

# STAT 509: Statistics for Engineers

## Chapter 2: Probability

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

## Chapter 2: Probability

### Learning Objectives:

1. Understand and describe sample spaces and events
2. Interpret probabilities and calculate probabilities of events
3. Use permutations and combinations to count outcomes
4. Calculate the probabilities of joint events
5. Interpret and calculate conditional probabilities
6. Determine independence and use independence to calculate probabilities
7. Understand Bayes' theorem and when to use it

## Random Experiment

An **experiment** is a procedure that is

- ▶ carried out under controlled conditions, and
- ▶ executed to discover an unknown result.

An experiment that results in different outcomes even when repeated in the same manner every time is a **random experiment**; e.g.,

- ▶ Flip a coin
- ▶ Toss a dice
- ▶ Measure the recycle time of a flash

How to describe the likelihood of observing a possible outcome from a random experiment? What is the probability of a "head" from a coin flipping?

The **set** of all possible outcomes of a random experiment is called the **sample space**, denoted by  $S$ .

- ▶  $S$  is **discrete** if it consists of a finite or countable infinite set of outcomes.
- ▶  $S$  is **continuous** if it contains an interval of real numbers.

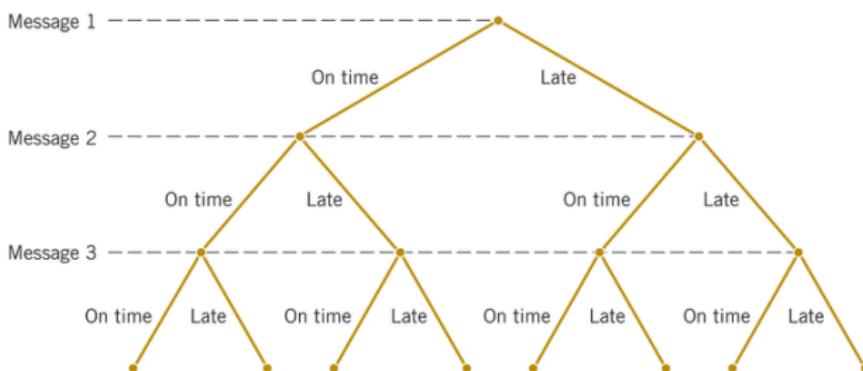
Examples:

1. Randomly select a camera and record the recycle time of a flash:  $S = \mathbb{R}^+ = (0, \infty)$ , all the positive real numbers, is continuous.
2. Suppose we know all the recycle times are between 1.5 and 5 seconds. Then  $S = (1.5, 5)$  is continuous.
3. It is known that the recycle time has only three values(low, medium or high). Then  $S = \{\text{low, medium, high}\}$  is discrete.
4. Does the camera conform to minimum recycle time specifications?  $S = \{\text{yes, no}\}$  is discrete.

# Tree diagram to list a discrete sample space

Messages are classified as on-time(o) or late(l). Classify the next 3 messages.

$$S = \{ooo, ool, olo, oll, loo, lol, llo, lll\}.$$



This only works for small sample spaces. Think we have 30 messages, the size of  $S$  is  $2^{30} = 1,073,741,824$ .

# Counting Techniques

There are three special rules, or counting techniques, used to determine the number of outcomes in events:

1. Multiplication rule
2. Permutation rule
3. Combination rule

Each has its special purpose that must be applied properly – the right tool for the right job.

## Multiplication Rule

Let an operation consists of  $k$  steps and there are

- ▶  $n_1$  ways of completing step 1,
- ▶  $n_2$  ways of completing step 2, ..., and
- ▶  $n_k$  ways of completing step  $k$ .

Then, the total number of ways to perform this operation is

$$n_1 \cdot n_2 \cdots n_k.$$

### Example: Web Site Design

In the design for a website, we can choose to use among: 4 colors, 3 fonts, and 3 positions for an image. How many designs are possible?

Answer via the multiplication rule:  $4 \cdot 3 \cdot 3 = 36$ .

## Permutation Rule

A permutation is a unique sequence (**order matters**) of distinct items. For example, if  $S = \{a, b, c\}$ , there are  $6 = 3 \times 2 \times 1$  permutations:

$$abc, acb, bac, bca, cab, cba.$$

How many different ways to permute  $n$  different items? Answer is

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

by definition  $0! = 1$ .

## Subset Permutations

How many different ways to permute  $r$  items from a set of  $n$  distinct items?

$$P_r^n = n(n-1)(n-2) \cdots (n-r+1) = \frac{n!}{(n-r)!}$$

$nPr(n,r)$

### Example

A printed circuit board has eight different locations in which a component can be placed. If four different components are to be placed on the board, how many designs are possible?

**Answer:** Order is important! Using the permutation formula with  $n = 8$ ,  $r = 4$ :

$$P_4^8 = \frac{8!}{(8-4)!} = 8 \cdot 7 \cdot 6 \cdot 5 = 1680.$$

$nPr(8,4)$

## Similar Item (not distinct) Permutations

Suppose the  $n$  items are not totally distinct. We have

- ▶  $n = n_1 + n_2 + \cdots + n_r$  items of which
- ▶  $n_1, n_2, \dots, n_r$  are identical.

The number of permutations of these  $n$  items is

$$\frac{n!}{n_1!n_2!\cdots n_r!}$$

SimPerm(c(n1,n2,...,nr))

### Example

In a hospital, an operating room needs to schedule 2 (identical) brain surgeries, 3 (identical) knee surgeries and 2 (identical) hip surgeries in a day. How many schedules are there?

$$\frac{(2+3+2)!}{2!3!2!} = 210.$$

SimPerm(c(2,3,2))

## Combination Rule

A combination is a selection of  $r$  items from a set of  $n$  where **order does not matter**.

### Example

If  $S = \{a, b, c\}$ ,  $n = 3$ . Then

- ▶ If pick  $r = 3$  out, we have 1 combination:  $abc$  (the same as  $acb, bca, \dots$ )
- ▶ If pick  $r = 2$  out, we have 3 combinations:  $ab, bc, ac$ .

The number of permutations (where order matters) is always larger or equal to the number of combinations (where order does not matter).

The number of combinations of  $r$  times out of  $n$  is

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

$nCr(n,r)$

## Example: Combination Rule

A bin of 50 parts contains 3 defectives and 47 non-defective parts.  
A sample of 6 parts is selected from the 50 without replacement.  
How many ways to get a sample of size 6 which contains 2 defective parts?

**Answer:**

Step 1: We need to sample 2 defectives out of the 3 defectives, which has  $C_2^3 = 3$  different ways.

Step 2: To sample the remaining 4 non-defective parts out of the total 47 ones, which has  $C_4^{47} = 178,365$  different ways.

Thus, in total, there are  $C_2^3 \times C_4^{47} = 3 \times 178,365 = 535,095$  different ways.

$$nCr(3,2) * nCr(47,4)$$

## Events and Set Operations

An event ( $E$ ) is a subset of the sample space of a random experiment.

Event combinations (set operations)

- ▶ The **Union** of two events,  $E_1$  and  $E_2$ , consists of all outcomes that are contained in one event **or** the other, denoted as  $E_1 \cup E_2$ .
- ▶ The **Intersection** of two events  $E_1$  and  $E_2$ , consists of all outcomes that are contained in one event **and** the other, denoted as  $E_1 \cap E_2$ .
- ▶ The **Complement** of an event  $E$  is the set of outcomes in the sample space that are **not** contained in the event, denoted as  $E^c$ .

## Example: Discrete Events

Suppose that the recycle times of two cameras are recorded. Consider only whether or not the cameras conform to the manufacturing specifications. We abbreviate yes and no as  $y$  and  $n$ . The sample space is  $S = \{yy, yn, ny, nn\}$ . Let

- ▶  $E_1$  denote an event that at least one camera conforms to specifications, then  $E_1 = \{yy, yn, ny\}$ ,
- ▶  $E_2$  an event that no camera conforms to specifications, then  $E_2 = \{nn\}$ ,
- ▶ and  $E_3$  an event that at least one camera does not conform, then  $E_3 = \{yn, ny, nn\}$ .

We have

- ▶  $E_1 \cup E_3 = S$
- ▶  $E_1 \cap E_3 = \{yn, ny\}$
- ▶  $E_1^c = \{nn\}$

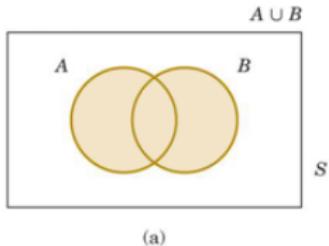
## Example: Continuous Events

Measurements of the thickness of a part are modeled with the sample space:  $S = (0, \infty)$ . Let  $E_1 = [10, 12)$  and  $E_2 = (11, 15)$ . Then

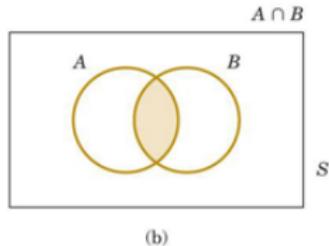
- ▶  $E_1 \cup E_2 = [10, 15)$
- ▶  $E_1 \cap E_2 = (11, 12)$
- ▶  $E_1^c = (0, 10) \cup [12, \infty)$
- ▶  $E_1^c \cap E_2 = [12, 15)$

# Venn Diagrams

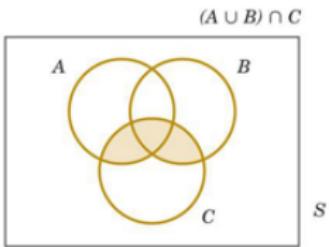
Events  $A$  and  $B$  contain their respective outcomes. The shaded regions indicate the event relation of each diagram.



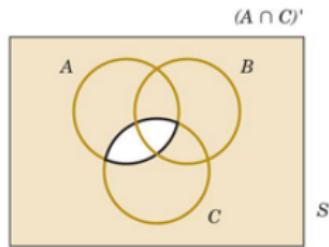
(a)



(b)



(c)



(d)

## Mutually Exclusive Events

Events  $A$  and  $B$  are **mutually exclusive** because they share no common outcomes. The occurrence of one event precludes the occurrence of the other (not independent at all, strongly dependent). Symbolically,  $A \cap B = \emptyset$  (the emptyset set).



## Some laws of set operations

- Commutative law:

$$A \cap B = B \cap A$$

$$A \cup B = B \cup A.$$

- Distributive law:

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

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

- Associative law:

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

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

- Complement law:  $(A^c)^c = A$

- De Morgan's law:

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

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

# Probability

Probability is the likelihood or chance that a particular outcome or event from a random experiment will occur.

Denote by  $P(E)$  the probability of event  $E$  will occur.

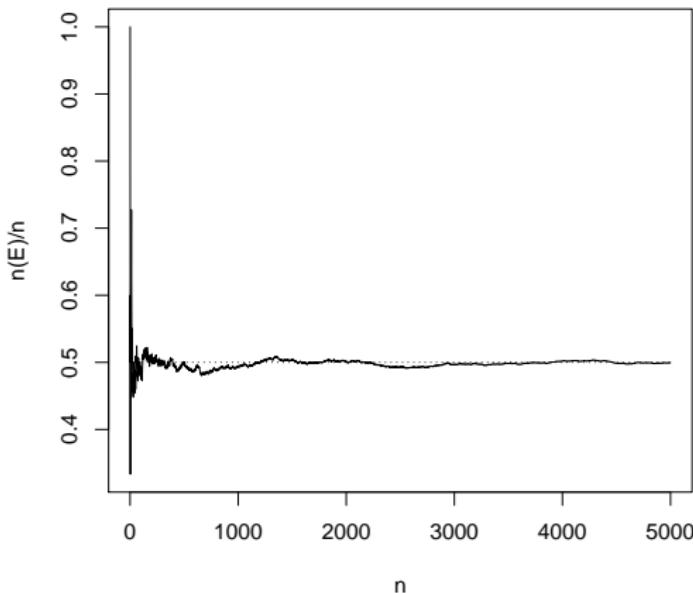
Mathematically, probability  $P(E)$  is a number between 0 and 1 that is assigned to the event  $E$  from a random experiment.

## How to assign probabilities?

- ▶ Subjective probability: a "degree of belief." (e.g., There is a 50% chance that I will study tonight.)
- ▶ Relative frequency probability: based on how often an event occurs over a very large sample space; i.e.,  
$$P(E) = \lim_{n \rightarrow \infty} n(A)/n.$$
- ▶ Equally-likely rule: probability of each member of the sample space is the same.
- ▶ ...

## Relative frequency probability

Flip a fair coin repeatedly, the relative frequency of observing "head" approaches the probability  $P(\text{" head"}) = 0.5$ .



However, using this to assign probability is NOT applicable in real applications. This is merely for interpretation.

## Random: Equally-likely Outcomes

Whenever a sample space consists of  $N$  possible outcomes that are equally likely, the probability of each outcome is  $1/N$ .

### Example

In a batch of 100 diodes, 1 is laser diode. A diode is **randomly** selected from the batch. **Random** means each diode has an equal chance of being selected. The probability of choosing the laser diode is  $1/100$  or  $0.01$ , because each outcome in the sample space is equally likely.

## Example

Again, from a bin of 50 parts, 6 parts are selected **randomly** without replacement. The bin contains 3 defective parts and 47 nondefective parts. What is the probability that exactly 2 defective parts are selected in the sample?

**Answer:** when **randomly** appears, it means equally-likely rule!

$$\begin{aligned} & P(\text{exactly 2 defective parts}) \\ &= \frac{\# \text{ of ways to select 6 parts of which 2 are defective}}{\# \text{ of ways to select 6 parts}} \\ &= \frac{C_2^3 C_4^{47}}{C_6^{50}} \\ &= \frac{nCr(3, 2) \times nCr(47, 4)}{nCr(50, 6)} \\ &= 0.03367347 \end{aligned}$$

## Probability of an Event (Discrete)

We now restrict our attention to a discrete sample space. By discrete, it means the sample sapce may be

- ▶ A finite set of outcomes; (e.g., number of winnings Gamecock can achieve in the next season)
- ▶ A countably infinite set of outcomes. (e.g., number of emails one receives on one day)

For a discrete sample space, the probability of an event  $E$  equals the sum of the probabilities of the outcomes in  $E$ .

## Example

A random experiment has a sample space  $S = \{a, b, c, d\}$ . These outcomes are not equally-likely; their probabilities are: 0.1, 0.3, 0.5, 0.1. Let event  $A = \{a, b\}$ ,  $B = \{b, c, d\}$ , and  $C = \{d\}$ . Then

- ▶  $P(A) = P(a) + P(b) = 0.1 + 0.3 = 0.4$
- ▶  $P(B) = 0.3 + 0.5 + 0.1 = 0.9$
- ▶  $P(C) = P(d) = 0.1$
- ▶  $P(A^c) = P(\{c, d\}) = P(c) + P(d) = 0.5 + 0.1 = 0.6 = 1 - P(A)$ ;  $P(B^c) = 1 - P(B) = 0.1$ ;  $P(C^c) = 1 - 0.1 = 0.9$ .
- ▶  $P(A \cap B) = P(b) = 0.3$ ,  
 $P(A \cup B) = P(\{a, b, c, d\}) = P(S) = 1$ , and  
 $P(A \cap C) = P(\emptyset) = 0$ .

We observe  $P(S) = 1$ ,  $P(\emptyset) = 0$ ,  $P(A^c) = 1 - P(A)$ .

## Example

A wafer is randomly selected from a batch that is classified by contamination and location.

		Location in Sputtering Tool	
Contamination	Center	Edge	Total
Low	514	68	582
High	112	246	358
Total	626	314	

Let  $H$  be the event of high concentrations of contaminants. Let  $C$  be the event of the wafer being located at the center of a sputtering tool.

- ▶  $P(H) = 358/940$
- ▶  $P(C) = 626/940$
- ▶  $P(H \cap C) = 112/940$
- ▶  $P(H \cup C) = (358+626-112)/940 = P(H)+P(C)-P(H \cap C)$

# Axioms of Probability

The assignment of probability to events from a random experiment must satisfy the following properties:

## Axioms

If  $S$  is the sample space and  $E$  is any event from the random experiment,

1.  $P(S) = 1$
2.  $0 \leq P(E) \leq 1$  (0 means impossible; 1 mean certainty)
3. For any two events  $E_1$  and  $E_2$  with  $E_1 \cap E_2 = \emptyset$  (mutually exclusive),

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

The axioms imply that

- $P(\emptyset) = 0$  and  $P(E^c) = 1 - P(E)$
- If  $E_1 \subset E_2$ , then  $P(E_1) \leq P(E_2)$ .

## Addition Rules

For any two events  $A$  and  $B$ , the probability of union is given by

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

If  $A$  and  $B$  are mutually exclusive, then  $P(A \cap B) = P(\emptyset) = 0$  and

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

## Addition Rules: 3 or more events

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

If a collection of events  $E_i$  are pairwise mutually exclusive; i.e.,  $E_i \cap E_j = \emptyset$  for  $i \neq j$ , then

$$P(E_1 \cup E_2 \cup \dots \cup E_k) = \sum_{i=1}^k P(E_i).$$

### Example

Let  $X$  denote the pH of a sample. Consider the event that  $X$  is greater than 6.5 but less than or equal to 7.8. Then  $P(6.5 < X \leq 7.8) = P(6.5 < X \leq 7) + P(7 < X \leq 7.5) + P(7.5 < X \leq 7.8)$ .

## Conditional Probability

$P(B|A)$  is the probability of event  $B$  occurring, given that event  $A$  has already occurred.

		Surface Flaws		
		Yes (event $F$ )	No	Total
Defective	Yes (event $D$ )	10	18	28
	No	30	342	372
Total		40	360	400

We have 400 parts classified by surface flaws and as (functionally) defective.

Let  $D$  denote the event that a part is defective, and  $F$  the event that a part has a surface flaw.

The probability of  $D$  given that a part has a flaw, as  $P(D|F)$ .

25% of the parts with flaws are defective,  $P(D|F) = 0.25$ .

5% of the parts without flaws are defective,  $P(D|F^c) = 0.05$ .

What are  $P(D^c|F)$  and  $P(D^c|F^c)$ ?

## Conditional Probability Rule and Multiplication Rule

The conditional probability of an event  $B$  given an event  $A$ , denoted as  $P(B|A)$ , is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \text{ for } P(A) > 0.$$

Consequently, we have the Multiplication Rule:

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

## Example

A batch of 50 parts contains 10 made by Tool 1 and 40 made by Tool 2. If 2 parts are selected **randomly**.

- (a) What is the probability that the 1st part came from Tool 1 and the 2nd part came from Tool 2?
- (b) What is the probability that the 2nd part came from Tool 2, given that the 1st part came from Tool 1?

**Answer:** Let  $E_1$  denote the event that the 1st part came from Tool 1;  $E_2$  the 2nd part came from Tool 2.

$$(a): P(E_2 \cap E_1) = \frac{10}{50} \times \frac{40}{49} = 8/49$$

$$(b): P(E_2|E_1) = P(E_2 \cap E_1)/P(E_1) = (8/49)/(10/50) = 40/49,$$

where  $P(E_1) = 10/50$ .

## Example

The probability that the first stage of a numerically controlled machining operation for high-rpm pistons meets specifications is 0.90. Failures are due to metal variations, fixture alignment, cutting blade condition, vibration, and ambient environmental conditions. Given that the first stage meets specifications, the probability that a second stage of machining meets specifications is 0.95. What is the probability that both stages meet specifications?

**Answer:** Let  $A$  and  $B$  denote the events that the first and second stages meet specifications, respectively. The probability requested is

$$P(A \cap B) = P(B|A)P(A) = 0.95 * 0.9 = 0.855.$$

Although it is also true that  $P(A \cap B) = P(A|B)P(B)$ , the information provided in the problem does not match this second formulation.

## Total Probability Rule

For any two events  $A$  and  $B$ :

$$P(B) = P(B \cap A) + P(B \cap A^c) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

For more than 2 events:

Assume  $E_1, E_2, \dots, E_k$  are  $k$  mutually exclusive and exhaustive sets;  
i.e.,

- ▶  $E_i \cap E_j = \emptyset$  for  $i \neq j$  (mutually exclusive)
- ▶  $E_1 \cup E_2 \cup \dots \cup E_k = S$  (exhaustive)

Then

$$\begin{aligned} P(B) &= P(B \cap E_1) + P(B \cap E_2) + \dots + P(B \cap E_k) \\ &= P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \dots + P(B|E_k)P(E_k). \end{aligned}$$

## Example

Let  $F$  denote the event that the product fails, and  $H$  the event that the chip is exposed to high levels of contamination. Find  $P(F)$ .

Probability of Failure	Level of Contamination	Probability of Level
0.1	High	0.2
0.005	Not high	0.8

**Answer:** The third column tells us that  $P(H) = 0.2$  and  $P(H^c) = 0.8$ . The first column tells  $P(F|H) = 0.1$  and  $P(F|H^c) = 0.005$ . We can use total probability rule to find  $P(F)$ :

$$\begin{aligned}P(F) &= P(F|H)P(H) + P(F|H^c)P(H^c) \\&= 0.1 \times 0.2 + 0.005 \times 0.8 = 0.024.\end{aligned}$$

## Example

Find  $P(F)$  based on the following information.

Probability of Failure	Level of Contamination	Probability of Level
0.100	High	0.2
0.010	Medium	0.3
0.001	Low	0.5

**Answer:** The third column tells us that  $P(H) = 0.2$ ,  $P(M) = 0.3$  and  $P(L) = 0.8$ . We see that  $H, M, L$  are mutually exclusive and  $P(H) + P(M) + P(L) = 1$  indicating they are also exhaustive.

The first column tells  $P(F|H) = 0.1$ ,  $P(F|M) = 0.01$ , and  $P(F|L) = 0.001$ . We can use total probability rule to find  $P(F)$ :

$$\begin{aligned}P(F) &= P(F|H)P(H) + P(F|M)P(M) + P(F|L)P(L) \\&= 0.1 \times 0.2 + 0.01 \times 0.3 + 0.001 \times 0.5 \\&= 0.0235.\end{aligned}$$

# Independence

Table 1 provides an example of 400 parts classified by surface flaws and as (functionally) defective. Suppose that the situation is different and follows Table 2. Let F denote the event that the part has surface flaws. Let D denote the event that the part is defective.

TABLE 1 Parts Classified				TABLE 2 Parts Classified (data chg'd)							
	Surface Flaws		Total		Surface Flaws		Total				
Defective	Yes ( $F$ )	No ( $F'$ )	Total	Defective	Yes ( $F$ )	No ( $F'$ )	Total				
Yes ( $D$ )	10	18	28	Yes ( $D$ )	2	18	20				
No ( $D'$ )	30	342	372	No ( $D'$ )	38	342	380				
Total	40	360	400	Total	40	360	400				
$P(D F) = 10/40 = 0.25$		$P(D F) = 2/40 = 0.05$									
$P(D) = 28/400 = 0.10$		$P(D) = 20/400 = 0.05$									
not same				same							
Events $D$ & $F$ are <b>dependent</b>				Events $D$ & $F$ are <b>independent</b>							

# Independence

Two events are independent if any one of the following equivalent statements is true:

1.  $P(A|B) = P(A)$
2.  $P(B|A) = P(B)$
3.  $P(A \cap B) = P(A) \cdot P(B)$

This means that occurrence of one event has no impact on the probability of occurrence of the other event.

- If  $A$  and  $B$  are mutually exclusive, are they independent?
- If  $(A$  and  $B)$  are independent, so are  $(A$  and  $B^c)$ ,  $(A^c$  and  $B)$ ,  $(A^c$  and  $B^c)$ .

## Independence with multiple events

The events  $E_1, E_2, \dots, E_n$  are independent, if and only if, for any subsets of these events:

$$P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}) = P(E_{i_1}) \cdot P(E_{i_2}) \cdots P(E_{i_k}).$$

## Circuit Operation

The following circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown on the graph. Assume that devices fail **independently**. What is the probability that the circuit operates?



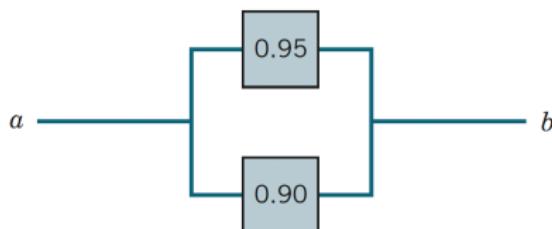
**Answer:** The circuit operates if and only if the two parts operate together.

$$P(L \cap R) = P(L) \cdot P(R) = 0.8 \times 0.9 = 0.72.$$

**Practical Interpretation:** Notice that the probability that the circuit operates degrades to approximately 0.7 when all devices are required to be functional. The probability that each device is functional needs to be large for a circuit to operate when many devices are connected in series.

## Circuit Operation

Assume that devices fail **independently**. What is the probability that the circuit operates?



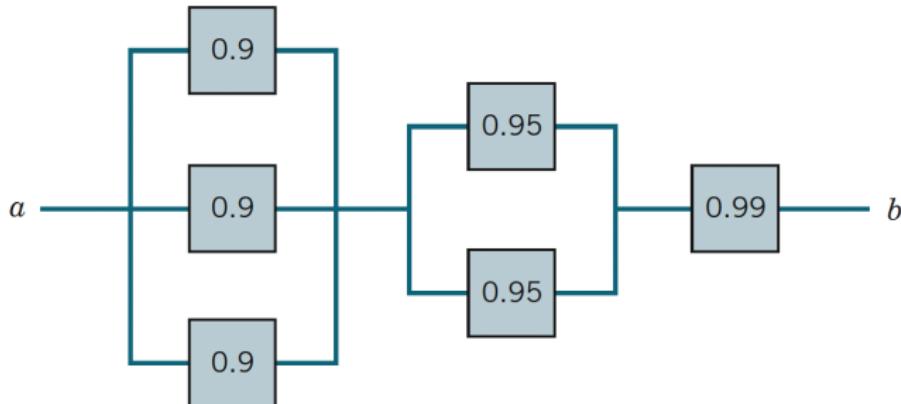
**Answer:** The circuit operates if at least one device operates.

$$\begin{aligned} P(T \cup B) &= 1 - P\{(T \cup B)^c\} = 1 - P(T^c \cap B^c) \\ &= 1 - P(T^c)P(B^c) = 1 - (1 - 0.95)(1 - 0.9) = 0.995 \end{aligned}$$

**Practical Interpretation:** Notice that the probability that the circuit operates is larger than the probability that either device is functional. This is an advantage of a parallel architecture.

# Circuit Operation

Assume that devices fail **independently**. What is the probability that the circuit operates?



**Answer:**

$$\begin{aligned} P(L \cap M \cap R) &= P(L)P(M)P(R) \\ &= (1 - 0.1^3)(1 - 0.05^2)0.99 = 0.987. \end{aligned}$$

# Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ for } P(B) > 0.$$

## Example

Let  $F$  denote the event that the product fails, and  $H$  the event that the chip is exposed to high levels of contamination. Find  $P(H|F)$ , the conditional probability that a high level of contamination was present when a failure occurred is to be determined.

Probability of Failure	Level of Contamination	Probability of Level
0.1	High	0.2
0.005	Not high	0.8

$$P(H|F) = \frac{P(F|H)P(H)}{P(F)} = \frac{0.1 \cdot 0.2}{0.24} = 0.83.$$

## Example: Medical Diagnostic

Because a new medical procedure has been shown to be effective in the early detection of an illness, a medical screening of the population is proposed. The probability that the test correctly identifies someone with the illness as positive (known as the sensitivity) is 0.95, and the probability that the test correctly identifies someone without the illness as negative (known as the specificity) is 0.99. The incidence of the illness in the general population is 0.0001. You take the test, and the result is positive. What is the probability that you have the illness?

**Answer:** Let  $I$  denote the event that you have the illness, and let  $T$  denote the event that the test signals positive. Then  $P(T|I) = 0.95$ ,  $P(T^c|I^c) = 0.99$ , and  $P(I) = 0.0001$ .

$$\begin{aligned}P(I|T) &= \frac{P(T|I)P(I)}{P(T|I)P(I) + P(T|I^c)P(I^c)} \\&= \frac{0.99(0.0001)}{0.99(0.0001) + (1 - 0.95)(1 - 0.0001)} = 0.002\end{aligned}$$

## Bayes' Theorem with total probability rule

If  $E_1, E_2, \dots, E_k$  are  $k$  mutually exclusive and exhaustive events and  $B$  is any event with  $P(B) > 0$ , then

$$\begin{aligned} P(E_1|B) &= \frac{P(B|E_1)P(E_1)}{P(B)} \\ &= \frac{P(B|E_1)P(E_1)}{P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \cdots + P(B|E_k)P(E_k)}. \end{aligned}$$

## Example: Bayesian Network

A printer manufacturer obtained the following three types of printer failure probabilities. Hardware  $P(H) = 0.3$ , software  $P(S) = 0.6$ , and other  $P(O) = 0.1$ . Also,  $P(F|H) = 0.9$ ,  $P(F|S) = 0.2$ , and  $P(F|O) = 0.5$ . If a failure occurs, determine if it's most likely due to hardware, software, or other.

**Answer:** We need to find out which of  $P(H|F)$ ,  $P(S|F)$ ,  $P(O|F)$  is the largest. We also note  $H, S, O$  are mutually exclusive and exhaustive events.

$$P(H|F) = \frac{P(F|H)P(H)}{P(F)} = \frac{0.9 \cdot 0.3}{0.44} = 0.6136.$$

where  $P(F) = P(F|H)P(H) + P(F|S)P(S) + P(F|O)P(O) = 0.9(0.3) + 0.2(0.6) + 0.5(0.1) = 0.44$ . Similarly,  $P(S|F) = 0.12/0.44 = 0.2727$  and  $P(O|F) = 0.05/0.44 = 0.1136$ .

# STAT 509: Statistics for Engineers

## Chapter 3: Discrete Random Variables and Probability Distributions

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

# Chapter 3: Discrete Random Variables and Probability Distributions

## Learning Objectives:

1. Understand random variables
2. Determine probabilities from probability mass functions and the reverse
3. Determine probabilities and probability mass functions from cumulative distribution functions and the reverse.
4. Calculate means and variances for discrete random variables.
5. Understand the assumptions for discrete probability distributions.
6. Select an appropriate discrete probability distribution to calculate probabilities.
7. Calculate probabilities, means and variances for discrete probability distributions.

## Random Variable and its Notation

A variable that associates a number with the outcome of a random experiment is called a **random variable**; e.g., flip a coin,

$$X = \begin{cases} 1 & \text{head} \\ 0 & \text{tail} \end{cases}$$

A random variable is a **function** that assigns a real number to each outcome in the sample space of a random experiment; e.g., the sample space of flipping a coin is  $\mathcal{S} = \{\text{head}, \text{tail}\}$

$$X : \mathcal{S} \mapsto \{0, 1\} : X(\text{head}) = 1 \quad X(\text{tail}) = 0.$$

A random variable is denoted by an uppercase letter such as  $X$ . After the experiment is conducted, the measured value of the random variable is denoted by a lowercase letter such as  $x = 70$  milliamperes.

## Discrete and Continuous Random Variables

A **discrete random variable** is a random variable with a finite or countably infinite range. Its values are obtained by counting.

- ▶ Number of scratches on a surface
- ▶ Proportion of defective parts among 100 tested
- ▶ Number of transmitted bits received in error
- ▶ Number of common stock shares traded per day

A **continuous random variable** is a random variable with an interval (either finite or infinite) of real numbers for its range. Its values are obtained by measuring.

- ▶ Electrical current and voltage
- ▶ Physical measurements, e.g., length, weight, time, temperature, pressure

# Discrete Random Variables

Many physical systems can be modeled by the same or similar random experiments and random variables. The distribution of the random variables involved in each of these common systems can be analyzed. We will learn many classical random variables that can be used in different applications and examples.

We start with **discrete random variables**.

## Notation

Let  $X$  be a discrete random variable. Denote its possible values by  $x_1, x_2, \dots, x_n, \dots$  (could be infinite).

We would like to know the probability  $P(X = x_i)$  for an  $x_i$  or  $P(X \leq x)$  for any  $x$ .

# Probability Distributions and Probability Mass Functions

The time to recharge the flash is tested in three cell-phone cameras. The probability that a camera passes the test is 0.8, and the cameras perform independently. The sample space for this experiment and associated probabilities are:

TABLE 3.1 Camera Flash Tests

Camera 1	Camera 2	Camera 3	Probability	$X$
Pass	Pass	Pass	0.512	3
Fail	Pass	Pass	0.128	2
Pass	Fail	Pass	0.128	2
Fail	Fail	Pass	0.032	1
Pass	Pass	Fail	0.128	2
Fail	Pass	Fail	0.032	1
Pass	Fail	Fail	0.032	1
Fail	Fail	Fail	0.008	0

where the random variable  $X$  denotes the number of cameras that pass the test. Then the possible value

## Probability Mass Function

For a discrete random variable  $X$  with possible values  $x_1, \dots, x_n$  (if finite) or  $x_1, x_2, \dots, x_n, \dots$  (if infinite), its **probability mass function** (pmf) is a function such that:

1.  $f(x_i) = P(X = x_i)$
2.  $f(x_i) \geq 0$
3.  $\sum_{i=1}^n f(x_i) = 1$  (if finite) or  $\sum_{i=1}^{\infty} f(x_i) = 1$  (if infinite).

We often call  $f(x)$  as the probability density function (pdf) of  $X$  as well, though rigorously, pdf is for a continuous random variable.

## Example: infinite possible values

Let the random variable  $X$  denote the number of wafers that need to be analyzed to detect a large particle of contamination. Assume that the probability that a wafer contains a large particle is 0.01, and that the wafers are independent. Determine the probability distribution of  $X$ .

**Answer:** Let  $p$  denote a wafer in which a large particle is present let  $a$  denote a wafer in which it is absent.

- ▶ The sample space is  $S = \{p, ap, aap, aaap, aaaap, \dots\}$
- ▶ The associated range of  $X$  is  $x_1, x_2, x_3, x_4, x_5, \dots$  where  $x_i = i$ .

We then have

$$P(X = 1) = 0.01, \quad P(X = 2) = 0.99 \cdot 0.01, \quad P(X = 3) = 0.99^2 \cdot 0.01,$$

so and so on; i.e., for  $i \geq 1$ ,

$$P(X = x_i) = P(X = i) = 0.99^{i-1} \cdot 0.01.$$

## Cumulative Distribution Function

The **cumulative distribution function** (cdf), is the probability that a random variable  $X$  will be found at a value less than or equal to  $x$ . Symbolically, the cdf is

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

For a discrete random variable  $X$ ,  $F(x)$  satisfies the following properties:

1.  $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$
2.  $0 \leq F(x) \leq 1$
3. If  $x \leq y$ , then  $F(x) \leq F(y)$ .

Note that pmf  $f$  is defined on all the possible values of  $X$  while cdf  $F$  is defined on the entire real line.

## Example: cdf

Suppose  $X$  is a discrete random variable with pmf being

$$f(x) = \begin{cases} 0.2 & x = -2 \\ 0.5 & x = 0 \\ c & x = 2. \end{cases}$$

Determine and draw the cdf of  $X$ .

**Answer:** the definition of  $f(x)$  suggests that  $X$  has only three possible values  $\{-2, 0, 2\}$ . Known from the definition of a pmf; i.e.,  $\sum_{i=1}^n f(x_i) = 1$ , we have

$$1 = f(-2) + f(0) + f(2) = 0.2 + 0.5 + c$$

which implies  $c = 0.3$ .

## Example: cdf continued

Now we find the cdf  $F(x)$  of  $X$ . Because  $F(x)$  is defined on the entire real line, and  $X$  only has three possible values which partition the entire real line into four regions:

$$(-\infty, \underbrace{-2)} \cup \underbrace{[-2, 0)} \cup \underbrace{[0, 2)} \cup \underbrace{[2, \infty)}.$$

$x_1$                      $x_2$                      $x_3$

- ▶ For  $x \in (-\infty, -2)$ ,  $F(x) = P(X \leq x) = 0$
- ▶ For  $x \in [-2, 0)$ ,  
 $F(x) = P(X \leq x) = P(X = -2) = f(-2) = 0.2$
- ▶ For  $x \in [0, 2)$ ,  $F(x) = P(X \leq x) = P(X = -2) + P(X = 0) = f(-2) + f(0) = 0.7$
- ▶ For  $x \in [2, \infty)$ ,  $F(x) = P(X \leq x) = P(X = -2) + P(X = 0) + P(X = 2) = f(-2) + f(0) + f(2) = 1$

## Example: cdf continued

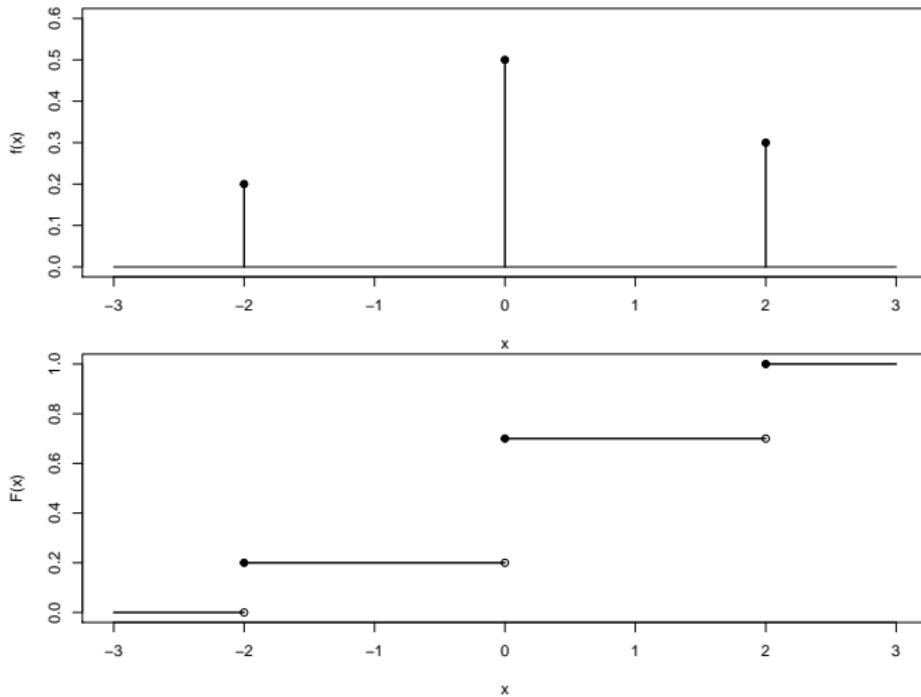
Now we have

$$F(x) = \begin{cases} 0 & x < -2 \\ 0.2 & -2 \leq x < 0 \\ 0.7 & 0 \leq x < 2 \\ 1 & 2 \leq x. \end{cases}$$

```
#StatEngine
x=c(-2,0,2);fx=c(0.2,0.5,0.3)
discrete.plotcdf(x,fx)
# or
discrete.summary(x,fx,plotpdf=FALSE,plotcdf=TRUE)
```

## Example: pmf/cdf plot

```
#StatEngine  
x=c(-2,0,2);fx=c(0.2,0.5,0.3)  
discrete.summary(x,fx)
```



## Mean and Variance of a Discrete Random Variable

Two numbers are often used to summarize a probability distribution for a random variable  $X$ .

- ▶ The **mean** is a measure of the center or middle of the probability distribution.
- ▶ The **variance** is a measure of the dispersion, or variability in the distribution.

These two measures **do not** uniquely identify a probability distribution. That is, two different distributions can have the same mean and variance. Still, these measures are simple, useful summaries of the probability distribution of  $X$ .

## Mean and Variance of a Discrete Random Variable

The **mean** or **expected value** of the discrete random variable  $X$ , denoted as  $\mu$  or  $E(X)$ , is

$$\mu = E(X) = \sum_i x_i f(x_i).$$

The **variance** of  $X$ , denoted by  $\sigma^2$  or  $V(X)$ , is

$$\sigma^2 = V(X) = E(X - \mu)^2 = \sum_i (x_i - \mu)^2 f(x_i) = \sum_i x_i^2 f(x_i) - \mu^2.$$

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

$$\sigma = \sqrt{\sigma^2}$$

```
# StatEngine  
discrete.summary(x,fx)
```

# Mean and Variance of a Discrete Random Variable

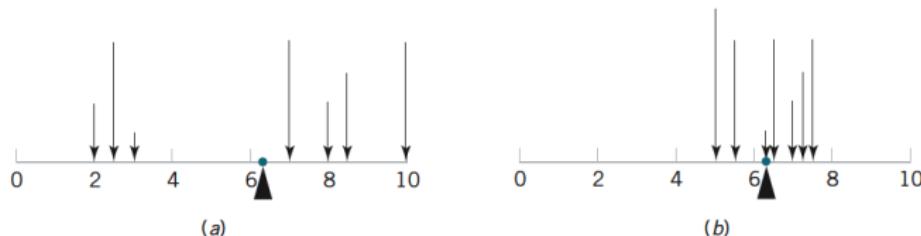


FIGURE 3.4

A probability distribution can be viewed as a loading with the mean equal to the balance point. Parts (a) and (b) illustrate equal means, but part (a) illustrates a larger variance.

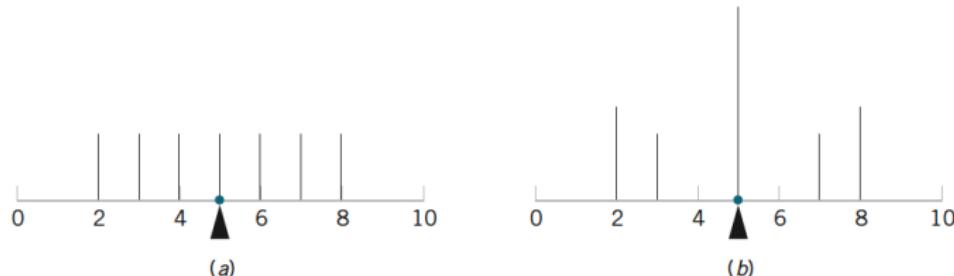


FIGURE 3.5

The probability distributions illustrated in parts (a) and (b) differ even though they have equal means and equal variances.

## Example: Mean and Variance

Suppose  $X$  is a discrete random variable with pmf being

$$f(x) = \begin{cases} 0.2 & x = -2 \\ 0.5 & x = 0 \\ 0.3 & x = 2. \end{cases}$$

```
#StatEngine
x=c(-2,0,2);fx=c(0.2,0.5,0.3);
discrete.summary(x,fx)
$mean
[1] 0.2
$variance
[1] 1.96
$standard.deviation
[1] 1.4
```

## Expected Value of a Function of a Discrete $X$

If  $X$  is a discrete random variable with pmf  $f(x)$  on possible values  $x_i$ 's, then for any real function  $h$ ,

$$E[h(X)] = \sum_i h(x_i)f(x_i).$$

Note that  $V(X) = E(X - \mu)^2 = E[h(X)]$  where  $h(x) = (x - \mu)^2$ .

If  $h(x)$  is linear in  $x$ ; i.e.,

$$h(x) = ax + b, \text{ where } a, b \text{ are two constants.}$$

Then

- $E[h(X)] = E(ax + b) = aE(X) + b$
- $V[h(X)] = V(ax + b) = a^2V(X).$

## Example

There is a chance that a bit transmitted through a digital transmission channel is received in error. Let  $X$  equal the number of bits in error in the next four bits transmitted. The possible values for  $X$  are 0, 1, 2, 3, 4. Based on a model for the errors that is presented in the following section, probabilities for these values will be determined. Suppose that the probabilities are

$$P(X = 0) = 0.6561 \quad P(X = 1) = 0.2916$$

$$P(X = 2) = 0.0486 \quad P(X = 3) = 0.0036$$

$$P(X = 4) = 0.0001$$

Find  $E(X)$ ,  $V(X)$ ,  $E(2X + 4)$ ,  $V(2X + 4)$ ,  $E\{\sin(e^X)\}$  and  $E(X^2)$

## Example (continued)

**Answer:**  $E(X) = 0.4$  and  $V(X) = 0.36$  can be found via

```
x=c(0,1,2,3,4);fx=c(0.6561,0.2916,0.0486,0.0036,0.0001);  
discrete.summary(x,fx)
```

$E(2X + 4) = 2E(X) + 4 = 4.8$  and  $V(2X + 4) = 4V(X) = 1.44$ .

To find  $E\{\sin(e^X)\}$ , you should use **R**. Herein,  $h(x) = \sin(e^x)$ . Define h in R:

```
h=function(x){sin(exp(x))};
```

Then  $E\{\sin(e^X)\}$  is

```
sum(h(x)*fx)
```

```
[1] 0.7186215
```

For  $E(X^2)$ , you can simply use

```
sum(x^2*fx)
```

```
[1] 0.52
```

## StatEngine: user-defined discrete distribution

```
discrete.summary(x,fx,plotpdf=c("TRUE","FALSE"),  
                 plotcdf=c("TRUE","FALSE"))  
  
discrete.prob(x,fx,lb)  
  
discrete.prob(x,fx,lb,ub,  
              inclusive=c("none","left","right","both"))
```

Example: Go back to the bit transmission example. The discrete distribution can be defined by

```
x=c(0,1,2,3,4);  
fx=c(0.6561,0.2916,0.0486,0.0036,0.0001);  
# Some calculation:  
discrete.summary(x,fx) # Mean, Variance, std  
discrete.prob(x,fx,2) # P(X=2)  
discrete.prob(x,fx,3,4,"right") # P(3<X<=4)
```

# Classical Discrete Distributions

In the rest of this chapter, we will learn several commonly used discrete distributions:

- ▶ Discrete Uniform Distribution
- ▶ Binomial Distribution
- ▶ Geometric Distribution
- ▶ Negative Binomial Distribution
- ▶ Hypergeometric Distribution
- ▶ Poisson Distribution

You must know:

1. Definition ([the right tool for the right job](#))
2. Probability calculation
3. Mean and variance calculation

## Discrete Uniform Distribution

The simplest discrete random variable is one that assumes only a finite number of possible values, each with equal probability.

### Discrete Uniform Distribution

A random variable  $X$  has a discrete uniform distribution if each of the  $n$  values in its range,  $x_1, x_2, \dots, x_n$ , has equal probability. Then

$$f(x_i) = \frac{1}{n}.$$

Suppose the range of  $X$  equals the consecutive integers  $a, a+1, a+2, \dots, b$  for  $a \leq b$ . The range of  $X$  contains  $b - a + 1$  values each with probability  $1/(b - a + 1)$ . Then the mean and variance  $X$  are

$$\mu = E(X) = \frac{b+a}{2} \text{ and } \sigma^2 = V(X) = \frac{(b-a+1)^2 - 1}{12}.$$

## Example

Suppose that the *discrete uniform* random variable  $Y$  has range  $5, 10, 15, \dots, 30$ . Find  $P(Y \leq 16)$ ,  $E(Y)$  and  $V(Y)$ .

**Answer:** If  $Y \leq 16$ ,  $Y$  can only be 5, 10 or 15. Thus  $P(Y \leq 16) = P(Y = 5) + P(Y = 10) + P(Y = 15) = 1/6 + 1/6 + 1/6 = 0.5$ .

```
range=seq(5,30,by=5);  
duniform.prob(range,-Inf,16,inclusive="right")
```

To find  $E(Y)$  and  $V(Y)$ , you can use

```
range=seq(5,30,by=5);duniform.summary(range)
```

Or you could do it in this way:

Obviously,  $Y = 5X$  where  $X$  is a discrete uniform random variable with range  $1, 2, \dots, 6$ . Known from the formula that  $a = 1$ ,  $b = 6$ ,

$$E(X) = \frac{a+b}{2} = 3.5 \text{ and } V(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{35}{12}.$$

Then using  $Y = 5X$ , we have

$$E(Y) = 5E(X) = 17.5 \text{ and } V(Y) = 25V(X) = 72.92.$$

## StatEngine: discrete uniform distribution

```
duniform.summary(range,plotpdf=c("TRUE","FALSE"),  
                  plotcdf=c("TRUE","FALSE"))  
  
duniform.prob(range,lb)  
  
duniform.prob(range,lb,ub,  
               inclusive=c("none","left","right","both"))
```

Example: Go back to the previous example. The uniform distribution of  $Y$  can be defined by

```
range=seq(5,30,by=5);  
# Some calculation:  
duniform.summary(range) # Mean, Variance, std  
duniform.prob(range,2) # P(X=2)  
duniform.prob(range,-Inf,16,"right") # P(X<=16)
```

## Bernoulli Distribution (Bernoulli Trials) $X \sim \text{Bernoulli}(p)$

The next three distributions (Binomial, Geometric, Negative Binomial) are from (a series of independent and identical) Bernoulli trials.

### Definition

A trial with only two possible outcomes (often labeled as "success" and "failure") called a **Bernoulli trial**. The associated distribution is called a **Bernoulli distribution**.

A random variable  $X$  is said to follow a Bernoulli distribution with probability of success  $p$ , denoted by  $X \sim \text{Bernoulli}(p)$ , if  $X$  only takes two values, 0 and 1 and

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

E.g., flip a coin; a part is defect or not; air contains contamination or not; the next birth is a boy or a girl; medication is effective or not.

## Binomial, Geometric, and Negative Binomial

Provided with a Bernoulli trial where the probability of success is  $p$ .

- ▶ Suppose we independently and identically repeat the trial for  $n$  times, let  $X$  be the **number of successes** from the  $n$  trials. Then  $X$  follows a Binomial distribution.
- ▶ Suppose we independently and identically repeat the trial, let  $X$  be the number of trial to observe the **first** success. Then  $X$  follows a Geometric distribution.
- ▶ Suppose we independently and identically repeat the trial, let  $X$  be the number of trial to observe the  **$r$ -th** success. Then  $X$  follows a Negative Binomial distribution.

## Binomial Distribution (Definition) $X \sim \text{Binomial}(n, p)$

A random experiment consists of  $n$  Bernoulli trials such that

- (1) The trials are independent.
- (2) Each trial results in only two possible outcomes, labeled as "success" and "failure."
- (3) The probability of a success in each trial, denoted as  $p$ , remains constant.

The random variable  $X$  that equals the number of trials that result in a success is a **binomial random variable** with parameters  $0 < p < 1$  and  $n$ , denoted by  $X \sim \text{Binomial}(n, p)$ . The probability mass function of  $X$  is

$$f(x) = C_x^n p^x (1 - p)^{n-x} \quad \text{for } x = 0, 1, \dots, n.$$

Its mean and variances are

$$\mu = E(X) = np \quad \text{and} \quad \sigma^2 = V(X) = np(1 - p).$$

## Binomial Distribution (Examples)

1. Flip a coin 10 times. Let  $X = \#$  of heads obtained.
2. A worn machine tool produces 1% defective parts.  $X = \#$  of defective parts in the next 25 parts produced.
3. Each sample of air has a 10% chance of containing a particular rare molecule.  $X = \#$  of air samples that contain the rare molecule in the next 18 samples analyzed.
4. Of all bits transmitted through a digital transmission channel, 10% are received in error.  $X = \#$  of bits in error in the next five bits transmitted.
5. A multiple-choice test contains 10 questions, each with four choices, and you guess at each question.  $X = \#$  of questions answered correctly.
6. In the next 20 births at a hospital,  $X = \#$  of female births.
7. Of all patients suffering a particular illness, 35% experience improvement from a medication. In the next 100 patients administered the medication,  $X = \#$  of patients who experience improvement.

## Binomial Distribution (Example)

Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant. Find the probability that in the next 18 samples, exactly 2 contain the pollutant.

**Answer:** Let  $X$  = the number of samples that contain the pollutant in the next 18 samples analyzed. We know that  $X \sim \text{Binomial}(n = 18, p = 0.1)$ . Thus

$$P(X = 2) = f(2) = C_2^{18} 0.1^2 (1 - 0.1)^{16} = 0.284.$$

In addition,  $P(X > 4) = \sum_{x=5}^{18} f(x) = \sum_{x=5}^{18} C_x^{18} 0.1^x (1 - 0.1)^{18-x}$   
or  $P(X > 4) = 1 - P(X \leq 4) = 1 - \sum_{x=0}^4 C_x^{18} 0.1^x (1 - 0.1)^{18-x}$ .  
Also,  $P(3 \leq X < 7) = \sum_{x=3}^6 C_x^{18} 0.1^x (1 - 0.1)^{18-x}$  or

$$P(3 \leq X < 7) = P(X \leq 6) - P(X \leq 2). \text{ Why?}$$

The mean and variance of  $X$  are  $\mu = np = 1.8$  and  $\sigma^2 = np(1-p) = 1.62$ .

## StatEngine: binomial distribution

```
binomial.summary(n,p,plotpdf=c("TRUE","FALSE"),  
                 plotcdf=c("TRUE","FALSE"))  
  
binomial.prob(n,p,lb)  
  
binomial.prob(n,p,lb,ub,  
              inclusive=c("none","left","right","both"))
```

Example: Go back to the previous example. The binomial distribution can be defined by

```
n=18;p=0.1;  
# Some calculation:  
binomial.summary(n,p) # Mean, Variance, std  
binomial.prob(n,p,2) # P(X=2)  
binomial.prob(n,p,3,7,"left") # P(3<=X<7)
```

## Geometric Distribution (Definition) $X \sim \text{Geometric}(p)$

In a series of Bernoulli trials (independent trials with constant probability  $p$  of a success), the random variable  $X$  that equals the number of trials until the **first** success is a **geometric random variable** with parameter  $0 < p < 1$ , denoted by  $X \sim \text{Geometric}(p)$ . The probability mass function of  $X$  is

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

Its cdf is

$$F(x) = 1 - (1 - p)^x \quad \text{for } x = 1, 2, \dots$$

Its mean and variances are

$$\mu = E(X) = 1/p \quad \text{and} \quad \sigma^2 = V(X) = (1 - p)/p^2.$$

## Geometric Distribution (Example)

The probability that a wafer contains a large particle of contamination is 0.01. If it is assumed that the wafers are independent, what is the probability that exactly 125 wafers need to be analyzed before a large particle is detected?

**Answer:** Let  $X$  denote the number of samples analyzed until a large particle is detected. Then  $X \sim \text{Geometric}(p = 0.01)$ . The requested probability is

$$P(X = 125) = 0.99^{124} 0.01 = 0.0029.$$

In addition, the  $\mu = 1/p = 100$  and  $\sigma^2 = (1 - p)/p^2 = 9900$ .

## StatEngine: geometric distribution

```
geometric.summary(p,plotpdf=c("TRUE","FALSE"),
                   plotcdf=c("TRUE","FALSE"))

geometric.prob(p,lb)

geometric.prob(p,lb,ub,
               inclusive=c("none","left","right","both"))
```

Example: Go back to the previous example. The geometric distribution can be defined by

```
p=0.01;
# Some calculation:
geometric.summary(p) # Mean, Variance, std
geometric.prob(p,125) # P(X=125)
geometric.prob(p,3,Inf,"none") # P(3<X)
```

## Geometric Distribution (Lack of Memory)

A geometric random variable has been defined as the number of trials until the first success. However, because the trials are independent, the count of the number of trials until the next success can be started at any trial without changing the probability distribution of the random variable.

For example, if 100 bits are transmitted, the probability that the first error, after bit 100, occurs on bit 106 is the probability that the next six outcomes are *O O O O O E* (*O* means okay, *E* means error). This probability is  $(0.9)^5(0.1) = 0.059$ , which is identical to the probability that the initial error occurs on bit 6.

The implication of using a geometric model is that the system presumably does not wear out. The probability of an error remains constant for all transmissions. In this sense, the geometric distribution is said to lack any memory. The **lack of memory property** is discussed again in the context of an exponential random variable in a later chapter.

## Negative Binomial Distribution (Definition)

$$X \sim \text{NegBinom}(r, p)$$

In a series of Bernoulli trials (independent trials with constant probability  $p$  of a success), the random variable  $X$  that equals the number of trials until the  $r$ -th success is a **negative binomial random variable** with parameters  $0 < p < 1$  and  $r$ , denoted by  $X \sim \text{NegBinom}(r, p)$ . The probability mass function of  $X$  is

$$f(x) = C_{r-1}^{x-1} (1-p)^{(x-r)} p^r \quad \text{for } x = r, r+1, r+2, \dots$$

Its mean and variances are

$$\mu = E(X) = r/p \quad \text{and} \quad \sigma^2 = V(X) = r(1-p)/p^2.$$

## Negative Binomial Distribution (Example)

The probability that a camera passes the test is 0.8, and the cameras perform independently. What is the probability that the third failure is obtained in five or fewer tests?

**Answer:** Let  $X$  denote the number of cameras tested until three failures have been obtained. Then  $X \sim \text{NegBinom}(r = 3, p = 0.2)$ . The requested probability is

$$P(X \leq 5) = \sum_{x=3}^5 C_2^{x-1} (1 - 0.2)^{(x-3)} 0.2^3 = 0.058.$$

Its mean and variance are  $\mu = r/p = 3/0.2 = 15$  and  $\sigma^2 = r(1-p)/p^2 = 3(0.8)/0.04 = 60$ .

## StatEngine: negative binomial distribution

```
negbinom.summary(r,p,plotpdf=c("TRUE","FALSE"),  
                 plotcdf=c("TRUE","FALSE"))  
  
negbinom.prob(r,p,lb)  
  
negbinom.prob(r,p,lb,ub,  
              inclusive=c("none","left","right","both"))
```

Example: Go back to the previous example. The negative binomial distribution can be defined by

```
r=3;p=0.2;  
# Some calculation:  
negbinom.summary(r,p) # Mean, Variance, std  
negbinom.prob(r,p,5) # P(X=5)  
negbinom.prob(r,p,-Inf,5,"right") # P(X<=5)
```

## Hypergeometric Distribution (Definition)

$$X \sim \text{HyperGeo}(N, K, n)$$

**Example:** A day's production of 850 manufactured parts contains 50 parts that do not conform to customer requirements. Two parts are selected at random without replacement from the day's production.

Let  $A$  and  $B$  denote the events that the first and second parts are non-conforming, respectively. From counting parts in the sample space,  $P(B|A) = 49/849$  and  $P(A) = 50/850$ . Consequently, knowledge that the first part is nonconforming suggests that it is less likely that the second part selected is nonconforming; i.e.,  $A$  and  $B$  are not independent. Let  $X$  equal the number of nonconforming parts in the sample of size 2. Then

$$P(X = 0) = \frac{800}{850} \frac{799}{849}, \quad P(X = 2) = \frac{50}{850} \frac{49}{849} = 0.003,$$

$$P(X = 1) = \frac{800}{850} \frac{50}{849} + \frac{50}{850} \frac{800}{849} = 0.111.$$

## Hypergeometric Distribution (Definition)

$$X \sim \text{HyperGeo}(N, K, n)$$

A set of  $N$  objects contains

- ▶  $K$  objects classified as successes
- ▶  $N - K$  objects classified as failures

A sample of size  $n$  objects is selected randomly (**without replacement**) from the  $N$  objects where  $K \leq N$  and  $n \leq N$ .

The random variable  $X$  that equals the number of successes in the sample of size  $n$  is a **hypergeometric random variable**, denoted by  $X \sim \text{HyperGeo}(N, K, n)$ , and its pmf is

$$f(x) = \frac{C_x^K C_{n-x}^{N-K}}{C_n^N}, \quad \text{from } x = \max\{0, n + K - N\} \text{ to } \min\{K, n\}.$$

The mean and variance are (similar to Binomial distribution)

$$\mu = E(X) = np \quad \text{and} \quad \sigma^2 = V(X) = np(1-p) \left( \frac{N-n}{N-1} \right),$$

where  $p = K/N$  and  $\frac{N-n}{N-1}$  is the finite Population Correction Factor.

## Hypergeometric Distribution (Example)

A batch of parts contains 100 from a local supplier of circuit boards and 200 from a supplier in the next state. If four parts are selected randomly and without replacement, what is the probability they are all from the local supplier?

**Answer:** Let  $X$  equal the number of parts in the sample from the local supplier. Then  $X \sim \text{HyperGeo}(N = 300, K = 100, n = 4)$ . The requested probability is

$$P(X = 4) = \frac{C_4^{100} C_0^{200}}{C_4^{300}} = 0.0119.$$

The mean and variance are  $\mu = nK/N = 4/3$  and  $\sigma^2 = n(K/N)(1 - K/N)(N - n)/(N - 1) = 4(100/300)(200/300)296/299 = 0.88$

## StatEngine: hypergeometric distribution

```
hypergeo.summary(N,K,n,plotpdf=c("TRUE","FALSE"),  
                 plotcdf=c("TRUE","FALSE"))  
  
hypergeo.prob(N,K,n,lb)  
  
hypergeo.prob(N,K,n,lb,ub,  
              inclusive=c("none","left","right","both"))
```

Example: Go back to the previous example. The hypergeometric distribution can be defined by

```
N=300;K=100;n=4;  
# Some calculation:  
hypergeo.summary(N,K,n) # Mean, Variance, std  
hypergeo.prob(N,K,n,4) # P(X=4)  
hypergeo.prob(N,K,n,1,2,"right") # P(1<X<=2)
```

## Poisson Distribution (Definition) $X \sim \text{Poisson}(\lambda, L)$

A Poisson random variable is the number of events occurs in an interval of length  $T$  (time, area, volume), where the occurrence of the events follows a Poisson process.

### Example

1. the number of flaws in a length of 10 millimeters of wire
2. the number of customers who enter a bank in an hour
3. the number of stars in a given volume of space

A Poisson Process is a random experiment with properties:

In general, consider subintervals of small length  $\Delta t \approx 0$ :

1. The probability of more than one event in a subinterval tends to zero.
2. The probability of one event in a subinterval tends to  $\Delta t$ .
3. The event in each subinterval is independent of other subintervals.

## Poisson Distribution (Definition) $X \sim \text{Poisson}(\lambda, L)$

The random variable  $X$  that equals the number of events in an interval of length  $L$ , where the occurrence follows a Poisson process with a mean rate  $\lambda$  (events per unit length). Then  $X$  is a **Poisson random variable** with parameter  $\lambda > 0$  and  $L > 0$ , denoted by  $X \sim \text{Poisson}(\lambda, L)$ , and the pmf is

$$f(x) = \frac{e^{-\lambda L} (\lambda L)^x}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

Its mean and variance are  $\mu = E(X) = \lambda L$  and  $\sigma^2 = V(X) = \lambda L s$ .

## Poisson Distribution (Example)

Flaws occur at random along the length of a thin copper wire. Suppose that the number of flaws follows a Poisson distribution with a mean of 2.3 flaws per millimeter (Thus, we know  $\lambda = 2.3$  flaws per millimeter).

Determine the probability of 10 flaws in 5 millimeters of wire.

**Answer:** Let  $X$  denote the number of flaws in 5 millimeters of wire. Then  $X \sim \text{Poisson}(\lambda = 2.3, L = 5)$  and  $\lambda L = 11.5$ .

$$P(X = 10) = e^{-11.5} \frac{11.5^{10}}{10!} = 0.113.$$

Determine the probability of at least one flaw in 2 millimeters.

**Answer:** Let  $X$  denote the number of flaws in 2 millimeters of wire. Then  $X \sim \text{Poisson}(\lambda = 2.3, L = 2)$  and  $\lambda L = 4.6$ .  $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-4.6} = 0.9899$ .

Determine the mean and variance of flaws in 10 millimeters of wire.

**Answer:** Let  $X$  denote the number of flaws in 10 millimeters of wire. Then  $X \sim \text{Poisson}(\lambda = 2.3, L = 10)$ . Thus,  $E(X) = V(X) = \lambda L = 23$

## StatEngine: Poisson distribution

```
poisson.summary(lambda,L,plotpdf=c("TRUE","FALSE"),  
                 plotcdf=c("TRUE","FALSE"))  
  
poisson.prob(lambda,L,lb)  
  
poisson.prob(lambda,L,lb,ub,  
              inclusive=c("none","left","right","both"))
```

Example: Go back to the previous example. if  $L = 5$ , the Poisson distribution can be defined by

```
lambda=2.3;L=5  
poisson.prob(lambda,L,10) # P(X=10)  
# If L=2  
poisson.prob(lambda,2,1,Inf,"left") # P(1<=X)  
# If L=10  
poisson.summary(lambda,10)
```

# STAT 509: Statistics for Engineers

## Chapter 4: Continuous Random Variables and Probability Distributions

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

# Chapter 4: Continuous Random Variables and Probability Distributions

## Learning Objectives:

1. Determine probabilities from probability density functions.
2. Determine probabilities from cumulative distribution functions, and cumulative distribution functions from probability density functions, and the reverse.
3. Calculate means and variances for continuous random variables.
4. Understand the assumptions for continuous probability distributions.
5. Select an appropriate continuous probability distribution to calculate probabilities for specific applications.
6. Calculate probabilities, means and variances for continuous probability distributions.
7. Standardize normal random variables.

# Continuous Random Variables

A continuous random variable is one which takes values in an uncountable set.

They are used to measure physical characteristics such as height, weight, time, volume, position, etc...

## Examples

1. Let  $Y$  be the height of a person (a real number).
2. Let  $X$  be the volume of juice in a can.
3. Let  $Y$  be the waiting time until the next person arrives at the server.

# Probability Density Functions

Because the number of possible values of  $X$  is uncountably infinite,  $X$  has a distinctly different distribution from the discrete random variables studied previously.

## How to assign probabilities?

- ▶ Discrete: probability mass function (pmf)
- ▶ Continuous: probability density function (pdf)

# Probability Density Functions

Density functions are commonly used in engineering to describe physical systems; e.g., the density of a loading on a long, thin beam.

- ▶ For any point  $x$  along the beam, the density can be described by a function (in grams/cm).
- ▶ Intervals with large loadings correspond to large values for the function. The total loading between points  $a$  and  $b$  is determined as the integral of the density function from  $a$  to  $b$ .
- ▶ This integral is the area under the density function over this interval, and it can be loosely interpreted as the sum of all the loadings over this interval.

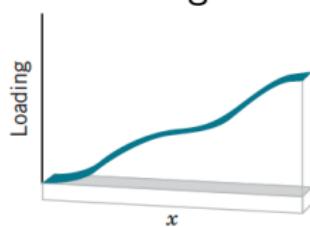


FIGURE 4.1

**Density function of a loading on a long, thin beam.**

# Probability Density Functions

Similarly, a probability density function  $f(x)$  can be used to describe the probability distribution of a continuous random variable  $X$ .

- ▶ If an interval is likely to contain a value for  $X$ , its probability is large and it corresponds to large values for  $f(x)$ .
- ▶ The probability that  $X$  is between  $a$  and  $b$  is determined as the integral of  $f(x)$  from  $a$  to  $b$ .

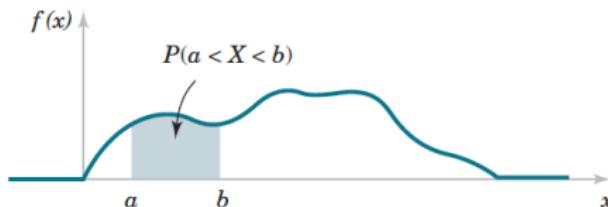


FIGURE 4.2

Probability determined from the area under  $f(x)$ .

## Probability Density Functions

For a continuous random variable  $X$ , a **probability density function** is a function such that

$$(1) \ f(x) \geq 0$$

$$(2) \ \int_{-\infty}^{+\infty} f(x)dx = 1$$

$$(3) \ P(a \leq X \leq b) = \int_a^b f(x)dx = \\ \text{area under } f(x) \text{ from } a \text{ to } b \text{ for any } a \text{ and } b$$

Note that, for a continuous random variable  $X$  and *any* value  $x$ ,

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

Interpretation: e.g., for the density function of a loading on a long, thin beam, because every point has zero width, the loading at any point is zero. Consequently,

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

# Probability Density Functions

Review of integral:

- ▶ polynomial:

$$\int_a^b x^r dx = \frac{x^{r+1}}{r+1} \Big|_a^b = \frac{b^{r+1}}{r+1} - \frac{a^{r+1}}{r+1}$$

for  $r \neq -1$ .

$$\int_a^b \frac{1}{x} dx = \log x \Big|_a^b = \log b - \log a.$$

- ▶ exponential:

$$\int_a^b e^{-x} dx = \{-e^{-x}\} \Big|_a^b = e^{-a} - e^{-b}.$$

- ▶  $\int_a^b xe^{-x} dx = \int_a^b (-x) de^{-x} = (-x)e^{-x} \Big|_a^b - \int_a^b (-e^{-x}) dx$

# Probability Density Functions

## Example 4.1

Let the continuous random variable  $X$  denote the current measured in a thin copper wire in milliamperes. Assume that the range of  $X$  is  $[4.9, 5.1]$  mA, and assume that the probability density function of  $X$  is  $f(x) = 5$  for  $4.9 \leq x \leq 5.1$ . What is the probability that a current measurement is less than 5 milliamperes?

A common rule: assume that  $f(x) = 0$  wherever it is not specifically defined; i.e.,  $f(x) = 0$  if  $x > 5.1$  or  $x < 4.9$ . Then

$$\begin{aligned} P(X < 5) &= P(4.9 \leq X < 5) = \int_{4.9}^5 f(x)dx \\ &= \int_{4.9}^5 5dx = (5x)\Big|_{4.9}^5 \\ &= 0.5 \end{aligned}$$

# Probability Density Functions

## Example 4.2

Let the continuous random variable  $X$  denote the diameter of a hole drilled in a sheet metal component. The target diameter is 12.5 millimeters. Most random disturbances to the process result in larger diameters. Historical data show that the distribution of  $X$  can be modeled by a probability density function  $f(x) = 20e^{-20(x-12.5)}$ , for  $x \geq 12.5$ . If a part with a diameter greater than 12.60 mm is scrapped, what proportion of parts is scrapped?

$$\begin{aligned} P(X > 12.60) &= \int_{12.6}^{\infty} f(x)dx = \int_{12.6}^{\infty} 20e^{-20(x-12.5)} dx \\ &= -e^{-20(x-12.5)} \Big|_{12.6}^{\infty} = e^{-20(12.6-12.5)} - 0 \\ &= 0.1353. \end{aligned}$$

```
integrand=function(x){20*exp(-20*(x-12.5))}  
integrate(integrand,12.6,Inf)  
0.1353353 with absolute error < 3.4e-05
```

# Probability Density Functions

## Example 4.2, continued

What proportion of parts is between 12.5 and 12.6 millimeters?

$$\begin{aligned} P(12.5 < X < 12.6) &= \int_{12.5}^{12.6} f(x)dx = \int_{12.5}^{12.6} 20e^{-20(x-12.5)}dx \\ &= -e^{-20(x-12.5)} \Big|_{12.5}^{12.6} = 0.8647. \end{aligned}$$

In fact,  $P(12.5 < X < 12.6) = 1 - P(X > 12.6) = 1 - 0.1353 = 0.8647$ .

```
integrand=function(x){20*exp(-20*(x-12.5))}  
integrate(integrand,12.5,12.6)  
0.8646647 with absolute error < 9.6e-15
```

# Cumulative Distribution Functions

An alternative method to describe the distribution of a discrete random variable can also be used for continuous random variables.

## Cumulative Distribution Function

The **cumulative distribution function** (cdf) of a continuous random variable  $X$  is

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du$$

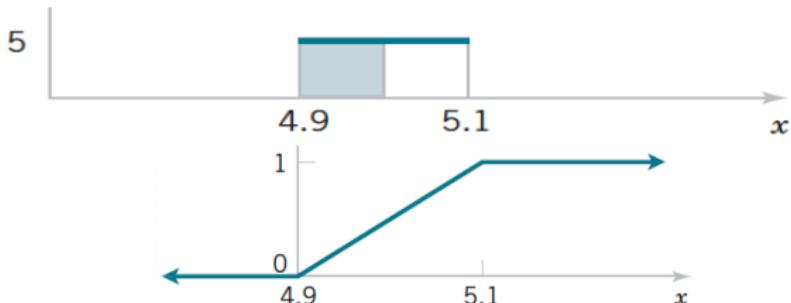
for  $-\infty < x < \infty$ .

The cumulative distribution function is defined for all real numbers.

# Cumulative Distribution Functions

## Example 4.1, continued

For the copper current measurement in Example 4.1, Find the cumulative distribution function of the random variable  $X$ .



If  $x < 4.9$ , no way to get  $X \leq x$ , then  $F(x) = 0$  for  $x < 4.9$ .

If  $x > 5.1$ ,  $X$  is definitely less than  $x$ , then  $F(x) = 1$  for  $x > 5.1$ .

If  $4.9 \leq x \leq 5.1$ , then  $F(x) = \int_{4.9}^x f(u)du = 5(x - 4.9)$ .

$$F(x) = \begin{cases} 0 & x < 4.9 \\ 5x - 24.5 & 4.9 \leq x \leq 5.1 \\ 1 & x > 5.1 \end{cases}$$

## pdf from cdf

Probability Density Function from the Cumulative Distribution Function

Given  $F(x)$

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

as long as the derivative exists.

### Example 4.4, Reaction Time

The time until a chemical reaction is complete (in milliseconds) is approximated by the cumulative distribution function

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-0.01x} & x \geq 0 \end{cases}$$

Determine the probability density function of  $X$ . What proportion of reactions is complete within 200 milliseconds?

$$f(x) = \begin{cases} 0 & x < 0 \\ 0.01e^{-0.01x} & x \geq 0 \end{cases}$$

and the probability is  $P(X \leq 200) = F(200) = 1 - e^{-2} = 0.8647$ . 14/41

## Mean and Variance of a Continuous Random Variable

Suppose that  $X$  is a continuous random variable with probability density function  $f(x)$ . The mean or expected value of  $X$ , denoted as  $\mu$  or  $E(X)$ , is

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

The variance of  $X$ , denoted as  $V(X)$  or  $\sigma^2$ , is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2.$$

The standard deviation of  $X$  is  $\sigma = \sqrt{\sigma^2}$ .

For any function of  $X$ , say  $Y = h(X)$ , then

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

Special case:  $h(X) = aX + b$ :  $E[h(X)] = aE(X) + b$  and  $V(X) = a^2V(X)$ .

## Mean and Variance of a Continuous Random Variable

Example 4.1, continued

$$E(X) = \int_{4.9}^{5.1} xf(x)dx = \int_{4.9}^{5.1} 5xdx = \frac{5x^2}{2} \Big|_{4.9}^{5.1} = 5$$

$$\begin{aligned}V(X) &= \int_{4.9}^{5.1} x^2 f(x)dx - 5^2 = \int_{4.9}^{5.1} 5x^2 dx - 25 \\&= \frac{5x^3}{3} \Big|_{4.9}^{5.1} - 25 \\&= 0.0033.\end{aligned}$$

```
integrand=function(x){5*x}  
integrate(integrand,4.9,5.1)  
5 with absolute error < 5.6e-14
```

```
integrand=function(x){5*x^2}  
integrate(integrand,4.9,5.1)  
25.00333 with absolute error < 2.8e-13
```

## Mean and Variance of a Continuous Random Variable

### Example 4.1, continued

$X$  is the current measured in milliamperes. What is the expected value of power when the resistance is 100 ohms?

**Solution:** We have  $P = 10^{-6}RI^2$  where  $I$  is the current in milliamperes and  $R$  is the resistance in ohms. Thus we define  $Y = 10^{-6}100X^2$  where  $h(X) = 10^{-6}100X^2$ .

$$E(Y) = E[h(X)] = 10^{-4} \int_{4.9}^{5.1} 5x^2 dx = 0.0025 \text{ watts.}$$

```
integrand=function(x){5*x^2*10^(-4)}
integrate(integrand,4.9,5.1)
0.002500333 with absolute error < 2.8e-17
```

## Continuous **Uniform** Distribution $X \sim U(a, b)$

- ▶ Keyword: Uniform
- ▶ pdf:  $f(x) = 1/(b - a)$  for  $a \leq x \leq b$ .
- ▶ Mean and variance:

$$\mu = E(X) = (a + b)/2, \quad \sigma^2 = V(X) = (b - a)^2/12.$$

- ▶ cdf:

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

## StatEngine: Continuous Uniform Distribution $X \sim U(a, b)$

```
# Find Mean and variance / plot the pdf/cdf:  
uniform.summary(a,b,plotpdf=c("TRUE","FALSE"),  
                 plotcdf=c("TRUE","FALSE"))  
  
# Find  $P(lb < X < ub)$ :  
uniform.prob(a,b,lb,ub)  
  
# Find  $x$  such that  $P(X < x) = q$  for a given  $q$ :  
uniform.quantile(a,b,q)
```

Example: In Example 4.1, the random variable  $X$  has a continuous uniform distribution on  $[4.9, 5.1]$ . The probability density function of  $X$  is  $f(x) = 5, 4.9 \leq x \leq 5.1$ . Find the probability  $P(4.95 < X < 5)$  and the value  $x$  such that  $P(X > x) = 0.1$

```
a=4.9;b=5.1;uniform.summary(a,b)  
uniform.prob(a,b,4.95,5.0)  
0.25  
uniform.quantile(a,b,1-0.1)  
5.08
```

## Normal Distribution $X \sim N(\mu, \sigma^2)$

- ▶ Keyword: Normal. Notation:  $X \sim N(\mu, \sigma^2)$ .
- ▶ pdf:

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

- ▶ Mean and variance:  $E(X) = \mu$ ,  $V(X) = \sigma^2$ .
- ▶ Standardization: If  $X \sim N(\mu, \sigma^2)$ , then

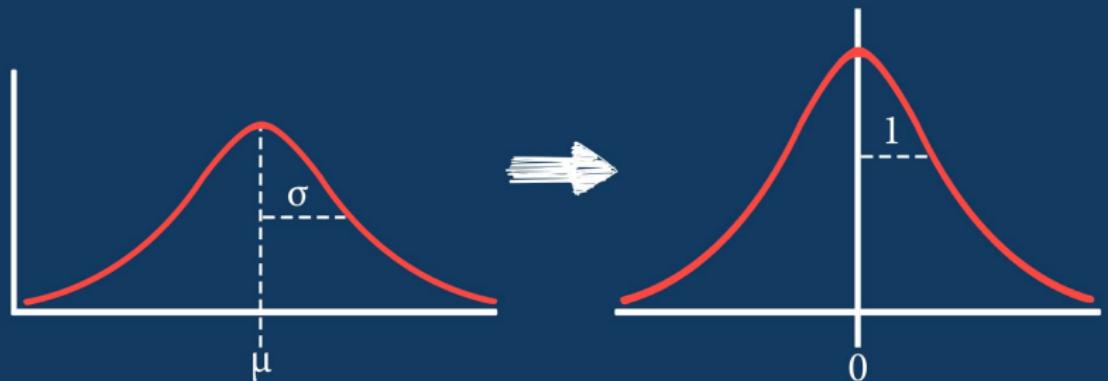
$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1); \text{standard normal distribution}$$

- ▶ cdf:  $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$  has no closed-form expression, where  $\Phi(x)$  is the cdf of  $N(0, 1)$ .

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx$$

Normal Distribution  $X \sim N(\mu, \sigma^2)$

# STANDARDIZATION



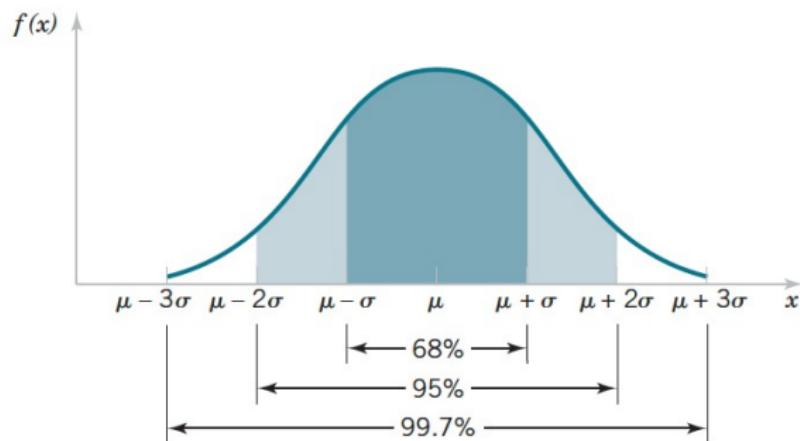
# Normal Distribution $X \sim N(\mu, \sigma^2)$

## Empirical Rule

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$$



## StatEngine: **Normal** Distribution $X \sim N(\mu, \sigma^2)$

```
# Find Mean and variance / plot the pdf/cdf:  
normal.summary(mu,sigma,plotpdf=c("TRUE","FALSE"),  
                plotcdf=c("TRUE","FALSE"))  
  
# Find P(lb<X<ub):  
normal.prob(mu,sigma,lb,ub)  
  
# Find x such that P(X<x)=q for a given q:  
normal.quantile(mu,sigma,q)
```

## Normal Distribution: Example

Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with  $\mu = 10$  and  $\sigma = 2$  mA.

1. What is the probability that the current measurement is between 9 and 11 mA?

**Solution:** Let  $X$  be the current measurement.

$$P(9 < X < 11) = \text{normal.prob}(10, 2, 9, 11) = 0.3829$$

2. Determine the value for which the probability that a current measurement is below 0.98.

**Solution:** Let  $X$  be the current measurement. We need to find  $x$  such that  $P(X < x) = 0.98$

$$x = \text{normal.quantile}(10, 2, 0.98) = 14.1075$$

## Normal Distribution: Example, continued

3. Suppose the value  $\mu$  can be adjusted while  $\sigma = 2\text{mA}$  is fixed. At which  $\mu$ , we can have  $P(X > 13) = 0.2$ ?

**Solution:**  $X \sim N(\mu, 2^2)$  implies  $Z = (X - \mu)/2 \sim N(0, 1)$ .  
Thus

$$0.2 = P(X > 13) = P\left(\frac{X - \mu}{2} > \frac{13 - \mu}{2}\right) = P\left(Z > \frac{13 - \mu}{2}\right),$$

Thus

$$\frac{13 - \mu}{2} = \text{normal.quantile}(0, 1, 1 - 0.2) = 0.8416$$

which leads to

$$\mu = 11.3168.$$

4. Now suppose  $\mu = 10\text{mA}$  is fixed while  $\sigma$  can be adjusted. At which  $\sigma$ , we can have  $P(X > 14) = 0.1$ ? (Answer: 3.1212)

## Exponential Distribution $X \sim \text{Exp}(\lambda)$

The random variable  $X$  that equals the distance between successive events from a Poisson process with mean number of events  $\lambda > 0$  per unit interval is an exponential random variable with parameter  $\lambda$ .

- ▶ Keyword: Exponential, Poisson process. Notation:  $X \sim \text{Exp}(\lambda)$ .
- ▶ Usage: model time (time to next event, lifetime of a product).
- ▶ pdf:

$$f(x) = \lambda e^{-\lambda x}, \quad 0 < x < \infty.$$

- ▶ Mean and variance:  $\mu = E(X) = \lambda^{-1}$ ,  $\sigma^2 = V(X) = \lambda^{-2}$ .
- ▶ cdf:  $F(x) = 1 - e^{-\lambda x}$ ,  $0 < x < \infty$ .
- ▶ Lack of Memory Property:

$$P(X < t_1 + t_2 | X > t_1) = P(X < t_2).$$

## StatEngine: Exponential Distribution $X \sim \text{Exp}(\lambda)$

```
# Find Mean and variance / plot the pdf/cdf:  
exponential.summary(lambda,plotpdf=c("TRUE","FALSE"),  
                     plotcdf=c("TRUE","FALSE"))  
  
# Find  $P(lb < X < ub)$ :  
exponential.prob(lambda,lb,ub)  
  
# Find x such that  $P(X < x) = q$  for a given q:  
exponential.quantile(lambda,q)
```

## Exponential Distribution: Example

In a large corporate computer network, user log-ons to the system can be modeled as a Poisson process with a mean of 25 log-ons per hour.

1. What is the probability that there are no log-ons in an interval of 6 minutes?

**Solution:** Let  $X$  denote the time in hours from the start of the interval until the first log-on. Then  $X \sim \text{Exp}(\lambda = 25)$ .

Known 6 minuets=0.1 hour!

$$P(X > 0.1) = \text{exponential.prob}(25, 0.1, Inf) = 0.0821$$

$$\text{or } = 1 - P(0 < X < 0.1) = 1 - \text{exponential.prob}(25, 0, 0.1)$$

2. Determine the interval of time such that the probability that no log-on occurs in the interval is 0.90.

**Solution:** Find  $x$  such that  $P(X > x) = 0.9$

$$x = \text{exponential.quantile}(25, 1 - 0.9) = 0.0042$$

## Exponential Distribution: Example (lack of memory)

Let  $X$  denote the time between detections of a particle with a Geiger counter and assume that  $X$  has an exponential distribution with  $E(X) = 1.4$  minutes (i.e.,  $\lambda = 1/1.4$ ). The probability that we detect a particle within 0.5 minute of starting the counter is

$$P(X < 0.5) = F(0.5) = 1 - e^{-0.5 \cdot 1.4} = 0.30.$$

Now, suppose that we turn on the Geiger counter and wait 3 minutes without detecting a particle. What is the probability that a particle is detected in the next 30 seconds?

**Solution:** It asks for the conditional property  $P(X < 3.5 | X > 3)$ . Because we have already been waiting for 3 minutes, we feel that a detection is “due.” That is, the probability of a detection in the next 30 seconds ( $= 0.5$  minutes) should be higher than 0.3. However, for an exponential distribution, this is not true because of the lack of memory property. The fact that we have waited 3 minutes without a detection does not change the probability of a detection in the next 30 seconds. We have  $P(X < 3.5 | X > 3) = P(X < 0.5) = 0.3$ .

## Gamma Distribution $X \sim \text{Gamma}(r, \lambda)$

- ▶ Keyword: Gamma, Poisson process. Notation:  $X \sim \text{Gamma}(r, \lambda)$ .  $r$ : shape,  $\lambda$ : scale.
- ▶ Usage: model time (time to next event, lifetime of a product).
- ▶ pdf:

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, \quad 0 < x < \infty,$$

where  $\Gamma(r)$  is the gamma function  $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$  for  $r > 0$ . It can be shown that  $\Gamma(r) = (r - 1)\Gamma(r - 1)$  and  $\Gamma(r) = (r - 1)!$ ,  $\Gamma(1) = 0! = 1$ ,  $\Gamma(0.5) = \sqrt{\pi}$ .

- ▶ Mean and variance:  $\mu = E(X) = r/\lambda$ ,  $\sigma^2 = V(X) = r/\lambda^2$ .
- ▶ cdf:  $F(x)$  no closed-form expression.

When  $r$  is an integer, it is also called an Erlang distribution.  $X$  can be interpreted as the distance until the next  $r$ th event in a Poisson process with mean number of events  $\lambda > 0$  per unit interval.

## StatEngine: **Gamma** Distribution $X \sim \text{Gamma}(r, \lambda)$

```
# Find Mean and variance / plot the pdf/cdf:  
gamma.summary(r,lambda,plotpdf=c("TRUE","FALSE"),  
               plotcdf=c("TRUE","FALSE"))  
  
# Find P(lb<X<ub):  
gamma.prob(r,lambda,lb,ub)  
  
# Find x such that P(X<x)=q for a given q:  
gamma.quantile(r,lambda,q)
```

## Gamma Distribution: Example

The time to prepare a slide for high-throughput genomics is a Poisson process with a mean of two hours per slide.

1. What is the probability that 10 slides require more than 25 hours to prepare?

**Solution:** Let  $X$  denote the time to prepare 10 slides.

Because of the assumption of a Poisson process,

$$X \sim \text{Gamma}(r = 10, \lambda = 1/2).$$

`P(X>25)=gamma.prob(10,0.5,25,Inf)=0.2014`

2. What are the mean and standard deviation of the time to prepare 10 slides?

`gamma.summary(10,0.5) #: \mu = 20, \sigma = 6.3246.`

3. The slides will be completed by what length of time with probability equal to 0.95?

**Solution:** Find  $x$  such that  $P(X \leq x) = 0.95$ .

`x=gamma.quantile(10,0.5,0.95)=31.4104`

## Weibull Distribution $X \sim \text{Weibull}(\beta, \delta)$

- ▶ Keyword: Weibull. Notation:  $X \sim \text{Weibull}(\beta, \delta)$ .  $\beta > 0$ : shape,  $\delta > 0$ : scale.
- ▶ Usage: model time (time to next event, lifetime of a product).
- ▶ pdf:

$$f(x) = \frac{\beta}{\delta} \left(\frac{x}{\delta}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\delta}\right)^\beta\right], \quad 0 < x < \infty,$$

- ▶ Mean and variance:

$$E(X) = \delta\Gamma(1+\beta^{-1}), \quad V(X) = \delta^2\Gamma(1+2/\beta) - \delta^2[\Gamma(1+1/\beta)]^2.$$

- ▶ cdf:

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\delta}\right)^\beta\right], \quad 0 < x < \infty.$$

## StatEngine: Weibull Distribution $X \sim \text{Weibull}(\beta, \delta)$

```
# Find Mean and variance / plot the pdf/cdf:  
weibull.summary(beta,delta,plotpdf=c("TRUE","FALSE"),  
                 plotcdf=c("TRUE","FALSE"))  
  
# Find P(lb<X<ub):  
weibull.prob(beta,delta,lb,ub)  
  
# Find x such that P(X<x)=q for a given q:  
weibull.quantile(beta,delta,q)
```

## Weibull Distribution: Example

The time to failure (in hours) of a bearing in a mechanical shaft is satisfactorily modeled as a Weibull random variable with  $\beta = 2$  and  $\delta = 5000$  hours.

1. Determine the mean and standard deviation of the time until failure.

**Solution:** Let  $X$  denote the time to failure.

$X \sim \text{Weibull}(\beta = 2, \delta = 5000)$ .

`weibull.summary(2,5000) #:  $\mu = 4431.135$ ,  $\sigma = 2316.257$`

2. Determine the probability that a bearing lasts at least 6000 hours.

$P(X \geq 6000) = \text{weibull.prob}(2, 5000, 6000, \text{Inf}) = 0.2369$

3. Find  $x$  such that  $P(X > x) = 0.05$ .

`x=weibull.quantile(2,5000,1-0.05)=8654.092`

## Lognormal Distribution $X \sim \log N(\theta, \omega^2)$

Let  $W \sim N(\theta, \omega^2)$ , then  $X = \exp(W)$  is a **lognormal** random variable; that is,  $\log X$  is a normal random variable.

- ▶ Keyword: Lognormal. Notation:  $X \sim \log N(\theta, \omega^2)$  means  $\log X \sim N(\theta, \omega^2)$ .  $\theta$ : mean of  $\log X$ .  $\omega^2$ : variance of  $\log X$ .
- ▶ Usage: model time (time to next event, lifetime of a product).
- ▶ pdf:

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} \exp\left[-\frac{(\log x - \theta)^2}{2\omega^2}\right], \quad 0 < x < \infty,$$

- ▶ Mean and variance:

$$E(X) = \exp\{\theta + \omega^2/2\}, \quad V(X) = e^{2\theta + \omega^2}(e^{\omega^2} - 1).$$

- ▶ cdf: no closed form.

## StatEngine: **Lognormal** Distribution $X \sim \log\mathcal{N}(\theta, \omega^2)$

```
# Find Mean and variance / plot the pdf/cdf:  
lognormal.summary(theta,omega,plotpdf=c("TRUE","FALSE")  
                    plotcdf=c("TRUE","FALSE"))  
  
# Find P(lb<X<ub):  
lognormal.prob(theta,omega,lb,ub)  
  
# Find x such that P(X<x)=q for a given q:  
lognormal.quantile(theta,omega,q)
```

## Lognormal Distribution: Example

The lifetime (in hours) of a semiconductor laser has a lognormal distribution with  $\theta = 10$  and  $\omega = 1.5$ .

1. Determine the mean and standard deviation of the lifetime.

**Solution:** Let  $X$  denote the lifetime.

$$X \sim \text{LogN}(\theta = 10, \omega = 1.5).$$

`lognormal.summary(10, 1.5) #:  $\mu = 67846.29$ ,  $\sigma = 197661.5$`

2. What is the probability that the lifetime exceeds 10,000 hours?

$$P(X > 10000) = \text{lognormal.prob}(10, 1.5, 10000, \text{Inf}) = 0.7007$$

3. What lifetime is exceeded by 99% of lasers?

**Solution:** Find  $x$  such that  $P(X > x) = 0.99$ .

$$x = \text{lognormal.quantile}(10, 1.5, 1 - 0.99) = 672.1478$$

## Beta Distribution $X \sim \text{Beta}(\alpha, \beta)$

- ▶ Keyword: Beta. Notation:  $X \sim \text{Beta}(\alpha, \beta)$ . Both  $\alpha > 0$  and  $\beta > 0$  are shape parameters.
- ▶ Usage: model proportion.
- ▶ pdf:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1,$$

- ▶ Mean and variance:

$$E(X) = \frac{\alpha}{\alpha + \beta}, \quad V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

- ▶ cdf: no closed form.

## StatEngine: **Beta** Distribution $X \sim \text{Beta}(\alpha, \beta)$

```
# Find Mean and variance / plot the pdf/cdf:  
beta.summary(alpha,beta,plotpdf=c("TRUE","FALSE"),  
             plotcdf=c("TRUE","FALSE"))  
  
# Find P(lb<X<ub):  
beta.prob(alpha,beta,lb,ub)  
  
# Find x such that P(X<x)=q for a given q:  
beta.quantile(alpha,beta,q)
```

## Beta Distribution: Example

The service of a constant-velocity joint in an automobile requires disassembly, boot replacement, and assembly. Suppose that the proportion of the total service time for disassembly follows a beta distribution with  $\alpha = 2.5$  and  $\beta = 1$ .

1. Determine the mean and standard deviation of the proportion.

**Solution:** Let  $X$  denote the proportion of service time for disassembly.  $X \sim \text{Beta}(\alpha = 2.5, \beta = 1)$ .

`beta.summary(2.5,1) #:  $\mu = 0.7143, \sigma = 0.2130$`

2. What is the probability that a disassembly proportion exceeds 0.7?

`P(X>0.7)=beta.prob(2.5,1,0.7,Inf)=0.59`

3. What lifetime is exceeded by 99% of lasers?

**Solution:** Find  $x$  such that  $P(X \leq x) = 0.99$ .

`x=beta.quantile(2.5,1,0.99)=0.996`

# STAT 509: Statistics for Engineers

## Chapter 6: Descriptive Statistics

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

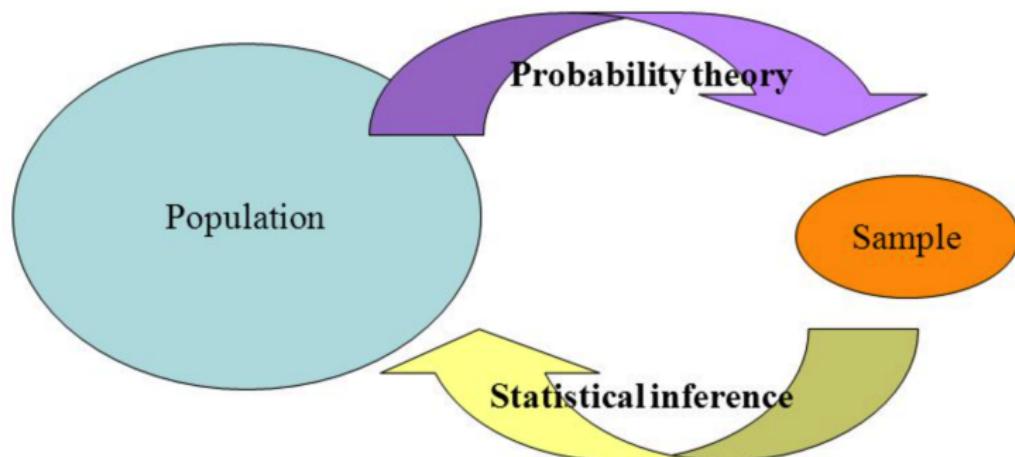
Fall 2020

## Chapter 6: Descriptive Statistics

### Learning Objectives:

1. Compute and interpret the sample mean, sample variance, sample standard deviation, sample median, and sample range
2. Explain the concepts of sample mean, sample variance, population mean, and population variance
3. Construct and interpret visual data displays, including the stem-and-leaf display, the histogram, and the box plot
4. Explain the concept of random sampling
5. Construct and interpret normal probability plots
6. Explain how to use box plots and other data displays to visually compare two or more samples of data

# Statistical Inference



**Population:** Our goal; i.e., time to next failure, lifetime of a product, salary after graduation, COVID-19 prevalence...

**Sample:** (Randomly) selected from the population. We analyze sample to learn about the population.

## Numerical Summaries of Data

Well-constructed data summaries and displays are essential to good statistical thinking because they can focus the engineer on important features of the data or provide insight about the type of model that should be used in solving the problem.

We often find it useful to describe data features

numerically and/or visually.

Now we learn some commonly-used descriptive statistics.

## Population mean/variance/standard deviation

In previous chapters, we have introduced the mean/variance/standard deviation of a probability distribution, denoted by  $\mu/\sigma^2/\sigma$ .

- ▶  $\mu$  characterizes the **center** of the distribution
- ▶  $\sigma^2$  and  $\sigma$  characterize the **variability** of the distribution.

If we think of a probability distribution as a model for the population. Then the mean/standard deviation are the center/variability of the population. We call

- ▶  $\mu$ : *population* mean
- ▶  $\sigma^2$ : *population* variance
- ▶  $\sigma$ : *population* standard deviation

## Sample mean/variance/standard deviation

Suppose we observe a sample of size  $n$  from the population. We denote the observed data by  $x_1, \dots, x_n$  (lower case  $x$ ). From these data, we can calculate

- ▶ the sample mean:

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n},$$

characterizing the **center** of the observed sample;

- ▶ the sample variance:

$$s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n(\bar{x}_n)^2}{n-1},$$

- ▶ and the sample standard deviation:

$$s_n = \sqrt{s_n^2},$$

characterizing the **variability** of the observed sample.

## Sample range & sensitive to outlier

- Another useful measure of variability (**spread**) is the sample range:

$$r_n = \max_i x_i - \min_i x_i.$$

Remark: all these statistics  $\bar{x}_n$ ,  $s_n^2$ , and  $r_n$  are **sensitive** to outliers.

### Example

Say we have observed  $x_i = i$  for  $i = 1, \dots, 20$ . Then

$$\bar{x}_n = 10.5, s_n^2 = 35, s_n = 5.9161, r_n = 19.$$

Suppose **One** careless input brought  $x_{20} = 20$  to  $x_{20} = 2000$ . Then

$$\bar{x}_n = 109.5, s_n^2 = 198035, s_n = 4455.0112, r_n = 1999.$$

$\frac{19}{20} = 95\%$  data are unchanged, but these statistics are changed dramatically.

## Robust statistics

- ▶ Sample percentiles:
  - ▶ 1st sample quartile (Q1): The 25th percentile of the sample, denoted by  $q_1$ .
  - ▶ Sample median (Q2): The 50th percentile of the sample, denoted by  $q_2$ .
  - ▶ 3rd sample quartile (Q3): The 75th percentile of the sample, denoted by  $q_3$ .
- ▶ Interquartile range (IQR):  $IQR = q_3 - q_1$ .

### Same example

Say we have observed  $x_i = i$  for  $i = 1, \dots, 20$ . Then

$$q_1 = 5.75, q_2 = 10.5, q_3 = 15.25, IQR = 9.5.$$

Suppose **One** careless input brought  $x_{20} = 20$  to  $x_{20} = 2000$ . Then

$$q_1 = 5.75, q_2 = 10.5, q_3 = 15.25, IQR = 9.5.$$

The outlier does not affect these statistics (robustness).

## data.summary in StatEngine

Name your data: e.g,

```
IornMan=1:20
```

```
Mulan=seq(2,50,by=3)
```

```
Tenet=seq(1,40,by=20)
```

```
x=c(12.6,12.9,13.4,12.3,13.6,13.5,12.6,13.1)
```

We use the textbook Table 6.2. These data are the compressive strengths in pounds per square inch (psi) of 80 specimens of a new aluminum-lithium alloy undergoing evaluation as a possible material for aircraft structural elements. The data can be read by

```
x=scan("https://raw.githubusercontent.com/Harrindy  
/StatEngine/master/Data/CompressiveStrength.csv")
```

Then run

```
data.summary(x,plot=TRUE)
```

# data.summary in StatEngine

A stem and leaf diagram is

The decimal point is 1 digit(s) to the right of the |

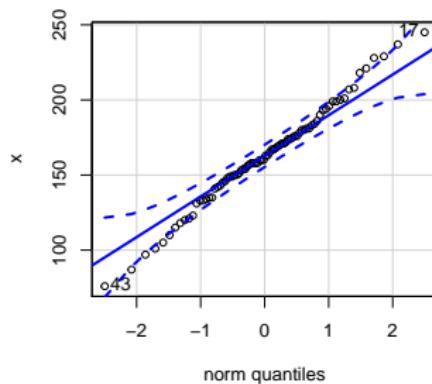
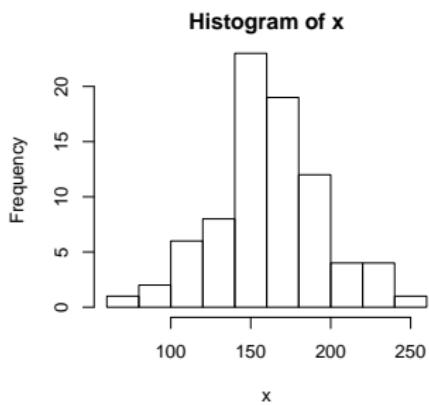
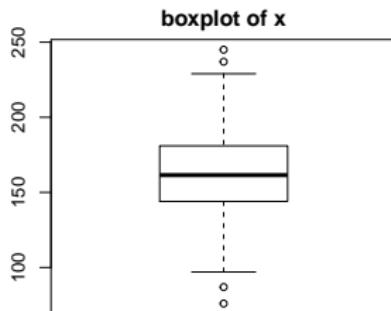
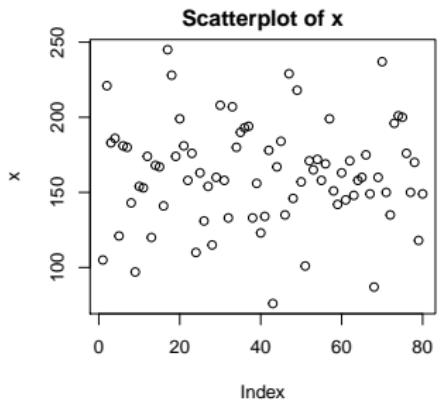
6		6
8		77
10		15058
12		013133455
14		12356899001344678888
16		00033577890112445668
18		0011346034699
20		01788
22		1897
24		5

There is no missing value.

Summary:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
statistics	min	mean	variance	std	max	range	Q1	Median	Q3	IRQ
result	76	162.6625	1140.6315	33.7732	245	169	144.5	161.5	181	36.5

# data.summary in StatEngine



## Stem and Leaf Diagram

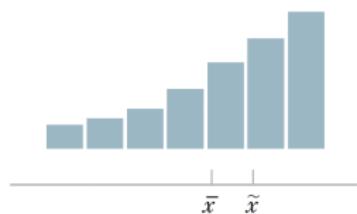
Steps to Construct a Stem-and-Leaf Diagram:

- (1) Divide each number  $x_i$  into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.
- (2) List the stem values in a vertical column.
- (3) Record the leaf for each observation beside its stem.
- (4) Write the units for stems and leaves on the display.

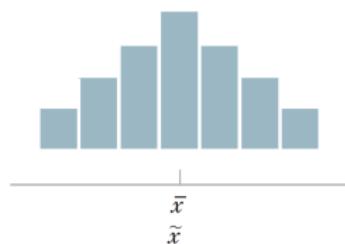
# Histogram

## Constructing a Histogram (Equal Bin Widths)

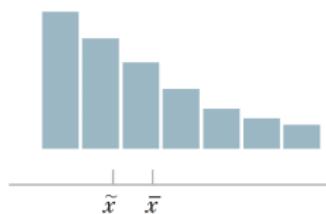
- (1) Label the bin (class interval) boundaries on a horizontal scale.
- (2) Mark and label the vertical scale with the frequencies or the relative frequencies.
- (3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.



Negative or left skew



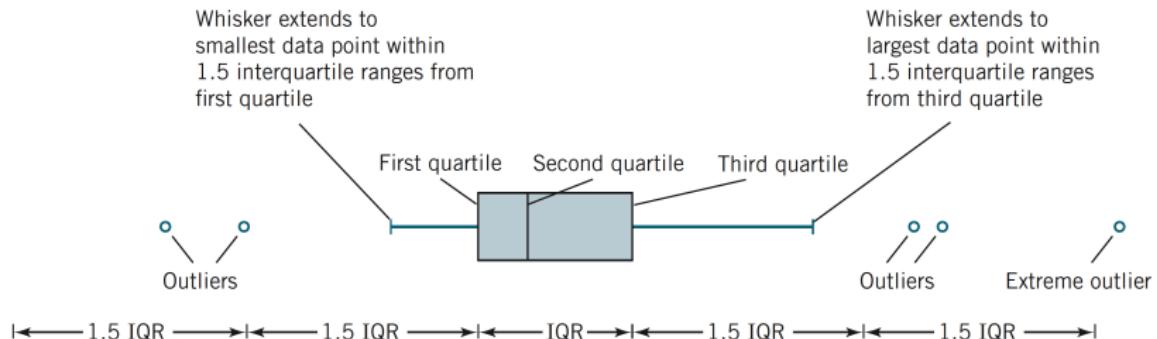
Symmetric



Positive or right skew

where  $\bar{x}$  is the sample mean and  $\tilde{x}$  is the sample median.

# Box Plots



# Box Plots

Description of a box plot.

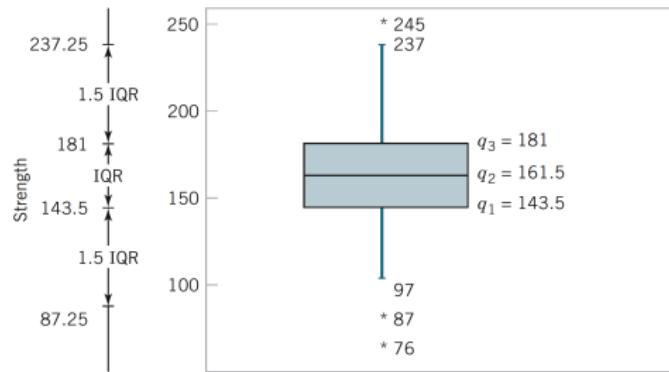


FIGURE 6.14

Box plot for compressive strength data in Table 6.2.

Box plots facilitate comparison between samples (populations).

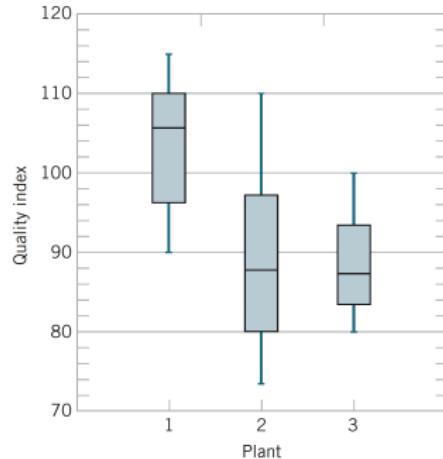


FIGURE 6.15

Comparative box plots of a quality index at three plants.

## Probability Plot (Q-Q plot)

In later chapters, we often assume a normal distribution for the population. To verify this assumption or to check whether the data are from a normal distribution, we often use Q-Q plot:

- (i) Rearrange the sample to the order statistics

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

- (ii) For each  $j$ , calculate  $p_j = (j - 0.5)/n$ .
- (iii) Find  $z_j$  such that  $P(Z \leq z_j) = p_j$ ; i.e.,  
 $p_j = \text{normal.quantile}(0, 1, p_j)$
- (iv) Plot  $(z_j, x_{(j)})$ .

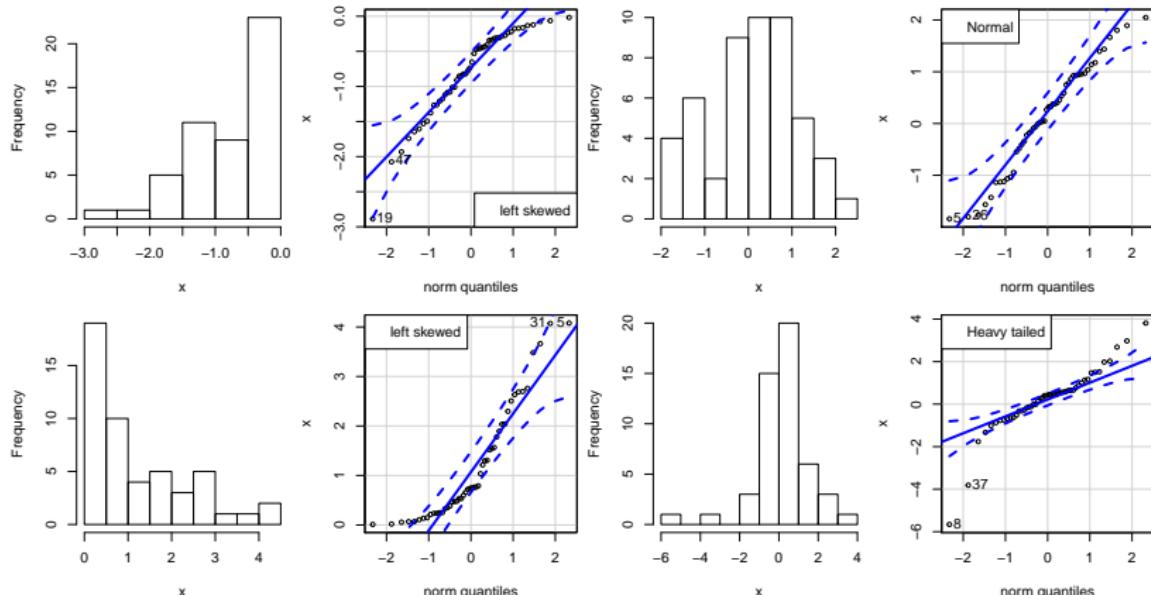
## Probability Plot (Q-Q plot) checks normality

TABLE 6.6

Calculation for Constructing a  
Normal Probability Plot

$j$	$x_{(j)}$	$(j - 0.5)/10$	$z_j$
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

# Probability Plot (Q-Q plot)



If the two dashed lines cover all the dots, we could conclude that the samples are from a normal distribution; otherwise, we might not be able to make such a conclusion.

# STAT 509: Statistics for Engineers

## Chapter 7: Point Estimation of Parameters and Sampling Distributions

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

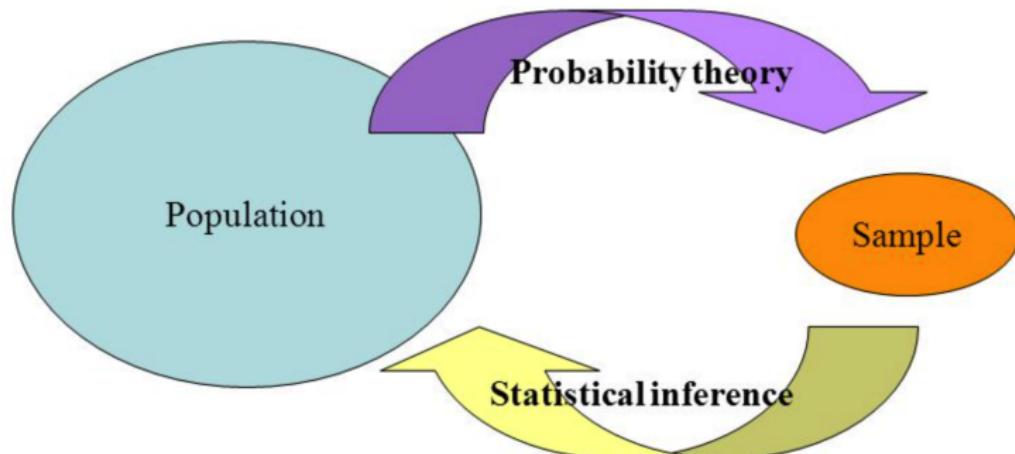
Fall 2020

# Chapter 7: Point Estimation of Parameters and Sampling Distributions

## Learning Objectives:

1. Explain the general concepts of estimating the parameters of a population or a probability distribution
2. Explain the important role of the normal distribution as a sampling distribution and the central limit theorem
3. Explain important properties of point estimators, including bias, variance, and mean square error

# Statistical Inference



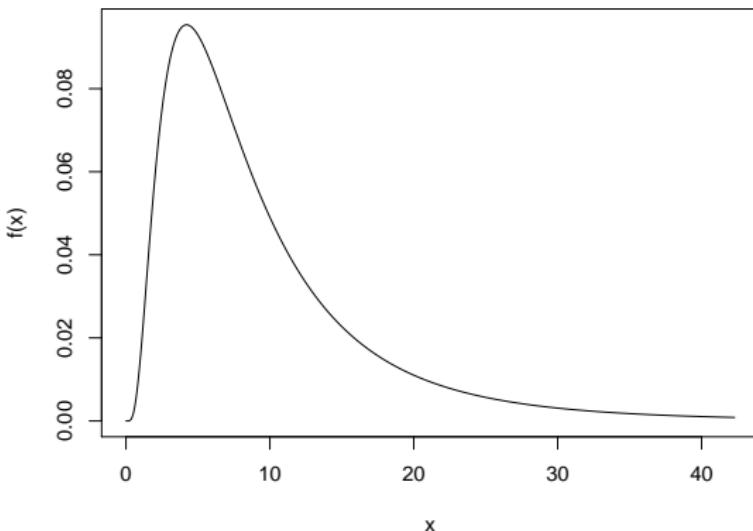
Point estimation is the first statistics inference covered in this course.

# Sampling error

In statistics, sampling error is the error caused by observing a sample instead of the whole population. A direct result is that two samples of the same size very likely give you totally different observations.

## Example

Suppose the population can be modeled by a lognormal distribution with  $\theta = 2$  and  $\omega = 0.75$ .



# Sampling error

We have twenty samples of size  $n = 10$  from the lognormal distribution on the right.

They are all different, caused by the sampling error.

Consequently, each sample has its own sample mean.

Obs	Sample									
	1	2	3	4	5	6	7	8	9	10
1	3.9950	8.2220	4.1893	15.0907	12.8233	15.2285	5.6319	7.5504	2.1503	3.1390
2	7.8452	13.8194	2.6186	4.5107	3.1392	16.3821	3.3469	1.4393	46.3631	1.8314
3	1.8858	4.0513	8.7829	7.1955	7.1819	12.0456	8.1139	6.0995	2.4787	3.7612
4	16.3041	7.5223	2.5766	18.9189	4.2923	13.4837	13.6444	8.0837	19.7610	15.7647
5	9.7061	6.7623	4.4940	11.1338	3.1460	13.7345	9.3532	2.1988	3.8142	3.6519
6	7.6146	5.3355	10.8979	3.6718	21.1501	1.6469	4.9919	13.6334	2.8456	14.5579
7	6.2978	6.7051	6.0570	8.5411	3.9089	11.0555	6.2107	7.9361	11.4422	9.7823
8	19.3613	15.6610	10.9201	5.9469	8.5416	19.7158	11.3562	3.9083	12.8958	2.2788
9	7.2275	3.7706	38.3312	6.0463	10.1081	2.2129	11.2097	3.7184	28.2844	26.0186
10	16.2093	3.4991	6.6584	4.2594	6.1328	9.2619	4.1761	5.2093	10.0632	17.9411
$\bar{x}$	9.6447	7.5348	9.5526	8.5315	8.0424	11.4767	7.8035	5.9777	14.0098	9.8727
Obs	11	12	13	14	15	16	17	18	19	20
1	7.5528	8.4998	2.5299	2.3115	6.1115	3.9102	2.3593	9.6420	5.0707	6.8075
2	4.9644	3.9780	11.0097	18.8265	3.1343	11.0269	7.3140	37.4338	5.5860	8.7372
3	16.7181	6.2696	21.9326	7.9053	2.3187	12.0887	5.1996	3.6109	3.6879	19.2486
4	8.2167	8.1599	15.5126	7.4145	6.7088	8.3312	11.9890	11.0013	5.6657	5.3550
5	9.0399	15.9189	7.9941	22.9887	8.0867	2.7181	5.7980	4.4095	12.1895	16.9185
6	4.0417	2.8099	7.1098	1.4794	14.5747	8.6157	7.8752	7.5667	32.7319	8.2588
7	4.9550	40.1865	5.1538	8.1568	4.8331	14.4199	4.3802	33.0634	11.9011	4.8917
8	7.5029	10.1408	2.6880	1.5977	7.2705	5.8623	2.0234	6.4656	12.8903	3.3929
9	8.4102	6.4106	7.6495	7.2551	3.9539	16.4997	1.8237	8.1360	7.4377	15.2643
10	7.2316	11.5961	4.4851	23.0760	10.3469	9.9330	8.6515	1.6852	3.6678	2.9765
$\bar{x}$	7.8633	11.3970	8.6065	10.1011	6.7339	9.3406	5.7415	12.3014	10.0828	9.1851

# Random sample

## Definition

To acknowledge the sampling error, we denote a **random sample** of size  $n$  by  $X_1, \dots, X_n$  (capital letters):

- (a) the  $X_i$ 's are **independent** random variables
- (b) every  $X_i$  has the **identical** probability distribution.

We also call  $X_1, \dots, X_n$  are **iid** samples from the population.

The observed sample values are denoted by (lower cast)  $x_1, \dots, x_n$ .

## Example

In the previous example, we have a random sample of size  $n = 10$ :

$X_1, \dots, X_{10}$ , where

- (a) the  $X_i$ 's are independent and
- (b) having the sample lognormal distribution.

In the first observed sample, we have

$x_1 = 3.9950, \dots, x_{10} = 16.2093, \dots$ , in the 20th observed sample,  
 $x_1 = 6.8075, \dots, x_{10} = 2.9765$ .

# Statistic

## Definition

A **statistic** is any function of a random sample.

## Example

of a random sample  $X_1, \dots, X_n$ , commonly used statistics are

- ▶ the **sample mean**:  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$   
(observed value denoted by  $\bar{x}_n$ )
- ▶ the **sample variance**:  $S_n^2 = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$   
(observed value denoted by  $s_n^2$ )
- ▶ the **sample standard deviation**:  $S_n = \sqrt{S_n^2}$   
(observed value denoted by  $s_n$ )

# Sampling distribution

## Definition

The probability distribution of a statistic is called a **sampling distribution**.

Recall that, a statistic is a function of a random sample which consists of  $n$  random variables. Thus the statistic itself is also a random variable. Hence, it has its own distribution, which is now called its **sampling distribution**

- ▶ Sampling distribution is different than the population distribution
- ▶ Sampling distribution depends on the sample size  $n$

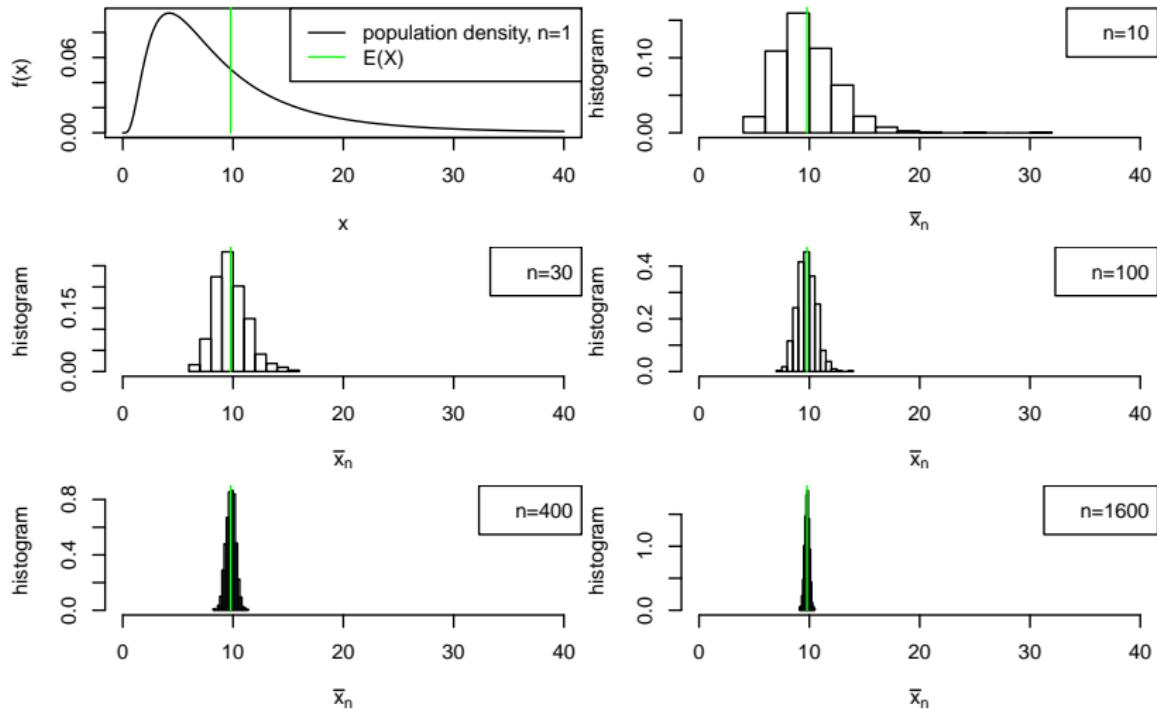
## Example

The sample mean  $\bar{X}_n$  is a statistic. Now we use simulation and histogram to illustrate the sampling distribution of  $\bar{X}_n$ :

- ▶ Suppose the population distribution is  $\log N(\theta = 2, \omega^2 = 0.75^2)$ .
- ▶ For each  $n \in \{10, 30, 100, 500, 1000\}$ , we generate 1000 samples of size  $n$ .
- ▶ From each of the 1000 samples, we calculate the sample mean  $\bar{x}_n$ . In total, we have 1000 observed sample means, each is an average of 1000 observed values from the population distribution.
- ▶ We plot the histogram of the 1000 sample means, which approximates the sampling distribution of the statistic  $\bar{X}_n$ .

## Example

Result: the distribution of  $\bar{X}_n$  becomes more and more centered at the population mean  $\mu = E(X) = 9.7889$  as  $n$  increases (i.e.,  $\bar{X}_n$  is a better and better estimator of  $\mu$  as  $n$  increases).



## Point estimator

As we just see, the sample mean  $\bar{X}_n$  is a (point) estimator of the population mean  $\mu = E(X)$ . In general,

- ▶ let  $\theta$  denote a parameter of the population distribution, which is unknown and of interest to us.
- ▶ From a random sample  $X_1, \dots, X_n$ , we calculate a statistic  $\hat{\Theta}_n = h(X_1, \dots, X_n)$  to estimate  $\theta$  and call  $\hat{\Theta}_n$  a **point estimator** of  $\theta$ .

### Example

- ▶  $\bar{X}_n = h(X_1, \dots, X_n) = n^{-1} \sum_{i=1}^n X_i$  is a point estimator of the population parameter (population mean)  $\mu = E(X)$ .
- ▶  $S_n^2 = h(X_1, \dots, X_n) = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  is a point estimator of the population variance  $\sigma^2 = V(X)$ .
- ▶ Later, we will learn how to estimate population parameters:  $p$  (population proportion),  $\mu_1 - \mu_2$  (difference in means of two population),  $p_1 - p_2$  (different in two population proportions).

## Point estimate

A **point estimate** of some population parameter  $\theta$  is a single numerical value  $\hat{\theta}_n$  of a statistic  $\hat{\Theta}_n$  (e.g., point estimator:  $\bar{X}_n$  and point estimate  $\bar{x}_n = 9.6447$ ).

Upper (random variable)	Lower (observed value)
Random variable $X$	an observed value $x$
Random sample: $X_1, \dots, X_n$	an observed sample: $x_1, \dots, x_n$
An estimator: $\hat{\Theta}_n$	an estimate: $\hat{\theta}_n$
Sample mean: $\bar{X}_n$	mean of an observed sample: $\bar{x}_n$
Sample variance: $S_n^2$	variance of an observed sample: $s_n^2$

Table 1: Notation

## Sampling distribution

Let  $X_1, \dots, X_n$  be a random (or equivalently, an iid) sample from a population with pdf  $f(x)$ . The iid means

- ▶ these  $X_i$ 's are identically distributed and have the same pdf  $f(x)$ ,
- ▶ and these  $X_i$ 's are mutually independent.

Let  $\widehat{\Theta}_n$  be a point estimator of  $\theta$ , a parameter in  $f(x)$ . For example  $\theta = \mu = E(X) = \int xf(x)dx$ . We know that

$$\widehat{\Theta}_n = h(X_1, \dots, X_n)$$

is also a random variable has its own distribution which is called the **sampling distribution**.

- ▶ The sampling distribution usually does not have the density function  $f(x)$ ,
- ▶ The sample distribution changes with  $n$  while  $f(x)$  does not change with  $n$ .

## Central limit theorem

What is the sampling distribution of a point estimator  $\hat{\Theta}_n$ ?

Often it **does not** have an analytic form.

For example, suppose the population distribution is  $\log N(\theta, \omega^2)$ . We have a random sample of size  $n$  and the sample mean is  $\bar{X}_n$ . What is the (sampling) distribution of  $\bar{X}_n$ ? No analytic form! But we can approximate it when  $n$  is large.

### Central limit theorem

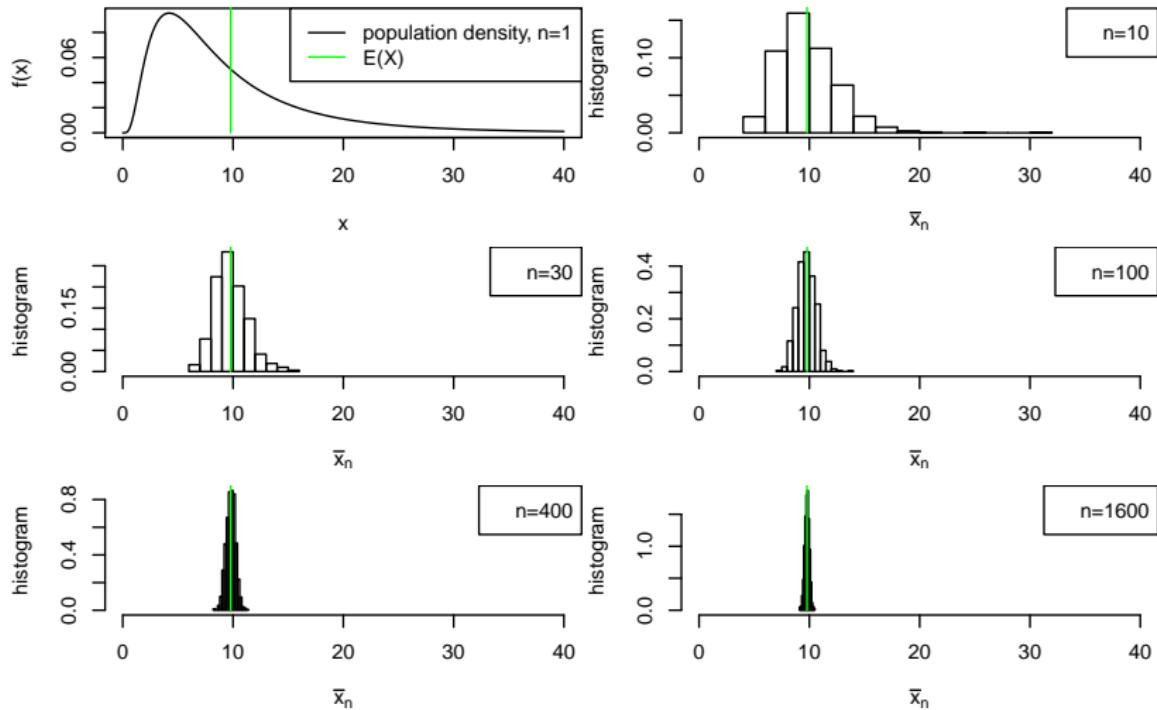
If  $X_1, \dots, X_n$  is a random sample of size  $n$  taken from a population (either finite or infinite) with mean  $\mu$  and variance  $\sigma^2$  and if  $\bar{X}_n$  is the sample mean, then

$$E(\bar{X}) = \mu, V(\bar{X}_n) = \frac{\sigma^2}{n}, \text{ and as } n \rightarrow \infty, \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \underbrace{N(0, 1)}_{\text{standard normal}};$$

or  $\bar{X}_n \sim AN(\mu, \sigma^2/n)$ , AN stands for Approximately Normal.

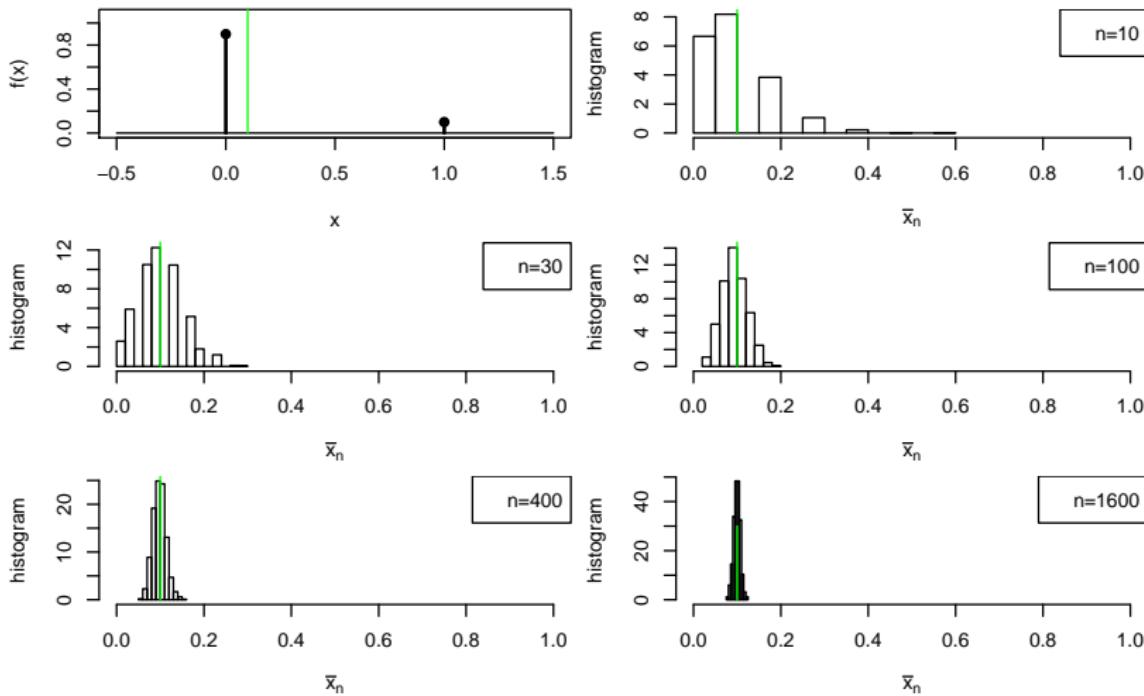
## Example

We have seen these histograms of observed sample means from the population distribution  $\log N(\theta = 2, \omega^2 = 0.75^2)$ . We know  $\mu = 9.7889, \sigma^2 = 72.3514$ . Thus from CLT,  $\bar{X}_n \sim AN(9.7889, 72.3514/n^2)$



## Example

Suppose the population distribution is  $\text{Bernoulli}(p = 0.1)$ . Then  $\mu = 0.1$  and  $\sigma^2 = p(1 - p) = 0.09$ . From CLT,  $\bar{X}_n \sim N(0.1, 0.09/n^2)$ .



## The use of CLT

- ▶ Do we have a random sample  $X_1, \dots, X_n$ ?
- ▶ Is it about  $\bar{X}_n$ ?
- ▶ What are  $\mu$  and  $\sigma^2$ ?
- ▶ CLT:  $\bar{X}_n \sim AN\left(\mu, \frac{\sigma^2}{n}\right)$
- ▶ Be careful that CLT may not work well if  $n$  is small (rule of thumb:  $n \geq 25$ ).

### A special case

If the population distribution is exactly  $N(\mu, \sigma^2)$ , then

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

holds exactly (not approximately).

## Example

Suppose a random variable  $X \sim \text{Uniform}(4, 7)$ . Find the distribution of the sample mean  $\bar{X}_n$  of a random sample of size  $n = 40$ . Find/approximate  $P(\bar{X}_{40} \geq 5.6)$ ?

**Solution:** This example asks for the distribution and a probability of the sample mean  $\bar{X}_n$ . Thus, we need to find/approximate the sampling distribution of  $\bar{X}_n$ . CLT tells  $\bar{X}_n \sim \text{AN}(\mu, \sigma^2/n)$ .

Using the information  $X \sim \text{Uniform}(4, 6)$ , we can find  $\mu = 5.5$  and  $\sigma^2 = 0.75$  by formula or StatEngine `uniform.summary(4, 7)`. Thus

$$\bar{X}_n \sim \text{AN}(5.5, 0.75^2/40).$$

Finally, we can use this normality to approximate  $P(\bar{X}_{40} \geq 5.6)$ . Using StatEngine

$$\text{normal.prob}(5.5, \sqrt{0.75^2/40}, 5.6, \text{Inf}) = 0.1995$$

## Expectation of a point estimator

An estimator  $\hat{\Theta}_n$  should be close in some sense to the true value of the unknown parameter  $\theta$ .

### Bias

The point estimator  $\hat{\Theta}_n$  is an **unbiased estimator** for the parameter  $\theta$  if

$$E(\hat{\Theta}_n) = \theta.$$

If the estimator is not unbiased (or biased), then the difference

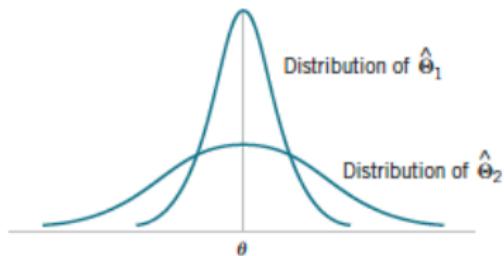
$$E(\hat{\Theta}_n) - \theta$$

is called the **bias** of the estimator  $\hat{\Theta}_n$  and denoted by  $\text{bias}(\hat{\Theta}_n)$

**For example**, we know from CLT,  $E(\bar{X}_n) = \mu$  always holds. Thus the sample mean  $\bar{X}_n$  is always an unbiased estimator of the population mean parameter  $\mu$ .

## Variance of a point estimator

Suppose that  $\hat{\Theta}_{n1}$  and  $\hat{\Theta}_{n2}$  are unbiased estimators of  $\theta$ . This indicates that the distribution of each estimator is centered at the true value of zero. However, the variance of these distributions may be different.



Though both estimators are unbiased, because  $\hat{\Theta}_{n1}$  has a smaller variance than  $\hat{\Theta}_{n2}$ , the estimator  $\hat{\Theta}_{n1}$  is more likely to produce an estimate close to the true value of  $\theta$ .

## Standard error of a point estimator

The standard error of an estimator  $\hat{\Theta}_n$  is its standard deviation given by

$$\sigma_{\hat{\Theta}_n} = \sqrt{V(\hat{\Theta}_n)}.$$

If the standard error involves unknown parameters that can be estimated, substitution of those values into  $\sigma_{\hat{\Theta}_n}$  produces an **estimated standard error**, denoted by  $\hat{\sigma}_{\hat{\Theta}_n}$  or  $SE(\hat{\Theta}_n)$ .

For example, Suppose that we are sampling from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , the standard error of  $\bar{X}_n$  is

$$\sigma_{\bar{X}_n} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

which involves the unknown parameter  $\sigma$ . We can estimate  $\sigma$  by the sample standard deviation  $S_n$  and substitute it into the preceding equation. The estimated standard error of  $\bar{X}_n$  would be

$$SE(\bar{X}_n) = \frac{S_n}{\sqrt{n}}.$$

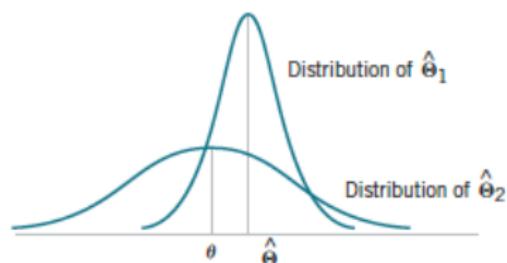
## Mean squared error of a point estimator

Sometimes it is necessary to use a biased estimator. In such cases, the mean squared error of the estimator can be important.

### MSE

The mean squared error of an estimator  $\hat{\Theta}_n$  of the parameter  $\theta$  is defined as

$$\text{MSE}(\hat{\Theta}_n) = E(\hat{\Theta}_n - \theta)^2 = \text{bias}^2(\hat{\Theta}_n) + V(\Theta_n).$$



Though  $\hat{\Theta}_{n1}$  is a biased estimator while  $\hat{\Theta}_{n2}$  is unbiased, the mse of  $\hat{\Theta}_{n1}$  is smaller than the one of  $\hat{\Theta}_{n2}$ , thus  $\hat{\Theta}_{n1}$  is more likely to produce an estimate close to the true value of  $\theta$ .

## Mean squared error of a point estimator

The mean squared error is an important criterion for comparing two estimators  $\hat{\Theta}_{n2}$  and  $\hat{\Theta}_{n1}$ . Define the **relative efficiency** of  $\hat{\Theta}_{n2}$  to  $\hat{\Theta}_{n1}$  by

$$\text{RE}(\hat{\Theta}_{n2} \text{ to } \hat{\Theta}_{n1}) = \frac{\text{MSE}(\hat{\Theta}_{n1})}{\text{MSE}(\hat{\Theta}_{n2})}.$$

If  $\text{RE}(\hat{\Theta}_{n2} \text{ to } \hat{\Theta}_{n1}) < 1$ , then  $\hat{\Theta}_{n1}$  is a more efficient estimator of  $\theta$  than  $\hat{\Theta}_{n2}$  in the sense that it has a smaller mean squared error.

For example, we estimate the population mean  $\mu$  by  $\bar{X}_n$ . Let  $\hat{\mu}_{n1} = \bar{X}_{n_1}$  and  $\hat{\mu}_{n2} = \bar{X}_{n_2}$ , where  $n_1 < n_2$ . Then

$$\text{RE}(\hat{\mu}_{n2} \text{ to } \hat{\mu}_{n1}) = \frac{\text{MSE}(\hat{\mu}_{n1})}{\text{MSE}(\hat{\mu}_{n2})} = \frac{\sigma^2/n_1}{\sigma^2/n_2} = \frac{n_2}{n_1} > 1.$$

Thus  $\hat{\mu}_{n2} = \bar{X}_{n_2}$  is a more efficient estimator of  $\mu$  than  $\hat{\mu}_{n1} = \bar{X}_{n_1}$ , which makes sense because  $n_2 > n_1$ ; i.e., more samples bring more information and thus lead to more accurate estimation.

## Different types of point estimation

Data on the oxide thickness of semiconductor wafers are as follows:  
425, 431, 416, 419, 421, 436, 418, 410, 431, 433, 423, 426, 410,  
435, 436, 428, 411, 426, 409, 437, 422, 428, 413, 416.

```
#Define your data
```

```
ot=c(425, 431, 416, 419, 421, 436, 418, 410, 431, 433,  
    423, 426, 410, 435, 436, 428, 411, 426, 409, 437,  
    422, 428, 413, 416)
```

```
#Summarize it
```

```
data.summary(ot)
```

Summary:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
statistics	min	mean	variance	std	max	range	Q1	Median	Q3	IRQ
result	409	423.3333	82.4928	9.0826	437	28	416	424	431	15

## Different types of point estimation

Summary:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
statistics	min	mean	variance	std	max	range	Q1	Median	Q3	IRQ
result	409	423.3333	82.4928	9.0826	437	28	416	424	431	15

- (a) Calculate a point estimate of the mean oxide thickness for all wafers in the population.

**Solution:** We use the sample mean:  $\hat{\mu} = \bar{x}_n = 423.3333$ .

- (b) Calculate a point estimate of the standard deviation of oxide thickness for all wafers in the population.

**Solution:** We use the sample standard deviation:  $\hat{\sigma} = 9.0826$ .

- (c) Calculate the standard error of the point estimate from part (a).

**Solution:**  $SE(\bar{X}_n) = \frac{S_n}{\sqrt{n}}$ . Thus we have it as  $\frac{9.0826}{\sqrt{24}} = 1.854$ .

## Different types of point estimation

Summary:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
statistics	min	mean	variance	std	max	range	Q1	Median	Q3	IRQ
result	409	423.3333	82.4928	9.0826	437	28	416	424	431	15

- (d) Calculate a point estimate of the median oxide thickness for all wafers in the population.

**Solution:** We use the sample median: 424.

- (e) Calculate a point estimate of the proportion of wafers in the population that have oxide thickness of more than 430 angstroms.

**Solution:** We use the sample proportion: 29.17%, which is the proportion of wafers in the sample that have oxide thickness of more than 430 angstroms.

`mean(ot>430)`

# STAT 509: Statistics for Engineers

## Chapter 8: Statistical Intervals for a Single Sample

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

## Chapter 8: Statistical Intervals for a Single Sample

### Learning Objectives:

1. Construct confidence intervals on the mean of a normal distribution, using either the normal distribution or the  $t$  distribution method
2. Construct confidence intervals on the variance and standard deviation of a normal distribution
3. Construct confidence intervals on a population proportion
4. Construct a prediction interval for a future observation

## An interval estimator

Estimating an unknown parameter  $\theta$  by a point estimator  $\hat{\Theta}_n$  is useful. However, it is like shooting a bird with a pistol. Often  $\hat{\Theta}_n$  has a continuous distribution, and if so, from Chapter 4,  $P(\hat{\Theta}_n = \theta) = 0$ ; i.e., we never capture the true parameter by using a point estimator even though it is an unbiased estimator.

Why not shoot a bird using a shotgun or capture it using a net? Translating to statistical language, why not use an interval to capture the true parameter? This motivates the consideration of interval estimators.

We estimate  $\theta$  by an interval  $[L_n, U_n]$ , where  $L_n$  and  $U_n$  are two statistics computed from a random sample of size  $n$  such that

$$P[L_n \leq \theta \leq U_n] = 1 - \alpha,$$

for some pre-specified  $\alpha \in (0, 1)$ . We call  $[L_n, U_n]$  a  $100(1 - \alpha)\%$  confidence interval estimator of  $\theta$  and  $100(1 - \alpha)\%$  the confidence level of this interval estimator.

## Confidence Interval on the Mean of a Normal Distribution, Variance Known

A confidence interval estimator is often built from a point estimator and the sampling distribution of the point estimator.

We consider building a confidence interval estimator of  $\mu$  in the normal distribution  $N(\mu, \sigma^2)$  where  $\sigma^2$  is known (based on historical data).

From Chapter 7, a good point estimator of  $\mu$  is  $\bar{X}_n$ , and the sampling distribution of  $\bar{X}_n$  is  $N(\mu, \sigma^2/n)$ . Thus

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

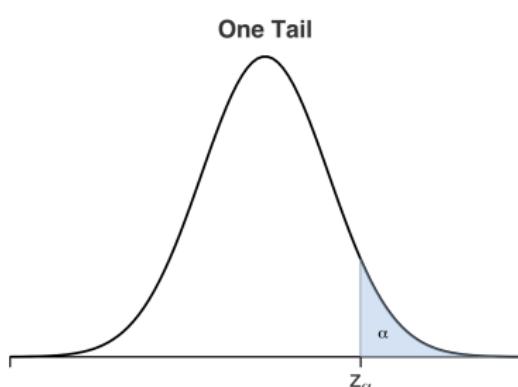
# Confidence Interval on the Mean of a Normal Distribution, Variance Known

For  $a \in (0, 1)$ , define  $z_a$  to be a quantile value of  $Z \sim N(0, 1)$  such that

$$P(Z > z_a) = a$$

Using StatEngine:

$$z_a = \text{normal.quantile}(0, 1, 1 - a).$$



Often used  $z_a$  values:

$$z_{0.005} = 2.5758$$

$$z_{0.025} = 1.96$$

$$z_{0.05} = 1.6449$$

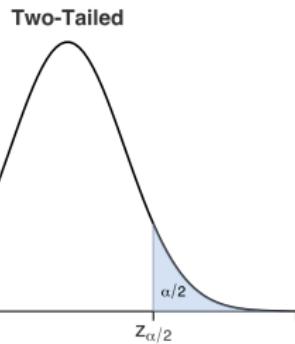
$$z_{0.1} = 1.2816.$$

# Confidence Interval on the Mean of a Normal Distribution, Variance Known

For any  $\alpha < 0.5$ , we know that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Thus



$$\begin{aligned}1 - \alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\&= P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\&= P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\&= P\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

## Confidence Interval on the Mean of a Normal distribution, variance known

Two-sided confidence interval

If  $\bar{X}_n$  is the sample mean of size  $n$  from a **normal** population with **known variance**  $\sigma^2$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimator on  $\mu$  is given by

$$\left[ L_n = \bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, U_n = \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

where  $z_{\alpha/2} = \text{normal.quantile}(0, 1, 1 - \alpha/2)$ .

Based on an observed sample  $x_1, \dots, x_n$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimate on  $\mu$  is

$$\left[ \bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Its length is  $2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ , which becomes smaller if either  $n$  is larger (more data) or  $\alpha$  is larger (less confidence level).

## One-sided confidence bounds for the Mean of a Normal distribution, variance known

Similarly, we have  $1 - \alpha = P(Z \leq z_\alpha)$  and  $1 - \alpha = P(-z_\alpha \leq Z)$ .

Plugging  $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ , we obtain

$$1 - \alpha = P\left(\mu \leq \bar{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = P\left(\bar{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu\right)$$

One-sided confidence bound

A  $100(1 - \alpha)\%$  upper-confidence bound for  $\mu$  is

$$\bar{x}_n + z_\alpha \frac{\sigma}{\sqrt{n}}$$

and a  $100(1 - \alpha)\%$  lower-confidence bound for  $\mu$  is

$$\bar{x}_n - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

## Example

ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of A238 steel cut at  $60^{\circ}\text{C}$  are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3. Assume that impact energy is normally distributed with  $\sigma = 1\text{J}$ . We want to find 95% and 99% CIs for  $\mu$ , the mean impact energy.

```
x=c(64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3)
Zinterval(level=0.95,sigma=1,sample=x)
Zinterval(level=0.99,sigma=1,sample=x)
```

A 95% CI for  $\mu$  is [63.8402, 65.0798]

A 99% CI for  $\mu$  is [63.6455, 65.2746]

When confidence level increases (or  $\alpha$  decreases), confidence interval becomes wider! (A larger net captures a bird more easily.)

## Interpreting a confidence interval

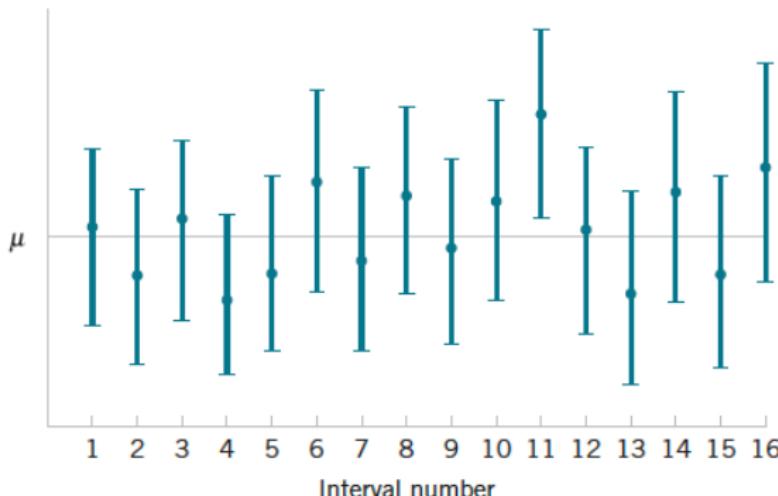
How does one interpret a confidence interval? In the previous example, a 95% CI is  $63.8402 \leq \mu \leq 65.0798$ , so it is tempting to conclude that  $\mu$  is within this interval with probability 0.95.

However, with a little reflection, it is easy to see that this **cannot be correct**; the true value of  $\mu$  is unknown, and the statement  $63.8402 \leq \mu \leq 65.0798$  is either correct (true with probability 1) or incorrect (false with probability 1).

The correct interpretation lies in the realization that a CI is a random interval because in the probability statement defining the end-points of the interval, both  $L_n$  and  $U_n$  are random variables. Consequently, the correct interpretation of a  $100(1 - \alpha)\%$  CI depends on the **relative frequency view of probability**. Specifically, if an infinite number of random samples are collected and a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  is computed from each sample,  $100(1 - \alpha)\%$  of these intervals will contain the true value of  $\mu$ .

## Interpreting a confidence interval

The situation is illustrated in the following figure, which shows several  $100(1 - \alpha)\%$  confidence intervals for the mean  $\mu$  of a normal distribution. The dots at the center of the intervals indicate the point estimate of  $\mu$  (that is,  $\bar{x}_n$ ). Notice that one of the intervals fails to contain the true value of  $\mu$ . If this were a 95% confidence interval, in the long run only 5% of the intervals would fail to contain  $\mu$ .



## Choice of sample size

A  $100(1 - \alpha)\%$  CI takes the form of

$$\left[ \bar{x}_n - \underbrace{z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\text{margin of error}}, \bar{x}_n + \underbrace{z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\text{margin of error}} \right].$$

When  $\alpha$  or the confidence level is fixed, the margin of error becomes smaller if  $n$  increases. It means if we have more data, the  $100(1-\alpha)\%$  CI has more precision of estimation. Suppose we specify the margin of error to be  $E$ , what is the smallest amount sample to collect?

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \implies n = \left\lceil \left( \frac{z_{\alpha/2} \sigma}{E} \right)^2 \right\rceil,$$

where  $\lceil a \rceil$  means the ceiling of  $a$ .

## Choice of sample size

Consider the previous example and suppose that we want to determine how many specimens must be tested to ensure that the 95% CI on  $\mu$  for A238 steel cut at  $60^\circ C$  has a length of at most 1.0 J.

**Solution:** The length is at most 1J, meaning the margin of error  $E$  is at most 0.5J. Thus ( $\alpha = 0.05$ ,  $\sigma = 1$ )

$$n = \left\lceil \left( \frac{z_{\alpha/2}\sigma}{E} \right)^2 \right\rceil = \left\lceil \left( \frac{z_{0.025} \times 1}{0.5} \right)^2 \right\rceil = 16.$$

```
sample.size.Zinterval(level=0.95,sigma=1,E=0.5)
```

CI for  $\mu$  when  $\sigma$  is known and the distribution is normal:

`Zinterval(level=?, sigma=?, sample=? )`

`Zinterval(level=?, sigma=?, n=?, barx=? )`

`sample.size.Zinterval(level=?, sigma=?, E=? )`

What if we do not know the variance  $\sigma^2$ ?

What if it is not normal?

## Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

Let  $X_1, \dots, X_n$  be a random sample from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . We know that

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

When  $\sigma$  is unknown, we replace  $\sigma$  by its estimator  $S_n$  and obtain

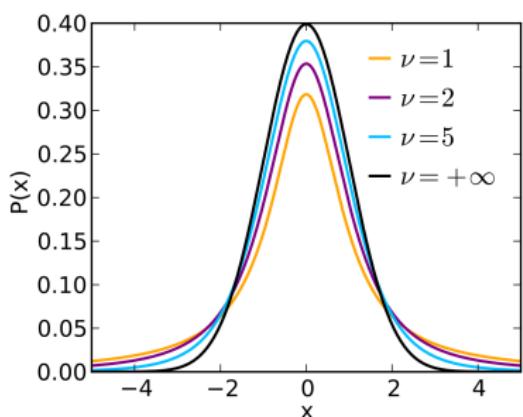
$$T_{n-1} = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

where  $t(n-1)$  stands for the **student  $t$  distribution** with degree of freedom  $n-1$ .

## Student $t(\nu)$ distribution

It is a continuous distribution with one parameter  $\nu$ , the degree of freedom. Its pdf is

$$f(x) = \frac{\Gamma(\{\nu + 1\}/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \frac{1}{[(x^2/\nu) + 1]^{(\nu+1)/2}}, -\infty < x < \infty.$$



- ▶ Similar to normal distributions: bell shape, symmetric with respect to 0
- ▶ Heavier tails than  $N(0, 1)$  when  $\nu$  is small
- ▶ When  $\nu \rightarrow \infty$ ,  $t(\nu)$  converges to  $N(0, 1)$

# Confidence Interval on the Mean of a Normal Distribution, Variance unknown

For  $a \in (0, 1)$ , define  $t_{n-1,a}$  to be a quantile value of  $t(n-1)$  such that

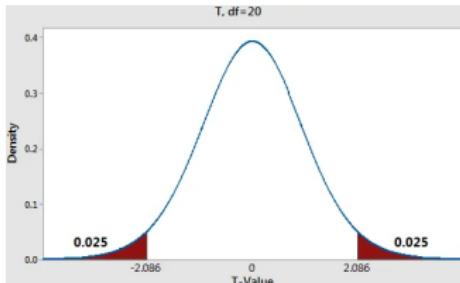
$$P(T_{n-1} > t_{n-1,a}) = a$$

Using StatEngine:

$$t_{n-1,a} = t.quantile(df = n - 1, 1 - a).$$

For any  $\alpha < 0.5$ , we know that

$$P(-t_{n-1,\alpha/2} \leq T_{n-1} \leq t_{n-1,\alpha/2}) = 1 - \alpha$$



$$\begin{aligned} &= P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) \\ &= P\left(\bar{X}_n - t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}}\right) \end{aligned}$$

## Confidence Interval on the Mean of a Normal distribution, variance unknown

Two-sided confidence interval

If  $\bar{X}_n$  is the sample mean of size  $n$  from a **normal** population with **unknown variance**  $\sigma^2$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimator on  $\mu$  is given by

$$\left[ L_n = \bar{X}_n - t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}}, U_n = \bar{X}_n + t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

where  $t_{n-1,\alpha/2} = t.quantile(df = n - 1, 1 - \alpha/2)$ .

Based on an observed sample  $x_1, \dots, x_n$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimate on  $\mu$  is

$$\left[ \bar{x}_n - t_{n-1,\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,\alpha/2} \frac{s_n}{\sqrt{n}} \right].$$

Its length is  $2 \times t_{n-1,\alpha/2} \frac{s_n}{\sqrt{n}}$ , which becomes smaller if either  $n$  is larger (more data) or  $\alpha$  is larger (less confidence level).

# One-sided confidence bounds for the Mean of a Normal distribution, variance unknown

One-sided confidence bound

A  $100(1 - \alpha)\%$  upper-confidence bound for  $\mu$  is

$$\bar{x}_n + t_{n-1,\alpha} \frac{s_n}{\sqrt{n}}$$

and a  $100(1 - \alpha)\%$  lower-confidence bound for  $\mu$  is

$$\bar{x}_n - t_{n-1,\alpha} \frac{s_n}{\sqrt{n}}.$$

StatEngine:

`Tinterval(level=?, sample=?)`

`Tinterval(level=?, n=?, barx=?, s=? )`

Remark: T-intervals are quite robust to the normality assumption when  $n$  is small. Thus, in practice, even if we do not have normality, one can still use T-intervals.

## Large-Sample Confidence Interval on the Mean of a population

### Two-sided confidence interval

If  $\bar{X}_n$  is the sample mean of size  $n$  from a population with mean  $\mu$  and a finite variance  $\sigma^2$ . When  $n$  is large ( $n \geq 25$ ), the CLT (plus the Slutsky theorem) tells

$$\frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim AN(0, 1).$$

Thus, a large-sample confidence interval estimator for  $\mu$  with confidence level of approximately  $100(1 - \alpha)\%$  is given by

$$\left[ L_n = \bar{X}_n - z_{\alpha/2} \frac{s_n}{\sqrt{n}}, U_n = \bar{X}_n + z_{\alpha/2} \frac{s_n}{\sqrt{n}} \right].$$

## One-sided confidence bounds for the Mean of a Normal distribution, variance known

Based on an observed sample  $x_1, \dots, x_n$ , a large-sample confidence interval estimator for  $\mu$  with confidence level of approximately  $100(1 - \alpha)\%$  is given by

$$\left[ \bar{x}_n - z_{\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{s_n}{\sqrt{n}} \right].$$

One-sided large-sample confidence bound

A  $100(1 - \alpha)\%$  large-sample upper-confidence bound for  $\mu$  is

$$\bar{x}_n + z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

and a  $100(1 - \alpha)\%$  large-sample lower-confidence bound for  $\mu$  is

$$\bar{x}_n - z_{\alpha} \frac{\sigma}{\sqrt{n}}.$$

## StatEngine summary of CIs on the population mean

CI for  $\mu$  when  $\sigma$  is known and the distribution is normal:

Zinterval(level=?, sigma=?, sample=? )

Zinterval(level=?, sigma=?, n=?, barx=? )

sample.size.Zinterval(level=?, sigma=?, E=? )

CI for  $\mu$  when  $\sigma$  is unknown and the distribution is normal (or for any distribution, but in this case, it provides approximated CIs):

Tinterval(level=?, sample=? )

Tinterval(level=?, n=?, barx=? , s=? )

Large-sample CI ( $n \geq 25$ ) for  $\mu$  under any distribution:

AZinterval(level=?, sample=? )

AZinterval(level=?, n=?, barx=? , s=? )

Both T-interval and AZ-interval are approximated CIs when normality does not hold. The T-intervals are more conservative (wider).

## Practice 1

An article in the Journal of Materials Engineering [“Instrumented Tensile Adhesion Tests on Plasma Sprayed Thermal Barrier Coatings” (1989, Vol. 11(4), pp. 275–282)] describes the results of tensile adhesion tests on 22 U-700 alloy specimens. The load at specimen failure is as follows (in megapascals):

```
x=c(19.8, 10.1, 14.9, 7.5, 15.4, 15.4, 15.4, 18.5, 7.9, 12.7, 11.9, 11.4, 11.4, 14.1, 17.6, 16.7, 15.8, 19.5, 8.8, 13.6, 11.9, 11.4)
```

Find a 95% CI on  $\mu$ , the population mean load at specimen failure.

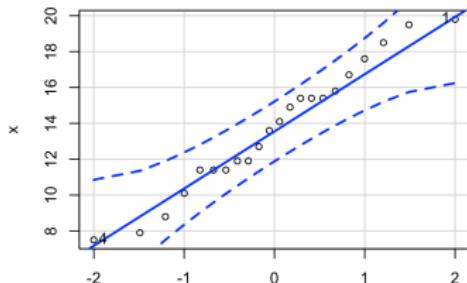
**Solution:** The sample size is  $n = 22 < 25$ , large-sample inference might not work. We do not know the population variance  $\sigma$  nor the type of the population distribution (did not say normal). But we can check normality first using the QQ plot.

```
data.summary(x)
```

It appears that the sample follows a normal distribution. Thus use the T-interval.

```
Tinterval(level=0.95,sample=x)
```

Conclusion: based on the data, we are 95% confident that the population mean load at specimen failure falls between 12.1381 and 15.2892.



## Practice 2

An article in the 1993 volume of the Transactions of the American Fisheries Society reports the results of a study to investigate the mercury contamination in large mouth bass. A sample of fish was selected from 53 Florida lakes, and mercury concentration in the muscle tissue was measured (ppm). The mercury concentration values were

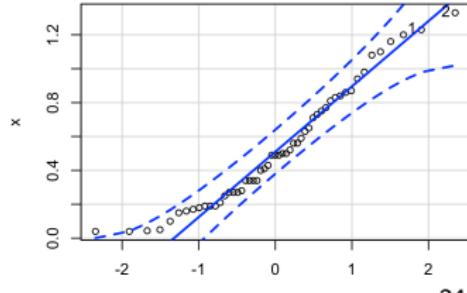
```
x=scan("https://raw.githubusercontent.com/Harrindy/StatEngine/master/Data/Mercury.csv")
```

Find a 95% confidence interval estimate for  $\mu$ , the population mean mercury concentration.

**Solution:** The dashed lines do not cover all dots; i.e., the sample might not be from a normal distribution. But  $n = 53 > 25$ , we could use a large-sample CI.

```
data.summary(x)
AZinterval(level=0.95,sample=x)
Tinterval(level=0.95,sample=x) #Try this!
```

Conclusion: based on the data, we are 95% confident that the population mean mercury concentration falls between 0.4311 and 0.6188.



## Practice 3

Past experience has indicated that the breaking strength of yarn used in manufacturing drapery material is normally distributed and that  $\sigma = 2$  psi. A random sample of nine specimens is tested, and the average breaking strength is found to be 98 psi. Find a 95% two-sided confidence interval on the true mean breaking strength.

**Solution:** Normality and known  $\sigma = 2$ .

```
Zinterval(level=0.95,sigma=2,n=9,barx=98)
```

Conclusion: based on the data, we are 95% confident that the true mean breaking strength falls between 96.6934 and 99.3066.

## Practice 4

A confidence interval estimate is desired for the gain in a circuit on a semiconductor device. Assume that gain is normally distributed. Consider the following cases where we suppose the sample standard deviation  $s_n$  is always 20.

- (a) Find a 95% CI for  $\mu$  when  $n = 10$  and  $\bar{x}_n = 1000$ .
- (b) Find a 95% CI for  $\mu$  when  $n = 25$  and  $\bar{x}_n = 1000$ .
- (c) Find a 99% CI for  $\mu$  when  $n = 10$  and  $\bar{x}_n = 1000$ .
- (d) Find a 99% CI for  $\mu$  when  $n = 25$  and  $\bar{x}_n = 1000$ .

**Solution:** Normality and  $\sigma$  unknown.

```
Tinterval(level=0.95,n=10,barx=1000,s=20)
```

```
[ 985.6929 , 1014.307 ]
```

```
Tinterval(level=0.95,n=25,barx=1000,s=20)
```

```
[ 991.7444 , 1008.256 ]
```

```
Tinterval(level=0.99,n=10,barx=1000,s=20)
```

```
[ 979.4462 , 1020.554 ]
```

```
Tinterval(level=0.99,n=25,barx=1000,s=20)
```

```
[ 988.8122 , 1011.188 ]
```

## Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

StatEngine:

Chi2interval(level=?,sample=?)

Chi2interval(level=?,n=?,s=? )

Reasoning: Let  $X_1, \dots, X_n$  be a random sample from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , and let  $S_n^2$  be the sample variance. Then the random variable

$$X_{n-1}^2 = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

where  $\chi^2(n-1)$  stands for the chi-square distribution with  $n-1$  degrees of freedom.

# Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Let  $\chi_{n-1,\alpha/2}^2$  and  $\chi_{n-1,1-\alpha/2}^2$  be the value such that

$P(\chi_{n-1}^2 > \chi_{n-1,\alpha/2}^2) = \alpha/2$ , and  $P(\chi_{n-1}^2 > \chi_{n-1,1-\alpha/2}^2) = 1-\alpha/2$ ,

respectively, where  $\chi_{n-1,\alpha/2}^2 = \text{Chi2.quantile}(df = n - 1, 1 - \alpha/2)$  and  $\chi_{n-1,1-\alpha/2}^2 = \text{Chi2.quantile}(df = n - 1, \alpha/2)$ .

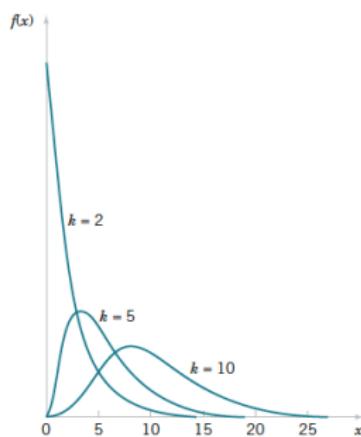
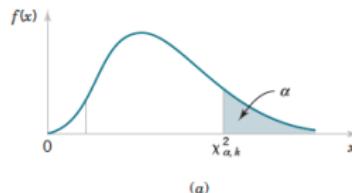
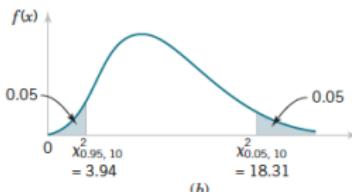


FIGURE 8.8

Probability density functions of several  $\chi^2$  distributions.



(a)



(b)

FIGURE 8.9

Percentage point of the  $\chi^2$  distribution. (a) The percentage point  $\chi_{\alpha,k}^2$ . (b) The upper percentage point  $\chi_{0.05,10}^2 = 18.31$  and the lower percentage point  $\chi_{0.95,10}^2 = 3.94$ .

## Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$$\begin{aligned}1 - \alpha &= P(\chi_{n-1,1-\alpha/2}^2 \leq X_{n-1}^2 \leq \chi_{n-1,\alpha/2}^2) \\&= P\left(\chi_{n-1,1-\alpha/2}^2 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}^2\right) \\&= P\left(\frac{\chi_{n-1,\alpha/2}^2}{(n-1)S_n^2} \leq \sigma^2 \leq \frac{\chi_{n-1,1-\alpha/2}^2}{(n-1)S_n^2}\right)\end{aligned}$$

If  $s_n^2$  is the sample variance from a random sample of  $n$  observations from a **normal** distribution with unknown variance  $\sigma^2$ , then a  $100(1-\alpha)\%$  confidence interval on  $\sigma^2$  is

$$\frac{(n-1)s_n^2}{\chi_{n-1,\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha/2}^2}$$

## One-sided confidence bounds on the Variance and Standard Deviation of a Normal Distribution

The  $100(1 - \alpha)\%$  lower and upper confidence bounds on  $\sigma^2$  are

$$\sigma^2 \geq \frac{(n-1)s_n^2}{\chi_{n-1,\alpha}^2}, \text{ and } \sigma^2 \leq \frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha}^2}.$$

respectively. If the parameter of interest is the population standard deviation  $\sigma$  instead of the population variance  $\sigma^2$ , one can take square root of the above results:

$$\sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,\alpha/2}^2}} \leq \sigma \leq \sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha/2}^2}},$$

$$\sigma \geq \sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,\alpha}^2}}, \text{ and } \sigma \leq \sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha}^2}}.$$

Remark: Chi2-intervals are **not** robust to the normality assumption.

## Example

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of  $s_n^2 = 0.01532^2$  (fluid ounce). If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. We will assume that the fill volume is approximately **normally distributed**. Find a 95% upper confidence bound for  $\sigma$ , the population standard deviation of fill volume.

**Solution:** Normality checked (with data, use QQ plot).

Chi2interval(level=0.95,n=20,s=0.01532)

The sample standard variance is 0.0002347024 and sample size is 20

A 95% two-sided confidence interval for the population variance is [ 0.0001357391 , 0.0005006835 ]

A 95% upper-confidence bound for the population variance is 0.0004407769

A 95% lower-confidence bound for the population variance is 0.0001479371

The sample standard deviation is 0.01532 and sample size is 20

A 95% two-sided confidence interval for the population standard deviation is [ 0.01165071 , 0.02237596 ]

A 95% upper-confidence bound for the population standard deviation is 0.02099469

A 95% lower-confidence bound for the population standard deviation is 0.01216294

Conclusion: based on the data, we are 95% confident that the population standard deviation of fill volume  $\sigma$  is bounded above by 0.021.

## Large-Sample Confidence Interval for a Population Proportion

It is often necessary to construct confidence intervals on a population proportion. For example, suppose that a random sample of size  $n$  has been taken from a large (possibly infinite) population and that  $X(\leq n)$  observations in this sample belong to a class of interest. Then  $\hat{p} = \frac{X}{n}$  is a point estimator of the proportion of the population  $p$  that belongs to this class. Note that  $X \sim \text{Binomial}(n, p)$ .

When  $n$  is large (rule of thumb :  $n\hat{p} \geq 5, n(1 - \hat{p}) \geq 5$ ), we have

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim AN(0, 1).$$

Thus

$$1 - \alpha \approx P \left[ -z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\alpha/2} \right]$$

## Large-Sample Confidence Interval for a Population Proportion

After some algebra and approximation, we have

### Approximate Confidence Interval on a Population Proportion

If  $\hat{p}$  is the proportion of observation in a random sample of size  $n$  that belongs to a class of interest, an approximate  $100(1 - \alpha)\%$  confidence interval on the proportion  $p$  of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Approximate  $100(1 - \alpha)\%$  One-Sided lower and upper Confidence Bounds are

$$p \geq \hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \text{ and } p \leq \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

StatEngine: *Propinterval(level =?, n =?, X =?)*.

## Choice of Sample Size

Suppose we want to choose  $n$  such that  $100(1 - \alpha)\%$  confident that the error is less than some specified value  $E$ , we have

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 p(1 - p).$$

Now we have a question, if we know  $p$ , we can calculate  $n$ . However, if we know  $p$ , why do we need to estimate  $p$ ? Two solutions:

1. Suppose we have an initial estimate of  $p$ , denoted by  $\tilde{p}$ :

$$n = \left\lceil \left( \frac{z_{\alpha/2}}{E} \right)^2 \tilde{p}(1 - \tilde{p}) \right\rceil.$$

2. If no information about  $p$  is available, then we use a conservative approach (because  $p(1 - p) \leq 0.25$ )

$$n = \left\lceil \left( \frac{z_{\alpha/2}}{E} \right)^2 0.25 \right\rceil.$$

StatEngine:

`sample.size.Propinterval(level =?, ini.p =?, E =?)`

## Example

In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. Find a 95% two-sided confidence interval for  $p$ , the proportion of bearings in the population that exceeds the roughness specification.

**Solution:**  $x = 10$ ,  $n = 85$ ,  $\hat{p} = 10/85$ ,  $n\hat{p} = 10 \geq 5$ ,  $n(1 - \hat{p}) = 75 \geq 5$ . Condition checked!

```
Propintervall(level=0.95,n=85,X=10)
```

Conclusion: based on the data, we are 95% confidence that  $p$  falls between 0.0492 and 0.1861.

Now using  $\tilde{p} = 0.12$  as an initial estimate of  $p$ , how large a sample is required if we want to be 95% confident that the error in using  $\hat{p}$  to estimate  $p$  is less than 0.05? Then redo this problem using the conservative approach (answer: 163 and 385).

```
sample.size.Propintervall(level=0.95,ini.p=0.12,E=0.05)  
sample.size.Propintervall(level=0.95,ini.p=0.5,E=0.05)
```

## Prediction Interval

Suppose that  $X_1, \dots, X_n$  is a random sample from a normal population. The sample mean and sample variance are  $\bar{X}_n$  and  $S_n^2$ , respectively. We wish to predict the value  $X_{n+1}$ , a single future observation. A point prediction of  $X_{n+1}$  is  $\bar{X}_n$ , the prediction error is  $X_{n+1} - \bar{X}_n$  and the variance of the prediction error is

$$V(X_{n+1} - \bar{X}_n) = \sigma^2 + \frac{\sigma^2}{n}, \text{ and } X_{n+1} - \bar{X}_n \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n}\right)\right)$$

Estimate  $\sigma$  by  $S_n$ , we have

$$T_{n-1} = \frac{X_{n+1} - \bar{X}_n}{S_n \sqrt{1 + \frac{1}{n}}} \sim t(n-1).$$

## Prediction Interval

Based on a random sample  $x_1, \dots, x_n$ , A  $100(1-\alpha)\%$  prediction interval (PI) on a single future observation from a normal distribution is given by

$$\bar{x}_n - t_{n-1,\alpha/2} s_n \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x}_n + t_{n-1,\alpha/2} s_n \sqrt{1 + \frac{1}{n}}$$

One could also compute the upper- and lower- prediction bounded.  
Use StatEngin:

```
Predinterval(level=? ,sample=? )
```

```
Predinterval(level=? ,n=? ,barx=? ,s=? )
```

The prediction interval for  $X_{n+1}$  will always be longer than the confidence interval for  $\mu$  because more variability is associated with the prediction error than with the error of estimation.

## Example

Consider the tensile adhesion tests on specimens of U-700 alloy described in Practice 1:

$x=c(19.8, 10.1, 14.9, 7.5, 15.4, 15.4, 15.4, 18.5, 7.9, 12.7, 11.9, 11.4, 11.4, 14.1, 17.6, 16.7, 15.8, 19.5, 8.8, 13.6, 11.9, 11.4)$

A 95% confidence interval on  $\mu$  is  $Tinterval(level = 0.95, sample = x)$  which gives

$$12.1381 \leq \mu \leq 15.2892.$$

We plan to test a 23rd specimen. A 95% prediction interval on the load at failure for this specimen is

$Predinterval(level = 0.95, sample = x)$  which gives

$$6.1575 \leq X_{23} \leq 21.2698$$

Conclusion, we are 95% confident that the next observation will be between 6.1575 and 21.2698.

## StatEngine Summary of One-Sample CIs

CI for  $\mu$  when  $\sigma$  is known and the distribution is normal:

Zinterval(level=?, sigma=?, sample=? )

Zinterval(level=?, sigma=?, n=?, barx=? )

sample.size.Zinterval(level=?, sigma=?, E=? )

CI for  $\mu$  when  $\sigma$  is unknown and the distribution is normal (or for any distribution, but in this case, it provides approximated CIs):

Tinterval(level=?, sample=? )

Tinterval(level=?, n=?, barx=? , s=? )

Large-sample CI ( $n \geq 25$ ) for  $\mu$  under any distribution:

AZinterval(level=?, sample=? )

AZinterval(level=?, n=?, barx=? , s=? )

Both T-interval and AZ-interval are approximated CIs when normality does not hold. The T-intervals are more conservative (wider).

## StatEngine Summary of One-Sample CIs

CI for  $\sigma^2$  (or  $\sigma$ ) when the distribution is **normal** (does not work well if lack of normality)

`Chi2interval(level=?, sample=?)`

`Chi2interval(level=?, n=?, s=? )`

Large-sample CI ( $n\hat{p}, n(1 - \hat{p}) \geq 5$ ) for a population proportion  $p$ :

`Propinterval(level=?, n=?, X=? )`

`sample.size.Propinterval(level=?, ini.p=?, E=? )`

Prediction interval of a future observation from **normal** distribution.

`Predinterval(level=?, sample=? )`

`Predinterval(level=?, n=?, barx=?, s=? )`

Common patterns:

- ▶ Based on the same sample, an interval estimate becomes wider if its confidence level increases.
- ▶ When confidence level is fixed, a larger sample size leads to a narrower (more precise) interval estimate.

# STAT 509: Statistics for Engineers

## Chapter 9: Tests of Hypotheses for a Single Sample

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

# Chapter 9: Tests of Hypotheses for a Single Sample

## Learning Objectives:

1. Structure engineering decision-making problems as hypothesis tests
2. Test hypotheses on the mean of a normal distribution using either a Z-test or a t-test procedure
3. Test hypotheses on the variance or standard deviation of a normal distribution
4. Test hypotheses on a population proportion
5. Use the *P*-value approach for making decisions in hypothesis tests
6. Compute power and type II error probability, and make sample size selection decisions for tests on means, variances, and proportions
7. Explain and use the relationship between confidence intervals and hypothesis tests

## Motivating example

Suppose that an engineer is designing an air crew escape system that consists of an ejection seat and a rocket motor that powers the seat. The rocket motor contains a propellant, and for the ejection seat to function properly, the propellant should have a mean burning rate of 50 cm/sec.

If the burning rate is too low, the ejection seat may not function properly, leading to an unsafe ejection and possible injury of the pilot. Higher burning rates may imply instability in the propellant or an ejection seat that is too powerful, again leading to possible pilot injury.

So the practical engineering question that must be answered is: Does the mean burning rate of the propellant equal 50 cm/sec, or is it some other value (either higher or lower)?

This type of question can be answered using a statistical technique called **hypothesis testing**.

# Hypothesis Testing

In the previous chapter, we illustrated how to construct a confidence interval estimate of a parameter from sample data. However, many problems in engineering require that we decide which of two competing claims or statements about some parameter is true. The statements are called **hypotheses**, and the decision-making procedure is called **hypothesis testing**.

## Statistical Hypothesis

A **statistical hypothesis** is a statement about the parameters of one or more populations

## Hypothesis Testing

Consider the air crew escape system. Suppose that we are interested in the burning rate of the solid propellant. Burning rate is a random variable that can be described by a probability distribution. Suppose that our interest focuses on the mean burning rate (a parameter of this distribution). Specifically, we are interested in deciding whether or not the mean burning rate is 50 centimeters per second. We may express this formally as

$$H_0 : \mu = 50 \text{ cm/sec} \text{ versus } H_1 : \mu \neq 50 \text{ cm/sec.}$$

The statement  $H_0$  is called the **null hypothesis**. This is a claim that is initially assumed to be true. The statement  $H_1$  is called the **two-tailed (or two-sided) alternative hypothesis** which contradicts the null hypothesis.

# Hypothesis Testing

In some situations, we may wish to formulate a **one-sided alternative hypothesis**, as in

$H_0 : \mu = 50$  cm/sec versus  $H_1 : \mu < 50$  cm/sec  
(left-tailed alternative hypothesis)

or  $H_0 : \mu = 50$  cm/sec versus  $H_1 : \mu > 50$  cm/sec  
(right-tailed alternative hypothesis)

We will always state the null hypothesis as an equality claim in this course. However, when the alternative hypothesis is stated with the  $<$  sign, the implicit claim in the null hypothesis can be taken as  $\geq$  and when the alternative hypothesis is stated with the  $>$  sign, the implicit claim in the null hypothesis can be taken as  $\leq$ .

## Tests of Statistical Hypotheses

To illustrate the general concepts, consider the propellant burning rate problem introduced earlier:

$$H_0 : \mu = 50 \text{ cm/sec} \text{ versus } H_1 : \mu \neq 50 \text{ cm/sec.}$$

Suppose that a sample of  $n = 10$  specimens is tested and that the sample mean burning rate  $\bar{x}_n$  is observed. The sample mean is an estimate of the true population mean  $\mu$ .

- ▶ A value of the sample mean  $\bar{x}_n$  that falls close to the hypothesized value of  $\mu = 50$  cm/sec does not conflict with the null hypothesis that the true mean  $\mu$  is really 50 cm/sec.
- ▶ On the other hand, a sample mean that is considerably different from 50 cm/sec is evidence in support of the alternative hypothesis  $H_1$ .

Thus, the sample mean is the test statistic in this case.

## Tests of Statistical Hypotheses (Critical region, acceptance region, critical values)

The sample mean can take on many different values. Suppose that

- ▶ if  $48.5 \leq \bar{x}_n \leq 51.5$ , we will not reject the null hypothesis  
 $H_0 : \mu = 50$
- ▶ if either  $\bar{x}_n < 48.5$  or  $\bar{x}_n > 51.5$ , we will reject  $H_0$  in favor of the alternative hypothesis  $H_1 : \mu \neq 50$ .

The values of  $\bar{x}_n$  that are less than 48.5 and greater than 51.5 constitute the **critical region** (or **rejection region**) for the test; all values that are outside the critical region form a region for which we will fail to reject  $H_0$ . By convention, this is usually called the **acceptance region**. The boundaries between the critical regions and the acceptance region are called the **critical values**.

In our example, the critical values are 48.5 and 51.5. It is customary to state conclusions relative to the null hypothesis  $H_0$ . Therefore, we reject  $H_0$  in favor of  $H_1$  if the test statistic falls in the critical region and fails to reject  $H_0$  otherwise.

## Type I, II errors

This decision procedure can lead to either of two wrong conclusions.

- ▶ The true mean burning rate **could** be equal to 50 cm/sec. However, for the randomly selected propellant specimens that are tested, we **could** observe a value of the test statistic  $\bar{x}_n$  that falls into the critical region. We would then reject the null hypothesis  $H_0$  in favor of the alternate  $H_1$  when, in fact,  $H_0$  is really true. This type of wrong conclusion is called a **type I error**.
- ▶ The true mean burning rate **could** be different from 50 cm/sec, yet the sample mean  $\bar{x}_n$  **could** fall in the acceptance region. In this case, we would fail to reject  $H_0$  when  $H_0$  is false, and this leads to the other type of error, called a **type II error**.

Decision	$H_0$ Is True	$H_0$ Is False
Fail to reject $H_0$	No error	Type II error
Reject $H_0$	Type I error	No error

## Probability of Type I Error

Because our decision is based on random variables (e.g.,  $\bar{X}_n$ ), probabilities can be associated with the type I and type II errors. The probability of making a type I error is denoted by the Greek letter  $\alpha$ .

### Probability of Type I Error

$$\alpha = P(\text{Type I Error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}).$$

Back to the previous example, suppose the population is normal and the population standard deviation is  $\sigma = 2.5$  cm/sec, then based on a sample of size  $n = 10$ ,

$$\begin{aligned}\alpha &= P(\bar{X}_n < 48.5 \text{ or } \bar{X}_n > 51.5 \text{ when } \mu = 50) \\ &= \text{normal}.\text{probability}(-\text{Inf}, 48.5, 50, 2.5/\sqrt{10}) \\ &\quad + \text{normal}.\text{probability}(51.5, \text{Inf}, 50, 2.5/\sqrt{10}) = 0.0574.\end{aligned}$$

where by the normality and  $\mu = 50$ ,  $\bar{X}_n \sim N(\mu = 50, \sigma^2/n = 2.5^2/10)$ .

## Probability of Type II Error and Power

Probability of Type II Error (or known as  $\beta$ -error)

$$\beta = P(\text{Type II Error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}).$$

The **power** of a statistical test is the probability of rejecting the null hypothesis  $H_0$  when the alternative hypothesis is true; i.e.,  $1 - \beta$ .

Back to the previous example, suppose the population is normal and the population standard deviation is  $\sigma = 2.5$  cm/sec, then based on a sample of size  $n = 10$ , the probability of Type II Error when  $\mu = 52$  is

$$\begin{aligned}\beta &= P(48.5 \leq \bar{X}_n \leq 51.5 \text{ when } \mu = 52) \\ &= \text{normal.probability}(48.5, 51.5, 52, 2.5/\sqrt{10}) = 0.2643.\end{aligned}$$

where by the normality and  $\mu = 52$ ,  $\bar{X}_n \sim N(\mu = 52, \sigma^2/n = 2.5^2/10)$ . The **power** of the test at  $\mu = 52$  is  $1 - 0.2643 = 0.7357$ .

# Summary

Acceptance Region	Sample Size	$\alpha$	$\beta$ at $\mu = 52$	$\beta$ at $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10	0.0576	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	0.5000	0.9705
$48.81 < \bar{x} < 51.19$	16	0.0576	0.0966	0.8606
$48.42 < \bar{x} < 51.58$	16	0.0114	0.2515	0.9578

1. The size of the critical region, and consequently the probability of a type I error  $\alpha$ , can always be reduced by appropriate selection of the critical values.
2. Type I and type II errors are related. A decrease in one always results in an increase in the other provided that the sample size  $n$  does not change.
3. An increase in  $n$  reduces  $\beta$  provided that  $\alpha$  is held constant.
4. When  $H_0$  is false,  $\beta$  increases as the true value of the parameter approaches the value hypothesized in the null hypothesis. The value of  $\beta$  decreases as the difference between the true mean and the hypothesized value increases.

## Summary (continued)

Generally, we control/fix the type I error probability  $\alpha$  to select the critical region. In this way, we can directly control the probability of wrongly rejecting  $H_0$ . We always think of rejection of the null hypothesis  $H_0$  as a **strong conclusion**. We call the fixed  $\alpha$  as the **significance level** of the test.

Below are the common steps (of a **critical value approach**)

1. Find a test statistics
2. Given the probability of type I error  $\alpha$ , find a critical region.
3. If the test statistics falls in the critical region, reject  $H_0$ ; otherwise, fail to reject.
4. Conclude the result.

## Summary (continued)

We could also use a **confidence-interval approach**:

1. Based on the alternative hypothesis  $H_1$  (two-tailed or left/right-tailed), compute a (two-sided or one-sided) confidence interval estimate of confidence level  $100(1 - \alpha)\%$ .
2. If the confidence interval estimate covers the hypothesized value in  $H_0$ , we fail to reject  $H_0$ ; otherwise, reject  $H_0$ .
3. Conclude the result.

Lastly, we could use a **P-value approach**. The  $P$ -value is the smallest level of significance that would lead to rejection of the null hypothesis  $H_0$  with the given data.

1. Compute  $P$ -value based on the data.
2. If  $P$ -value less than  $\alpha$ , reject  $H_0$ ; otherwise, fail to reject  $H_0$ .
3. Conclude the result.

All these can be done using StatEngine!

## Tests on the Mean of a Normal Distribution, Variance Known

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Test statistic:  $Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$

Based on sample  $x_1, \dots, x_n$ , the observed test statistic is  $z_0 = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$

If the alternative hypothesis is  $H_1 : \mu \neq \mu_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $|z_0| > z_{\alpha/2}$  (critical-value approach),
- ▶  $\mu_0 \notin [\bar{x}_n \pm z_{\alpha/2}\sigma/\sqrt{n}]$  (confidence-interval approach),
- ▶ the  $P$ -value  $= 2[1 - P(Z \leq |z_0|)] < \alpha$ , where  $Z \sim N(0, 1)$  ( $P$ -value approach).

## Tests on the Mean of a Normal Distribution, Variance Known

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Test statistic:  $Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$

Based on sample  $x_1, \dots, x_n$ , the observed test statistic is  $z_0 = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$

If the alternative hypothesis is  $H_1 : \mu > \mu_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $z_0 > z_\alpha$  (critical-value approach),
- ▶  $\mu_0 \notin [\bar{x}_n - z_\alpha \sigma / \sqrt{n}, +\infty)$  (confidence-interval approach) or  $\mu_0 < \bar{x}_n - z_\alpha \sigma / \sqrt{n}$ , the  $100(1 - \alpha)\%$  lower bound on  $\mu$  ( $\mu_0$  is even smaller than the smallest confident guess of  $\mu$ ).
- ▶ the  $P$ -value  $= P(Z > z_0) < \alpha$ , where  $Z \sim N(0, 1)$  ( $P$ -value approach).

## Tests on the Mean of a Normal Distribution, Variance Known

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Test statistic:  $Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$

Based on sample  $x_1, \dots, x_n$ , the observed test statistic is  $z_0 = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$

If the alternative hypothesis is  $H_1 : \mu < \mu_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $z_0 < -z_\alpha$  (critical-value approach),
- ▶  $\mu_0 \notin (-\infty, \bar{x}_n + z_\alpha \sigma / \sqrt{n}]$  (confidence-interval approach) or  $\mu_0 > \bar{x}_n + z_\alpha \sigma / \sqrt{n}$ , the  $100(1 - \alpha)\%$  upper bound on  $\mu$  ( $\mu_0$  is even larger than the largest confident guess of  $\mu$ ),
- ▶ the  $P$ -value  $= P(Z \leq z_0) < \alpha$ , where  $Z \sim N(0, 1)$  ( $P$ -value approach).

## Probability of a Type II Error for Tests on the Mean, Variance Known

Type II error occurs when  $H_1$  is true. Suppose the true mean value is  $\mu = \mu_1$ . Let  $\delta = \mu_1 - \mu_0$ .

If the alternative hypothesis is  $H_1 : \mu \neq \mu_0$ . The probability of the type II error of the previous two-tailed test is

$$\beta = P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2}, \text{ under } H_1 : \mu = \mu_1\right)$$

Under  $H_1 : \bar{X}_n \sim N(\mu = \mu_1, \sigma^2/n)$

$$= P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu_1 + \mu_1 - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right)$$

$$= P\left(-z_{\alpha/2} \leq Z + \frac{\delta}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right)$$

$$= P\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \leq Z \leq z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right).$$

## Probability of a Type II Error for Tests on the Mean, Variance Known

If the alternative hypothesis is  $H_1 : \mu > \mu_0$ . The probability of the type II error of the previous right-tailed test is

$$\beta = P\left(Z \leq z_\alpha - \frac{\delta\sqrt{n}}{\sigma}\right).$$

If the alternative hypothesis is  $H_1 : \mu < \mu_0$ . The probability of the type II error of the previous left-tailed test is

$$\beta = P\left(-z_\alpha - \frac{\delta\sqrt{n}}{\sigma} \leq Z\right).$$

## Choice of Sample Size for Tests on the Mean, Variance Known

What is an appropriate sample size to control the probability of type II error  $\beta$  (or the power  $1 - \beta$ ) for given  $\delta$  and  $\alpha$ ?

If the alternative hypothesis is  $H_1 : \mu \neq \mu_0$ . The sample size should be at least the smallest positive integer  $n$  such that

$$\beta \geq P\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \leq Z \leq z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right).$$

If the alternative hypothesis is  $H_1 : \mu > \mu_0$ . The sample size should be at least the smallest positive integer  $n$  such that

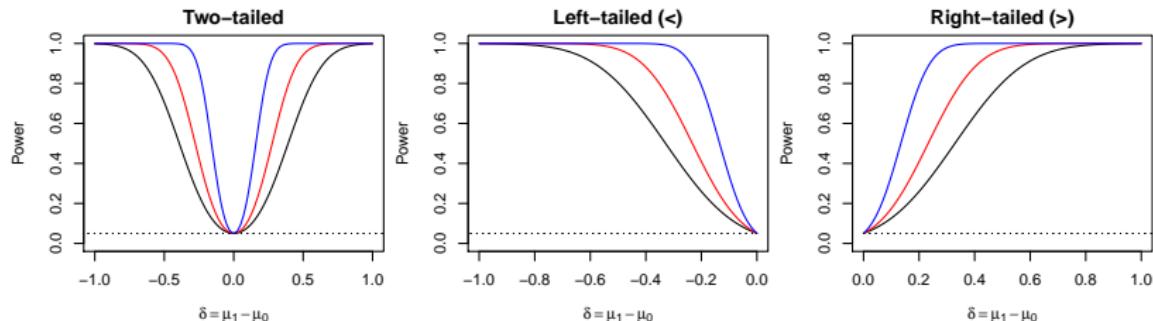
$$\beta \geq P\left(Z \leq z_\alpha - \frac{\delta\sqrt{n}}{\sigma}\right).$$

If the alternative hypothesis is  $H_1 : \mu < \mu_0$ . The sample size should be at least the smallest positive integer  $n$  such that

$$\beta \geq P\left(-z_\alpha - \frac{\delta\sqrt{n}}{\sigma} \leq Z\right).$$

We can find these numerically (StatEngine).

# Summary of Tests on the Mean, Variance Known



Herein,  $\alpha = 0.05$ ,  $\sigma = 1$ . Power curves: black  $n = 25$ , red  $n = 50$ , blue  $n = 200$ .

## summary of the patterns

- ▶ Probability of Type I Error is always  $\alpha$ , user-chosen.
- ▶ When the sample size  $n$  is fixed, the power of test ( $1 - \beta$ , where  $\beta$  is the Type II Error) increases as the difference ( $\delta$ ) between the true parameter value and the hypothesized value increases in favor of  $H_1$ . Easier to detect a stronger signal!
- ▶ When a non-zero difference ( $\delta \neq 0$ ) between the true parameter value and the hypothesized value in favor of  $H_1$  is fixed, the power of test ( $1 - \beta$ ) increases as  $n$  increases. Easier to detect a signal with more samples!
- ▶ If  $\delta = 0$ , the hypothesized value is the true parameter value, the power of the test is always  $\alpha$  no matter how large  $n$  is. Probability of Type I Error does not change with  $n$ .

# Summary of Tests on the Mean, Variance Known

## Z-test

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Three types of  $H_1 (\mu \neq, <, > \mu_0)$

Observed test statistic:  $z_0 = \frac{\bar{x}_n - \mu_0}{\sigma / \sqrt{n}}$

- ▶ Critical-value approach
- ▶ Confidence-interval approach
- ▶  $P$ -value approach
- ▶ Power calculation
- ▶ Choice of sample size

# StatEngine Summary of Tests on the Mean, Variance Known

Conduct the test:

```
#If data are available
```

```
Ztest(mu0=? ,H1=? ,alpha=? ,sigma=? ,sample=? )
```

```
#If statistics are provided
```

```
Ztest(mu0=? ,H1=? ,alpha=? ,sigma=? ,n=? ,barx=? )
```

Calculate power  $(1 - \beta)$  of the test for given  $\delta$  and  $n$ :

```
Ztest.power(H1=? ,alpha=? ,sigma=? ,n=? ,delta=? )
```

Calculate the minimum sample size to control  $\beta$  for given  $\delta$  and  $\alpha$ :

```
sample.size.Ztest(H1=? ,sigma=? ,alpha=? ,beta=? ,delta=? )
```

## Example

A manufacturer produces crankshafts for an automobile engine. The crankshafts wear after 100,000 miles (0.0001 inch) is of interest because it is likely to have an impact on warranty claims. A random sample of  $n = 15$  shafts is tested and  $\bar{x}_n = 2.78$ . It is known that  $\sigma = 0.9$  and that wear is normally distributed.

- (a) Test  $H_0 : \mu = 3$  versus  $H_1 : \mu \neq 3$  using  $\alpha = 0.05$ .
- (b) What is the power of this test if  $\mu = 3.25$ ?
- (c) What sample size would be required to detect a true mean of 3.75 if we wanted the power to be at least 0.9?
- (d) Explain how the question in part (a) could be answered by constructing a confidence interval on the  $\mu$ .

**Solution:** 1. It is about  $\mu$ ; 2. Normal distribution; 3.  $\sigma = 0.9$  is known. Bingo: Z-test!

- (a) No data but statistics are provided.

```
Ztest(mu0=3,H1="two",alpha=0.05,sigma=0.9,n=15,barx=2.78)
```

## Example (continued)

- (a) No data but statistics are provided.

```
Ztest(mu0=3,H1="two",alpha=0.05,sigma=0.9,n=15,barx=2.78)
```

It also answers part (d).

The sample mean is 2.78 and sample size is 15

H1 is two-tailed: mu does not equal to mu0=3 . The results are:

1. Test statistic z0 is -0.9467293 , z\_(alpha/2) is 1.959964, Because -z\_(alpha/2)<=z0<=z\_(alpha/2), we fail to reject H0 at significance level 0.05
2. A 95% two-tailed confidence interval for the population mean is [2.324546 , 3.235454] which contains the hypothesized value mu0=3, we fail to reject H0 at significance level 0.05
3. The P-value is 0.3437768 which is not smaller than alpha= 0.05, so we fail to reject H0 at significance level 0.05

Conclusion: at significance level  $\alpha = 0.05$ , the data do not provide sufficient evidence to reject the null hypothesis.

## Example (continued)

(b) Power calculation for  $n = 15$  and  $\delta = 3.25 - 3 = 0.25$ :

```
Ztest.power(H1="two",alpha=0.05,sigma=0.9,n=15,delta=3.25-3)
```

H1 is two-tailed

The probability of the Type II error of this test at  $\delta = \mu_1 - \mu_0 = 0.25$  is 0.8104889 and the associated power is 0.1895111

(c) Sample size calculation for  $\beta = 1 - 0.9 = 0.1$ ,  $\delta = 3.75 - 3$ , and  $\alpha = 0.05$ .

```
sample.size.Ztest(H1="two",sigma=0.9,alpha=0.05,beta=1-0.9,delta=3.75-3)
```

At significance level  $\alpha = 0.05$ , we need at least  $n = 16$  to make the power of this test at  $\delta = 0.75$  be at least 0.9

(d) See the solution to part (a).

## Tests on the Mean of a Normal Distribution, Variance Unknown

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Test statistic:  $T_0 = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$

Based on sample  $x_1, \dots, x_n$ , the observed test statistic is  $t_0 = \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}}$

If the alternative hypothesis is  $H_1 : \mu \neq \mu_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $|t_0| > t_{n-1, \alpha/2}$  (critical-value approach),
- ▶  $\mu_0 \notin [\bar{x}_n \pm t_{n-1, \alpha/2} s_n / \sqrt{n}]$  (confidence-interval approach),
- ▶ the  $P$ -value  $= 2[1 - P(T_{n-1} \leq |t_0|)] < \alpha$ , where  $T_{n-1} \sim t(n-1)$  ( $P$ -value approach).

## Tests on the Mean of a Normal Distribution, Variance Unknown

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Test statistic:  $T_0 = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$

Based on sample  $x_1, \dots, x_n$ , the observed test statistic is  $t_0 = \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}}$

If the alternative hypothesis is  $H_1 : \mu > \mu_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $t_0 > t_{n-1,\alpha}$  (critical-value approach),
- ▶  $\mu_0 \notin [\bar{x}_n - t_{n-1,\alpha} s_n / \sqrt{n}, +\infty)$  (confidence-interval approach)  
or  $\mu_0 < \bar{x}_n - t_{n-1,\alpha} s_n / \sqrt{n}$ , the  $100(1 - \alpha)\%$  lower bound on  $\mu$ ,
- ▶ the  $P$ -value =  $P(T_{n-1} > t_0) < \alpha$  ( $P$ -value approach).

## Tests on the Mean of a Normal Distribution, Variance Unknown

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Test statistic:  $T_0 = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}$

Based on sample  $x_1, \dots, x_n$ , the observed test statistic is  $t_0 = \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}}$

If the alternative hypothesis is  $H_1 : \mu < \mu_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $t_0 < -t_{n-1, \alpha}$  (critical-value approach),
- ▶  $\mu_0 \notin (-\infty, \bar{x}_n + t_{n-1, \alpha} s_n / \sqrt{n}]$  (confidence-interval approach)  
or  $\mu_0 > \bar{x}_n + t_{n-1, \alpha} s_n / \sqrt{n}$ , the  $100(1 - \alpha)\%$  upper bound on  $\mu$ ,
- ▶ the  $P$ -value  $= P(T_{n-1} \leq t_0) < \alpha$  ( $P$ -value approach).

# Power and Choice of Sample Size for Tests on the Mean of a Normal Distribution, Variance Unknown

## T-test

Null hypothesis:  $H_0 : \mu = \mu_0$  for a hypothesized value  $\mu_0$ .

Three types of  $H_1 (\mu \neq, <, > \mu_0)$

- ▶ Power calculation
- ▶ Choice of sample size

To exactly calculate the power and the minimum sample size, we must know  $\sigma$  which is unknown in the use of a T-test. Therefore, we use the sample standard deviation  $s_n$  as an estimate to approximate the results.

In addition, like T-interval, T-test is also robust to the normality assumption. In other words, if the distribution is not normal, one can still use a T-test. But the result are all approximation; e.g., the probability of Type I error is controlled to be approximately  $\alpha$ .

# StatEngine Summary of Tests on the Mean, Variance Unknown

Conduct the test:

```
#If data are available  
Ttest(mu0=? ,H1=? ,alpha=? ,sample=? )
```

```
#If statistics are provided  
Ttest(mu0=? ,H1=? ,alpha=? ,n=? ,barx=? ,s=? )
```

Calculate power  $(1 - \beta)$  of the test for given  $\delta$  and  $n$ :

```
sn=sd(data)  
Ttest.power(H1=? ,est.sigma=sn, alpha=? ,n=? ,delta=? )
```

Calculate the minimum sample size to control  $\beta$  for given  $\delta$  and  $\alpha$ :

```
sample.size.Ttest(H1=? ,est.sigma=sn, beta=? ,delta=? ,alpha=? )
```

## Example

The sodium content of twenty 300-gram boxes of organic cornflakes was determined. The data (in milligrams) are as follows:

$x=c(131.15, 130.69, 130.91, 129.54, 129.64, 128.77, 130.72, 128.33, 128.24, 129.65, 130.14, 129.29, 128.71, 129.00, 129.39, 130.42, 129.53, 130.12, 129.78, 130.92)$

- (a) Can you support a claim that mean sodium content of this brand of cornflakes differs from 130 milligrams? Use  $\alpha = 0.05$ . Find the  $P$ -value.
- (b) Check that sodium content is normally distributed.
- (c) Compute the power of the test if the true mean sodium content is 130.5 milligrams.
- (d) What sample size would be required to detect a true mean sodium content of 130.1 milligrams if you wanted the power of the test to be at least 0.75?
- (e) Explain how the question in part (a) could be answered by constructing a two-sided confidence interval on the mean sodium content.

## Example (solution)

(a) Consider  $H_0 : \mu = 130$  versus  $H_1 : \mu \neq 130$ . We do not know the variance of the population. Known T-test is robust to distributional assumption. We will use T-test:

```
Ttest(mu0=130,H1="two",alpha=0.05,sample=x)
```

The sample mean is 129.747 sample standard deviation is 0.8764288 , and sample

H1 is two-sided: mu does not equal to mu0= 130 . The results are:

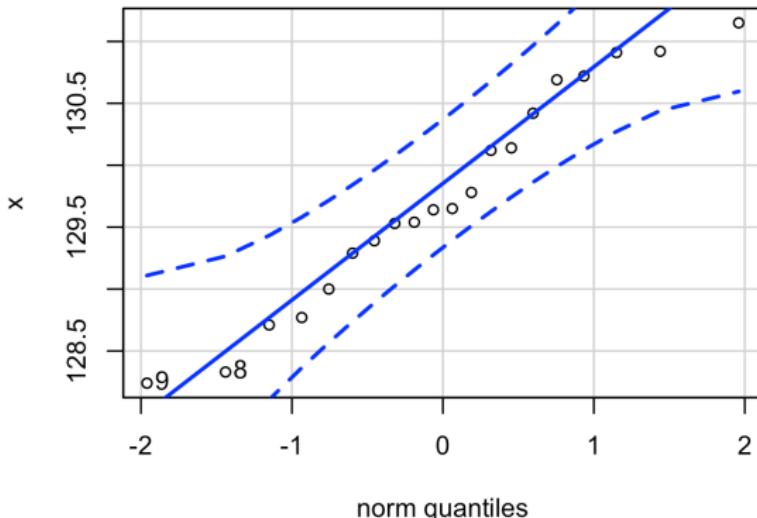
1. Test statistic t0 is -1.290978 ,  $t_{(n-1,\alpha/2)}$  is 2.093024,  
Because  $|t_0| \leq t_{(n-1,\alpha/2)}$ , we fail to reject  $H_0$  at significance level 0.05
2. A 95% two-sided confidence interval for the population mean is  
[ 129.3368 , 130.1572 ] which contains the hypothesized value  $\mu_0 = 130$  ,  
so we fail to reject  $H_0$  at significance level 0.05
3. The P-value is 0.2121988 which is not smaller than alpha= 0.05 ,  
so we fail to reject  $H_0$  at significance level 0.05

Conclusion: at significance level 0.05, the data do not provide sufficient evidence to reject  $H_0$ .

## Example (solution continued)

(b) We use QQ plot to check normality

```
data.summary(x)
```



Most dots are fitted well by the line. Thus, we could conclude that the data are from a normal distribution, which further supports our use of the T-test in (a).

## Example (solution continued)

(c) Find the power when  $\mu_1 = 130.5$ . From (a), we know  $\mu_0 = 130$  yielding  $\delta = \mu_1 - \mu_0 = 0.5$

```
# The power calculation requires an estimate of  
# the population standard deviation, we use  
# the sample standard deviation which can be obtained  
# by sd(x). The sample size n=length(x)
```

```
Ttest.power(H1="two",est.sigma=sd(x),alpha=0.05,  
n=length(x),delta=0.5)
```

H1 is two-sided

The probability of the Type II error of this test at  
 $\delta = \mu_1 - \mu_0 = 0.5$  is 0.3224236  
and the associated power is 0.6775764

## Example (solution continued)

(d) Now we have  $\mu_1 = 130.1$  and  $\beta = 1 - 0.75 = 0.25$ . From (a), we know  $\mu_0 = 130$  yielding  $\delta = \mu_1 - \mu_0 = 0.1$

```
# The sample size calculation requires an estimate of  
# the population standard deviation, we use  
# the sample standard deviation  
# which can be obtained by sd(x).
```

```
sample.size.Ttest(H1="two",est.sigma =sd(x),  
beta=0.25,delta=0.1,alpha=0.05)
```

At significance level alpha= 0.05 , we need at least  
n= 536 to achieve a power >= 0.75 of this test  
at delta= 0.1

When n= 535 , the power is 0.7499781

When n= 536 , the power is 0.7507622

When n= 537 , the power is 0.7515444

(e) See the StatEigine output in part (a).

## Large-sample tests on the Mean of a Distribution

Similarly to the large-sample Z-interval, we can use the central limit theorem to develop a large-sample Z-test for  $H_0 : \mu = \mu_0$  versus the three types of alternative hypotheses.

- ▶ Advantages are to lift the normality assumption and to allow  $\sigma$  to be unknown.
- ▶ A disadvantage is the need of a large sample size (e.g.,  $n \geq 25$ ) and everything is approximated.

However, we know T-test is robust to the normality assumption and is designed for unknown  $\sigma$ . Thus, we can always use T-test whenever large-sample Z-test can be used. One must acknowledge that T-test is always more conservative (less powerful) than large-sample Z-test; i.e., T-test often requires a larger sample size to reach the same power when compared to large-sample Z-test; or when the sample size is the same, the power of T-test is smaller than the one of large-sample Z-test.

## Tests on the Variance and Standard Deviation of a Normal Distribution

- ▶ An automated filling machine is used to fill bottles with liquid detergent. If the variance of fill volume exceeds 0.01 (fluid ounces)<sup>2</sup>, an unacceptable proportion of bottles will be underfilled or overfilled.
- ▶ If the standard deviation of hole diameter exceeds 0.01 millimeters, there is an unacceptably high probability that the rivet will not fit.

In these applications, we want to make inference about the population variance  $\sigma^2$ /population standard deviation  $\sigma$ . In particular, we now focus on these hypotheses:

$$H_0 : \sigma^2 = \sigma_0^2 \text{ versus } H_1 : \sigma^2 \neq \sigma_0^2 \text{ (two-tailed)}$$

$$H_0 : \sigma^2 = \sigma_0^2 \text{ versus } H_1 : \sigma^2 > \sigma_0^2 \text{ (right-tailed)}$$

$$H_0 : \sigma^2 = \sigma_0^2 \text{ versus } H_1 : \sigma^2 < \sigma_0^2 \text{ (left-tailed)}$$

# Tests on the Variance and Standard Deviation of a Normal Distribution

## Basic reasoning:

We know the sample variance  $S_n^2$  is a good estimator of the true population variance  $\sigma^2$ . Thus, if based on the observed sample,  $s_n^2$  is close to  $\sigma_0^2$  or  $\frac{s_n^2}{\sigma^2}$  is close to 1, then the data suggests the hypothesized value  $\sigma_0^2$  is close to the true value  $\sigma^2$ , and thus we do not reject  $H_0$ . This reasoning leads to our consideration of test statistic

$$X_0^2 = \frac{(n - 1)S_n^2}{\sigma_0^2}$$

When  $H_0$  is true (control the type I error),  $X_0^2$  follows the chi-square distribution with  $n - 1$  degrees of freedom. Thus:

# Tests on the Variance and Standard Deviation of a Normal Distribution

Chi-square test (need normality):

Null hypothesis:  $H_0 : \sigma^2 = \sigma_0^2$  for a hypothesized value  $\sigma_0^2$ .

Test statistic:  $X_0 = \frac{(n-1)S_n^2}{\sigma_0^2}$

Based on sample  $x_1, \dots, x_n$ , the observed test statistic is

$$x_0^2 = \frac{(n-1)s_n^2}{\sigma_0^2}$$

- If  $H_1 : \sigma^2 \neq \sigma_0^2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $x_0^2 < \chi_{n-1, 1-\alpha/2}^2$  or  $x_0^2 > \chi_{n-1, \alpha/2}^2$ ;
- if  $H_1 : \sigma^2 < \sigma_0^2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $x_0^2 < \chi_{n-1, 1-\alpha}^2$ ;
- if  $H_1 : \sigma^2 > \sigma_0^2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $x_0^2 > \chi_{n-1, \alpha}^2$ .

# Tests on the Variance and Standard Deviation of a Normal Distribution

## Chi-square test:

- ▶ Besides the above critical value approach,
- ▶ we also have a confidence-interval approach,
- ▶ and the  $P$ -value approach.
- ▶ Furthermore, we can compute the power
- ▶ and determine the minimum required sample size.

A power is 1 minus the probability of a Type II error. A Type II error occurs when  $H_1$  is true. Suppose the true population variance is  $\sigma_1^2 \neq \sigma_0^2$ , we denote

$$\lambda = \frac{\sigma_1}{\sigma_0},$$

and will use this  $\lambda$  to compute the power and determine the minimum required sample size.

# StatEngine Summary of Tests on the Variance and Standard Deviation of a Normal Distribution

Conduct the test:

```
#If data are available
```

```
Chi2test(sigma0=? ,H1=? ,alpha=? ,sample=? )
```

```
#If statistics are provided
```

```
Chi2test(sigma0=? ,H1=? ,alpha=? ,n=? ,s=? )
```

Calculate power  $(1 - \beta)$  of the test for given  $\lambda$  and  $n$ :

```
Chi2test.power(H1=? ,alpha=? ,n=? ,lambda=? )
```

Calculate the minimum sample size to control  $\beta$  for given  $\lambda$  and  $\alpha$ :

```
sample.size.Chi2test(H1=? ,beta=? ,lambda=? ,alpha=? )
```

Remark: Normality is important. If normality does not hold, using these might give you wrong inference. If possible, use QQ-plot to check normality (data.summary(.) in StatEngine).

## Example

An automated filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of  $s_n^2 = 0.0153$  (fluid ounces)<sup>2</sup>. If the variance of fill volume exceeds 0.01 (fluid ounces)<sup>2</sup>, an unacceptable proportion of bottles will be underfilled or overfilled. Is there evidence in the sample data to suggest that the manufacturer has a problem with underfilled or overfilled bottles? Use  $\alpha = 0.05$ , and assume that fill volume has a normal distribution.

**Solution:** It is about the population variance  $\sigma^2$ . We use Chi-square test, which requires normality. Fortunately, normality is assumed by the question. No data are provided, we do have  $s_n^2 = 0.0153$  where  $n = 20$ . The hypotheses are

$$H_0 : \sigma^2 = 0.01 \text{ versus } H_1 : \sigma^2 > 0.01$$

where  $\sigma_0^2 = 0.01$ , and the alternative is right-tailed. We set  $\alpha = 0.05$ .

## Example (continued)

### Using StatEngin

```
Chi2test(sigma0=sqrt(0.01),H1="right",alpha=0.05,n=20,s=sqrt(0.0153))
```

```
df= 19 , sample variance is 0.0153 , sample standard deviation  
is 0.1236932 , and sample size is 20
```

H1 is right-tailed:  $\sigma^2$  is more than  $\sigma_0^2 = 0.01$  .

The results are:

1. Test statistic  $x_0^2$  is 29.07 ,  $\text{Chi2}_{(n-1,\alpha)}$  is 30.14353 ,  
Because  $x_0^2 \leq \text{Chi2}_{(n-1,\alpha)}$ ,  
we fail to reject  $H_0$  at significance level 0.05

2. A 95% one-sided confidence interval for the population  
variance is [ 0.009643861 , Inf )  
which contains the hypothesized value  $\sigma_0^2 = 0.01$  ,  
so we fail to reject  $H_0$  at significance level 0.05

3. The P-value is 0.064892 which is not smaller than  $\alpha = 0.05$  ,  
so we fail to reject  $H_0$  at significance level 0.05

Conclusion: at significance level  $\alpha = 0.05$ , the data do not provide sufficient evidence  
to reject  $H_0$ .

## Example (continued)

Suppose that if the true standard deviation of the filling process exceeds the hypothesized value  $\sigma_0$  by 25%, what is the power of the previous test?

**Solution:** This asks for a power calculation of the chi.square test we had conducted. In the previous test, we have a sample of size  $n = 20$ . Now the true parameter is  $\sigma_1 = (1 + 0.25)\sigma$ , what is the power? We use StatEngine, which needs the type of  $H_1$  (right-tailed), the significance level  $\alpha = 0.05$ , the sample size  $n = 20$ , and the value of  $\lambda = \sigma_1/\sigma_0 = 1.25$ .

```
Chi2test.power(H1="right",alpha=0.05,n=20,lambda=1.25)
```

```
H1 is right-tailed (>)
```

The probability of the Type II error of this test at  
 $\lambda = \sigma_1/\sigma_0 = 1.25$  is 0.5617379  
and the associated power is 0.4382621

The power of the right-tailed test we have conducted when  $n = 20$  is 0.4383.

## Example (continued)

Suppose that if the true standard deviation of the filling process exceeds the hypothesized value  $\sigma_0$  by 25%, we would like to detect this with probability at least 0.8. Is the sample size of  $n = 20$  adequate?

**Solution:** This asks for a sample size calculation. It says if  $\sigma_1 = (1+0.25)\sigma_0$ , we want to control the Type II error  $\beta$  by  $1-0.8 = 0.2$ . Is  $n = 20$  adequate? We use StatEngine to find the minimum sample size. The StatEngine needs the type of  $H_1$  (right-tailed), the required value of  $\beta = 0.2$ , the value of  $\lambda = \sigma_1/\sigma_0 = 1.25$ , and the significance level  $\alpha = 0.05$ .

```
sample.size.Chi2test(H1="right",beta=0.2,lambda=1.25,alpha=0.05)
```

At significance level alpha= 0.05 , we need at least n= 61 to achieve a power >= 0.8 of this test

at lambda=sigma1/sigma0= 1.25

When n= 60 , the power is 0.7954305

When n= 61 , the power is 0.8008053

When n= 62 , the power is 0.8060504

We need at least 61 samples. Thus  $n = 20$  is not adequate!

## Example 2

If the standard deviation of hole diameter exceeds  $0.01$  millimeters, there is an unacceptably high probability that the rivet will not fit. Suppose that  $n = 15$  and  $s_n = 0.008$  millimeter.

- (a) Is there strong evidence to indicate that the standard deviation of hole diameter exceeds  $0.01$  millimeter? Use  $\alpha = 0.01$ . State any necessary assumptions about the underlying distribution of the data. Find the P-value for this test.
- (b) Suppose that the actual standard deviation of hole diameter exceeds the hypothesized value by  $50\%$ . What is the probability that this difference will be detected by the test described in part (a)?
- (c) If  $\sigma$  is really as large as  $0.0125$  millimeters, what sample size will be required to detect this with power of at least  $0.8$ ?

## Example 2 (Solution)

- (a) Right-tailed alternative with  $\sigma_0 = 0.01^2$

```
Chi2test(sigma0=0.01,H1="right",alpha=0.01,n=15,s=0.008)
```

We need to assume the samples are from a normal distribution.

Conclusion:...

- (b) You should be able to identify  $\lambda = 1.5$

```
Chi2test.power(H1="right",alpha=0.01,n=15,lambda=1.5)
```

- (c) Now  $\lambda = 0.0125/0.01 = 1.25$ ,  $\beta = 1 - 0.8 = 0.2$ .

```
sample.size.Chi2test(H1="right",beta=0.2,lambda=1.25,alpha=0.01)
```

Try these yourself!

## Tests on a Population Proportion

It is often necessary to test hypotheses on a population proportion:

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0 \text{ (two-tailed)}$$

$$H_0 : p = p_0 \text{ versus } H_1 : p < p_0 \text{ (left-tailed)}$$

$$H_0 : p = p_0 \text{ versus } H_1 : p > p_0 \text{ (right-tailed)}$$

suppose that a random sample of size  $n$  has been taken from a large (possibly infinite) population and that  $X(\leq n)$  observations in this sample belong to a class of interest.

We know the sample proportion  $\hat{p} = \frac{X}{n}$  is a good estimator of the true population proportion  $p$ . Thus, if based on the observed sample,  $\hat{p}$  is close to  $p_0$ , then the data suggests the hypothesized value  $p_0$  is close to the true value  $p$ , and thus we do not reject  $H_0$ . This reasoning leads to our consideration of test statistic

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

## Tests on a Population Proportion

**One-Sample Proportion Z-test (a large-sample test):**

Require:  $n\hat{p} \geq 5$  and  $n(1 - \hat{p}) \geq 5$ .

Null hypothesis:  $H_0 : p = p_0$  for a hypothesized value  $p_0$ .

The observed test statistic:  $z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ .

- ▶ If  $H_1 : p \neq p_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $|z_0| > z_{\alpha/2}$ ;
- ▶ if  $H_1 : p < p_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $z_0 < -z_\alpha$ ;
- ▶ if  $H_1 : p > p_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $z_0 > z_\alpha$ .

# Tests on the Variance and Standard Deviation of a Normal Distribution

## One-Sample Proportion Z-test:

- ▶ Besides the above critical value approach,
- ▶ we also have a confidence-interval approach,
- ▶ and the  $P$ -value approach.
- ▶ Furthermore, we can **approximate** the power
- ▶ and **approximate** the minimum required sample size.

In this test, the confidence-interval approach **might** give a different conclusion than the other two approaches. But this difference is very minor, especially when sample size  $n$  is large.

When approximating the power, which is 1 minus the probability of a Type II error and a Type II error occurs when  $H_1$  is true, we suppose the true population proportion is  $p_1$ . The  $p_0$ ,  $p_1$ , and  $n$  are needed when computing a power and minimum required sample size for this test.

# StatEngine Summary of Tests on the Variance and Standard Deviation of a Normal Distribution

Conduct the test:

```
Proptest(p0=? ,H1=? ,alpha=? ,n=? ,X=? )
```

Approximate power  $(1 - \beta)$  of the test for given  $\delta$  and  $n$ :

```
Proptest.power(H1=? ,alpha=? ,p0=? ,p1=? ,n=? )
```

Approximate the minimum sample size to control  $\beta$  for given  $p_1$ ,  $p_0$ , and  $\alpha$ :

```
sample.size.Proptest(H1=? ,beta=? ,alpha=? ,p0=? ,p1=? )
```

## Example

A semiconductor manufacturer produces controllers used in automobile engine applications. The customer requires that the process fallout or fraction defective at a critical manufacturing step not exceed 0.05 and that the manufacturer demonstrate process capability at this level of quality using  $\alpha = 0.05$ . The semiconductor manufacturer takes a random sample of 300 devices and finds that 7 of them are defective. Can the manufacturer demonstrate process capability for the customer?

**Solution:**  $p$  is the population defective rate. We want to test  $H_0 : p \geq 0.05$  versus  $H_1 : p < 0.05$ . Thus  $p_0 = 0.05$ . We also know  $n = 300$ ,  $X = 7$  and  $\alpha = 0.05$ . Then  $\hat{p} = 7/300$  and we can verify that  $n\hat{p} \geq 5$  and  $n(1 - \hat{p}) \geq 5$ . Now we can perform the left-tailed One-Sample Proportion Z-test:

```
Proptest(p0=0.05, H1="left", alpha=0.05, n=300, X=7)
```

## Example (continued)

Perform the left-tailed One-Sample Proportion Z-test:

```
Proptest(p0=0.05, H1="left",alpha=0.05,n=300,X=7)
```

The sample proportion is 0.02333333 and sample size is 300

H1 is left-tailed: p is less than p0= 0.05 . The results are:

1. Test statistic z0 is -2.119252 , z\_alpha is 1.644854 .

Because z0<-z\_alpha, we reject H0 at significance level 0.05

2. A 95% one-sided confidence interval for the population mean is  
(-Inf , 0.03766934 ] which does not contain the hypothesized  
value p0= 0.05 , we reject H0 at significance level 0.05

3. The P-value is 0.0170346 which is smaller than alpha= 0.05 ,  
so we reject H0 at significance level 0.05

Conclusion: at significance level  $\alpha = 0.05$ , the data provide sufficient evidence  
to reject  $H_0$ ; i.e., the process is capable.

## Example (continued)

Suppose that its process fallout rate is really  $p = 0.03$ . What is the  $\beta$ -error for a test of process capability that uses  $n = 300$  and  $\alpha = 0.05$ ?

**Solution:** This is a Type II Error (1-Power) calculation, where  $p_1 = 0.03$ ,  $p_0 = 0.05$ ,  $n = 300$ , left-tailed alternative at  $\alpha = 0.05$ .

```
Proptest.power(H1="left",alpha=0.05,p0=0.05,p1=0.03,n=300)
```

H1 is left-tailed

The probability of the Type II error of this test at  $p_1-p_0 = -0.02$  is 0.5282211 and the associated power is 0.4717789

## Example (continued)

Suppose that the semiconductor manufacturer was willing to accept a  $\beta$ -error as large as 0.10 if the true value of the process fraction defective was  $p = 0.03$ . If the manufacturer continues to use  $\alpha = 0.05$ , what sample size would be required?

**Solution:** This is a sample size calculation, where  $p_1 = 0.03$ ,  $p_0 = 0.05$ ,  $\beta = 0.1$ , left-tailed alternative at  $\alpha = 0.05$ .

```
sample.size.Proptest(H1="left",beta=0.1,alpha=0.05,p0=0.05,p1=0.03)
```

At significance level alpha= 0.05 , we need at least n= 833  
to achieve a power >= 0.9 of this test at p1-p0= -0.02

When n= 832 , the power is 0.899778

When n= 833 , the power is 0.9001346

When n= 834 , the power is 0.9004902

# StatEngine Summary of One-Sample Tests

Z-tests on  $\mu$  (Normality, known  $\sigma$ )

```
Ztest(mu0=?,H1=?,alpha=?,sigma=?,sample=?) #If data are available  
Ztest(mu0=?,H1=?,alpha=?,sigma=?,n=?,barx=?) #If statistics are provided  
Ztest.power(H1=?,alpha=?,sigma=?,n=?,delta=?)  
sample.size.Ztest(H1=?,sigma=?,alpha=?,beta=?,delta=?)
```

T-tests on  $\mu$  (unknown  $\sigma$ )

```
Ttest(mu0=?,H1=?,alpha=?,sample=?) #If data are available  
Ttest(mu0=?,H1=?,alpha=?,n=?,barx=?,s=?) #If statistics are provided  
sn=sd(data)  
Ttest.power(H1=?,est.sigma=sn,alpha=?,n=?,delta=?)  
sample.size.Ttest(H1=?,est.sigma=sn, beta=?,delta=?,alpha=?)
```

# StatEngine Summary of One-Sample Tests

Chi-square tests on  $\sigma^2$  or  $\sigma$  (Normality)

```
Chi2test(sigma0=?,H1=?,alpha=?,sample=?) #If data are available  
Chi2test(sigma0=?,H1=?,alpha=?,n=?,s=?) #If statistics are provided  
Chi2test.power(H1=?,alpha=?,n=?,lambda=?)  
sample.size.Chi2test(H1=?,beta=?,lambda=?,alpha=?)
```

One-Sample Proportion Z-tests on  $p$  ( $n\hat{p} \geq 5$ ,  $n(1 - \hat{p}) \geq 5$ )

```
Proptest(p0=?,H1=?,alpha=?,n=?,X=?)  
Proptest.power(H1=?,alpha=?,p0=?,p1=?,n=?)  
sample.size.Proptest(H1=?,beta=?,alpha=?,p0=?,p1=?)
```

## A Summary of Pattern of Hypothesis Testing

- ▶ Probability of Type I Error is always  $\alpha$ , user-chosen.
- ▶ When the sample size  $n$  is fixed, the power of test increases as the difference between the true parameter value and the hypothesized value increases in favor of  $H_1$ . Easier to detect a stronger signal!
- ▶ When a non-zero difference between the true parameter value and the hypothesized value in favor of  $H_1$  is fixed, the power of test increases as  $n$  increases. Easier to detect a signal with more samples!
- ▶ If the difference is zero or the hypothesized value is the true parameter value, the power of the test reduces to the Probability of Type I Error which is always  $\alpha$  no matter how large  $n$  is. Probability of Type I Error does not change with  $n$ .

# STAT 509: Statistics for Engineers

## Chapter 10: Statistical Inference for Two Samples

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

# Chapter 10: Statistical Inference for Two Samples

## Learning Objectives:

1. Structure comparative experiments involving two samples as hypothesis tests
2. Test hypotheses and construct confidence intervals on the difference in means of two normal distributions
3. Test hypotheses and construct confidence intervals on the ratio of the variances or standard deviations of two normal distributions
4. Test hypotheses and construct confidence intervals on the difference in two population proportions
5. Use the *P*-value approach for making decisions in hypotheses tests
6. Explain and use the relationship between confidence intervals and hypothesis tests

## Inference on the Difference in Means of Two Normal Distributions, Variances Known

The previous two chapters presented hypothesis tests and confidence intervals for a single population parameter (the mean  $\mu$ , the variance  $\sigma^2$ , or a proportion  $p$ ). This chapter extends those results to the case of two independent populations.

Most of the practical applications of the procedures in this chapter arise in the context of simple **comparative experiments** in which the objective is to study the difference in the parameters of the two populations.

## Inference on the Difference in Means of Two Normal Distributions, Variances Known

Engineers and scientists are often interested in comparing two different conditions to determine whether either condition produces a significant effect on the response that is observed. These conditions are sometimes called **treatments**.

For example, a product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. A study was conducted to determine whether the new formulation results in a significant effect—reducing drying time. In this situation, the product developer (the experimenter) randomly assigned 10 test specimens to one formulation and 10 test specimens to the other formulation. Then the paints were applied to the test specimens in random order until all 20 specimens were painted. This is an example of a **completely randomized experiment**.

## Inference on the Difference in Means of Two Normal Distributions, Variances Known

When statistical significance is observed in a randomized experiment, the experimenter can be confident in the conclusion that the difference in treatments resulted in the difference in response. That is, we can be confident that a **cause-and-effect** relationship has been found.

Sometimes the objects to be used in the comparison are not assigned at random to the treatments. A study, done in Finland, tracked 1931 men for 5 years and showed a statistically significant effect of increasing iron levels on the incidence of heart attacks. In this study, the comparison was not performed by randomly selecting a sample of men and then assigning some to a “low iron level” treatment and the others to a “high iron level” treatment. The researchers just tracked the subjects over time. This type of study is called an **observational study**.

## Inference on the Difference in Means of Two Normal Distributions, Variances Known

It is difficult to identify causality in observational studies because the observed statistically significant difference in response for the two groups may be due to some other underlying factor (or group of factors) that was not equalized by randomization and not due to the treatments. For example, the difference in heart attack risk could be attributable to the difference in iron levels or to other underlying factors that form a reasonable explanation for the observed results—such as cholesterol levels or hypertension.

In this chapter, we assume that we have two populations. Based on a random sample from each, we conduct two-sample inference:

1.  $X_{11}, \dots, X_{1n_1}$  is a random sample of size  $n_1$  from population 1.
2.  $X_{21}, \dots, X_{2n_2}$  is a random sample of size  $n_2$  from population 2.

The  $n_1$  and  $n_2$  could be different.

## Inference on the Difference in Means of Two Normal Distributions, Variances Known

We start with the **difference in means**  $\mu_1 - \mu_2$  of two **normal** distributions where the variances  $\sigma_1^2$  and  $\sigma_2^2$  are **known**. We consider two types of inferences:

(1) Constructing a  $100(1 - \alpha)\%$  confidence interval

- ▶ `twosample.Zinterval(level=?,  
sigma1=? , sigma2=? , sample1=? , sample2=? )`
- ▶ `twosample.Zinterval(level=? ,  
sigma1=? , sigma2=? , barx1=? , barx2=? , n1=? , n2=? )`

(2) and testing the null hypothesis  $H_0 : \mu_1 - \mu_0 = \Delta_0$  for a hypothesized value  $\Delta_0$  against a two-tailed ( $\neq$ ), left-tailed ( $<$ ), or right-tailed ( $>$ ) alternative hypothesis.

- ▶ `twosample.Ztest(Delta0=? , H1=? , alpha=? , sigma1=? ,  
sigma2=? , sample1=? , sample2=? )`
- ▶ `twosample.Ztest(Delta0=? , H1=? , alpha=? , sigma1=? ,  
sigma2=? , barx1=? , barx2=? , n1=? , n2=? )`

## Inference on the Difference in Means of Two Normal Distributions, Variances Known

Same as the one-sample case, the types of inferences both start with a point estimator of  $\mu_1 - \mu_2$ , which is

$$\bar{X}_{1n_1} - \bar{X}_{2n_2} = n_1^{-1} \sum_{i=1}^{n_1} X_{1i} - n_2^{-1} \sum_{i=1}^{n_2} X_{2i}.$$

And we know that

$$Z = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

Thus, the two-tailed  $100(1 - \alpha)\%$  CI for  $\mu_1 - \mu_2$  comes from

$$-z_{\alpha/2} \leq Z = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}.$$

## Inference on the Difference in Means of Two Normal Distributions, Variances Known

When testing  $H_0 : \mu_1 - \mu_2 = \Delta_0$  for a hypothesized value  $\mu_0$ ,

The test statistic is  $Z_0 = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ .

Based on samples  $x_{11}, \dots, x_{1n_1}$  and  $x_{21}, \dots, x_{2n_2}$ , the observed test statistic is  $z_0 = \frac{\bar{x}_{1n_1} - \bar{x}_{2n_2} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ .

If the alternative hypothesis is  $H_1 : \mu_1 - \mu_2 \neq \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $|z_0| > z_{\alpha/2}$  (critical-value approach),
- ▶  $\mu_0 \notin \left[ \bar{x}_{1n_1} - \bar{x}_{2n_2} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$  (confidence-interval approach),
- ▶ the  $P$ -value  $= 2[1 - P(Z \leq |z_0|)] < \alpha$ , where  $Z \sim N(0, 1)$  ( $P$ -value approach).

## Tests on the Mean of a Normal Distribution, Variance Known

If the alternative hypothesis is  $H_1 : \mu_1 - \mu_2 > \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $z_0 > z_\alpha$  (critical-value approach),
- ▶  $\mu_0 \notin \left[ \bar{x}_{1n_1} - \bar{x}_{2n_2} - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \infty \right)$  (confidence-interval approach),
- ▶ the  $P$ -value =  $P(Z > z_0) < \alpha$ , where  $Z \sim N(0, 1)$  ( $P$ -value approach).

If the alternative hypothesis is  $H_1 : \mu_1 - \mu_2 < \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if one of the following holds:

- ▶  $z_0 < -z_\alpha$  (critical-value approach),
- ▶  $\mu_0 \notin \left( -\infty, \bar{x}_{1n_1} - \bar{x}_{2n_2} + z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$  (confidence-interval approach),
- ▶ the  $P$ -value =  $P(Z \leq z_0) < \alpha$ , where  $Z \sim N(0, 1)$  ( $P$ -value approach).

## Example 1

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows:  $n_1 = 10$ ,  $\bar{x}_1 = 87.6$ ,  $\sigma_1 = 1$ ,  $n_2 = 12$ ,  $\bar{x}_2 = 74.5$ , and  $\sigma_2 = 1.5$ . If  $\mu_1$  and  $\mu_2$  denote the true mean tensile strengths for the two grades of spars, find a 90% confidence interval on the difference in mean strength. Assume normality.

**solution:** Normality and known variances.

```
twosample.Zinterval(level=0.9, sigma1=1,sigma2=1.5,
                     barx1=87.6,barx2=74.5,n1=10,n2=12)
A 90% two-sided confidence interval for
the difference in population means is [ 12.21805 , 13.98195 ]
A 90% upper-confidence bound for the population mean is 13.78716
A 90% lower-confidence bound for the population mean is 12.41284
```

Conclusion: Based on the data, a 90% confidence interval on  $\mu_1 - \mu_2$  is [12.2181, 13.982].

## Example 2

A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are  $\bar{x}_1 = 121$  minutes and  $\bar{x}_2 = 112$  minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using  $\alpha = 0.05$ ?

## Example 2

**solution:** Normality and known variances with  $\sigma_1 = \sigma_2 = 8$ . We are testing  $H_0 : \mu_1 - \mu_2 = 0$  versus  $H_1 : \mu_1 > \mu_2$ . Then  $\Delta_0 = 0$  and we have a right-tailed alternative.

```
twosample.Ztest(Delta0=0,H1="right",alpha=0.05, sigma1=8,sigma2=8,  
                 barx1=121,barx2=112,n1=10,n2=10)
```

H1 is right-tailed. The results are:

1. Test statistic z0 is 2.515576 , z\_alpha is 1.644854 .  
Because z0>z\_alpha, we reject H0 at significance level 0.05

2. A 95% one-sided confidence interval for the population mean is  
[ 3.115193 , Inf ) which does not contain the hypothesized value Delta0= 0 ,  
so we reject H0 at significance level 0.05

3. The P-value is 0.005941895 which is smaller than alpha= 0.05 ,  
so we reject H0 at significance level 0.05

Conclusion: at significance level 0.05, the data provide sufficient evidence to reject  $H_0$ .

## Inference on the Difference in Means of Two Normal Distributions, Variances Unknown, pooled=yes or no

We then consider the **difference in means**  $\mu_1 - \mu_2$  of two **normal distributions** where the variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown**. In order to have a better power, we separate the unknown variances into two cases

Case I: pooled=yes: We do not know  $\sigma_1^2$  or  $\sigma_2^2$ , but there is information for us to conclude  $\sigma_1^2 = \sigma_2^2$ .

Case II: pooled=no: We do not know  $\sigma_1^2$  or  $\sigma_2^2$ , and there is no information for us to conclude  $\sigma_1^2 = \sigma_2^2$ .

If we are able to conclude  $\sigma_1^2 = \sigma_2^2$  (Case I), this extra piece of information can help us gain a better power for testing whether  $H_0 : \mu_1 - \mu_2 = \Delta_0$  holds.

## Inference on the Difference in Means of Two Normal Distributions, Variances Unknown, pooled=yes

We then consider the **difference in means**  $\mu_1 - \mu_2$  of two **normal** distributions where the variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown** but we can conclude  $\sigma_1^2 = \sigma_2^2$  (Pooled=yes). We consider two types of inferences:

(1) Constructing a  $100(1 - \alpha)\%$  confidence interval

- ▶ `twosample.Tinterval(level=?, pooled=yes,  
sample1=? ,sample2=? )`
- ▶ `twosample.Tinterval(level=?, pooled=yes,  
barx1=? ,barx2=? ,n1=? ,n2=? ,s1=? ,s2=? )`

(2) and testing the null hypothesis  $H_0 : \mu_1 - \mu_0 = \Delta_0$  for a hypothesized value  $\Delta_0$  against a two-tailed ( $\neq$ ), left-tailed ( $<$ ), or right-tailed ( $>$ ) alternative hypothesis.

- ▶ `twosample.Ttest(Delta0=? ,H1=? ,alpha=? ,pooled=yes,  
sample1=? ,sample2=? )`
- ▶ `twosample.Ttest(Delta0=? ,H1=? ,alpha=? ,pooled=yes,  
barx1=? ,barx2=? ,n1=? ,n2=? ,s1=? ,s2=? )`

## Inference on the Difference in Means of Two Normal Distributions, Variances Unknown, pooled=yes

The point estimator of  $\mu_1 - \mu_2$  is still  $\bar{X}_{1n_1} - \bar{X}_{2n_2}$ . If  $\sigma_1^2 = \sigma_2^2$  holds, we let the common variance be  $\sigma^2$ . Then we can combine the two samples together to get a pooled estimator of  $\sigma^2$ , denoted by  $S_p^2$ ,

$$S_p^2 = \frac{(n_1 - 1)S_{1n_1}^2 + (n_2 - 1)S_{2n_2}^2}{n_1 + n_2 - 2}.$$

Then

$$T = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

Thus, the two-tailed  $100(1 - \alpha)\%$  CI for  $\mu_1 - \mu_2$  comes from

$$-t_{n_1+n_2-2,\alpha/2} \leq T = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2,\alpha/2}.$$

## Inference on the Difference in Means of Two Normal Distributions, Variances Unknown, pooled=yes

When testing  $H_0 : \mu_1 - \mu_2 = \Delta_0$  for a hypothesized value  $\mu_0$ ,

The test statistic is  $T_0 = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ .

Based on samples  $x_{11}, \dots, x_{1n_1}$  and  $x_{21}, \dots, x_{2n_2}$ , the observed test statistic is  $t_0 = \frac{\bar{x}_{1n_1} - \bar{x}_{2n_2} - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ .

- ▶ If  $H_1 : \mu_1 - \mu_2 \neq \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $|t_0| > t_{n_1+n_2-2, \alpha/2}$  (critical-value approach),
- ▶ If  $H_1 : \mu_1 - \mu_2 < \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $t_0 < -t_{n_1+n_2-2, \alpha}$  (critical-value approach),
- ▶ If  $H_1 : \mu_1 - \mu_2 > \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $t_0 > t_{n_1+n_2-2, \alpha}$  (critical-value approach).

We also have the confidence-interval approach and the  $P$ -value approach.

## Inference on the Difference in Means of Two Normal Distributions, Variances Unknown, pooled=no

We then consider the **difference in means**  $\mu_1 - \mu_2$  of two **normal** distributions where the variances  $\sigma_1^2$  and  $\sigma_2^2$  are **unknown** but we cannot conclude  $\sigma_1^2 = \sigma_2^2$  (Pooled=no). We consider two types of inferences:

(1) Constructing a  $100(1 - \alpha)\%$  confidence interval

- ▶ `twosample.Tinterval(level=?, pooled=no,  
sample1=? ,sample2=? )`
- ▶ `twosample.Tinterval(level=?, pooled=no,  
barx1=? ,barx2=? ,n1=? ,n2=? ,s1=? ,s2=? )`

(2) and testing the null hypothesis  $H_0 : \mu_1 - \mu_0 = \Delta_0$  for a hypothesized value  $\Delta_0$  against a two-tailed ( $\neq$ ), left-tailed ( $<$ ), or right-tailed ( $>$ ) alternative hypothesis.

- ▶ `twosample.Ttest(Delta0=? ,H1=? ,alpha=? ,pooled=no,  
sample1=? ,sample2=? )`
- ▶ `twosample.Ttest(Delta0=? ,H1=? ,alpha=? ,pooled=no,  
barx1=? ,barx2=? ,n1=? ,n2=? ,s1=? ,s2=? )`

## Inference on the Difference in Means of Two Normal Distributions, Variances Unknown, pooled=no

The point estimator of  $\mu_1 - \mu_2$  is still  $\bar{X}_{1n_1} - \bar{X}_{2n_2}$ . If  $\sigma_1^2 = \sigma_2^2$  does not hold, we estimate them separately by the sample variances  $S_1^2$  and  $S_2^2$ . Then

$$T = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(v),$$

where

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

Thus, the two-tailed  $100(1 - \alpha)\%$  CI for  $\mu_1 - \mu_2$  comes from

$$-t_{v,\alpha/2} \leq T = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{v,\alpha/2}.$$

## Inference on the Difference in Means of Two Normal Distributions, Variances Unknown, pooled=no

When testing  $H_0 : \mu_1 - \mu_2 = \Delta_0$  for a hypothesized value  $\mu_0$ ,

The test statistic is  $T_0 = \frac{\bar{x}_{1n_1} - \bar{x}_{2n_2} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ .

Based on samples  $x_{11}, \dots, x_{1n_1}$  and  $x_{21}, \dots, x_{2n_2}$ , the observed test statistic is  $t_0 = \frac{\bar{x}_{1n_1} - \bar{x}_{2n_2} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ .

- ▶ If  $H_1 : \mu_1 - \mu_2 \neq \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $|t_0| > t_{v,\alpha/2}$  (critical-value approach),
- ▶ If  $H_1 : \mu_1 - \mu_2 < \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $t_0 < -t_{v,\alpha}$  (critical-value approach),
- ▶ If  $H_1 : \mu_1 - \mu_2 > \Delta_0$ . At significance level  $\alpha$ , we reject  $H_0$  if  $t_0 > t_{v,\alpha}$  (critical-value approach).

We also have the confidence-interval approach and the  $P$ -value approach.

## Example 3

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently used; but catalyst 2 is acceptable. Because catalyst 2 is cheaper, it should be adopted, if it does not change the process yield. A test is run in the pilot plant and results in the data.

```
catalyst1=c(91.50,94.18,92.18,95.39,91.79,89.07,94.72,89.21)  
catalyst2=c(89.19,90.95,90.46,93.21,97.19,97.04,91.07,92.75)
```

Use QQ-plot to check normality assumption. Then answer: Is there any difference in the mean yields? Use  $\alpha = 0.05$ , and assume equal variances.

**Solution:** We first check normality using QQ-plot, which can be done using

```
data.summary(catalyst1)  
data.summary(catalyst2)
```

## Example 3

Then conduct the test  $H_0 : \mu_1 = \mu_2$  versus  $H_1 : \mu_1 \neq \mu_2$  in which  $\Delta_0 = 0$ . We have normality checked, but we do not know the population variances. Thus we use two-sample T-tests. Furthermore, the question assumes equal variances, which requires the use of "pooled=yes"

```
twosample.Ttest(Delta0=0,H1="two",alpha=0.05, pooled="yes",
                 sample1=catalyst1,sample2=catalyst2)
```

H1 is two-tailed. The results are:

1. Test statistic t0 is -0.3535909 ,  $t_{(v,\alpha/2)}$  is 2.144787 .  
Because  $|t_0| <= t_{(v,\alpha/2)}$ , we fail to reject  $H_0$  at significance level 0.05
2. A 95% two-tailed confidence interval for the population mean is [ -3.373886 , 2.418886 ] which contains the hypothesized value Delta0= 0 , so we fail to reject  $H_0$  at significance level 0.05
3. The P-value is 0.7289136 which is not smaller than alpha= 0.05 , so we fail to reject  $H_0$  at significance level 0.05

Conclusion: at significance level 0.05, the data do not provide sufficient evidence to reject  $H_0$ .

## Example 4

An article in Polymer Degradation and Stability (2006, Vol. 91) presented data from a nine-year aging study on S537 foam. Foam samples were compressed to 50% of their original thickness and stored at different temperatures for nine years. At the start of the experiment as well as during each year, sample thickness was measured, and the thicknesses of the eight samples at each storage condition were recorded. The data for two storage conditions follow.

$C50=c(0.047, 0.060, 0.061, 0.064, 0.080, 0.090, 0.118, 0.165, 0.183)$   
 $C60=c(0.062, 0.105, 0.118, 0.137, 0.153, 0.197, 0.210, 0.250, 0.335)$

- (a) Is there evidence to claim that mean compression increases with the temperature at the storage condition? ( $\alpha = 0.05$ ).
- (b) Find a 95% confidence interval for the difference in the mean compression for the two temperatures.
- (c) Is the value zero contained in the 95% confidence interval? Explain the connection with the conclusion reached in part (a).
- (d) Do normal probability plots of compression indicate any violations of the assumptions for the tests and confidence interval that you performed?

## Example 4

Solution: (a) Let  $\mu_1$  be mean compression at 50 Celsius and  $\mu_2$  be the one at 60 Celsius. We are testing  $H_0 : \mu_1 = \mu_2$  versus  $H_1 : \mu_1 < \mu_2$  ( $\Delta = 0$  and left-tailed).

```
twosample.Ttest(Delta0=0,H1="left",alpha=0.05, pooled="no",
                 sample1=C50,sample2=C60)
```

H1 is left-tailed. The results are:

1. Test statistic t0 is -2.409256 ,  $t_{(v,\alpha)}$  is 1.772046 .  
Because  $t_0 < t_{(v,\alpha)}$ , we reject  $H_0$  at significance level 0.05
2. A 95% one-sided confidence interval for the population mean is  
 $(-\infty, -0.0205416]$  which does not contain the hypothesized value  $\Delta_0 = 0$  ,  
so we reject  $H_0$  at significance level 0.05
3. The P-value is 0.01583664 which is smaller than  $\alpha = 0.05$  ,  
so we reject  $H_0$  at significance level 0.05

Conclusion: at significance level 0.05, the data provide sufficient evidence to reject  $H_0$ .  
Thus, there is evidence to claim that mean compression increases with the temperature  
at the storage condition.

## Example 4

Solution: (b,c) see result 2 of part (a). Because we are focusing on a one-side comparison. We use the one-sided confidence interval  $(-\infty, -0.0205]$  for  $\mu_1 - \mu_2$ . Because 0 is not contained it, we would reject  $H_0$  in (a).

(d) Check normality:

```
data.summary(C50)  
data.summary(C60)
```

We can see that normality are checked. We could not conclude whether this is pooled or not. It is important to learn how to test whether two population variances are equal (later).

## Paired T-Test and T-interval

A special case of the two-sample T-tests occurs when the observations on the two populations of interest are collected in pairs.

For example, suppose that we are interested in comparing two different types of tips for a hardness-testing machine. This machine presses the tip into a metal specimen with a known force. By measuring the depth of the depression caused by the tip, the hardness of the specimen can be determined.

If several specimens were selected at random, half tested with tip 1, half tested with tip 2, and a two sample t-test (pooled or not) was applied, the results of the test could be erroneous.

The metal specimens could have been cut from bar stock that was produced in different heats, or they might not be homogeneous in some other way that might affect hardness. Then **the observed difference in mean hardness readings for the two tip types also includes hardness differences in specimens.**

## Paired T-Test and T-interval

A more powerful experimental procedure is to collect the data in pairs—that is, to make two hardness readings on each specimen, one with each tip. The test procedure would then consist of analyzing the differences in hardness readings on each specimen. If there is no difference between tips, the mean of the differences should be zero. This test procedure is called the paired T-test.

Let  $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$  be a set of  $n$  paired observations for which we assume that the mean and variance of the population represented by  $X_1$  are  $\mu_1$  and  $\sigma_1^2$ , the mean and variance of the population represented by  $X_2$  are  $\mu_2$  and  $\sigma_2^2$ . Define the difference for each pair of observations as  $D_j = X_{1j} - X_{2j}$ ,  $j = 1, 2, \dots, n$ . Then  $D_j$ 's are assumed to be **normally** distributed with mean

$$\mu_D = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

and some variance  $\sigma_D^2$  (remains **unknown**).

## Paired T-Test and T-interval

So testing  $H_0 : \mu_1 - \mu_2 = \Delta_0$  is equivalent to testing

$$H_0 : \mu_D = \Delta_0,$$

which can be done using the one-sample T-test:

```
Ttest(mu0=Delta0, H1=? ,alpha=? ,sample=sample1-sample2)  
Ttest(mu0=Delta0, H1=? ,alpha=? ,n=? ,  
      barx=mean(sample1-sample2) ,s=sd(sample1-sample2))
```

Confidence intervals for  $\mu_D = \mu_1 - \mu_2$  from paired samples can be done using the one-sample T-interval:

```
Tinterval(level=? ,sample=? )  
Tinterval(level=? ,n=? ,  
      barx=mean(sample1-sample2) ,s=sd(sample1-sample2))
```

## Example 5

An article in the Journal of Strain Analysis [1983, Vol. 18(2)] reports a comparison of several methods for predicting the shear strength for steel plate girders. Data for two methods, the Karlsruhe and Lehigh procedures, when applied to nine specific girders, are shown as

```
Girder=c(1, 2, 3, 4, 5, 6, 7, 8, 9)
Karlsrube=c(1.186,1.151,1.322,1.229,1.200,1.402,1.365,1.537,1.559)
Lehigh=c(1.061,0.992,1.063,1.062,1.065,1.178,1.037,1.086,1.052)
```

We wish to determine whether there is any difference (on the average) for the two methods.

**Solution:** Obviously, data were taken in pairs. We should use paired T-test to determine whether there is any difference on the average for the two methods. The hypotheses are  $H_0 : \mu_1 = \mu_2$  vs  $H_1 : \mu_1 \neq \mu_2$ , which equal to  $H_0 : \mu_D = 0$  vs  $H_1 : \mu_D \neq 0$ . The *alpha* is not told, we can simply choose 0.05 to compute the *P*-value.

```
data.summary(Karlsrube-Lehigh) #check normality
Ttest(mu0=0, H1="two", alpha=0.05, sample=Karlsrube-Lehigh)
3. The P-value is 0.000498523
```

Conclusion: at significance level  $\alpha > 0.0005$ , data provide sufficient evidence to reject  $H_0$ ; i.e., there very likely exists difference (on the average) between the two methods. 29 / 45

## Inference on the Variances of Two Normal Distributions

Suppose that two independent normal populations are of interest when the population means and variances, say,  $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ , are unknown. Assume that two random samples of size  $n_1$  from population 1 and of size  $n_2$  from population 2 are available, and let  $S_1^2$  and  $S_2^2$  be the sample variances. We wish to build confidence intervals (F-interval) on the ratio between two population variances

$$\frac{\sigma_1^2}{\sigma_2^2},$$

and test (F-test) the hypotheses

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_1 : \sigma_1^2 \neq \sigma_2^2 \text{ (two-tailed)}$$

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_1 : \sigma_1^2 < \sigma_2^2 \text{ (left-tailed)}$$

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_1 : \sigma_1^2 > \sigma_2^2 \text{ (right-tailed)}$$

## Inference on the Variances of Two Normal Distributions

More specifically, we consider two types of inferences on the ratio  $\sigma_1^2/\sigma_2^2$  of the variances of two **normal** distributions.

(1) Constructing a  $100(1 - \alpha)\%$  confidence interval

- ▶ Finterval(level=?, sample1=?,sample2=?)
- ▶ Finterval(level=?, n1=?,n2=?,s1=?,s2=?)

(2) and testing the null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2$  against a two-tailed ( $\neq$ ), left-tailed ( $<$ ), or right-tailed ( $>$ ) alternative hypothesis.

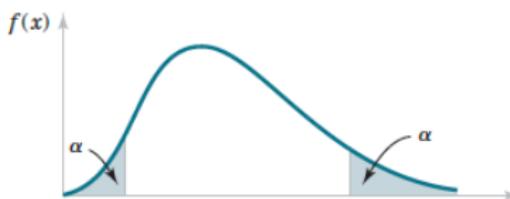
- ▶ Ftest(H1=?,alpha=?,sample1=?,sample2=?)
- ▶ Ftest(H1=?,alpha=?,n1=?,n2=?,s1=?,s2=?)

# Inference on the Variances of Two Normal Distributions

The CI starts with a fact that

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \div \frac{\sigma_1^2}{\sigma_2^2} \sim F(n_1 - 1, n_2 - 1),$$

where  $F(n_1 - 1, n_2 - 1)$  stands for an  $F$ -distribution with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom.



Then a  $100(1 - \alpha)\%$  two-tailed CI for  $\sigma_1^2/\sigma_2^2$  comes from

$$f_{n_1-1, n_2-1, 1-\alpha/2} \leq F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \div \frac{\sigma_1^2}{\sigma_2^2} \leq f_{n_1-1, n_2-1, \alpha/2}$$

## Inference on the Variances of Two Normal Distributions

When testing  $H_0 : \sigma_1^2 = \sigma_2^2$ ,

The test statistic is  $F = \frac{s_1^2}{s_2^2}$ .

Based on samples  $x_{11}, \dots, x_{1n_1}$  and  $x_{21}, \dots, x_{2n_2}$ , the observed test statistic is  $F_0 = \frac{s_1^2}{s_2^2}$ .

- ▶ If  $H_1 : \sigma_1^2 \neq \sigma_2^2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $F_0 > f_{n_1-1, n_2-1, \alpha/2}$  or  $F_0 < f_{n_1-1, n_2-1, 1-\alpha/2}$  (critical-value approach),
- ▶ If  $H_1 : \sigma_1^2 < \sigma_2^2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $F_0 < f_{n_1-1, n_2-1, 1-\alpha}$  (critical-value approach),
- ▶ If  $H_1 : \sigma_1^2 > \sigma_2^2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $F_0 > f_{n_1-1, n_2-1, \alpha}$  (critical-value approach).

We also have the confidence-interval approach and the  $P$ -value approach.

## Example 6

A company manufactures impellers for use in jet-turbine engines. One of the operations involves grinding a particular surface finish on a titanium alloy component. Two different grinding processes can be used, and both processes can produce parts at identical mean surface roughness. The manufacturing engineer would like to select the process having the least variability in surface roughness. A random sample of  $n_1 = 11$  parts from the first process results in a sample standard deviation  $s_1 = 5.1$  microinches, and a random sample of  $n_2 = 16$  parts from the second process results in a sample standard deviation of  $s_2 = 4.7$  microinches. Is there any difference between  $\sigma_1$  and  $\sigma_2$  at  $\alpha = 0.1$ ?

## Example 6

**Solution:** We are testing  $H_0 : \sigma_1 = \sigma_2$  vs  $H_1 : \sigma_1 \neq \sigma_2$  (F-test at  $\alpha = 0.1$ )

```
Ftest(H1="two",alpha=0.1,n1=11,n2=16,s1=5.1,s2=4.7)
```

H1 is two-tailed. The results are:

1. Test statistic F0 is 1.177456 ,  
 $f_{(n_1-1, n_2-1, 1-\alpha/2)}$  is 0.3514918 ,  
 $f_{(n_1-1, n_2-1, \alpha/2)}$  is 2.543719 .

Because  $f_{(n_1-1, n_2-1, 1-\alpha/2)} \leq F_0 \leq f_{(n_1-1, n_2-1, \alpha/2)}$ ,  
we fail to reject  $H_0$  at significance level 0.1

2. A 90% two-tailed confidence interval for the ratio  
between two population variances is [ 0.4628876 , 3.349881 ]  
which contains 1, so we fail to reject  $H_0$  at significance level 0.1

3. The P-value is 0.7508517 which is not smaller than  $\alpha = 0.1$  ,  
so we fail to reject  $H_0$  at significance level 0.1

Conclusion: at significance level  $\alpha = 0.1$ , data do not provide sufficient evidence to  
reject  $H_0$ .

## Large-sample Inference on Two Population Proportions

Suppose that two independent random samples of sizes  $n_1$  and  $n_2$  are taken from two populations, and let  $X_1$  and  $X_2$  represent the number of observations that belong to the class of interest in samples 1 and 2, respectively. Denote the population proportions of interest by  $p_1$  and  $p_2$ .

We wish to build confidence intervals on the **difference in proportions** ( $p_1 - p_2$ ) and test the hypotheses

$$H_0 : p_1 = p_2 \text{ versus } H_1 : p_1 \neq p_2 \text{ (two-tailed)}$$

$$H_0 : p_1 = p_2 \text{ versus } H_1 : p_1 < p_2 \text{ (left-tailed)}$$

$$H_0 : p_1 = p_2 \text{ versus } H_1 : p_1 > p_2 \text{ (right-tailed)}$$

# Large-sample Inference on the Variances of Two Normal Distributions

More specifically, we consider two types of inferences on  $p_1 - p_2$ .

(1) Constructing a  $100(1 - \alpha)\%$  confidence interval

► `twosample.Propinterval(level=? , n1=? , n2=? , X1=? , X2=? )`

(2) and testing the null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2$  against a two-tailed ( $\neq$ ), left-tailed ( $<$ ), or right-tailed ( $>$ ) alternative hypothesis.

► `twosample.Proptest(H1=? , alpha=? , n1=? , n2=? , X1=? , X2=? )`

## Large-sample Inference on the Variances of Two Normal Distributions

We estimate  $p_1$  by  $\hat{p}_1 = X_1/n_1$  and  $p_2$  by  $\hat{p}_2 = X_2/n_2$ . The CI starts with a fact that

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim AN(0, 1).$$

Then an **approximate**  $100(1 - \alpha)\%$  two-tailed CI for  $p_1 - p_2$  comes from

$$-z_{\alpha/2} \leq Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_{\alpha/2}.$$

It is valid if all  $n_1\hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2, n_2(1 - \hat{p}_2) \geq 5$ .

## Large-sample Inference on the Variances of Two Normal Distributions

Require:  $n_1\hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2, n_2(1 - \hat{p}_2) \geq 5$

When testing  $H_0 : p_1 = p_2$ ,

The observed test statistic is  $z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

where  $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ .

- ▶ If  $H_1 : p_1 \neq p_2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $|z_0| > z_{\alpha/2}$  (critical-value approach),
- ▶ If  $H_1 : p_1 < p_2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $z_0 < -z_\alpha$  (critical-value approach),
- ▶ If  $H_1 : p_1 > p_2$ . At significance level  $\alpha$ , we reject  $H_0$  if  $z_0 > z_\alpha$  (critical-value approach).

We also have the confidence-interval approach and the  $P$ -value approach. The confidence-interval approach **might** give a different conclusion than the other two approaches. But this difference is very minor, especially when sample sizes  $n_1$  and  $n_2$  are large.

## Example 7

Extracts of St. John's Wort are widely used to treat depression. An article in the April 18, 2001, issue of the Journal of the American Medical Association ("Effectiveness of St. John's Wort on Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. John's Wort with a placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo. After eight weeks, 19 of the placebo-treated patients showed improvement, and 27 of those treated with St. John's Wort improved. Is there any reason to believe that St. John's Wort is effective in treating major depression? Use  $\alpha = 0.05$ .

## Example 7

**Solution:** The parameters of interest are  $p_1$  and  $p_2$ , the proportion of patients who improve following treatment with St. John's Wort ( $p_1$ ) or the placebo ( $p_2$ ). We are testing  $H_0 : p_1 = p_2$  vs  $H_1 : p_1 > p_2$  (two-sample proportion Z-test, right-tailed, at  $\alpha = 0.05$ )

```
twosample.Proptest(H1="right",alpha=0.05,n1=100,n2=100,X1=27,X2=19)  
H1 is right-tailed. The results are:
```

1. Test statistic  $z_0$  is 1.344206 ,  $z_{\alpha}$  is 1.644854 .  
Because  $z_0 \leq z_{\alpha}$ , we fail to reject  $H_0$  at significance level 0.05
2. A 95% one-sided confidence interval for the difference  
in population proportions is [ 0.1774498 , Inf )  
which contains 0, so we fail to reject  $H_0$  at significance level 0.05
3. The P-value is 0.08944095 which is not smaller than  $\alpha = 0.05$  ,  
so we fail to reject  $H_0$  at significance level 0.05

Conclusion: at significance level  $\alpha = 0.05$ , data do not provide sufficient evidence to  
reject  $H_0$ .

## StatEngine Summary of Two-Sample Inferences

Z-intervals and Z-tests On  $\mu_1 - \mu_2$  (Normality, known  $\sigma_1^2$  and  $\sigma_2^2$ ).

```
twosample.Zinterval(level=?, sigma1=?, sigma2=?,
                     sample1=?, sample2=?)
```

```
twosample.Zinterval(level=?, sigma1=?, sigma2=?,
                     barx1=?, barx2=? , n1=? , n2=?)
```

```
twosample.Ztest(Delta0=?, H1=?, alpha=?, sigma1=?, sigma2=? ,
                  sample1=?, sample2=?)
```

```
twosample.Ztest(Delta0=?, H1=?, alpha=?, sigma1=?, sigma2=? ,
                  barx1=?, barx2=? , n1=? , n2=?)
```

## StatEngine Summary of Two-Sample Inferences

T-intervals and T-tests On  $\mu_1 - \mu_2$  (Normality, unknown  $\sigma_1^2$  and  $\sigma_2^2$ , pooled=yes or no).

```
twosample.Tinterval(level=?, pooled=?, sample1=?, sample2=?)
```

```
twosample.Tinterval(level=?, pooled=?, barx1=?, barx2=?,
                     n1=?, n2=?, s1=?, s2=?)
```

```
twosample.Ttest(Delta0=?, H1=?, alpha=?, pooled=yes,
                  sample1=?, sample2=?)
```

```
twosample.Ttest(Delta0=?, H1=?, alpha=?, pooled=yes,
                  barx1=?, barx2=? ,n1=?, n2=?, s1=?, s2=?)
```

## StatEngine Summary of Two-Sample Inferences

Paired T-intervals and T-tests On  $\mu_1 - \mu_2$  (Paires, Normality, unknown varaince  $\sigma_D^2$ ).

```
Ttest(mu0=Delta0, H1=? ,alpha=? ,sample=sample1-sample2)  
Ttest(mu0=Delta0, H1=? ,alpha=? ,n=? ,  
      barx=mean(sample1-sample2) ,s=sd(sample1-sample2))
```

```
Tinterval(level=? ,sample=sample1-sample2)  
Tinterval(level=? ,n=? ,  
      barx=mean(sample1-sample2) ,s=sd(sample1-sample2))
```

## StatEngine Summary of Two-Sample Inferences

F-intervals and F-tests On  $\sigma_1^2/\sigma_2^2$  (Normality).

`Finterval(level=? , sample1=? , sample2=? )`

`Finterval(level=? , n1=? , n2=? , s1=? , s2=? )`

`Ftest(H1=? , alpha=? , sample1=? , sample2=? )`

`Ftest(H1=? , alpha=? , n1=? , n2=? , s1=? , s2=? )`

Two-Sample Proportion Z-intervals and Z-tests On  $p_1 - p_2$  ( $n_1 \hat{p}_1$ ,  $n_1(1 - \hat{p}_1)$ ,  $n_2 \hat{p}_2$ ,  $n_2(1 - \hat{p}_2) \geq 5$ ).

`twosample.Propinterval(level=? , n1=? , n2=? , X1=? , X2=? )`

`twosample.Proptest(H1=? , alpha=? , n1=? , n2=? , X1=? , X2=? )`

# STAT 509: Statistics for Engineers

## Chapters 11-12: Linear Regression

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

## Chapters 11-12: Linear Regression

### Learning Objectives:

1. Use linear regression for building empirical models to engineering and scientific data
2. Understand how the method of least squares is used to estimate the parameters in a linear regression model
3. Test statistical hypotheses and construct confidence intervals on regression model parameters
4. Use the regression model to predict a future observation and to construct an appropriate prediction interval on the future observation
5. Build regression models with polynomial terms
6. Use indicator variables to model categorical regressors
7. Analyze residuals to determine whether the regression model is an adequate fit to the data or whether any underlying assumptions are violated

## Introduction

In previous chapters, we used the normal distribution  $Y \sim N(\mu, \sigma^2)$  to model data. Suppose  $Y$  represents the salary of your first job after undergraduate study,  $\mu = 35K$ , and  $\sigma = 5K$ . It is saying that, about 95% students should expect a salary between  $[25K, 45K]$ . Now let  $x$  be your GPA. It is possible to think  $x$  could be an important factor determining your salary level. But the above normal model does not account for the factor  $x$ .

A simple improvement is to let  $\mu$  be a linear function of  $x$ ; i.e.,

$$Y|x \sim N(\mu_x, \sigma^2), \text{ where } \mu_x = \beta_0 + \beta_1 x,$$

or we write (simple linear regression)

$$Y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

If  $\beta_0 = 10K$ ,  $\beta_1 = 10K$ ,  $\sigma = 3K$ , then  $Y|x = 1 \sim N(20K, 3K^2)$ ,  $Y|x = 3 \sim N(40K, 3K^2)$ . If  $x = 3$ , with  $\approx 95\%$  chance, salary will be between  $[34K, 46K]$ . The  $\epsilon$  accounts for a **non-deterministic** relationship between  $Y$  (salary) and  $x$  (GPA).

## Deterministic Models

Many problems in engineering and the sciences involve a study or analysis of the relationship between two or more variables. For example, the velocity of water in an open channel is related to the width of the channel, and the displacement of a particle at a certain time is related to its velocity. In this last example, if we let  $d_0$  be the displacement of the particle from the origin at time  $x = 0$  and  $v$  be the velocity, the displacement at time  $x$  is  $d_x = d_0 + vx$ . This is an example of a **deterministic** linear relationship because (apart from measurement errors) the model predicts displacement perfectly.

## Deterministic Models

However, in many situations, the relationship between variables is not deterministic. For example, the electrical energy consumption of a house ( $y$ ) is related to the size of the house ( $x$ , in square feet), but it is unlikely to be a deterministic relationship. Similarly, the fuel usage of an automobile ( $y$ ) is related to the vehicle weight  $x$ , but the relationship is not deterministic. In both of these examples, the value of  $y$  cannot be predicted perfectly from knowledge of the corresponding  $x$ . It is possible for different automobiles to have different fuel usage even if they weigh the same, and it is possible for different houses to use different amounts of electricity even if they are the same size.

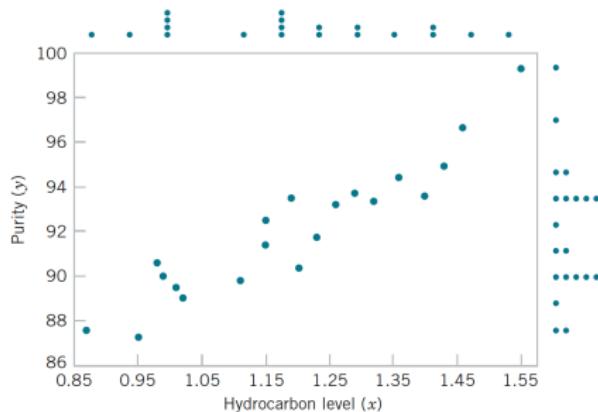
## Empirical Models

The collection of statistical tools that are used to model and explore relationships between variables that are related in a **non-deterministic** manner is called **regression analysis**. Because problems of this type occur so frequently in many branches of engineering and science, regression analysis is one of the most widely used statistical tools.

For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature. Regression analysis can be used to build a model to predict yield at a given temperature level. This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes

## Example 1

Observation Number	Hydrocarbon Level $x$ (%)	Purity $y$ (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



# Simple Linear Regression

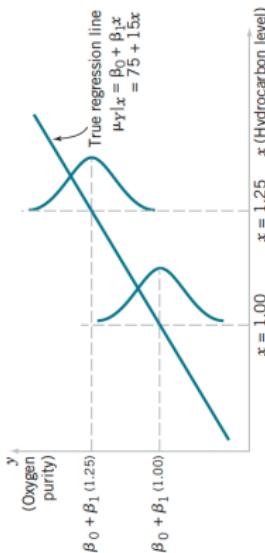
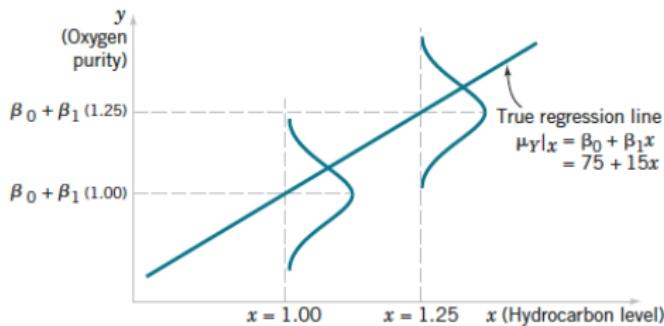
We explain the relationship between  $Y = y$  and  $x$  by a empirical model (simple linear regression)

$$Y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

The  $\epsilon$  is a random error term which collects all the variation that cannot be explained by the linear relationship  $Y = \beta_0 + \beta_1 x$ .

At a fixed  $x$ , the distribution of  $Y$  is

$$N(\beta_0 + \beta_1 x, \sigma^2).$$



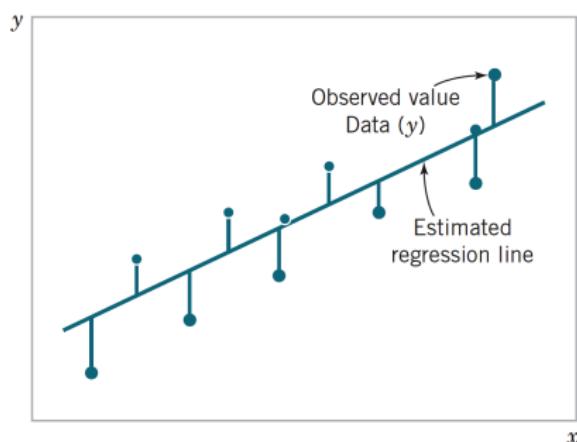
# Least Squares Estimates

In simple linear regression, we have

$$Y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

There are three unknown parameters  $\beta_0$  (intercept),  $\beta_1$  (slope), and  $\sigma^2$  (the variance of the random error) we wish to estimate based on  $n$  pairs of observations  $(x_i, y_i), i = 1, \dots, n$ .

The estimation of  $\beta_0$  and  $\beta_1$  comes from the method called Least Squares (by German scientist Karl Gauss, 1777–1855); i.e.,



estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  minimize the sum of the squares of the deviations of the observations from the true regression line:

$$L = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

## (Multiple) Linear Regression

In applications, a response  $y$  might depend on more than one factors. For example, the gasoline mileage performance ( $y$ ) of a vehicle depends on the vehicle weight ( $x_1$ ) and the engine displacement ( $x_2$ ). Then we need a little bit more complex linear regression:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

Or maybe even a “nonlinear” structure such as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon,$$

where  $\epsilon \sim N(0, \sigma^2)$ . If we let  $x_3 = x_1^2$ ,  $x_4 = x_2^2$ ,  $x_5 = x_1 x_2$ ,  $\beta_3 = \beta_{11}$ ,  $\beta_4 = \beta_{22}$ , and  $\beta_5 = \beta_{12}$ , then we can rewrite the nonlinear regression as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon,$$

where  $\epsilon \sim N(0, \sigma^2)$  (though it is a nonlinear structure, as long as it can be represented in a linear form, we view it as a linear regression).

# Linear Regression and Least Squares Estimate

The general linear regression takes the form

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$

Herein,  $Y$  is the response variable, and  $x_1, \dots, x_k$  are  $k$  predict variables (or regressor variables), and  $\epsilon$  is the random error which accounts for all the variation that cannot be explained by the linear structure.

The observations are

$$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i), i = 1, \dots, n \text{ and } n > k.$$

Each observation satisfies

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, i = 1, \dots, n. \quad (1)$$

Estimates  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  come from the minimization of the sum of square

$$L = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2.$$

# Linear Regression and Least Squares Estimate

If we write things in matrix form:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Then we have (1) written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

and the Least Squares estimate  $\hat{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta}$  has a closed from expression

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

## Fitted model and residuals

With the Least Squares estimates, the fitted regression model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

or

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}, i = 1, \dots, n.$$

And the residuals are

$$e_i = y_i - \hat{y}_i, i = 1, \dots, n,$$

the difference between observed values and the fitted values. These residuals represent the variation of the observed data that cannot be explained by the linear regression model. We collect these unexplained variation to a terms called the **error sum of squares**, which is the sum of squares of the residuals:

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

## Estimation of Variance

Our estimate of the unknown variance  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - p} = \frac{SS_E}{n - (k + 1)}.$$

Herein,  $p$  denotes the number of regression coefficients ( $\beta_0, \dots, \beta_k$ ) we have already estimated; i.e.,  $p = k + 1$ .

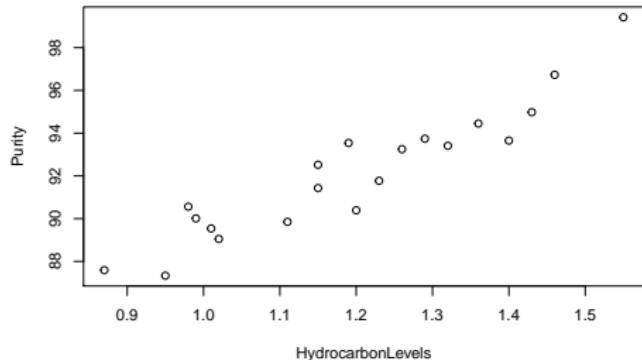
### Summary

By now, we have estimated the entire linear regression model which contains  $p = k + 1$  regression coefficients ( $\beta_0, \dots, \beta_k$ ) and the variance of the random error ( $\sigma^2$ ). We note that all the estimators are unbiased:  $E(\hat{\beta}_j) = \beta_j$  for  $j = 0, 1, \dots, k$ , and  $E(\hat{\sigma}^2) = \sigma^2$ .

## Calculation in R

Back to Example 1 on page 7,  $y$  is the purity of oxygen produced in a chemical distillation process, and  $x$  is the percentage of hydrocarbons present in the main condenser of the distillation unit. The data can be read and fitted as following.

```
> Example1=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
master/Data/HydrocarbonPurity.csv")
> head(Example1,2)
  HydrocarbonLevels Purity
1                  0.99  90.01
2                  1.02  89.05
> plot(Example1)
```



## Calculation in R

Back to Example 1 on page 7,  $y$  is the purity of oxygen produced in a chemical distillation process, and  $x$  is the percentage of hydrocarbons present in the main condenser of the distillation unit. The data can be read and fitted as following.

```
> x=Example1$HydrocarbonLevels # make sure letter cases are matched exactly!
> y=Example1$Purity
> fit=lm(y~x); fit
Call:
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x
74.28	14.95

```
> fit$fitted.values # compute the fitted values
> fit$residuals # compute all the residuals
> lm.est(fit)
      Estimate
beta0 74.283314
beta1 14.947480
sigma 1.086529
```

The estimated variance is  $\hat{\sigma}^2 = 1.0865^2$ . The estimated regression coefficients are  $\hat{\beta}_0 = 74.28$  and  $\hat{\beta}_1 = 14.95$ . The fitted model is

$$\hat{y} = 74.28 + 14.95x_1.$$

## Calculation in R (continued)

```
> summary(fit)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.83029	-0.73334	0.04497	0.69969	1.96809

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	74.283	1.593	46.62	< 2e-16 ***
x	14.947	1.317	11.35	1.23e-09 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.087 on 18 degrees of freedom

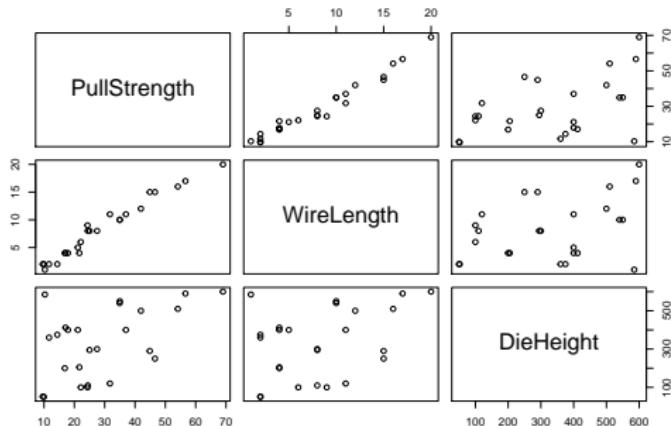
Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706

F-statistic: 128.9 on 1 and 18 DF, p-value: 1.227e-09

## Example 2

we used data on pull strength of a wire bond in a semiconductor manufacturing process, wire length, and die height to illustrate building an empirical model. The data can be loaded as

```
> Example2=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/  
master/Data/WireBond.csv")  
> head(Example2,2)  
PullStrength WireLength DieHeight  
1 9.95 2 50  
2 24.45 8 110  
  
> plot(Example2)
```



## Example 2

```
> y=Example2$PullStrength # make sure letter cases are matched exactly!
> x1=Example2$WireLength
> x2=Example2$DieHeight
> fit=lm(y~x1+x2)
> fit
Call:
lm(formula = y ~ x1 + x2)
```

Coefficients:

(Intercept)	x1	x2
2.26379	2.74427	0.01253

```
> fit$fitted.values # compute the fitted values
> fit$residuals # compute all the residuals
> lm.est(fit)
      Estimate
beta0 2.26379143
beta1 2.74426964
beta2 0.01252781
sigma 2.28804683
```

The estimated variance is  $\hat{\sigma}^2 = 2.288^2$ . The estimated regression coefficients are  $\hat{\beta}_0 = 2.26379$ ,  $\hat{\beta}_1 = 2.74427$ , and  $\hat{\beta}_2 = 0.01253$ . The fitted model is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2.$$

## Example 2 (continued)

```
> summary(fit)
```

Call:

```
lm(formula = y ~ x1 + x2)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.865	-1.542	-0.362	1.196	5.841

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.263791	1.060066	2.136	0.044099 *
x1	2.744270	0.093524	29.343	< 2e-16 ***
x2	0.012528	0.002798	4.477	0.000188 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.288 on 22 degrees of freedom

Multiple R-squared: 0.9811, Adjusted R-squared: 0.9794

F-statistic: 572.2 on 2 and 22 DF, p-value: < 2.2e-16

## Confidence Intervals in Multiple Linear Regression

Let  $\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1}$  which is a  $p \times p$  dimensional matrix. Denote by  $C_{jj}$  the  $j$ th diagonal entry of the matrix  $\mathbf{C}$ . Then the estimated standard error of  $\hat{\beta}_j$  is

$$se(\hat{\beta}_j) = \hat{\sigma} \sqrt{C_{jj}}.$$

More importantly, we have

$$T_0 = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{C_{jj}}} \sim t(n-p).$$

Thus a two-tailed  $100(1 - \alpha)\%$  CI of  $\beta_j$  can be derived from

$$-t_{n-p,\alpha/2} \leq T_0 = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \leq t_{n-p,\alpha/2}$$

which is

$$\left[ \hat{\beta}_j \pm t_{n-p,\alpha/2} se(\hat{\beta}_j) \right].$$

# Confidence Intervals in Multiple Linear Regression

A two-tailed  $100(1-\alpha)\%$  confidence interval of  $\beta_j$  is  $\left[ \hat{\beta}_j \pm t_{n-p,\alpha/2} se(\hat{\beta}_j) \right]$ .

The  $100(1 - \alpha)\%$  upper bound of  $\beta_j$  is  $\hat{\beta}_j + t_{n-p,\alpha} se(\hat{\beta}_j)$ .

The  $100(1 - \alpha)\%$  lower bound of  $\beta_j$  is  $\hat{\beta}_j - t_{n-p,\alpha} se(\hat{\beta}_j)$ .

## Example 2 continued

Find a  $95\%$  two-tailed confidence interval of  $\beta_1$ . We have already fitted a (multiple) linear regression model saved in "fit".

```
> lm.coef.CI(fit, level=0.95)
Two-sided 95 % confidence intervals of regression coefficients are
      CI.lb      CI.ub
beta0 0.065348613 4.46223426
beta1 2.550313061 2.93822623
beta2 0.006724246 0.01833138
One-sided 95 % (lower and upper) confidence bounds are
      lower.bound  upper.bound
beta0 0.443504657 4.0840782
beta1 2.583675700 2.9048636
beta2 0.007722522 0.0173331
```

Conclusion: based on the data, we are  $95\%$  confident that  $\beta_1$  is between 2.5503 and 2.9382.

## Hypothesis Tests in Multiple Linear Regression

We now consider test

$$H_0 : \beta_j = \beta_{j0} \text{ versus } H_1 : \beta_j \neq \beta_{j0}$$

at significance level  $\alpha$ . The test statistic is

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{se(\hat{\beta}_j)}.$$

Reject  $H_0$  if  $|T_0| > t_{n-p,\alpha/2}$ .

For one-tailed alternatives:

- ▶  $H_1 : \beta_j < \beta_{j0}$ , reject  $H_0$  if  $T_0 < -t_{n-p,\alpha}$ .
- ▶  $H_1 : \beta_j > \beta_{j0}$ , reject  $H_0$  if  $T_0 > t_{n-p,\alpha}$ .

We also have a confidence interval approach and a  $P$ -value approach. These three approaches can be done using StatEngin function `lm.coef.test`

## Example 2 (continued)

We now want to test  $H_0 : \beta_1 = 0$  versus  $H_1 : \beta_1 \neq 0$  at  $\alpha = 0.05$ .

```
> lm.coef.test(fit,alpha=0.05,H1="two")
   P_value t_alpha.2 T0.abs CI.lb hypo.beta CI.ub Reject
beta0  0.0441    2.0739  2.1355  0.0653      0  4.4622    Yes
beta1      0     2.0739  29.343  2.5503      0  2.9382    Yes
beta2  2e-04    2.0739  4.4767  0.0067      0  0.0183    Yes
```

This R program tests  $H_0 : \beta_j = 0$  versus  $H_1 : \beta_j \neq 0$  at  $\alpha = 0.05$  for all  $j$ 's separately.

Conclusion: at  $\alpha = 0.05$ , the data provide sufficient evidence to reject  $H_0 : \beta_1 = 0$ . It means that  $x_1$  is an important predictor variable for the response  $y$ .

For one-tailed alternatives:

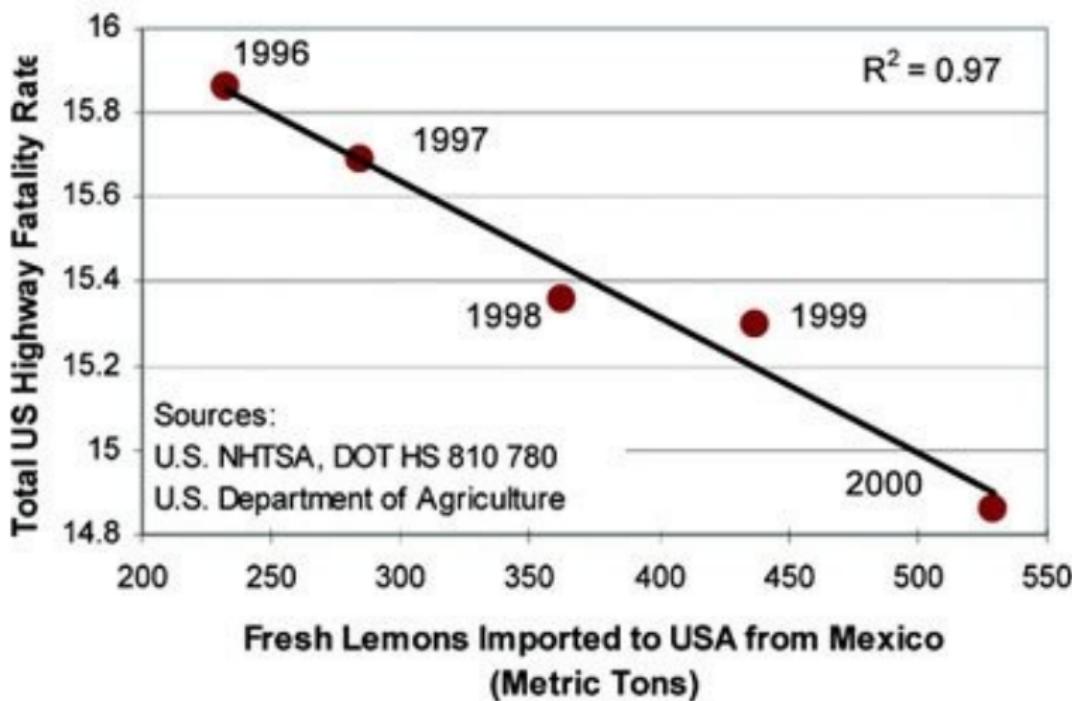
- ▶ `lm.coef.test(fit,alpha=0.05,H1="left")`  
tests  $H_0 : \beta_j = 0$  versus  $H_1 : \beta_j < 0$  at  $\alpha = 0.05$  for all  $j$ 's separately.
- ▶ `lm.coef.test(fit,alpha=0.05,H1="right")`  
tests  $H_0 : \beta_j = 0$  versus  $H_1 : \beta_j > 0$  at  $\alpha = 0.05$  for all  $j$ 's separately.

## Attention 1

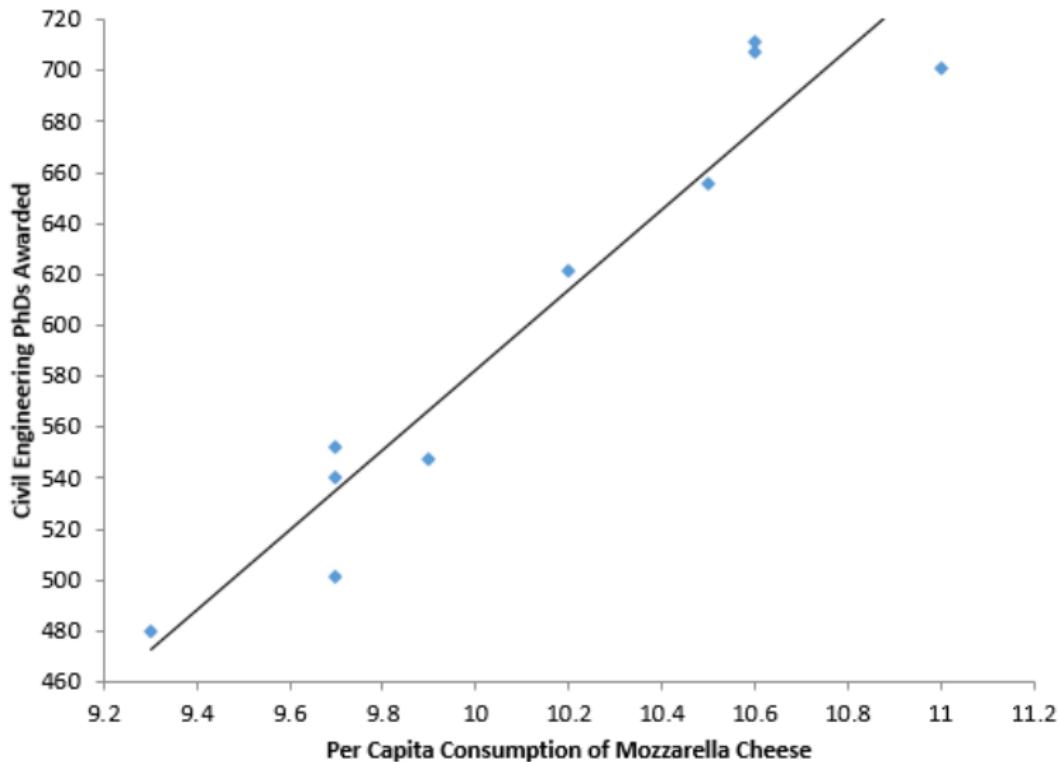
Abuses of Regression Regression is widely used and frequently misused; we mention several common abuses of regression briefly here. Care should be taken in selecting variables with which to construct regression equations and in determining the form of the model. It is possible to develop statistically significant relationships among variables that are completely unrelated in a **causal** sense.

For example, we might attempt to relate the shear strength of spot welds with the number of empty parking spaces in the visitor parking lot. A straight line may even appear to provide a good fit to the data, but the relationship is an unreasonable one on which to rely. We cannot increase the weld strength by blocking off parking spaces. A strong observed association between variables **does not necessarily imply** that a causal relationship exists between them. This type of effect is encountered fairly often in retrospective data analysis and even in **observational studies**. **Designed experiments** are the only way to determine cause-and-effect relationships.

## Attention 1



## Attention 1



## Example 2 (continued)

We now want to test  $H_0 : \beta_1 = 2$  versus  $H_1 : \beta_1 \neq 2$  at  $\alpha = 0.05$ .

```
> lm.coef.test(fit,alpha=0.05,H1="two",hypo.beta=c(0,2,0))
   P_value t_alpha 2 T0.abs CI.lb hypo.beta CI.ub Reject
beta0  0.0441    2.0739 2.1355 0.0653          0 4.4622    Yes
beta1      0     2.0739 7.9581 2.5503          2 2.9382    Yes
beta2  2e-04    2.0739 4.4767 0.0067          0 0.0183    Yes
```

Conclusion: at  $\alpha = 0.05$  the data provide sufficient evidence to reject  $H_0 : \beta_1 = 2$ .

We now want to test  $H_0 : \beta_0 = 1$  versus  $H_1 : \beta_0 < 1$  at  $\alpha = 0.05$ .

```
lm.coef.test(fit,alpha=0.05,H1="left",hypo.beta=c(1,0,0))
   P_value      T0 t_alpha CI.lb hypo.beta CI.ub Reject
beta0  0.8771 1.1922 -1.7171 -Inf          1 0.4435    No
beta1      1 29.343 -1.7171 -Inf          0 2.5837    No
beta2  0.9999 4.4767 -1.7171 -Inf          0 0.0077    No
```

Conclusion: at  $\alpha = 0.05$  data do not provide sufficient evidence to reject  $H_0 : \beta_0 = 1$ .

Test  $H_0 : \beta_2 = 0.02$  versus  $H_1 : \beta_2 > 0.02$  at  $\alpha = 0.05$ .

```
lm.coef.test(fit,alpha=0.05,H1="right",hypo.beta=c(0,0,0.02))
   P_value t_alpha      T0 CI.lb hypo.beta CI.ub Reject
beta0  0.022  1.7171  2.1355 0.4435          0 Inf    Yes
beta1      0  1.7171  29.343 2.5837          0 Inf    Yes
beta2  0.993  1.7171 -2.6701 0.0077         0.02 Inf    No
```

At  $\alpha = 0.05$  data do not provide sufficient evidence to reject  $H_0 : \beta_2 = 0.02$ .

## Prediction of New Observations

Now we have a fitted model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k.$$

Suppose we have a new

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ \vdots \\ x_{0k} \end{bmatrix},$$

we want to predict the future observation of  $Y$  at  $\mathbf{x} = \mathbf{x}_0$ .

We know that at this  $\mathbf{x} = \mathbf{x}_0$ , the response  $Y$  has a distribution as

$$Y \sim N(\mu_{Y|\mathbf{x}_0}, \sigma^2)$$

where  $\mu_{Y|\mathbf{x}_0} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k}$ .

## Prediction of New Observations

We have two types of prediction.

- ▶ Prediction of  $\mu_{Y|x_0}$ : we predict the averaged  $Y$ -value at  $x = x_0$ . For this goal, we have a statistical inference called **Confidence Interval on the Mean Response**.
- ▶ Prediction of  $Y$  at  $x = x_0$ : we predict a single observation of  $Y$  at  $x = x_0$ . For this goal, the statistical inference is **Prediction Interval on a Single Future Response**.

Both of them are built on the fitted value  $\hat{y}_0 = \mathbf{x}_0^\top \hat{\beta}$ , but account for different errors.

## Prediction of New Observations

We have two types of prediction, where  $\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1}$ :

- ▶ Confidence Interval on the Mean Response ( $\mu_{Y|\mathbf{x}_0}$ ) at  $\mathbf{x} = \mathbf{x}_0$ :

$$\hat{y}_0 - t_{n-p,\alpha/2} \hat{\sigma} \sqrt{\mathbf{x}_0^\top \mathbf{C} \mathbf{x}_0} \leq \mu_{Y|\mathbf{x}_0} \leq \hat{y}_0 + t_{n-p,\alpha/2} \hat{\sigma} \sqrt{\mathbf{x}_0^\top \mathbf{C} \mathbf{x}_0}$$

- ▶ Prediction Interval on a Single Future Response ( $Y_0$ ) at  $\mathbf{x} = \mathbf{x}_0$ :

$$\hat{y}_0 - t_{n-p,\alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{x}_0^\top \mathbf{C} \mathbf{x}_0} \leq Y_0 \leq \hat{y}_0 + t_{n-p,\alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{x}_0^\top \mathbf{C} \mathbf{x}_0}$$

The prediction interval is always wider than the confidence interval. The confidence interval expresses the error in estimating the mean of a distribution, and the prediction interval expresses the error in predicting a future observation from the distribution at the point  $\mathbf{x}_0$ . This must include the error in estimating the mean at that point as well as the inherent variability in the random variable  $Y$  at the same value  $\mathbf{x} = \mathbf{x}_0$ .

## Example 2 (continued)

Suppose that the engineer wishes to construct a 95% CI on the mean pull strength for a wire bond with wire length  $x_1 = 8$  and the die height is  $x_2 = 275$ , and a 95% prediction interval on the wire bond pull strength when the wire length is  $x_1 = 8$  and the die height is  $x_2 = 275$ .

**Solution:** continue to use "fit".

```
> fit
Call:
lm(formula = y ~ x1 + x2)
> predict.lm(fit,new=data.frame(x1=8,x2=275),interval="confidence")
      fit      lwr      upr
1 27.6631 26.66324 28.66296

> predict.lm(fit,new=data.frame(x1=8,x2=275),interval="prediction")
      fit      lwr      upr
1 27.6631 22.81378 32.51241
```

Based on the data, we are 95% confidence that the mean pull strength at  $x_1 = 8$  and  $x_2 = 275$  is between 26.6632 and 28.663, and the wire bond pull strength at  $x_1 = 8$  and  $x_2 = 275$  is between 22.8138 and 32.5124.

## Attention 2

Regression relationships are valid for values of the regressor variable only within the range of the original data. The linear relationship that we have tentatively assumed may be valid over the original range of  $x$ , but it may be unlikely to remain so as we extrapolate—that is, if we use values of  $x$  beyond that range. In other words, as we move beyond the range for which data were collected, we become less certain about the validity of the assumed model. Regression models are not necessarily valid for extrapolation purposes.

Now this does not mean do not ever extrapolate. For many problem situations in science and engineering, extrapolation of a regression model is the only way to even approach the problem. However, there is a strong warning to be careful. A modest extrapolation may be perfectly all right in many cases, but **a large extrapolation will almost never produce acceptable results.**

## Test for Significance of Regression

Testing  $H_0 : \beta_j = 0$  versus  $H_1 : \beta_j \neq 0$  for some  $j$  is already one type of significance tests. It investigates whether the predictor  $x_j$  is significant to the response  $y$ . If  $H_0$  is rejected, then it is; otherwise,  $x_j$  might not have a significant role to explain  $y$ .

Another significance test is (ANOVA test):

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$  v.s.  $H_1 : \beta_j \neq 0$  for at least one  $j$ .

Rejection of  $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$  implies that at least one of the regressor variables  $x_1, \dots, x_k$  contributes significantly to the model,

- ▶ but we do not know which one(s).
- ▶ This significance test is different than testing  $H_0 : \beta_j = 0$  versus  $H_1 : \beta_j \neq 0$  for all  $j$ 's separately and cannot be done using `lm.coef.test`

## Test for Significance of Regression

Test statistic for ANOVA is

$$F_0 = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E}$$

- ▶  $SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$ , the regression sum of squares.
- ▶  $SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ , the error sum of squares
- ▶  $MS_R$  and  $MS_E$  are called mean squares.

Let  $SS_T = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ , the total corrected sum of squares. We have

$$SS_T = SS_R + SS_E.$$

At significance level  $\alpha$ , we reject  $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$  if  $F_0 > f_{r,n-p,\alpha}$  (always use  $P$ -value for this one).

Note: Rejection of  $H_0$  does not necessarily imply that the linear relationship found is an appropriate model for predicting  $y$  as a function of  $x_1, \dots, x_k$ . Further tests of model adequacy are required before we can be comfortable using this model in practice.

## Example 2 (continued)

We test  $H_0 : \beta_1 = \beta_2 = 0$  versus  $H_1 : \beta_j \neq 0$  for at least one  $j$  at  $\alpha = 0.05$ .

**Solution:** Back to `summary(fit)`. Look for

F-statistic: 572.2 on 2 and 22 DF, p-value: < 2.2e-16

This result tells that  $F_0 = 572.2$ ,  $r = 2$ ,  $n - p = 22$ , and the  $P$ -value for the ANOVA test is less than  $2.2 \times 10^{-16}$  (thus less than  $\alpha$ , thus reject  $H_0$ ).

**Conclusion:** at  $\alpha = 0.05$ , the data provide sufficient evidence to reject  $H_0$ .

*Practical Interpretation:* Rejection of  $H_0$  does not necessarily imply that the relationship found is an appropriate model for predicting pull strength as a function of wire length and die height. Further tests of model adequacy are required before we can be comfortable using this model in practice.

## $R^2$ and Adjusted $R^2$

We may also use the coefficient of multiple determination  $R^2$  as a global statistic to assess the fit of the model. Computationally,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

In Example 2, back to `summary(fit)`. Look for

Multiple R-squared: 0.9811, Adjusted R-squared: 0.9794

This tells us  $R^2 = 0.9811$ .

**Interpretation:** the model accounts for about 98.11% of the variability in the pull strength response.

**Drawback:** The  $R^2$  statistic is somewhat problematic as a measure of the quality of the fit for a multiple regression model because it never decreases when a variable is added to a model. Solely based on  $R^2$  to select predictors could cause **overfitting**. A better statistic to use is the Adjusted  $R^2$ .

```
> set.seed(100)
```

```
> summary(lm(y~x1+x2+rnorm(25)))
```

Multiple R-squared: 0.9813, Adjusted R-squared: 0.9786

## Adjusted $R^2$

$$R_{adj}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}.$$

Because  $SS_E/(np)$  is the error or residual mean square and  $SS_T/(n1)$  is a constant,  $R_{adj}^2$  will only increase when a variable is added to the model if the new variable reduces the error mean square.

In Example 2,

```
> summary(fit)
Multiple R-squared:  0.9811,  Adjusted R-squared:  0.9794
```

If we only include one predictor:

```
> summary(lm(y~x1))
Multiple R-squared:  0.964,  Adjusted R-squared:  0.9624
> summary(lm(y~x2))
Multiple R-squared:  0.2429,  Adjusted R-squared:    0.21
```

Therefore, we would conclude that adding  $x_2$  or  $x_1$  to the model does result in a meaningful reduction in unexplained variability in the response.

## Partial F-test

We would like to determine whether a subset of regressor variables, say without loss of generality  $x_1, x_2, \dots, x_r$  ( $r \leq k$ ), as a whole contributes significantly to the regression model. This leads to test

$$\begin{aligned} H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0 &\text{ versus} \\ H_1 : \text{at least one of } \beta_1, \dots, \beta_r &\text{ is nonzero} \end{aligned}$$

We call this as a general regression test or a partial F-test. The ANOVE test is a special case when  $r = k$ .

How to do? The test statistic is also a F-statistic, call it  $F_0$  again. Then we reject  $H_0$  at  $\alpha$  if  $F_0 > f_{r,n-p,\alpha}$ . There is also a P-value approach.

StatEngine: First fit a linear regression model using all the rest predictors not involved in  $H_0$ , call the fit as lmH0. Then fit a full linear regression using all the predictors. Call the fit as lmALL.

```
lm.partialFtest(fit.H0=lmH0,fit.ALL=lmALL,alpha=?)
```

## Example 2 (continued)

Consider the wire bond pull-strength data. We investigate the contribution of two new variables,  $x_3 = x_1^2$  and  $x_4 = x_2^2$ , to the model using the partial F-test approach. That is we wish to test, at  $\alpha = 0.05$ ,

$$H_0 : \beta_3 = \beta_4 = 0 \quad H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0.$$

```
> x3=x1^2
> x4=x2^2
> lmH0=lm(y~x1+x2)
> lmALL=lm(y~x1+x2+x3+x4)
> lm.partialFtest(fit.H0=lmH0,fit.ALL=lmALL,alpha=0.05)
      F0  f_alpha  P_value Reject
PartialFtest 4.047007 3.492828 0.033432    Yes
```

Conclusion: we see the  $P$ -value is less than 0.05, thus conclude that at significance level  $\alpha = 0.05$ , the data provide sufficient evidence to reject  $H_0$ ; i.e., at least one of the new variables contributes significantly to the model. Further analysis and tests will be needed to refine the model and determine whether one or both of  $x_3$  and  $x_4$  are important.

## Example 2 (continued)

```
> summary(lm(y~x1+x2+x3))
Multiple R-squared:  0.9864, Adjusted R-squared:  0.9845
> summary(lm(y~x1+x2+x4))
Multiple R-squared:  0.9815, Adjusted R-squared:  0.9788
> summary(lm(y~x1+x2+x3+x4))
Multiple R-squared:  0.9866, Adjusted R-squared:  0.9839
```

Based on the Adjusted  $R^2$ , we see that the best choice is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

or equivalently,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \epsilon \text{ (Polynomial Regression)}$$

## Polynomial Regression

When data present a non-linear pattern, we could still use linear regression by adding high-degree polynomial term. For example, the second-degree polynomial regression in one variable is

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon,$$

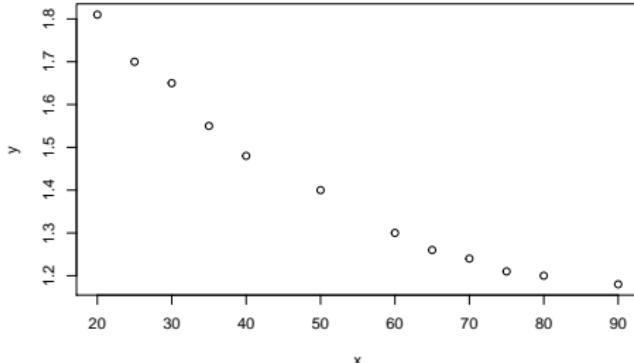
in two variables is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon.$$

## Example 3: Airplane Sidewall Panels

Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit manufacturing cost varies with the production lot size. The following data give the average cost per unit (in hundreds of dollars) for this product ( $y$ ) and the production lot size ( $x$ ).

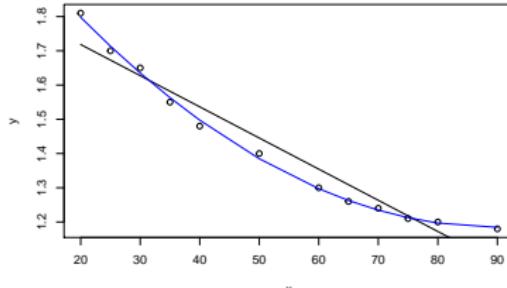
```
> Example3=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/  
master/Data/AirplaneSidewallPanels.csv")  
> head(Example3,2)  
> y=Example3$cost  
> x=Example3$lotsize  
> plot(x,y)
```



## Example 3: Airplane Sidewall Panels (continued)

We first fit  $Y = \beta_0 + \beta_1 x + \epsilon$ . Result is the black line. Then we fit a second-order polynomial regression:  $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ . Check the blue curve.

```
> plot(x,y)
> fit=lm(y~x)
> summary(fit)
Multiple R-squared:  0.9386, Adjusted R-squared:  0.9324
> lines(x,fit$fitted.values)
> x2=x^2
> fit.quad=lm(y~x+x2)
> summary(fit.quad)
Multiple R-squared:  0.9975, Adjusted R-squared:  0.9969
> lines(x,fit.quad$fitted.values,col="blue")
```



## Categorical Regressors and Indicator Variables

To handle categorical predictor variables, we use indicator variables. For example, to introduce the effect of two different operators into a regression model, we could define an indicator (or dummy) variable as follows:

$$x = \begin{cases} 0 & \text{if the observation is from operator 1} \\ 1 & \text{if the observation is from operator 2} \end{cases}$$

Or more than two categories:

$x_1$	$x_2$	
0	0	if the observation is from operator 1
1	0	if the observation is from operator 2
0	1	if the observation is from operator 3

## Example 4: Surface Finish

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown as

observation $i$	Surface Finish $y_i$	RPM $x_{i1}$	Type of Cutting Tool
1	45.44	225	302
11	33.50	224	416

$x_1$  is the lathe speed in revolutions per minute. We use

$$x_2 = \begin{cases} 0 & \text{for tool type 302} \\ 1 & \text{for tool type 416} \end{cases}$$

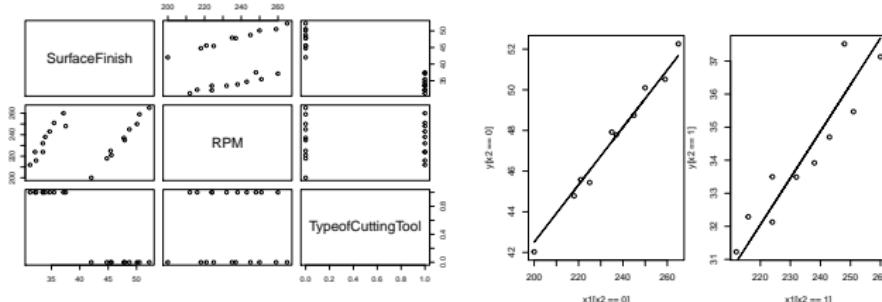
Then build

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{for tool type 302} \\ \beta_0 + \beta_1 x_1 + \beta_2 + \epsilon & \text{for tool type 416} \end{cases},$$

two straight lines with different intercepts to account for the different tool types.

## Example 4: Surface Finish

```
> Example4=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/  
master/Data/SurfaceFinishData.csv")  
> head(Example4,2)  
> plot(Example4)  
> y=Example4$SurfaceFinish  
> x1=Example4$RPM  
> x2=Example4>TypeofCuttingTool  
> fit=lm(y~x1+x2)  
> summary(fit)  
Multiple R-squared:  0.9924, Adjusted R-squared:  0.9915  
> par(mfrow=c(1,2))  
> plot(x1[x2==0],y[x2==0])  
> lines(x1[x2==0],fit$fitted.values[x2==0])  
> plot(x1[x2==1],y[x2==1])  
> lines(x1[x2==1],fit$fitted.values[x2==1])
```



## Example 4: Surface Finish

It is also possible to use indicator variables to investigate whether tool type affects both the slope and intercept. Let the model be  
Then build

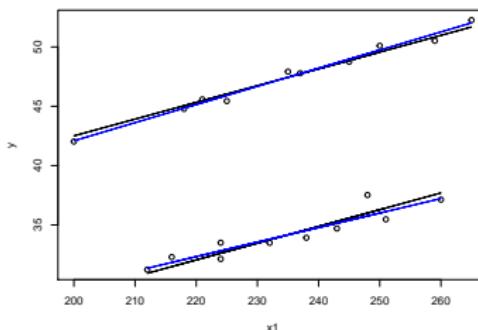
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \epsilon$$
$$= \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{for tool type 302} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon & \text{for tool type 416} \end{cases}$$

two straight lines with different intercepts and slopes.

## Example 4: Surface Finish

New model is in blue while the previous one is in black.

```
> x3=x1*x2  
> fit2=lm(y~x1+x2+x3)  
> summary(fit2)  
Multiple R-squared:  0.9936, Adjusted R-squared:  0.9924  
> par(mfrow=c(1,1))  
> plot(x1,y)  
> lines(x1[x2==0],fit$fitted.values[x2==0])  
> lines(x1[x2==1],fit$fitted.values[x2==1])  
> lines(x1[x2==0],fit2$fitted.values[x2==0],col="blue")  
> lines(x1[x2==1],fit2$fitted.values[x2==1],col="blue")
```



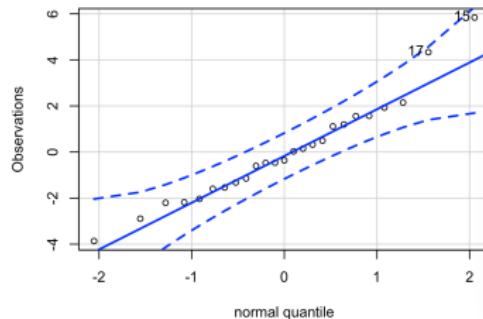
The Adjusted  $R^2$  indicates the new (more complex) model is slightly better.

# Model Adequacy Checking

The residuals from the multiple regression model, defined by  $e_i = y_i - \hat{y}_i$ , play an important role in judging model adequacy.

One model assumption is that  $\epsilon \sim N(0, \sigma^2)$ . Thus we can check the normality on  $e_i$ 's by using QQ-plot. Back to Example 2, we have

```
> Example2=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/  
master/Data/WireBond.csv")  
> y=Example2$PullStrength;x1=Example2$WireLength;x2=Example2$DieHeight  
> fit=lm(y~x1+x2)  
> data.summary(fit$residuals)
```



We see the normality seems to be satisfied thought we have two points near the boundary.

# Model Adequacy Checking

## The standardized residuals

$$d_i = \frac{e_i}{\hat{\sigma}}, \text{ for } i = 1, \dots, n,$$

are often more useful than the ordinary residuals when assessing residual magnitude. If a  $|d_i| > 3$ , then we can think the  $i$ th observation might be an outlier.

It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to be alert to the possibility of **influential observations** and to take them into consideration when interpreting the results. Cook's Distance is a good measure of the influence of an observation (Using `cook.distance`). Let  $D_i$  be the Cook's Distance of the  $i$ th observation, if  $D_i > 1$ , then the  $i$ th observation might be an influential observation.

# Model Adequacy Checking

To obtain our least squares estimates  $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ , it requires the matrix  $\mathbf{X}^\top \mathbf{X}$  is invertible. In many times, the predictors could present a very strong **multicollinearity** which makes  $\mathbf{X}^\top \mathbf{X}$  nearly singular and leads to unreliable inferences.

To check whether strong **multicollinearity** exists, we often use **Variance Inflation Factor (VIF)**, which can be done using `vif` function.

## Example 2

```
> fit=lm(y~x1+x2)
> vif(fit)
      x1      x2
1.167128 1.167128
```

For each predictor, we have a VIF, denoted by  $VIF(\hat{\beta}_i)$  for  $i = 1, \dots, k$ . A rule of thumb is that if  $VIF(\hat{\beta}_i) > 10$  then multicollinearity is high (a cutoff of 5 is also commonly used). In Example 2, the VIFs are all less than 5, thus we do not need to worry about multicollinearity.

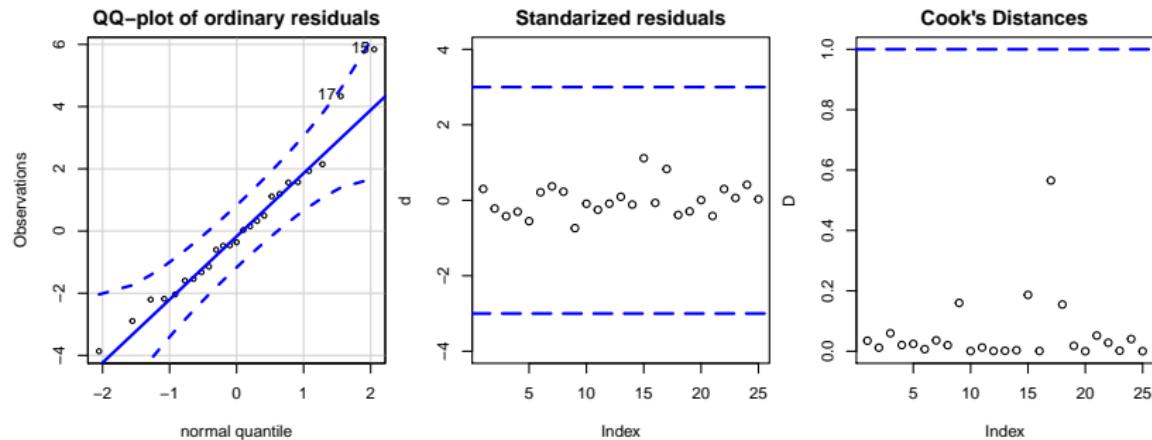
## Example 2 (continued)

StatEngine has the function `lm.modelcheck` to compute the QQ-plot of ordinary residuals ( $e_i$ 's), the standardized residuals ( $d_i$ 's), and the cook's distances ( $D_i$ 's)

```
> lm.modelcheck(fit)
```

VIFs are:

x1	x2
1.167128	1.167128



# Summary

First step, name  $y$ ,  $x_1, \dots, x_k$  in R. Then fit the linear regression using `lm` function. Below is a summary of related R functions.

```
summary() # Summary of results (ANOVA test, R-square, and Adjusted R-square)
lm.est() # all the estimates
lm.coef.CI() # Confidence intervals on the regression coefficients
lm.coef.test() # Hypothesis testing on the regression coefficients
predict.lm() # Confidence interval and prediction interval at a new x
lm.partialFtest() # Partial F-test
lm.modelcheck() # Some residual analysis to check model adequacy
```

# STAT 509: Statistics for Engineers

Chapters 13-14: Single-/Two-Factor Factorial Design

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

Fall 2020

## Chapters 13-14: Single-/Two-Factor Factorial Design

### Learning Objectives:

1. Design and conduct engineering experiments involving one factor or two factors with an arbitrary number of levels
2. Understand how the analysis of variance is used to analyze the data from these experiments
3. Use multiple comparison procedures to identify specific differences between means

## Introduction

Experiments are a natural part of the engineering and scientific decision-making process. Suppose, for example, that a civil engineer is investigating the effects of different curing methods on the mean compressive strength of concrete. The experiment would consist of making up several test specimens of concrete using each of the proposed curing methods and then testing the compressive strength of each specimen. The data from this experiment could be used to determine which curing method should be used to provide maximum mean compressive strength.

If there are only two curing methods of interest, this experiment could be designed and analyzed using the statistical hypothesis methods for two samples introduced in Chapter 10. That is, the experimenter has a single factor of interest—curing methods—and there are only two levels of the factor. The t-test can be used to decide if the two means differ.

## Introduction

Many single-factor experiments require that more than two levels of the factor be considered. For example, the civil engineer may want to investigate five different curing methods. In this chapter, we show how the **analysis of variance** (frequently abbreviated **ANOVA**) can be used for comparing means when there are more than two levels of a single factor. We also discuss randomization of the experimental runs and the important role this concept plays in the overall experimentation strategy.

## Example: Tensile Strength

A manufacturer of paper is interested in improving the product's tensile strength. Product engineering believes that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level by using a pilot plant. All 24 specimens are tested on a laboratory tensile tester in random order. The data from this experiment are

Hardwood Concentration (%)	Observations						Totals	Averages
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

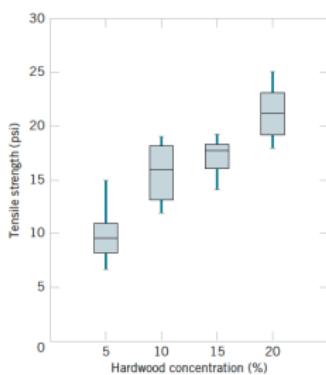
## Example: Tensile Strength

This is an example of a completely randomized **single-factor** experiment with **four** levels of the factor. The levels of the factor are sometimes called **treatments**, and each treatment has six observations or replicates. **The role of randomization in this experiment is extremely important.** By randomizing the order of the 24 runs, the effect of any nuisance variable that may influence the observed tensile strength is approximately balanced out.

For example, suppose that there is a warm-up effect on the tensile testing machine; that is, the longer the machine is on, the greater the observed tensile strength. If all **24** runs are made in order of increasing hardwood concentration (that is, all six 5% concentration specimens are tested first, followed by all six 10% concentration specimens, etc.), any observed differences in tensile strength could also be due to the warm-up effect.

## Example: Tensile Strength

If possible, the first step is always to graphically analyze the data from a designed experiment. The figure below presents box plots of tensile strength at the four hardwood concentration levels. This figure indicates that changing the hardwood concentration has an effect on tensile strength; specifically, higher hardwood concentrations produce higher observed tensile strength. Furthermore, the distribution of tensile strength at a particular hardwood level is reasonably symmetric, and the variability in tensile strength does not change dramatically as the hardwood concentration changes.



Graphical interpretation of the data is always useful. Box plots show the variability of the observations **within** a treatment (factor level) and the variability **between** treatments. We now discuss how the data from a single-factor randomized experiment can be analyzed statistically.

## One-Factor Factorial Experiments

Suppose that we have  $a$  different levels (or treatments) of a single factor that we wish to compare. Say  $Y_{ij}$  represents the  $j$ th observation taken under treatment  $i$  for  $j = 1, \dots, n$  and  $i = 1, \dots, a$ . We describe the observations by the linear statistical model

$$Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij},$$

$j = 1, \dots, n$  and  $i = 1, \dots, a$ , where

- ▶  $\mu$  is a parameter common to all treatments called the **overall mean**
- ▶  $\tau_i$  is a parameter associated with the  $i$ th treatment called the  **$i$ th treatment effect** and satisfies a constraint that

$$\sum_{i=1}^a \tau_i = 0,$$

- ▶  $\epsilon_{ij} \sim N(0, \sigma^2)$  is a random error component.
- , The unknown parameters are  $\mu_i = \mu + \tau_i$  for  $i = 1, \dots, a$  and  $\sigma^2$ .

## Analysis of Variance

We are interested in testing the equality of the  $a$  treatment means  $\mu_1, \mu_2, \dots, \mu_a$ ; i.e.,  $H_0 : \mu_1 = \mu_2 = \dots = \mu_a$  versus  $H_1$ : at least two of  $\mu_i$ 's are unequal. This is equivalent to testing the hypotheses

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0 \text{ versus } H_1 : \tau_i \neq 0 \text{ for at least one } i.$$

if the null hypothesis is true, each observation consists of the overall mean  $\mu$  plus a realization of the random error component  $\epsilon_{ij}$ . This is equivalent to saying that all  $N$  observations are taken from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Therefore, if the null hypothesis is true, changing the levels of the factor has no effect on the mean response.

We use an analysis of variance to conduct the test.

## Analysis of Variance

Let

$$y_{i\cdot} = \sum_{j=1}^n y_{ij}, \quad \bar{y}_{i\cdot} = y_{i\cdot}/n, \quad i = 1, 2, \dots, a$$

$$y_{\cdot\cdot} = \sum_{i=1}^a \sum_{j=1}^n y_{ij}, \quad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N$$

where  $N = an$  is the total number of observations and the "dot" subscript notation implies summation over the subscript that it replaces.

The total variability in the entire data is described by the **total sum of squares**

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{\cdot\cdot})^2.$$

Then we partition the total variability into two component parts.

# Analysis of Variance

## ANOVA Sum of Squares Identity

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2$$

or symbolically

$$SS_T = SS_{treatments} + SS_E$$

and degrees of freedom can be partitioned as

$$an - 1 = a - 1 + a(n - 1)$$

or

$$df_{Total} = df_{Treatments} + df_{Error}.$$

The  $SS_{treatments}$  is known as the **treatment sum of squares**. The  $SS_E$  is the **error sum of squares**.

# Analysis of Variance

The ratio

$$MS_{Treatments} = \frac{SS_{Treatments}}{a - 1}$$

is called the **mean square for treatments** and the **mean square for error** is

$$MS_{Error} = \frac{SS_{Error}}{a(n - 1)}.$$

We have

$$E[MS_{Treatments}] = \sigma^2 + \underbrace{\frac{n \sum_{i=1}^a \tau_i^2}{a - 1}}_{\geq 0},$$

$$E[MS_{Error}] = \sigma^2.$$

Only if  $H_0 : \tau_1 = \dots = \tau_a = 0$  is true,  $E[MS_{Treatments}] = E[MS_{Error}]$ ;  
otherwise,  $E[MS_{Treatments}] > E[MS_{Error}]$ .

# Analysis of Variance

To test

$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$  versus  $H_1 : \tau_i \neq 0$  for at least one  $i$ ,

our test statistic is

$$F_0 = \frac{MS_{Treatments}}{MS_{Error}}.$$

Reject  $H_0$  at significance level  $\alpha$  if

$$F_0 > f_{a-1, a(n-1), \text{alpha}},$$

or if  $P$ -value is less than  $\alpha$ .

TABLE 13.3 Analysis of Variance for a Single-Factor Experiment, Fixed-Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	$F_0$
Treatments	$SS_{Treatments}$	$a - 1$	$MS_{Treatments} = \frac{SS_{Treatments}}{a - 1}$	$\frac{MS_{Treatments}}{MS_E}$
Error	$SS_E$	$a(n - 1)$	$MS_{Error} = \frac{SS_{Error}}{a(n - 1)}$	
Total	$SS_T$	$an - 1$		

## A Multiple Comparison Following the ANOVA

When the null hypothesis  $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$  is rejected in the ANOVA, we know that some of the treatment or factor-level means are different. However, the ANOVA does not identify which means are different. Methods for investigating this issue are called **multiple comparisons methods**. Many of these procedures are available. R package provide a simple function TukeyHSD to identify which pairs of treatments have different means.

## Example: Tensile Strength

Consider the paper tensile strength experiment described previously. This experiment is a completely randomized design.

- (a) Test the hypothesis that different hardwood concentrations do not affect the mean tensile strength of the paper.
- (b) If different hardwood concentrations affects the mean tensile strength of the paper, identify which pairs of treatments have different means.

**Solution:** The four treatments are labeled by the concentration level. Let  $\tau_1, \dots, \tau_4$  denote the treatment effects. We wish to test

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_4 = 0 \text{ versus } H_1 : \tau_i \neq 0 \text{ for at least one } i,$$

## Example: Tensile Strength

**Solution (continued):** We read the data

```
> my_data=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
                     master/Data/TensileStrength.csv")
> head(my_data,2)
  Concentration Strength
1                  5       7
2                  5       8
# We need to tell R Concentration levels are the factors
> my_data$Concentration=as.factor(my_data$Concentration)

# Run ANOVA test
> res.aov=aov(Strength~Concentration,data=my_data)
> summary(res.aov)
    Df Sum Sq Mean Sq F value    Pr(>F)
Concentration   3  382.8  127.60   19.61 3.59e-06 ***
Residuals      20  130.2    6.51
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We see the  $P$ -value is significantly smaller than any commonly used  $\alpha$ . Thus, we reject  $H_0$ .

## Example: Tensile Strength

**Solution (continued):** Now Part (b):

```
> TukeyHSD(res.aov, conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Strength ~ Concentration, data = my_data)
```

\$Concentration	diff	lwr	upr	p adj
10-5	5.666667	1.54410408	9.789229	0.0051108
15-5	7.000000	2.87743741	11.122563	0.0006501
20-5	11.166667	7.04410408	15.289229	0.0000015
15-10	1.333333	-2.78922925	5.455896	0.8022275
20-10	5.500000	1.37743741	9.622563	0.0065966
20-15	4.166667	0.04410408	8.289229	0.0470251

For each pair, it gives a 95% confidence interval of  $\mu_i - \mu_j$  by considering multiple comparison. All we need to focus on is the adjusted  $P$ -value. Only the pair between the 10% and 15% concentration treatments does not have a significance mean difference.

## Two-Factor Factorial Experiments

In practice, the experiment may involve two factors, say  $A$  and  $B$ . There are  $a$  levels of factor  $A$  and  $b$  levels of factor  $B$  (shown as)

		Factor $B$				
		1	2	...	$b$	Totals
<b>Factor A</b>	1	$y_{111}, y_{112}, \dots, y_{11n}$	$y_{121}, y_{122}, \dots, y_{12n}$		$y_{1b1}, y_{1b2}, \dots, y_{1bn}$	$y_{1..}$
	2	$y_{211}, y_{212}, \dots, y_{21n}$	$y_{221}, y_{222}, \dots, y_{22n}$		$y_{2b1}, y_{2b2}, \dots, y_{2bn}$	$y_{2..}$
	:					
	$a$	$y_{a11}, y_{a12}, \dots, y_{a1n}$	$y_{a21}, y_{a22}, \dots, y_{a2n}$		$y_{ab1}, y_{ab2}, \dots, y_{abn}$	$y_{a..}$
Totals		$y_{..1}$	$y_{..2}$		$y_{..b}$	$y_{...}$
Averages		$\bar{y}_{..1}$	$\bar{y}_{..2}$		$\bar{y}_{..b}$	$\bar{y}_{...}$

The experiment has  $n$  replicates, and each replicate contains all  $ab$  treatment combinations. The observation in the  $ij$ -th cell for the  $k$ th replicate is denoted by  $y_{ijk}$ . In performing the experiment, the  $abn$  observations would be run in random order. Thus, like the single-factor experiment, the two-factor factorial is a *completely randomized design*.

## Two-Factor Factorial Experiments

The observations may be described by the linear statistical model

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$$

for  $i = 1, 2, \dots, a$ ,  $j = 1, 2, \dots, b$ , and  $k = 1, 2, \dots, n$ , where

- ▶  $\mu$  is the overall mean effect
- ▶  $\tau_i$  is the effect of the  $i$ th level of factor  $A$
- ▶  $\beta_j$  is the effect of the  $j$ th level of factor  $B$
- ▶  $(\tau\beta)_{ij}$  is the **effect of the interaction** between  $A$  and  $B$
- ▶  $\epsilon_{ijk} \sim N(0, \sigma^2)$  is a random error component.

We have a constraint:

$$\sum_{i=1}^a \tau_i = 0, \sum_{j=1}^b \beta_j = 0, \sum_{i=1}^a (\tau\beta)_{ij} = 0, \sum_{j=1}^b (\tau\beta)_{ij} = 0$$

We are interested in testing the hypotheses of no main effect for factor  $A$ , no main effect for  $B$ , and no  $AB$  interaction effect.

## Two-Factor Factorial Experiments

The hypotheses that we test are as follows:

1. No main effect of factor  $A$ :

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0 \text{ versus } H_1 : \text{at least one } \tau_i \neq 0.$$

2. No main effect of factor  $B$ :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0 \text{ versus } H_1 : \text{at least one } \beta_j \neq 0.$$

3. No interaction:

$$H_0 : (\tau\beta)_{11} = (\tau\beta)_{12} = \cdots = (\tau\beta)_{ab} = 0 \text{ versus}$$
$$H_1 : \text{at least one } (\tau\beta)_{ij} \neq 0.$$

# Analysis of Variance

The tests can be done via analysis of variance.

$$\begin{aligned}SS_T &(\text{total sum of squares}) \\&= SS_A (\text{sum of squares for factor } A) \\&\quad + SS_B (\text{sum of squares for factor } B) \\&\quad + SS_{AB} (\text{sum of squares for the interaction between } A \text{ and } B) \\&\quad + SS_E (\text{error sum of squares}).\end{aligned}$$

The degrees of freedom are

$$\underbrace{abn - 1}_{df_{Total}} = \underbrace{(a - 1)}_{df_A} + \underbrace{(b - 1)}_{df_B} + \underbrace{(a - 1)(b - 1)}_{df_{AB}} + \underbrace{ab(n - 1)}_{df_{Error}}.$$

# Analysis of Variance

The mean squares are

$$MS_A = \frac{SS_A}{a - 1}, \quad MS_B = \frac{SS_B}{b - 1},$$

$$MS_{AB} = \frac{SS_A}{(a - 1)(b - 1)}, \quad MS_E = \frac{SS_E}{ab(n - 1)}.$$

And their expectations are

$$E[MS_A] = \sigma^2 + \frac{bn \sum_{i=1}^a \tau_i^2}{a - 1}, \quad E[MS_B] = \sigma^2 + \frac{an \sum_{j=1}^b \beta_j^2}{b - 1}$$

$$E[MS_{AB}] = \sigma^2 + \frac{n \sum_{i=1}^a \sum_{j=1}^b (\tau \beta)_{ij}^2}{(a - 1)(b - 1)}, \quad E[MS_E] = \sigma^2.$$

## Test statistics

1.  $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$  versus  $H_1 : \text{at least one } \tau_i \neq 0$ ,  
we use

$$F_0 = \frac{MS_A}{MS_E}, \text{ reject } H_0 \text{ if } F_0 > f_{a-1, ab(n-1), \alpha}.$$

2.  $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0$  versus  $H_1 : \text{at least one } \beta_j \neq 0$ ,  
we use

$$F_0 = \frac{MS_B}{MS_E}, \text{ reject } H_0 \text{ if } F_0 > f_{b-1, ab(n-1), \alpha}.$$

3.  $H_0 : (\tau\beta)_{11} = (\tau\beta)_{12} = \cdots = (\tau\beta)_{ab} = 0$  versus  
 $H_1 : \text{at least one } (\tau\beta)_{ij} \neq 0$ , we use

$$F_0 = \frac{MS_{AB}}{MS_E}, \text{ reject } H_0 \text{ if } F_0 > f_{(a-1)(b-1), ab(n-1), \alpha}.$$

Always use the  $P$ -value approach!

# ANOVA Table

Here is the structure of an ANOVA table most statistical software produces:

**TABLE 14.4** ANOVA Table for a Two-Factor Factorial, Fixed-Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	$F_0$
A treatments	$SS_A$	$a - 1$	$MS_A = \frac{SS_A}{a - 1}$	$\frac{MS_A}{MS_E}$
B treatments	$SS_B$	$b - 1$	$MS_B = \frac{SS_B}{b - 1}$	$\frac{MS_B}{MS_E}$
Interaction	$SS_{AB}$	$(a - 1)(b - 1)$	$MS_{AB} = \frac{SS_{AB}}{(a - 1)(b - 1)}$	$\frac{MS_{AB}}{MS_E}$
Error	$SS_E$	$ab(n - 1)$	$MS_E = \frac{SS_E}{ab(n - 1)}$	
Total	$SS_T$	$abn - 1$		

Note: if  $H_0$  is rejected, we still do not know which pairs are significant different in means. Again, we can use TukeyHSD.

## Example: Aircraft Primer Paint

Aircraft primer paints are applied to aluminum surfaces by two methods: dipping and spraying. The purpose of using the primer is to improve paint adhesion, and some parts can be primed using either application method. The process engineering group responsible for this operation is interested in learning whether three different primers differ in their adhesion properties. A factorial experiment was performed to investigate the effect of paint primer type and application method on paint adhesion. For each combination of primer type and application method, three specimens were painted, then a finish paint was applied and the adhesion force was measured. The data from the experiment are shown

Primer Type	Dipping	Spraying	$y_{i..}$
1	4.0, 4.5, 4.3	5.4, 4.9, 5.6	28.7
2	5.6, 4.9, 5.4	5.8, 6.1, 6.3	34.1
3	3.8, 3.7, 4.0	5.5, 5.0, 5.0	27.0
$y_{.j}$	40.2	49.6	89.8 = y...

## Example: Aircraft Primer Paint

**Solution:** Run analysis of variance:

```
> my_data=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/
                     master/Data/AdhesionForce.csv")
> head(my_data,2)
  Primer Method Adhesion
1      1     Dip      4.0
2      1     Dip      4.5

# We need to tell R Primer and Method are the two factors
> my_data$Primer=as.factor(my_data$Primer)
> my_data$Method=as.factor(my_data$Method)

# Run ANOVA test
> res.aov=aov(Adhesion~Primer+Method+Primer:Method,data=my_data)
> summary(res.aov)
    Df Sum Sq Mean Sq F value    Pr(>F)
Primer        2  4.581   2.291  27.858 3.10e-05 ***
Method         1  4.909   4.909  59.703 5.36e-06 ***
Primer:Method 2  0.241   0.121   1.466    0.269
Residuals     12  0.987   0.082
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## Example: Aircraft Primer Paint

**Solution:** Based on the three  $P$ -values, we have

1. for the test on the factor  $A$  (Primer Type):

$H_0 : \tau_1 = \tau_2 = \tau_3 = 0$  versus  $H_1 : \text{at least one } \tau_i \neq 0$ , the  $P$ -value is  $3.10e - 05$  much smaller than any commonly used  $\alpha$ , we reject  $H_0$  (Primer Type affects the response).

2. for the test on the factor  $B$  (Method: Dipping or Spraying):

$H_0 : \beta_1 = \beta_2 = 0$  versus  $H_1 : \text{at least one } \beta_j \neq 0$ , the  $P$ -value is  $5.36e - 06$  much smaller than any commonly used  $\alpha$ , we reject  $H_0$  (Method affects the response).

3. For the test on the interaction, we have  $P$ -value  $0.269$  is quite large, we conclude that the data do not provide sufficient evidence to reject the null hypothesis (the interaction might not affect the response).

## Example: Aircraft Primer Paint

**Solution:** Because the data concluded Primer Type affects the response (at least two Primer Types produce different means), but we do not know which two. To find out, we can apply the TukeyHSD

```
> TukeyHSD(res.aov, 'Primer')
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Adhesion ~ Primer + Method + Primer:Method, data = my_data)

$Primer
      diff      lwr      upr      p adj
2-1  0.9000000  0.4583303 1.3416697 0.0004100
3-1 -0.2833333 -0.7250030  0.1583364 0.2409687
3-2 -1.1833333 -1.6250030 -0.7416636 0.0000323
```

We see that only the pair, Primer types 3 and 1, do not provide significantly different means.

## Summary

One can also apply the R program to more than 2-factors factorial designs! The steps are the same.

```
# Example 14.2 in the textbook, where we have three factors.  
# Each has two levels.  
# Read data  
> my_data=read.csv("https://raw.githubusercontent.com/Harrindy/StatEngine/  
    master/Data/CodedSurfaceRoughness.csv")  
  
# We need to tell R all the factors  
> head(my_data,2)  
> my_data$Feed=as.factor(my_data$Feed)  
> my_data$Depth=as.factor(my_data$Depth)  
> my_data$Angle=as.factor(my_data$Angle)  
  
# Run ANOVA to find P-value for the tests  
> res.aov=aov(Roughness~Feed+Depth+Angle+  
    Feed:Depth+Feed:Angle+Depth:Angle+  
    Feed:Depth:Angle,data=my_data)  
> summary(res.aov)  
  
# If rejecting H0, TukeyHSD finds which pairs caused the difference.  
> TukeyHSD(res.aov,'Feed')
```