y, \sigma_Y)$ , where  $\mu_X, \mu_Y$  are unknown and  $\sigma_X, \sigma_Y$  are known.

Let  $\bar{X}$  and  $\bar{Y}$  be the sample means of  $X$  and  $Y$  respectively. Then

$$(\bar{X} - \bar{Y}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

is a  $(1 - \alpha)100\%$  **confidence interval** for  $\mu_X - \mu_Y$ .

Note that  $z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2}) = \text{qnorm}(1 - \frac{\alpha}{2})$

# Confidence Interval for the difference of two means

## Theorem

Let  $X_1, X_2, \dots, X_n$  and  $Y_1, Y_2, \dots, Y_m$  be two independent random samples from two normal distributions  $N(\mu_X, \sigma_X)$  and  $N(\mu_Y, \sigma_Y)$ , where  $\mu_X, \mu_Y$  are unknown and  $\sigma_X, \sigma_Y$  are unknown but equal. Let  $\bar{X}, \bar{Y}$  be the sample means and  $S_X^2, S_Y^2$  be the sample variances of  $X$  and  $Y$  respectively; also let  $S_p = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$  be the pooled standard deviation. Then

$$(\bar{X} - \bar{Y}) \pm t_{\frac{\alpha}{2}, n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

is a  $(1 - \alpha)100\%$  **confidence interval for  $\mu_X - \mu_Y$** .

Note that  $t_{\frac{\alpha}{2}, n+m-2} = \text{qt}(1 - \frac{\alpha}{2}, \text{df} = n + m - 2)$

# Confidence Interval for the difference of two means

## Theorem

Let  $X_1, X_2, \dots, X_n$  and  $Y_1, Y_2, \dots, Y_m$  be two independent random samples from two normal distributions  $N(\mu_X, \sigma_X)$  and  $N(\mu_Y, \sigma_Y)$ , where  $\mu_X, \mu_Y$  are unknown and  $\sigma_X, \sigma_Y$  are unknown but unequal. Let  $\bar{X}, \bar{Y}$  be the sample means and  $S_X^2, S_Y^2$  be the sample variances of  $X$  and  $Y$  respectively. Then

$$(\bar{X} - \bar{Y}) \pm t_{\frac{\alpha}{2}, r} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}},$$

where  $r = \left\lfloor \frac{(\frac{S_X^2}{n} + \frac{S_Y^2}{m})^2}{\frac{1}{n-1}(\frac{S_X^2}{n})^2 + \frac{1}{m-1}(\frac{S_Y^2}{m})^2} \right\rfloor$  and  $\lfloor \cdot \rfloor$  denotes the floor function, is a  $(1 - \alpha)100\%$  **confidence interval** for  $\mu_X - \mu_Y$ .

Note that  $t_{\frac{\alpha}{2}, r} = \text{qt}(1 - \frac{\alpha}{2}, \text{df} = r)$



# Confidence Interval for the difference of two means

**Example.** To study the genetic stability of a virus strain, two nucleotide sequences of virus strains from rats isolated in two different years were compared, yielding the following results:

$X$  : 12, 14, 16, 18, 14, 9, 16, 13

$Y$  : 11, 11, 12, 13, 13, 11, 13, 11, 12, 13

Assuming that the two sets are independent random samples from normal populations with equal variances, compute a 95% confidence interval for the difference between the mean number of substitutions in the nucleotide sequences.

## Confidence Interval for the difference of two means

**Example.** Here is the solution using R:

```
> x <-c(12,14,16,18,14,9,16,13)
> y <-c(11,11,12,13,13,11,13,11,12,13)
> t.test(x,y,conf.level=0.95,var.equal=TRUE)
```

Two Sample t-test

data: x and y

t = 2.1419, df = 16, p-value = 0.04793

alternative hypothesis: true difference in means is  
not equal to 0

95 percent confidence interval:

0.02055439 3.97944561

sample estimates:

mean of x mean of y

14 12

## Confidence Interval for the difference of two means

**Example.** Here is the solution using R if we assume the variances to be unequal:

```
> x <-c(12,14,16,18,14,9,16,13)
> y <-c(11,11,12,13,13,11,13,11,12,13)
> t.test(x,y,conf.level=0.95,var.equal=FALSE)
```

Welch Two Sample t-test

data: x and y

t = 1.9489, df = 8.2952, p-value = 0.08586

alternative hypothesis: true difference in means is  
not equal to 0

95 percent confidence interval:

-0.3519416 4.3519416

sample estimates:

mean of x mean of y

14 12

# Confidence Interval for the variance of a normal distribution

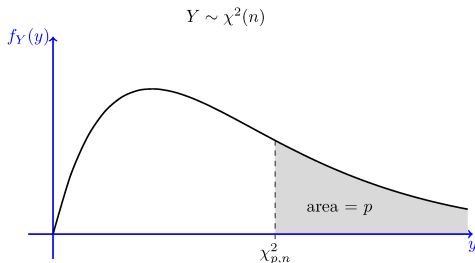
Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution  $N(\mu, \sigma^2)$ . Our goal is to find an interval estimator for  $\sigma^2$ . We assume that  $\mu$  is also unknown.

As noted above, the random variable

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

has a chi-squared distribution with  $n - 1$  degrees of freedom.

We define  $\chi_{p,n}^2$  as the real value for which  $P(Y > \chi_{p,n}^2) = p$ .



# Confidence Interval for the variance of a normal distribution

A  $(1 - \alpha)$  interval for  $Y = \frac{(n-1)S^2}{\sigma^2}$  is given by

$$P(\chi_{1-\frac{\alpha}{2}, n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\frac{\alpha}{2}, n-1}^2) = 1 - \alpha$$

which is equivalent to

$$P\left(\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2}\right) = 1 - \alpha$$

We conclude that

$$\left[ \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}, n-1}^2}, \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2} \right]$$

is a  $(1 - \alpha)100\%$  confidence interval for  $\sigma^2$ .

# Confidence Interval for the variance of a normal distribution

**Example.** A farmer weights 10 randomly chosen watermelons from his farm and he obtains the following values (in lbs):

7.72, 9.58, 12.38, 7.77, 11.27, 8.80, 11.10, 7.80, 10.17, 6.00

Assuming that the weight is normally distributed, find a 95% confidence interval for the variance.

**Solution.** Using the data, we obtain  $\bar{X} = 9.26$ ,  $S^2 = 3.96$ . Using the statistical tables or R, we find  $\chi^2_{0.025,9} = 19.02$ ,  $\chi^2_{0.975,9} = 2.70$ . Thus, we obtain the 95% confidence interval for the variance

$$\left[ \frac{(9)(3.96)}{19.02}, \frac{(9)(3.96)}{2.70} \right] = [1.87, 13.20]$$

# Confidence Interval for the variance of a normal distribution

## Example. Numerical solution using R.

```
> s2 <- 3.96
> n <- 10
> left <- s2*(n-1)/qchisq(0.975, n-1)
> right <- s2*(n-1)/qchisq(0.025, n-1)
> left
[1] 1.873544
> right
[1] 13.1981
```

Note:  $\chi^2_{0.025,9} = 19.02$ ,  $\chi^2_{0.975,9} = 2.70$

$\text{qchisq}(0.025, 9) = 2.700389$

$\text{qchisq}(0.975, 9) = 19.02277$

$\text{qchisq}(0.025, 9, \text{lower.tail} = \text{FALSE}) = 19.02277$

$\text{qchisq}(0.975, 9, \text{lower.tail} = \text{FALSE}) = 2.700389$

# Confidence Interval - Summary Tables

Table: Confidence Interval for the mean

Case	Confidence Interval
$X_i \sim N(\mu, \sigma^2), \sigma \text{ known}$	$[\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$
$n \text{ large}, \sigma \text{ known}$	$[\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$
$n \text{ large}, \sigma \text{ unknown}$	$[\bar{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}]$
$X_i \sim N(\mu, \sigma^2), \sigma \text{ unknown}$	$[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}]$

Table: Confidence Interval for the proportion

Case	Confidence Interval
$X_i \sim \text{Bernoulli}(p),$	$[\bar{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \bar{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}]$

Table: Confidence Interval for the variance

Case	Confidence Interval
$X_i \sim N(\mu, \sigma^2),$	$[\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}, n-1}^2}, \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2}]$



## 3.3 Hypothesis Testing

# Hypothesis Testing

Hypothesis testing is addressing the problem of testing whether a hypothesis is true or false.

For example, a pharmaceutical company might be interested in knowing if a new drug is effective in treating a disease. Here, there are two hypotheses:

- $H_0$  : the drug is not effective;
- $H_1$  : the drug is effective.

$H_0$  is called the **null hypothesis** and  $H_1$  is called the **alternative hypothesis**.

The null hypothesis,  $H_0$ , is usually referred to as the default hypothesis, i.e., the hypothesis that is initially assumed to be true. The alternative hypothesis,  $H_1$ , is the statement contradictory to  $H_0$ .

Based on the observed data, we need to decide either to **accept**  $H_0$ , or to **reject** it, in which case we say we accept  $H_1$ .

## Hypothesis Testing - Example

**Example.** You have a coin and you would like to check whether it is fair or not.

Let  $p$  be the probability of heads,  $p = P(H)$ . We set up the following problem:

$H_0$  : the coin is fair, i.e.,  $p = p_0 = \frac{1}{2}$ ;

$H_1$  : the coin is not fair, i.e.,  $p \neq \frac{1}{2}$ .

**Solution.** To check whether the coin is fair or not, we perform the following experiment. We toss the coin 100 times and record the number of heads. Let  $X$  be the number of heads that we observe, so  $X \sim \text{binom}(n = 100, p)$

If  $H_0$  is true, we expect the number of heads to be close to 50.

Hence we apply the following criteria: for a given **threshold**  $t$ ,

if  $|X - 50| \leq t$ , we accept  $H_0$ ;

if  $|X - 50| > t$ , we reject  $H_0$  and accept  $H_1$ .

How do we choose the threshold  $t$ ?

# Hypothesis Testing - Example

To determine the threshold  $t$ , we examine the probability of error.

**Type I error:** error incurred when we reject  $H_0$  while in fact it is true, that is  $|X - 50| > t$  when  $H_0$  is true.

Clearly, we want to control this error, so we want a test for which

$$P(\text{Type I error}) = P(|X - 50| > t \mid H_0 \text{ is true}) < \alpha,$$

where  $\alpha$  is the **significance level** (e.g.,  $\alpha = 0.05$ ).

Since we know the distribution of  $X$  under  $H_0$ , i.e.,  $X \mid H_0 \sim \text{binom}(n = 100, p_0 = 0.5)$ , we should be able to choose  $t$  such that the above condition holds.

## Hypothesis Testing - Example

By the Central Limit Theorem, we can approximate the binomial distribution using the normal distribution when  $n$  is large ( $n\hat{p}, n(1 - \hat{p})$  must be larger than 5). That is

$$Y = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{X - 50}{5} \sim N(0, 1).$$

Since  $P(\text{Type I error}) = P(|X - 50| > t \mid H_0 \text{ is true})$

$$= P\left(\frac{|X - 50|}{5} > \frac{t}{5} \mid H_0 \text{ is true}\right),$$

to determine  $t$ , we impose

$$\begin{aligned}\alpha &= P\left(\frac{|X - 50|}{5} > \frac{t}{5}\right) = 1 - P\left(\frac{|X - 50|}{5} \leq \frac{t}{5}\right) \\&= 1 - P\left(-\frac{t}{5} \leq \frac{X - 50}{5} \leq \frac{t}{5}\right) \\&= 2 - 2P\left(\frac{X - 50}{5} \leq \frac{t}{5}\right) \\&\approx 2 - 2\Phi\left(\frac{t}{5}\right)\end{aligned}$$

## Hypothesis Testing - Example

Hence, if we set the significance level  $\alpha = 0.05$ , we have

$$\begin{aligned}2 - 2\Phi\left(\frac{t}{5}\right) &= 0.05 &\Rightarrow &\Phi\left(\frac{t}{5}\right) = 0.975 \\&&\Rightarrow &\frac{t}{5} = \Phi^{-1}(0.975) = 1.960 \\&&\Rightarrow &t = 9.8\end{aligned}$$

It follows that

if  $|X - 50| \leq 9.8$ , we accept  $H_0$ ;

if  $|X - 50| > 9.8$ , we reject  $H_0$  and accept  $H_1$ .

That is, we reject  $H_0$  if  $X \geq 59.8$  or  $X \leq 40.2$ . Otherwise we do not reject  $H_0$ .

Note that failing to reject  $H_0$  does not imply that  $H_0$  is true. All we know is that our data are not statistically contradictory to  $H_0$ .

# Hypothesis Testing - General setting

Suppose that  $\theta$  is an unknown parameter of a distribution. Hypothesis testing is a method to decide between two contradictory hypotheses about  $\theta$  based on observed data.

There are 4 steps involved.

1. Specify the null and alternative hypotheses.

Letting  $S \in \mathbb{R}$  be the set of possible values for  $\theta$ , we partition  $S$  into two disjoint sets  $S_0$  and  $S_1$

$$H_0 : \theta \in S_0;$$

$$H_1 : \theta \in S_1.$$

# Hypothesis Testing - General setting

## 2. Compute the test statistic.

To decide between  $H_0$  and  $H_1$ , we look at a function of the observed data  $X_1, X_2, \dots, X_n$  (a random sample).

A **statistic** is a real-valued function  $W = W(X_1, X_2, \dots, X_n)$  of the data and a **test statistic** is a statistic based on which we build our test.

In our example above,  $W = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$  is a statistic, where  $X$  is the number of observed heads.

The sample mean  $W = \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$  or  $W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  are also statistics.



# Hypothesis Testing - General setting

## 3. Find the critical value of the test statistic.

We want to determine a set  $A$  of values of the test statistic  $W$  for which we would accept  $H_0$  (while assuming  $H_0$  is true).

$A$  is called the **acceptance region** and its complement  $R = A^c$  is the **rejection region**.

To determine  $A$ , we minimize the probability of **type I error**, i.e., the event that we reject  $H_0$  when  $H_0$  is true:

$$P(\text{type I error}) \leq \alpha \quad \text{for all } \theta \in S_0 \quad (1)$$

where  $\alpha$  is the significance level, typically  $\alpha = 0.10, 0.05$ , or  $0.01$ .

Condition (1) identifies a critical value  $c$  of the test statistic  $W$ .

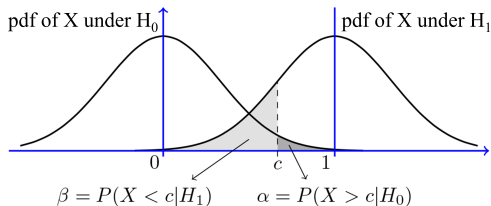
# Hypothesis Testing - General setting

## Remark.

In addition to the type I error, there is a **type II error** corresponding to the event that we accept  $H_0$  when  $H_0$  is false. The probability of type II error is

$$P(\text{type II error}) = \beta \quad \text{for } \theta \in S_1$$

We would ideally like both  $\alpha$  and  $\beta$  to be small. However, there is a trade-off between  $\alpha$  and  $\beta$ . That is, if we want to decrease the probability of type I error, then the probability of type II error increases, and vice versa.



# Hypothesis Testing - General setting

## 4. Compare the test statistic to the critical value.

If the test statistic is more extreme in the direction of the alternative than the critical value, reject the null hypothesis in favor of the alternative hypothesis.

If the test statistic is less extreme than the critical value, do not reject the null hypothesis.

The specific form of the comparison between the test statistic to the critical value depends on the type of problems as we show next.

# Hypothesis Testing for the Mean

We assume that we have a random sample  $X_1, X_2, \dots, X_n$  from a distribution and our goal is to make inference about the mean of the distribution  $\mu$ .

We consider three possible hypothesis testing problems.

- Case 1: two-sided test, two-tailed

$$H_0 : \mu = \mu_0;$$

$$H_1 : \mu \neq \mu_0;$$

- Case 2: one-sided test, left-tailed

$$H_0 : \mu \geq \mu_0;$$

$$H_1 : \mu < \mu_0;$$

- Case 3: one-sided test, right-tailed

$$H_0 : \mu \leq \mu_0;$$

$$H_1 : \mu > \mu_0;$$

# Hypothesis Testing for the Mean

In all of the three cases, we use the sample mean  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$  to define our statistic.

If we know the variance of the  $X_i$ 's,  $\text{Var}(X_i) = \sigma^2$ , then we define our test statistic as the normalized sample mean (assuming  $H_0$ ):

$$W = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \quad (1)$$

If the variance of the  $X_i$ 's is not known, then we define our test statistic as:

$$W = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \quad (2)$$

where  $S$  is the sample standard deviation.

Note: in case (1), if  $X_i$  is normal,  $W \sim N(0, 1)$ ; in case (2), if  $n$  is large,  $W \sim N(0, 1)$ ; if  $X_i$  is normal and  $n$  is small,  $W \sim T(n - 1)$ .

# Hypothesis Testing for the Mean

## Example (two-sided test)

The average adult male height in a certain country is 170 cm. We suspect that the men in a certain city in that country might have a different average height due to some environmental factors. We pick a random sample of size 9 from the adult males in the city and obtain the following values for their heights (in cm):

176.2, 157.9, 160.1, 180.9, 165.1, 167.2, 162.9, 155.7, 166.2

Assume that the height distribution in this population is normally distributed. Based on the observed data, is there enough evidence to support the hypothesis that the average population height is different from 170 cm? Assume a significance level  $\alpha = 0.05$ .

# Hypothesis Testing for the Mean

## Example - Solution.

(Step 1) We specify the hypothesis testing problem:

$$H_0 : \mu = 170;$$

$$H_1 : \mu \neq 170;$$

We need to determine if there enough evidence to reject  $H_0$  at significance level  $\alpha = 0.05$ ?

From the data, we obtain:

$$\bar{X} = \sum_{i=1}^9 X_i = 165.8, \quad S^2 = \frac{1}{8} \sum_{i=1}^9 (X_i - \bar{X})^2 = 68.01, \quad S = \sqrt{S^2} = 8.25.$$

(Step 2) We compute the test statistic:

$$W = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{165.8 - 170}{8.25/\sqrt{9}} = -1.52$$

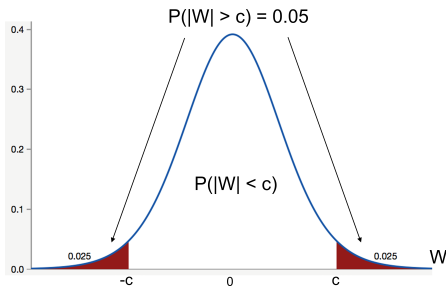
W follows a t-distribution with 8 degrees of freedom.

# Hypothesis Testing for the Mean

## Example - Solution.

(Step 3) To find the critical value  $c$  of the test statistic, we impose  $P(\text{type I error}) < 0.05$ . Hence we examine

$$P(|W| > c) = 0.05 \Rightarrow P(W > c) = 0.025$$



We find  $c = t_{0.025,8} = 2.31$ .

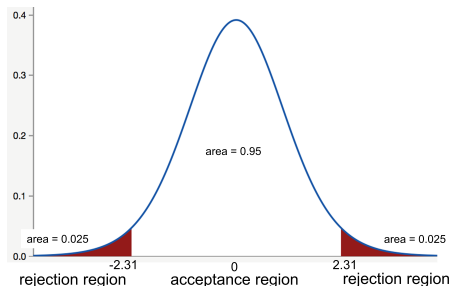
The rejection region is  $|W| > 2.31$ ,  
that is  $W < -2.31$  or  $W > 2.31$ .



# Hypothesis Testing for the Mean

## Example - Solution.

(Step 4) We compare the critical value  $c$  with the test statistic  $W$ .



The rejection region is  $W < -2.31$  or  $W > 2.31$ .

Since  $W = -1.52 > -2.31$ ,  $W$  is in the acceptance region and we cannot reject  $H_0$ .

That is, data do not support the hypothesis that the average population height under consideration is different from 170 cm.

# Hypothesis Testing for the Mean

## Remark. Relation to Confidence Intervals

Let us examine the acceptance region in the last Example.

We found that we accept  $H_0$  if

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| \leq t_{0.025,8} (= 2.31)$$

This is equivalent to say that

$$\mu_0 \in \left[ \bar{X} - t_{0.025,8} \frac{S}{\sqrt{n}}, \bar{X} + t_{0.025,8} \frac{S}{\sqrt{n}} \right]$$

This shows that the acceptance region is exactly the same as the confidence interval of the mean.

This is true in general for the two-sided hypothesis testing.

## Hypothesis Testing for the Mean

**Example (left-tailed test).** The average tar content (per cigarette) of a brand of cigarettes is 11.5 mg with  $\sigma = 0.6$  mg. A new filter is proposed which is claimed to reduce the average tar content. We consider a sample of  $n = 40$  randomly selected cigarettes with the new filter and found that the sample average tar content is  $\bar{x} = 11.4$ . Is there enough evidence to conclude that the new filter reduces the tar content? Assume a significance level  $\alpha = 0.10$ .

**Solution.**

(Step 1) We specify the hypothesis testing problem:

$$H_0 : \mu \geq 11.5;$$

$$H_1 : \mu < 11.5;$$

(Step 2) We compute the test statistic:

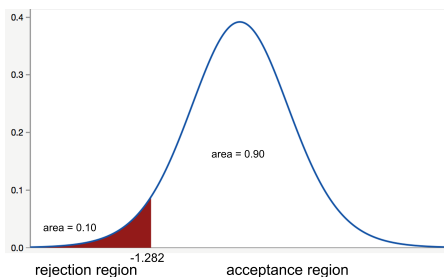
$$W = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{11.4 - 11.5}{0.6 / \sqrt{40}} = -1.05$$

W follows a standard normal distribution.

# Hypothesis Testing for the Mean

## Example - Solution.

(Step 3) To find the critical value  $c$  of the test statistic, we impose  $P(\text{type I error}) < 0.10$ . Hence we examine  $P(W < -c) = 0.10$ . We find  $-c = -z_{0.10} = -1.28$ .



(Step 4) The rejection region is  $W < -1.28$ . Since  $W = -1.05 > -1.28$ ,  $W$  is in the acceptance region and we cannot reject  $H_0$ .

# Hypothesis Testing for the Mean

**Example (right-tailed test).** A soft drink company claims that the average sodium content of a certain soda is 1.5 g per can with standard deviation  $\sigma = 0.20$ . From a random sample of 32 cans, we found that the sample average sodium content is  $\bar{x} = 1.6$  g. Is there enough evidence to conclude that the average sodium content per can is over 1.5g? Assume a significance level  $\alpha = 0.05$ .

**Solution.**

(Step 1) We specify the hypothesis testing problem:

$$H_0 : \mu \leq 1.5;$$

$$H_1 : \mu > 1.5;$$

(Step 2) We compute the test statistic:

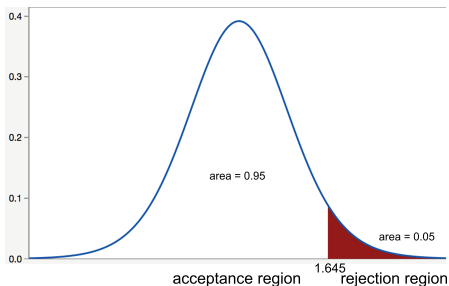
$$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.6 - 1.5}{0.2/\sqrt{32}} = 2.828$$

W follows a standard normal distribution.

# Hypothesis Testing for the Mean

## Example - Solution.

(Step 3) To find the critical value  $c$  of the test statistic, we impose  $P(\text{type I error}) < 0.05$ . Hence we examine  $P(W > c) = 0.05$ . We find  $c = z_{0.05} = 1.645$ .



(Step 4) The rejection region is  $W > 1.645$ . Since  $W = 2.282 > 1.645$ ,  $W$  is in the rejection region, hence we reject  $H_0$  and accept  $H_1$ .

# Hypothesis Testing for the Mean - Summary tables

**Table:** Two-sided hypothesis testing for the mean  $H_1 : \mu \neq \mu_0$

case	test statistic	rejection region
$X_i \sim N(\mu, \sigma^2), \sigma$ known	$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$ W  > z_{\frac{\alpha}{2}}$
$n$ large, $\sigma$ known or unknown	$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$ W  > z_{\frac{\alpha}{2}}$
$X_i \sim N(\mu, \sigma^2), \sigma$ unknown	$W = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$ W  > t_{\frac{\alpha}{2}, n-1}$

**Table:** One-sided hypothesis testing for the mean  $H_1 : \mu < \mu_0$

case	test statistic	rejection region
$X_i \sim N(\mu, \sigma^2), \sigma$ known	$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$W < -z_{\alpha}$
$n$ large, $\sigma$ known or unknown	$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$W < -z_{\alpha}$
$X_i \sim N(\mu, \sigma^2), \sigma$ unknown	$W = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$W < -t_{\alpha, n-1}$

**Table:** One-sided hypothesis testing for the mean  $H_1 : \mu > \mu_0$

case	test statistic	rejection region
$X_i \sim N(\mu, \sigma^2), \sigma$ known	$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$W > z_{\alpha}$
$n$ large, $\sigma$ known or unknown	$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$W > z_{\alpha}$
$X_i \sim N(\mu, \sigma^2), \sigma$ unknown	$W = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$W > t_{\alpha, n-1}$

# Hypothesis Testing for the proportion

The hypothesis testing for the proportion is carried out very similarly once we apply the Central limit theorem to approximate the binomial distribution using the normal distribution.

As we observed in the discussion of the sampling distribution, given  $Y \sim \text{binom}(n, p)$  with  $\mu_Y = np$  and  $\sigma_Y^2 = np(1 - p)$ , if  $np$  and  $n(1 - p) > 5$  we can approximate

$$Y \sim N(\mu = np, \sigma^2 = np(1 - p))$$

or

$$\frac{Y}{n} \sim N(\mu = p, \sigma^2 = p(1 - p)/n)$$



# Hypothesis Testing for the proportion

We have collected  $n$  Bernoulli trials and found  $y$  successes from a binomial pmf. Our goal to use  $\hat{p} = \frac{y}{n}$  to make an inference about the true proportion  $p$ .

As in the hypothesis testing for the mean, we have three cases

- Case 1: two-sided test, two-tailed

$$H_0 : p = p_0;$$

$$H_1 : p \neq p_0;$$

Rejection region

$$|W| > z_{\frac{\alpha}{2}}$$

- Case 2: one-sided test, left-tailed

$$H_0 : p \geq p_0;$$

$$H_1 : p < p_0;$$

$$W < -z_{\alpha}$$

- Case 3: one-sided test, right-tailed

$$H_0 : p \leq p_0;$$

$$H_1 : p > p_0.$$

$$W > z_{\alpha}$$

## Hypothesis Testing for the proportion

**Example [right-tailed].** We would like to check whether a coin is fair or biased. We toss the coin 100 times and observe 60 heads. Is there enough evidence to support the hypothesis that the coin is biased and that  $P(\text{head}) > \frac{1}{2}$ ? Use  $\alpha = 0.05$ .

**Solution.** We set up the upper-tailed hypothesis problem

$$H_0 : p \leq \frac{1}{2};$$

$$H_1 : p > \frac{1}{2}.$$

The observed proportion is  $\hat{p} = \frac{60}{100} = 0.6$ . Note that  $pn = 50$ ,  $(1 - p)n = 50 > 5$  so that we can apply the Central Limit Theorem. Hence, we apply the test statistic

$$W = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0, 1).$$

We find that  $W = \frac{0.6 - 0.5}{\sqrt{(0.5)(0.5)/100}} = 2$ .

Since  $W = 2 > z_{0.05} = 1.645$ , we reject the null hypothesis.

## Hypothesis Testing for the proportion

**Example [left-tailed].** Newborn babies are less likely to be girls than boys. A random sample found 12,295 girls were born among 25,468 newborn children. Is this sample evidence that the birth of girls is less common than the birth of boys in the entire population? Use  $\alpha = 0.001$ .

**Solution.** We set up the lower-tailed hypothesis problem

$$H_0 : p \geq \frac{1}{2};$$

$$H_1 : p < \frac{1}{2}.$$

The sample proportion of girls was 0.4828. Note that  $pn = (1 - p)n = 12734 > 5$  so that we can apply the Central Limit Theorem. Hence, we apply the test statistic

$$W = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.4828 - 0.5}{\sqrt{(0.5)(0.5)/25468}} = -5.490$$

Since  $W = -5.490 < -z_{0.001} = -3.090$ , we reject the null hypothesis.

# Hypothesis Testing for the Mean - Introduction to $p$ -value

**Example.** We want to test the hypothesis that the average weight of adult men (above 21 years old) in a certain state is more than 191 pounds. From a random sample data of size  $n = 100$ , we find the sample average weight  $\bar{x} = 197.1$  lb and  $s = 25.6$  lb. Assume a significance level  $\alpha = 0.05$ .

**Solution.** We test the hypothesis (right-tailed)

$$H_0 : \mu \leq 191;$$

$$H_1 : \mu > 191;$$

The test statistic is  $W = \frac{197.1 - 191}{25.6 / \sqrt{100}} = 2.38$ . Since  $n$  is large,  $W$  follows a standard normal distribution and  $z_{0.05} = 1.645$ .

According to the table,  $W = 2.38$  is contained in the rejection region  $W > 1.645$ , hence we reject  $H_0$  accept the alternative hypothesis that the average weight is above 191 lb.

# Hypothesis Testing for the Mean

In the last problem, we found that the value of the test statistic  $W = 2.38$  is in the rejection region as it satisfies  $W > z_{0.05} = 1.645$ .

Hence, we rejected  $H_0$  with confidence level  $\alpha = 0.05$

What if we set  $\alpha = 0.01$ ?

Since  $W > z_{0.01} = 2.326$ , again we can reject  $H_0$ .

What is the smallest  $\alpha$  for which we are able to reject  $H_0$ ?

For that we compute

$$P(Z > 2.38) = 1 - \Phi(2.38) = 1 - 0.99134 = 0.0087.$$

The last quantity is the  **$p$ -value** of the hypothesis problem.

The smaller the  $p$ -value, the more confident we are in rejecting the null hypothesis.

# Hypothesis Testing using $p$ -value

**Definition.** The  $p$ -value is the lowest significance level  $\alpha$  that results in rejecting the null hypothesis.

If the  $p$ -value is small, it means that the observed data is very unlikely to have occurred under  $H_0$ , so we are more confident in rejecting the null hypothesis. The smaller the  $p$ -value, the more confident we are in rejecting  $H_0$ .

To compute  $p$ -values:

- Let  $W$  be the test statistic of a hypothesis testing problem and  $w_1$  be the observed value of  $W$ .
- The  $p$ -value is  $P(\text{type I error})$  when the test threshold  $c$  is chosen to be  $c = w_1$ , under the assumption that  $H_0$  is true.

# Hypothesis Testing using $p$ -value

In practice, to compute the  $p$ -value we first compute  $w_1$ , the observed value of  $W$ .

Next, if  $W$  satisfies a standard normal distribution:

- 2-sided test:  $p\text{-value} = 2\Phi(-|w_1|) = 2 * \text{pnorm}(-|w_1|)$
- 1-sided, left-tailed test:  $p\text{-value} = \Phi(w_1) = \text{pnorm}(w_1)$
- 1-sided, right-tailed test:  $p\text{-value} = 1 - \Phi(w_1) = 1 - \text{pnorm}(w_1)$

If  $W$  satisfies a t-distribution with  $n - 1$  degrees of freedom (df):

- 2-sided test:  $p\text{-value} = 2P(T < -|w_1|) = 2 * \text{pt}(-|w_1|, \text{df})$
- 1-sided, left-tailed test:  $p\text{-value} = P(T < w_1) = \text{pt}(w_1, \text{df})$
- 1-sided, right-tailed test:  $p\text{-value} = 1 - P(T < w_1) = 1 - \text{pt}(w_1, \text{df})$

Recall that  $\Phi$  is the standard cumulative normal distribution:

$$\Phi(y) = P(Z \leq y)$$

# Hypothesis Testing - Summary

**Hypothesis testing:** method to make decisions about the values of parameters of a distribution, e.g., a **mean** or a **proportion**.

1) State the hypothesis.

Three cases:

- ①  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$  (two-side test)
- ②  $H_0 : \theta \geq \theta_0$  versus  $H_1 : \theta < \theta_0$  (one-side test, lower tail)
- ③  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$  (one-side test, upper tail)

2) Choose the test statistic.

For the hypothesis testing of the **mean**, it is

$$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \text{ or } W = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \quad (\text{depending if you have } \sigma \text{ or } s)$$

For the hypothesis testing of the **proportion**, it is

$$W = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$



# Hypothesis Testing - Summary

3) Compute the test statistic and identify the probability model.

- For the hypothesis testing of the **mean**,

- ① if  $\sigma$  is known and data are normal, then  $W \sim N(0, 1)$

- ② if  $\sigma$  is unknown and data are normal, then  $W \sim T(n - 1)$

- ③ if  $n > 30$ , then  $W \sim N(0, 1)$ .

- If  $W \sim N(0, 1)$ , you apply a "z-test", that is, the critical value ( $z_{\alpha/2}$  or  $z_{\alpha}$  or  $-z_{\alpha}$ ) is computed using the standard normal distribution.

- If  $W \sim T(n - 1)$ , you apply a "t-test", that is, the critical value ( $t_{\alpha/2, n-1}$  or  $t_{\alpha, n-1}$  or  $-t_{\alpha, n-1}$ ) is computed using the Student t distribution.

- For the hypothesis testing of the **proportion**,

- if  $np > 5$  and  $n(1 - p) > 5$ , then  $W \sim N(0, 1)$ .

# Hypothesis Testing - Summary

4) Compute the p value.

We first compute  $w_1$ , the observed value of  $W$ .

- If  $W \in N(0, 1)$  and

- 2-sided test:  $p\text{-value} = 2 * \text{pnorm}(-|w_1|)$
- 1-sided, lower tail:  $p\text{-value} = \text{pnorm}(w_1)$
- 1-sided, upper tail:  $p\text{-value} = 1 - \text{pnorm}(w_1)$

- If  $W \sim T(n - 1)$  and

- 2-sided test:  $p\text{-value} = 2 * \text{pt}(-|w_1|, n - 1)$
- 1-sided, lower tail:  $p\text{-value} = \text{pt}(w_1, n - 1)$
- 1-sided, upper tail:  $p\text{-value} = 1 - \text{pt}(w_1, n - 1)$

Conclusion:

If  $p\text{-value} < \alpha$ , then  $H_0$  is rejected. Otherwise,  $H_0$  is accepted.

## Hypothesis Testing using $p$ -value

**Example.** The average tar content (per cigarette) of a brand of cigarettes is 11.5 mg with  $\sigma = 0.6$  mg. A new filter is proposed which is claimed to reduce the average tar content. We consider a sample of  $n = 40$  randomly selected cigarettes with the new filter and found that the sample average tar content is  $\bar{x} = 11.4$ . Is there enough evidence to conclude that the new filter reduces the tar content at significance level  $\alpha = 0.05$ ?

**Solution.** We have the left-tailed problem

$$H_0 : \mu \geq 11.5;$$

$$H_1 : \mu < 11.5;$$

We compute the test statistic:  $W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{11.4 - 11.5}{0.6/\sqrt{40}} = -1.05$

Since  $n > 30$ , we can assume  $W \sim N(0, 1)$  and we apply a  $z$  test.

$p\text{-value} = P(Z < -1.05) = \text{pnorm}(-1.05) = 0.147$ .

Since  $p$  value is larger than  $\alpha$ , we cannot reject  $H_0$ .

## Hypothesis Testing using $p$ -value

**Example.** We want check whether a coin is fair or biased. We toss the coin 100 times and observe 60 heads. Is there enough evidence to support the hypothesis that the coin is biased and that  $P(\text{head}) > \frac{1}{2}$ ? Use  $\alpha = 0.05$

**Solution.** We set up the upper-tailed hypothesis problem

$$H_0 : p \leq \frac{1}{2};$$

$$H_1 : p > \frac{1}{2}.$$

The observed proportion is  $\hat{p} = \frac{60}{100} = 0.6$ . Note that  $pn = 50$ ,  $(1 - p)n = 50 > 5$  so that we can apply the Central Limit Theorem. Hence, we apply the test statistic

$$W = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0, 1).$$

We find that  $W = \frac{0.6 - 0.5}{\sqrt{(0.5)(0.5)/100}} = 2.$

Since  $p\text{-value} = 1 - \text{pnorm}(2) = 0.023 < \alpha$ , then we reject  $H_0$ .

# Hypothesis Testing using R

The commands to perform hypothesis testing in R are

```
t.test(y, mu = mu0, alternative = "greater",  
conf.level = 0.95)
```

```
prop.test(x = x0, n = n0, p = pNULL, alternative =  
"greater", correct = FALSE)
```

There are 3 possible alternatives: "two.sided", "less", or "greater" depending on the type of test

There is no built-in command to perform the one-sample z test. However you can load the library TeachingDemos to use the command

```
z.test(y, mu = mu0, stdev, alternative = "greater",  
conf.level = 0.95)
```

## Hypothesis Testing for the Mean using R

**Example.** We want to test the hypothesis that the average weight of adult men (above 21 years old) in a certain state is more than 191 lb. From a random sample data of size  $n = 100$ , we find the sample average weight  $\bar{x} = 197.1$  lb and  $s = 25.6$  lb. Set  $\alpha = 0.01$

Since  $n > 30$ , we can assume  $W \sim N(0, 1)$  and apply a z test.

```
# Set the values of sample mean and total sample size
```

```
xbar<-197.1
```

```
n<-100
```

```
# Set mean value under the null hypothesis
```

```
mu0<-191
```

```
# Set known S
```

```
S<-25.6
```

```
# Calculate test statistic and p-values
```

```
z<-sqrt(n)*(xbar-mu0)/S = 2.382812
```

```
# p-value for the right-tailed case
```

```
p_value_righttail=1-pnorm(z)= 0.008590471
```

Since  $p\text{-value} < \alpha$ , then we reject  $H_0$ .

# Hypothesis Testing for the Mean using R

**Example.** The average adult male height in a country is 170 cm. We pick a random sample of size 9: 176.2, 157.9, 160.1, 180.9, 165.1, 167.2, 162.9, 155.7, 166.2. Assuming the height is normally distributed, test the hypothesis that the average height is different from 170 cm using  $\alpha = 0.05$ .

Solution:

Since  $\sigma$  is unknown and data are normal, we will apply a t-test. Note that  $n < 30$  in this case, so we are not allowed to apply the CLT in this case.

# Hypothesis Testing for the Mean using R

```
> y = c(176.2,157.9,160.1,180.9,165.1,167.2,162.9,155.7,166.2)
> t.test(y, mu = 170, alternative = "two.sided",
conf.level = 0.95)
One Sample t-test
data:  y
t = -1.5278, df = 8, p-value = 0.1651
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
159.4608 172.1392
sample estimates:
mean of x
165.8
```

Since  $p\text{-value} > \alpha$ , then we cannot reject  $H_0$ .



# Hypothesis Testing for the Mean using R

**Example.** We want to test the hypothesis that the mean systolic blood pressure in a certain population equals 140 mmHg using  $\alpha = 0.01$ . We can assume that the population is normally distributed. We collect 55 random samples:

120,115,94,118,111,102,102,131,104,107,115,139,115,113,114,105,  
115,134,109,109,93,118,109,106,125,150,142,119,127,141,149,144,  
142,149,161,143,140,148,149,141,146,159,152,135,134,161,130,125,  
141,148,153,145,137,147,169.

Solution:

Since  $\sigma$  is unknown and data are normal, we will apply a t-test.

Note that  $n > 30$  in this case, so we are also allowed to apply a z-test in this case.

We start by creating a data vector in R

```
> x <- c(120,115,94,118,111,102,102,131,104,107,115,139,  
115,113,114,105,115,134,109,109,93,118,109,106,125,150,142,  
119,127,141,149,144,142,149,161,143,140,148,149,141,146,159,  
152,135,134,161,130,125,141,148,153,145,137,147,169)
```

# Hypothesis Testing for the Mean using R

We run a two-sided test:

```
> t.test(x, mu = 140, alternative = "two.sided",  
conf.level = 0.95)
```

One Sample t-test

data: x

t = -3.8693, df = 54, p-value = 0.0002961

alternative hypothesis: true mean is not equal to  
140

95 percent confidence interval:

124.8185 135.1815

sample estimates:

mean of x

130

Since  $p\text{-value} < \alpha$ , then we reject  $H_0$ .

# Hypothesis Testing for the Mean using R

If we want to test a one-sided lower-tailed test instead:

```
> t.test(x, mu = 140, alternative = "less",  
conf.level = 0.95)
```

One Sample t-test

data: x

t = -3.8693, df = 54, p-value = 0.0001481

alternative hypothesis: true mean is less than 140

95 percent confidence interval:

-Inf 134.3253

sample estimates:

mean of x

130

Again, since  $p\text{-value} < \alpha$ , then we reject  $H_0$ .

# Hypothesis Testing for the Mean using R

Suppose we know the standard deviation to be  $\sigma = 20$ .  
In this case, we need to run a z test.

```
# Calculate sample mean and total sample size
xbar<-mean(x)
n<-55
# Set mean value under the null hypothesis
mu0<-140
# Set known sigma
sigma<-20

# Calculate test statistic and p-values
z<-sqrt(n)*(xbar-mu0)/sigma = -3.708099
p_value_2tail=2*pnorm(-abs(z)) = 0.0002088208
p_value_lefttail=pnorm(z) = 0.0001044104
```

Again, since  $p\text{-value} < \alpha$ , then we reject  $H_0$ .

## Hypothesis Testing for the Mean using R

We can solve the last problem using `z.test` from the library `TeachingDemos`

```
> library(TeachingDemos)
> stdev=20
> z.test(x, mu = 140, stdev,alternative = "less",
conf.level = 0.95)
```

One Sample z-test

data: x

z = -3.7081, n = 55.0000, Std. Dev. = 20.0000, Std.  
Dev. of the sample

mean = 2.6968, p-value = 0.0001044

alternative hypothesis: true mean is less than 140

95 percent confidence interval:

-Inf 134.4358

sample estimates:

mean of x

130

## Hypothesis Testing for the proportion using R

**Example.** We would like to check whether a coin is fair or biased. We toss the coin 100 times and observe 60 heads. Is there enough evidence to support the hypothesis that the coin is biased and that  $P(\text{head}) > \frac{1}{2}$  (use  $\alpha = 0.05$ )?

```
> prop.test(x = 60, n = 100, p = 0.5, , alternative =  
"greater", correct = FALSE)
```

1-sample proportions test without continuity correction

data: 60 out of 100, null probability 0.5

X-squared = 4, df = 1, p-value = 0.02275

alternative hypothesis: true p is greater than 0.5

95 percent confidence interval:

0.5178095 1.0000000

sample estimates:

p

0.6

Since  $p\text{-value} < \alpha$ , then we reject  $H_0$ .

# Hypothesis Testing: Power of the test

Consider an hypothesis testing problem where the p-value computed from the data was 0.12. As a result, one would fail to reject the null hypothesis because  $0.12 > 0.05$ .

However, there still exist two possible cases for which we failed to reject the null hypothesis  $H_0$ :

- 1  $H_0$  is a reasonable conclusion;
- 2 the sample size is not large enough to either accept or reject  $H_0$ , i.e., additional samples might provide additional evidence.

**Power analysis** is a procedure to determine if the test contains enough power to make a reasonable conclusion.

The **power of a hypothesis test** is the probability that the test rejects the null hypothesis  $H_0$  when a specific alternative hypothesis  $H_1$  is true - i.e., it indicates the probability of avoiding a type II error.  $\text{Power} = 1 - P(\text{type II error}) = 1 - \beta$ .

Recall: the type II error corresponds to the event that we accept  $H_0$  when  $H_0$  is false and  $H_1$  is true.

# Hypothesis Testing: Power of the test

**Example.** Let  $X$  denote the height of a randomly selected UH student. Assume that  $X$  is normally distributed with unknown mean and standard deviation  $\sigma = 9$ .

We collect a random sample of  $n = 25$  students, so that, after setting the probability of committing a Type I error at  $\alpha = 0.05$ , we can test the hypothesis testing problem

$$H_0 : \mu = 170;$$

$$H_1 : \mu > 170.$$

What is the power of the hypothesis test if the true population mean were  $\mu = 175$ ?



## Hypothesis Testing: Power of the test

The 90% confidence interval of the mean about  $\bar{x} = 170$  is

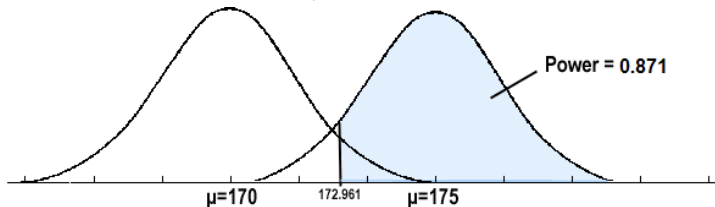
$$170 \pm 1.645 \frac{9}{\sqrt{25}} = 170 \pm 2.961$$

So, we should reject  $H_0$  when the observed sample mean is larger or equal than 172.961.

Power of the test

$$P(\bar{x} > 172.961 | \mu = 175) = P(Z \geq \frac{172.961 - 175}{9/\sqrt{5}}) = 0.8713$$

We have a 87.13% chance of rejecting  $H_0$  in favor of  $H_1$  if the true unknown population mean is  $\mu = 175$ .



# Hypothesis Testing: Power of the test

If the sample size is fixed, as shown by the plot, decreasing Type I error will increase Type II error.

To decrease both errors, then one has to increase the sample size.

To calculate the smallest sample size needed for specified  $\alpha$ ,  $\beta$ ,  $\mu_1$ , ( $\mu_1$  is the likely value of at which you want to evaluate the power) then we choose

$$\text{One-tailed test: } n = \frac{\sigma^2(z_\alpha + z_\beta)^2}{(\mu_0 - \mu_1)^2}$$

$$\text{Two-tailed test: } n = \frac{\sigma^2(z_{\alpha/2} + z_\beta)^2}{(\mu_0 - \mu_1)^2}$$

# Hypothesis Testing: Power of the test

**Example.** Let  $X$  denote the height of a randomly selected UH student. Assume that  $X$  is normally distributed with unknown mean and standard deviation  $\sigma = 9$ . We are interested in solving at significance level  $\alpha = 0.05$  the hypothesis testing problem

$$H_0 : \mu = 170;$$

$$H_1 : \mu > 170.$$

Find the sample size  $n$  that is necessary to achieve 0.90 power at the alternative  $\mu_1 = 175$ .

In this case,  $z_\alpha = z_{0.05} = 1.645$ ,  $z_\beta = z_{0.10} = 1.28$ . Then

$$n = \frac{9^2(1.645 + 1.28)^2}{(170 - 175)^2} = 27.72$$

Hence we need to choose  $n = 28$  to achieve the desired level of  $\alpha$  and  $\beta$  when we choose  $\mu_1 = 175$  as alternative value of  $\mu$ .

# Hypothesis Testing: Power of the test

We now compute the critical value  $C$  for the test, and state an appropriate decision rule. To find  $C$ , we may substitute known numerical values into either

$$C = 170 + 1.645 \frac{9}{\sqrt{28}} = 172.798$$

or

$$C = 175 - 1.280 \frac{9}{\sqrt{28}} = 172.823$$

The difference is due to rounding error.

The new decision rule is now as follows. Select a sample of size  $n = 28$  and compute  $\bar{X}$ . If  $\bar{X} \geq 172.798$ , then  $H_0$  is rejected. If  $\bar{X} < 172.798$ , we do not reject  $H_0$ .

# Hypothesis Testing: Two sample means

In many hypothesis testing problems, one wants to compare two treatments or populations and determine if there is a difference.

It is important to be able to distinguish between an independent sample or a dependent sample.

**Independent sample.** The samples from two populations are independent if the samples selected from one of the populations have no relationship with the samples selected from the other population.

**Dependent sample.** The samples are dependent (also called paired data) if each measurement in one sample is matched or paired with a particular measurement in the other sample. Another way to consider this is how many measurements are taken off of each subject. If only one measurement, then independent; if two measurements, then paired.

# Hypothesis Testing: Two sample means

The following are examples to illustrate the two types of samples.

**Example: Gas Mileage.** We want to compare the gas mileage of two brands of gasoline.

Independent samples: Randomly assign 12 cars to use Brand A and another 12 cars to use Brand B.

Dependent samples: Using 12 cars, have each car use Brand A and Brand B. Compare the differences in mileage for each car.

**Example: Soft drink comparison.** We want to compare whether people give a higher taste rating to Coke or Pepsi. To avoid a possible psychological effect, the subjects should taste the drinks blind (i.e., they don't know the identity of the drink)

Independent samples: Randomly assign half of the subjects to taste Coke and the other half to taste Pepsi.

Dependent samples: Allow each subject to rate both Coke and Pepsi with the drinks given in random order. The same subject's ratings of the Coke and the Pepsi form a paired data set.

# Hypothesis Testing: Two sample means (independent)

We assume that we have 2 independent random samples

$X_1, \dots, X_{n_1}$  from a distribution with mean  $\mu_1$  and variance  $\sigma_1^2$  and

$Y_1, \dots, Y_{n_2}$  from a distribution with mean  $\mu_2$  and variance  $\sigma_2^2$ .

Our goal is to make inference about the mean of the distributions  $\mu_1$  and  $\mu_2$ .

We consider three possible hypothesis testing problems.

- Case 1: two-sided test, two-tailed

$$H_0 : \mu_1 = \mu_2;$$

$$H_1 : \mu_1 \neq \mu_2;$$

- Case 2: one-sided test, left-tailed

$$H_0 : \mu_1 \geq \mu_2;$$

$$H_1 : \mu_1 < \mu_2;$$

- Case 3: one-sided test, right-tailed

$$H_0 : \mu_1 \leq \mu_2;$$

$$H_1 : \mu_1 > \mu_2;$$

# Hypothesis Testing: Two sample means (independent)

To solve the hypothesis testing problem, we proceed as in the one-sample case by introducing a test statistic.

If  $n_1, n_2$  are large (above 30), by the central limit theorem we have that

$$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \quad \bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

The statement above also holds for any  $n_1, n_2$  if the samples are taken from normal distributions.

We define the test statistic as

$$W = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under the null hypothesis that  $\mu_1 = \mu_2$ , we have that  $W \sim N(0, 1)$

Note: if  $\sigma_1$  and  $\sigma_2$  are unknown, they can be replaced by their estimates  $s_1$  and  $s_2$ .



# Hypothesis Testing: Two sample means (independent)

Once we have computed the value of test statistic  $W$ , we can apply the z-test method as in the one-sample case.

- Case 1:  $H_0 : \mu_1 = \mu_2$ ;  $H_1 : \mu_1 \neq \mu_2$

Reject  $H_0$  if  $|W| > z_{\frac{\alpha}{2}}$

- Case 2:  $H_0 : \mu_1 \geq \mu_2$ ;  $H_1 : \mu_1 < \mu_2$

Reject  $H_0$  if  $W < -z_{\alpha}$

- Case 3:  $H_0 : \mu_1 \leq \mu_2$ ;  $H_1 : \mu_1 > \mu_2$

Reject  $H_0$  if  $W > z_{\alpha}$

# Hypothesis Testing: Two sample means (independent)

**Example.** We want to compare the Systolic Blood Pressure in two random samples of adult men and adult women:

$$n_1 = 1,623, \bar{X} = 128.2, S_1 = 17.5; n_2 = 1,911, \bar{Y} = 126.5, S_2 = 20.1$$

Use  $\alpha = 0.05$ .

We test the hypothesis

$$H_0 : \mu_1 = \mu_2;$$

$$H_1 : \mu_1 \neq \mu_2;$$

Test statistic

$$W = \frac{128.2 - 126.5}{\sqrt{\frac{17.5^2}{1622} + \frac{20.1^2}{1910}}} = 2.688$$

Since  $W > z_{0.025} = 1.960$ , we reject  $H_0$ .

Also,  $p\text{-value} = 2\Phi(-2.688) = 0.00718814$

## Hypothesis Testing: Two sample means

If the sample size is small, we cannot apply the central limit theorem.

In this case, under the assumptions that the two populations are normally distributed, we introduce the test statistic

$$W = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where  $s_p$  is the pooled standard deviation

$$s_p = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

and  $W$  obeys a t-distribution with  $n_1 + n_2 - 2$  degrees of freedom.

When the sample sizes are small, the sample variances may not be that accurate and one gets a better estimate by pooling the data from both populations.

# Hypothesis Testing: Two sample means (independent)

In this case, we apply the t-test method as in the one-sample case.

- Case 1:  $H_0 : \mu_1 = \mu_2$ ;  $H_1 : \mu_1 \neq \mu_2$

Reject  $H_0$  if  $|W| > t_{\frac{\alpha}{2}, n_1+n_2-2}$

- Case 2:  $H_0 : \mu_1 \geq \mu_2$ ;  $H_1 : \mu_1 < \mu_2$

Reject  $H_0$  if  $W < -t_{\alpha, n_1+n_2-2}$

- Case 3:  $H_0 : \mu_1 \leq \mu_2$ ;  $H_1 : \mu_1 > \mu_2$

Reject  $H_0$  if  $W > t_{\alpha, n_1+n_2-2}$

# Hypothesis Testing: Two sample means

**Example.** A new drug is proposed to lower total cholesterol. A randomized controlled trial is designed to evaluate the efficacy of the medication in lowering cholesterol. Thirty participants are enrolled in the trial and are randomly assigned (without knowing) to receive either the new drug or a placebo. Each participant is asked to take the assigned treatment for 6 weeks. At the end of 6 weeks, each patient's total cholesterol level is measured and the sample statistics are as follows.

New Drug  $n_1 = 15$ ,  $\bar{x} = 195.9$ ,  $s_1 = 28.7$

Placebo  $n_2 = 15$ ,  $\bar{y} = 227.4$ ,  $s_2 = 30.3$

We set up hypotheses:  $H_0 : \mu_1 = \mu_2$ ;  $H_1 : \mu_1 < \mu_2$ . Set  $\alpha = 0.05$

We compute  $s_p = 29.5$  and  $W = \frac{195.9 - 227.4}{29.5 \sqrt{\frac{1}{15} + \frac{1}{15}}} = -2.92$

We find that  $W = -2.92 < -1.701 = -t_{0.05, 28}$  ( $-\text{qt}(0.95, 28)$ )

Thus we reject  $H_0$  and accept  $H_1$ .

$p\text{-value} = P(W < -2.92) = 0.0034$  ( $\text{pt}(-2.92, 28)$ )

# Hypothesis Testing: Two sample means

Solution using R.

```
t.test(x, y, alternative = "two.sided", paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

`alternative`: it specifies the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less".

`paired`: a logical indicating whether you want a paired t-test or not. If `paired = FALSE`, the two samples are assumed to be independent

`var.equal`: a logical variable indicating whether to treat the two variances as being equal. If `TRUE` then the pooled variance is used to estimate the variance otherwise the Welch approximation to the degrees of freedom is used

`conf.level`: confidence level of the interval.

# Hypothesis Testing: Two sample means

**Example.** Two rubber compounds were tested for tensile strength and the following values were found

$x$  :32; 30; 33; 32; 29; 34; 32

$y$  :33; 35; 36; 37; 35

Under the assumption that the data are drawn from normal distributions, test the hypothesis that the average tensile strength of the two rubber compounds is different.

We test the hypothesis

$$H_0 : \mu_x = \mu_y;$$

$$H_1 : \mu_x \neq \mu_y.$$

# Hypothesis Testing: Two sample means

Assume same variance.

```
> x <- c(32, 30, 33, 32, 29, 34, 32)
> y <- c(33, 35, 36, 37, 35)
> t.test(x,y,alternative = "two.sided", paired = FALSE,
var.equal = TRUE, conf.level = 0.95)
```

Two Sample t-test

data: x and y

t = -3.6758, df = 10, p-value = 0.004276

alternative hypothesis: true difference in means is not  
equal to 0

95 percent confidence interval:

-5.598649 -1.372779

sample estimates:

mean of x mean of y

31.71429 35.20000



# Hypothesis Testing: Two sample means

Assume different variance.

```
> x <- c(32, 30, 33, 32, 29, 34, 32)
> y <- c(33, 35, 36, 37, 35)
> t.test(x,y,alternative = "two.sided", paired = FALSE,
var.equal = FALSE, conf.level = 0.95)
```

Welch Two Sample t-test

data: x and y

t = -3.7698, df = 9.4808, p-value = 0.004025

alternative hypothesis: true difference in means is not  
equal to 0

95 percent confidence interval:

-5.561344 -1.410084

sample estimates:

mean of x mean of y

31.71429 35.20000

# Hypothesis Testing: Two sample means

Assume same variance.  $H_1 : \mu_x < \mu_y$

```
> x <- c(32, 30, 33, 32, 29, 34, 32)
```

```
> y <- c(33, 35, 36, 37, 35)
```

```
> t.test(x,y,alternative = "less", paired = FALSE,  
var.equal = TRUE, conf.level = 0.95)
```

Two Sample t-test

data: x and y

t = -3.6758, df = 10, p-value = 0.002138

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -1.766965

sample estimates:

mean of x mean of y

31.71429 35.20000

# Hypothesis Testing: Paired-sample t test

A paired t-test is used to compare two population means where you have two samples of the same size in which observations in one sample can be paired with observations in the other sample.

Examples of where this might occur are: Before-and-after observations on the same subjects, such as a medical evaluation or the performance of a student/worker before and after taking a course.

In a typical experiment, we select a random sample consisting of  $n$  pairs of observations  $(X_1, Y_1), \dots, (X_n, Y_n)$ , where each observation  $(X_i, Y_i)$  is a pair and the variables  $X_i, Y_i$  are not independent.

We define the differences  $D_i = X_i - Y_i$ ,  $i = 1, \dots, n$ . The random variables  $D_1, \dots, D_n$  are a random sample of size  $n$  from a distribution with mean  $\mu_1 - \mu_2$  and variance  $\sigma_D^2$ .

# Hypothesis Testing: Paired-sample t test

Under the assumption that the distribution of the variable  $D_i$  is normal, we define the test statistic

$$W = \frac{\bar{D}}{S_D / \sqrt{n}},$$

where  $\bar{D}$  is the sample mean  $\frac{1}{n} \sum_{i=1}^n D_i$  and  $S_D^2$  is the sample variance.  $W$  satisfies a t-distribution with  $n - 1$  degrees of freedom

We solve hypothesis testing problems about paired samples as we did in the case of independent samples:

- Case 1:  $H_0 : \mu_1 = \mu_2$ ;  $H_1 : \mu_1 \neq \mu_2$   
Reject  $H_0$  if  $|W| > t_{\frac{\alpha}{2}, n-1}$
- Case 2:  $H_0 : \mu_1 \geq \mu_2$ ;  $H_1 : \mu_1 < \mu_2$   
Reject  $H_0$  if  $W < -t_{\alpha, n-1}$
- Case 3:  $H_0 : \mu_1 \leq \mu_2$ ;  $H_1 : \mu_1 > \mu_2$   
Reject  $H_0$  if  $W > t_{\alpha, n-1}$

# Hypothesis Testing: Paired-sample t test

**Example.** 10 randomly selected students in an engineering class are tested about their knowledge on basic statistical concepts before and after attending a course in statistics. Knowledge is measured based on a scale  $[0, 100]$  and the experiment results in the following data:

Before  $x = 43, 82, 77, 39, 51, 66, 55, 61, 79, 43$

After  $y = 51, 84, 74, 48, 53, 61, 59, 75, 82, 48$

We want to test  $H_0 : \mu_x \geq \mu_y$  vs  $H_1 : \mu_x < \mu_y$  at significance level  $\alpha = 0.05$ .

**Solution.** We have  $\bar{d} = \bar{x} - \bar{y} = -3.9$ ,  $s_d^2 = 31.21$ .

We compute the test statistic  $W = \frac{3.9}{\sqrt{31.21/10}} = -2.21$

Since  $W = -2.21 < -t(0.05, 9) = -1.833$ , we reject  $H_0$ .

# Hypothesis Testing: Paired-sample t test

Here is the solution using R

```
> x <- c(43, 82, 77, 39, 51, 66, 55, 61, 79, 43)
> y <- c(51, 84, 74, 48, 53, 61, 59, 75, 82, 48)
> t.test(x,y,alternative = "less", paired = TRUE,
var.equal = TRUE, conf.level = 0.95)
```

Paired t-test

data: x and y

t = -2.2075, df = 9, p-value = 0.02733

alternative hypothesis: true difference in means is  
less than 0

95 percent confidence interval:

-Inf -0.6615005

sample estimates:

mean of the differences

-3.9

# Hypothesis Testing: two independent samples, proportions

Here we consider the situation where there are two independent comparison groups and the outcome of interest is binary, e.g., success/failure.

The goal of the analysis is to compare **proportions** of successes between the two groups.

The relevant sample data are the sample sizes in each comparison group ( $n_1$  and  $n_2$ ) and the sample proportions ( $\hat{p}_1$  and  $\hat{p}_2$ ) which are computed by taking the ratios of the numbers of successes to the sample sizes in each group

$$\hat{p}_1 = \frac{X_1}{n_1}, \quad \hat{p}_2 = \frac{X_2}{n_2}$$

## Hypothesis Testing: two independent samples, proportions

If the sample size is sufficiently large, the binomial random variables  $X_1$  and  $X_2$  can be approximated by normal random variables and  $\hat{p}_i = \frac{X_i}{n_i}$  are approximately  $N(p_i, p_i(1 - p_i)/n_i)$ ,  $i = 1, 2$ .

It follows that

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

follows approximately a  $N(0, 1)$  distribution.

Since  $p_1, p_2$  are unknown, we replace them with the sample proportions. If the null hypothesis is  $H_0 : p_1 = p_2 = p$ , we replace the common  $p$  by the pooled value  $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ . Hence we define the test statistic

$$W = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Hypothesis testing is then carried out using the usual z test.



# Hypothesis Testing: two independent samples, proportions

**Example:** A randomized trial is designed to evaluate the effectiveness of a newly developed pain reliever. The trial compares the new pain reliever to the pain reliever currently in use. A total of 100 patients agreed to participate in the trial. Patients were randomly assigned to receive either the new pain reliever or the standard pain reliever following a medical procedure and were blind to the treatment assignment. Before receiving the assigned treatment, patients were asked to rate their pain on a scale of 0-10 with higher scores indicative of more pain. Each patient was then given the assigned treatment and after 30 minutes was again asked to rate their pain on the same scale. We counted the number of patients indicating a pain reduction of 3+ points:

New Pain Reliever:  $n_1 = 50$ ,  $x_1 = 23$

Old Pain Reliever:  $n_2 = 50$ ,  $x_2 = 11$

Is there a statistically significant difference in the proportions of patients reporting a meaningful pain reduction? Use  $\alpha = 0.05$

# Hypothesis Testing: two independent samples, proportions

We test the hypothesis

$$H_0 : p_1 = p_2;$$

$$H_1 : p_1 \neq p_2.$$

From the data, we compute  $\hat{p}_1 = 0.46$ ,  $\hat{p}_2 = 0.22$  and

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{34}{100} = 0.34$$

To apply the central limit theorem, we check that the sample size is adequate, i.e.,  $\min(n_1\hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2, n_2(1 - \hat{p}_2)) \geq 5$ .

In this example, we have  $\min(50(0.46), 50(1-0.46), 50(0.22), 50(1-0.22)) = \min(23, 27, 11, 39) = 11$ .

Sample size is adequate so the following formula can be used

$$W = \frac{0.46 - 0.22}{\sqrt{0.34(1-0.34)(\frac{1}{50} + \frac{1}{50})}} = 2.533$$

Since  $W = 2.533 > z_{0.025} = 1.960$ , we reject  $H_0$ .

$$p\text{-value} = 2 * (1 - \text{pnorm}(2.533)) = 0.0113091$$

# Hypothesis Testing: two independent samples, proportions

Solution using R

```
> prop.test(x=c(23,11),n=c(50,50),alternative =  
"two.sided",conf.level = 0.95, correct = FALSE)
```

2-sample test for equality of proportions without  
continuity correction

data: c(23, 11) out of c(50, 50)

X-squared = 6.4171, df = 1, p-value = 0.0113

alternative hypothesis: two.sided

95 percent confidence interval:

0.06036633 0.41963367

sample estimates:

prop 1 prop 2

0.46 0.22

# Test for normality

As we observed above, many statistical tests require that the data follow a normal distribution or the result of the test will not be meaningful.

There are four commonly used methods to test that the data follow a normal distribution:

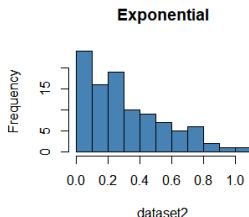
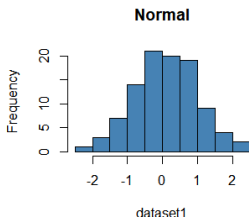
- ① Histogram (visual)
- ② Q-Q plot (visual)
- ③ Shapiro-Wilk Test
- ④ Kolmogorov-Smirnov Test

# Histogram plot

Histogram are useful to approximate probability density functions.

A visual inspection can separate a dataset that is normally distributed (hence, bell-shaped) from a dataset that is not.

```
# generate a normally distributed dataset
set.seed(1)
dataset1 <- rnorm(100)
# generate an exponentially distributed dataset
dataset2 <- rexp(100, rate=3)
hist(dataset1, col='steelblue', main='Normal')
hist(dataset2, col='steelblue', main='Exponential')
```



# Q-Q plot

A Q-Q (quantile-quantile) plot is a plot of the quantiles of two distributions against each other.

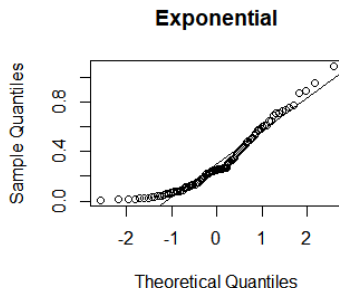
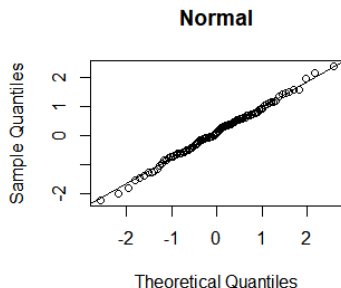
If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line  $y = x$ . As a result, the pattern of points in the plot is useful to compare the two distributions.

A Q-Q plot is generally a more powerful approach to compare distributions than the comparison of the histograms of two samples.

## Q-Q plot

Here we use R to generate the normal Q-Q plots comparing the dataset generated above to a standard normal population.

```
# create Q-Q plot for both datasets  
qqnorm(dataset1, main='Normal')  
qqline(dataset1)  
qqnorm(dataset2, main='Exponential')  
qqline(dataset2)
```



# Statistical Tests for normality

The two most common normality tests are the Shapiro-Wilk test and the Kolmogorov-Smirnov test.

Both tests address the same hypotheses, that is:

- $H_0$  : the data follow a normal distribution
- $H_1$  : the data do not follow a normal distribution

p-value > 0.05 implies that we do not reject the null hypothesis so that the data follow a normal distribution.

Shapiro-Wilk test is recommended for normality test as it provides better power than Kolmogorov-Smirnov test.



# Shapiro-Wilk Test

In R, the Shapiro-Wilk test is implemented with the command `shapiro.test`

```
shapiro.test(dataset1)
```

Shapiro-Wilk normality test

```
data:  dataset1
```

```
W = 0.9956, p-value = 0.9876
```

According to the test, since the p value is above 0.05 then the data is normally distributed.

# Shapiro-Wilk Test

```
shapiro.test(dataset2)
```

Shapiro-Wilk normality test

```
data: dataset2
```

```
W = 0.91505, p-value = 7.759e-06
```

According to the test, since the p value is less than 0.05 then the data is not normally distributed.

# Kolmogorov-Smirnov Test

In R, the Kolmogorov-Smirnov test is implemented with the command `ks.test`

```
ks.test(dataset1, 'pnorm')
```

One-sample Kolmogorov-Smirnov test

```
data: dataset1
```

```
D = 0.094659, p-value = 0.3317
```

```
alternative hypothesis: two-sided
```

According to the test, since the p value is above 0.05 then the data is normally distributed.

# Kolmogorov-Smirnov Test

```
ks.test(dataset2, 'pnorm')
```

One-sample Kolmogorov-Smirnov test

```
data: dataset2
```

```
D = 0.50269, p-value < 2.2e-16
```

```
alternative hypothesis: two-sided
```

According to the test, since the p value is less than 0.05 then the data is not normally distributed.

Note that the Kolmogorov-Smirnov test is not specific for the normal distribution but can be used more generally to decide if a sample comes from a population with any specific distribution. This explains the presence of parameter 'pnorm' in the R command `ks.test`.

# Statistical Tests for normality

**Remark.** Normality tests are often considered as too conservative in the sense that for large sample size ( $n > 50$ ), a small deviation from the normality may cause the normality condition to be violated.

Since a normality test is a hypothesis test, as the sample size increases their capacity of detecting smaller differences increases. So as the number of observations increases, the Shapiro-Wilk and the Kolmogorov-Smirnov tests become very sensitive even to a small deviation from normality. As a consequence, the normality test might indicate that data do not follow a normal distribution although the departures from the normal distribution is negligible. For this reason, it is often the case that the normality condition is verified based on a combination of multiple methods including visual inspections (with histograms and QQ-plots) and a formal inspection (with the Shapiro-Wilk test for instance).