EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

Worksheets for Statistics

Petra Schreiberová, Marcela Rabasová

Introduction

The study material is designed for students of VSB - Technical University of Ostrava.

The worksheets consist of several theoretical sheets, some solved problems and some sheets with unsolved problems for practicing. The materials should support classwork and they are not recommended for self-study or as a replacement for textbooks.

The worksheets are based on SCHREIBEROVÁ, P.; VOLNÝ, P.; KRČEK, J. et al.: Matematika III: Pracovní listy. Ostrava: VŠB - Technická Univerzita Ostrava, 2015. ISBN 978-80-248-3875-5.

Thanks

The study material was written with the financial support of the project Technology for the Future 2.0, CZ.02.2.69/0.0/0.0/18_058/0010212

ISBN 978-80-248-4489-3
DOI 10.31490/9788024844893

Contents

Worksheets for Statistics

Combinatorics

5 – Permutations

Definition

Permutations are arrangements of objects (with or without repetition) where the internal order is significant. **The number of permutations of n objects** is the number of all different arrangements in which n items can be placed.

Formulas:

- **number of permutations of n objects without repetition**

$$P(n) = n!$$

factorial

$$n! = n(n-1)(n-2)\cdots 1; \quad 0! = 1$$

- **number of permutations of n objects with repetitions**

$$P(n)^* = \frac{n!}{n_1!n_2!\cdots n_k!}$$

- used in case the group of n items has k groups of indistinguishable items where $n_1, n_2, \ldots, n_k$ are the numbers of the items in these groups

- **number of circular permutations of n objects**

$$P(n)^\circ = (n-1)!$$

- used in case the items are arranged in a circle

Remark

Excel:
$P(n)$ =PERMUTACE$(n;n)$
$n!$ =FAKTORIÁL(n)

Worksheets for Statistics

6 – Permutations

Example

a) In how many ways can a group of eight persons arrange themselves in a row?

b) In how many ways can a group of eight persons arrange themselves in a circle?

c) How many different numbers can be created by using all the digits of the number 11 212 251?

In all these cases, we work with a group of eight items and change the arrangement of all of them so we make permutations.

We work with a group of different items in examples a) and b) so we use the formula for the number of permutations without repetition.

Some items are repeated in example c) so we must use the formula for the number of permutations with repetition here.

a) $P(8) = 8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 40320$

Excel:

$$P(8) = \text{PERMUTACE}(8;8) = 40320$$

b) $P(8)^{\circ} = (8-1)! = 7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$

Excel:

$$P(7) = \text{PERMUTACE}(7;7) = 5040$$

c) $P(8)^{*} = \dfrac{8!}{4! \cdot 3! \cdot 1!} = 280$

Excel:

$$P(8)^{*} = \text{FAKTORIÁL}(8)/(\text{FAKTORIÁL}(4) * \text{FAKTORIÁL}(3)) = 280$$

Worksheets for Statistics

7 – Variations

Definition

Variations are arrangements of selections of objects (with or without repetition) where the internal order is significant. **The number of k-element variations of n objects** is the number of all different arrangements of all different k-item selections from a group of n items.

Formulas:

- **number of k-element variations of n objects without repetition**

$$V_k(n) = \frac{n!}{(n-k)!}$$

- **number of k-element variations of n objects with repetitions**

$$V_k^*(n) = n^k$$

  - used in case any of n items can be repeated in k-element selection arbitrarily

Remark

Excel:
$V_k(n)$ =PERMUTACE$(n;k)$
$V_k^*(n)$ =POWER$(n;k)$

Worksheets for Statistics

8 – Variations

Example

a) Specify the number of all possible three-digit codes in which the numbers can not be repeated.

b) Specify the number of all possible three-digit codes in which the numbers can be repeated.

c) In how many ways can be occupied medal positions by ten sprinters?

In all these cases, we choose three items from a group of ten items and change their arrangement so we make variations.

We can choose only different items in examples a) and c) so we use the formula for the number of variations without repetition.

Any item (from the group of ten digits) can be repeated in any selection of three items (in any three-digit code) in example b) so we use the formula for the number of variations with repetitions here.

a) $V_3(10) = \dfrac{10!}{(10-3)!} = \dfrac{10!}{7!} = \dfrac{10 \cdot 9 \cdot 8 \cdot 7!}{7!} = 720$

Excel:

$$V_3(10) = \text{PERMUTACE}(10;3) = 720$$

b) $V_3^*(10) = 10^3 = 1000$

Excel:

$$V_3^*(10) = \text{POWER}(10;3) = 1000$$

c) $V_3(10) = \dfrac{10!}{(10-3)!} = 720$

Excel:

$$V_3(10) = \text{PERMUTACE}(10;3) = 720$$

Worksheets for Statistics

9 – Combinations

Definition

Combinations are selections of objects (with or without repetition) where the internal order is not significant. **The number of k-element combinations of n objects** is the number of all ways of choosing k items from a group of n items where we do not take the order into account.

Formulas:

- **number of k-element combinations of n objects without repetition**

$$C_k(n) = \binom{n}{k}$$

combinatorial number

$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

- **number of k-element combinations of n objects with repetition**

$$C_k^*(n) = \binom{n+k-1}{k}$$

Remark

Excel:
$C_k(n)$ =KOMBINACE$(n;k)$
$C_k^*(n)$ =KOMBINACE$(n+k-1;k)$

10 – Combinations

Example

a) Ten kinds of cakes are sold in the cake shop. How many options do we have to order eight cakes?

b) Ten kinds of cakes are sold in the cake shop. How many options do we have to order eight different cakes?

c) There are ten people waiting for a lift. How many options do we have to choose eight of them to go there?

In all these cases, we make selections of eight items (from a group of ten items) where the order is not important so we make combinations.

We can choose only different items in examples b) and c) so we use the formula for the number of combinations without repetition.

Any item (from the group of ten kinds of cakes) can be repeated in any selection of eight items (in any order of eight cakes) in example a) so we use the formula for the number of combinations with repetitions here.

a) $C_8^*(10) = \binom{10+8-1}{8} = \binom{17}{8} = \frac{17!}{(17-8)! \cdot 8!} = 24310$

Excel:

$$C_8^*(10) = \text{KOMBINACE}(10+8-1;8) = 24310$$

b) $C_8(10) = \binom{10}{8} = \frac{10!}{(10-8)! \cdot 8!} = 45$

Excel:

$$C_8(10) = \text{KOMBINACE}(10;8) = 45$$

c) $C_8(10) = \binom{10}{8} = 45$

Excel:

$$C_8(10) = \text{KOMBINACE}(10;8) = 45$$

Worksheets for Statistics

11 – Combinatorics

Exercise

a) How many options do we have to seat 8 students to 8 different computers?

b) How many different anagrams (words) can be created using all the letters of the word STATISTICS?

c) How many games will be played in an eight-team hockey tournament in which every team will play with every other team just once?

d) Ten friends have sent holiday postcards to each other. How many postcards have been sent?

Hints

Permutations:

- $P(n) = n!$
- $P(n)^* = \frac{n!}{n_1! \cdot n_2! \cdots n_k!}$
- $P(n)^\circ = (n-1)!$

Variations:

- $V_k(n) = \frac{n!}{(n-k)!}$
- $V_k^*(n) = n^k$

Combinations:

- $C_k(n) = \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$
- $C_k^*(n) = \binom{n+k-1}{k}$

Factorial:

- $n! = n(n-1)(n-2)\ldots 1$
- $0! = 1$

Combinatorial number:

- $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$

12 – Combinatorics

Exercise

a) How many elements does a set of all five-digit natural numbers contain?

b) The shop offers seven kinds of postcards. In how many ways can be bought
   1) ten postcards,
   2) five postcards,
   3) five different postcards?

c) Specify the number of all possible tosses of
   1) two six-sided dice,
   2) three six-sided dice.

Hints

Permutations:

- $P(n) = n!$
- $P(n)^* = \dfrac{n!}{n_1! \cdot n_2! \cdots n_k!}$
- $P(n)^\circ = (n-1)!$

Variations:

- $V_k(n) = \dfrac{n!}{(n-k)!}$
- $V_k^*(n) = n^k$

Combinations:

- $C_k(n) = \binom{n}{k} = \dfrac{n!}{(n-k)! \cdot k!}$
- $C_k^*(n) = \binom{n+k-1}{k}$

Factorial:

- $n! = n(n-1)(n-2)\ldots 1$
- $0! = 1$

Combinatorial number:

- $\binom{n}{k} = \dfrac{n!}{(n-k)! \cdot k!}$

Worksheets for Statistics

Basic probability

Worksheets for Statistics

14 – Random trial, random event

Random trial - any action with a random result (its outcome is not known in advance)

Sample space Ω - the set of all possible outcomes of a random trial

Random event A - any subset of Ω, $A \subset \Omega$

Types of events:

Impossible event - event A that never occurs, $A = \emptyset$

Sure event - event A that always occurs, $A = \Omega$

Complementary (oposit) event - event $\overline{A}$ that occurs in case A does not occur, $\overline{A} = \Omega - A$

Example

Random trial: throwing a regular die

Sample space Ω:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Random event: A ... the result is even

$$A = \{2, 4, 6\}$$

Impossible event: B ... the result is larger than 7

$$B = \emptyset$$

Sure event: C ... the result is less than 7

$$C = \Omega$$

Complement of A: $\overline{A}$ ... the result is odd

$$\overline{A} = \{1, 3, 5\}$$

15 – Random trial, random event

Relationship between events

Event A is **a subset** of event B:

$$A \subset B \Leftrightarrow \{\forall \omega \in \Omega : (\omega \in A) \Rightarrow (\omega \in B)\}$$

- if A occurs then B occurs

Events A and B are **equal**:

$$A = B \Leftrightarrow \{\forall \omega \in \Omega : (\omega \in A) \Leftrightarrow (\omega \in B)\}$$

- A occurs if and only if B occurs

Events A and B are **mutually exclusive** or **disjoint** $\Leftrightarrow$ they cannot happen at the same time.

Events $A_1, A_2, \ldots, A_n$ are **mutually exclusive** $\Leftrightarrow$ the occurrence of any one of them implies the non-occurrence of the remaining $n - 1$ events.

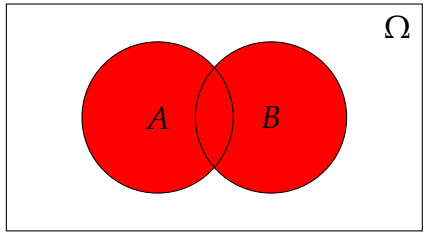
Events $A_1, A_2, \ldots, A_n$ are **jointly** or **collectively exhaustive** $\Leftrightarrow$ at least one of them must occur.

Events $A_1, A_2, \ldots, A_n$ **partition the sample space** $\Omega \Leftrightarrow$ they are mutually exclusive and collectively exhaustive.

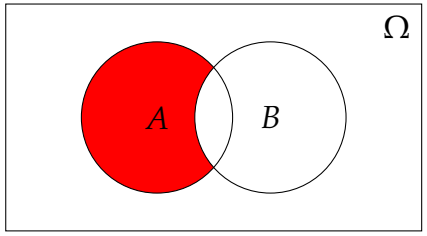
16 – Random trial, random event

Operations with events

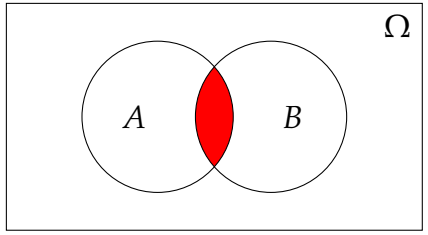
The union of A and B:
$A \cup B = \{\forall \omega \in \Omega : (\omega \in A) \vee (\omega \in B)\}$
- the occurrence of A or B or both
"A or B"



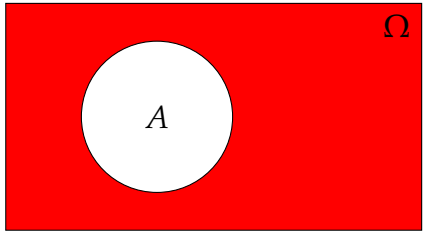
The difference of A and B:
$A - B = \{\forall \omega \in \Omega : (\omega \in A) \wedge (\omega \notin B)\}$
- the occurrence of A and the non-occurrence of B
"A but not B"



The intersection of A and B:
$A \cap B = \{\forall \omega \in \Omega : (\omega \in A) \wedge (\omega \in B)\}$
- the occurrence of both A and B
"A and B"



The complement of A:
$\overline{A} = \{\forall \omega \in \Omega : \omega \notin A\}$
- the non-occurrence of A
"not A"



17 – Random trial, random event

Example

A player throws two fair dice (the first one is red and the second one is blue). Write out all the elements of the sample space Ω and of the following random events:

A_1 ... the number of dots on the face of the red die equals six
A_2 ... the number of dots on the face of the blue die equals one
A_3 ... the number of dots on the face of the red die equals six and the number of dots on the face of the blue die equals one
A_4 ... the number of dots on the face of the red die equals six or the number of dots on the face of the blue die equals one
A_5 ... the number of dots on the face of the red die equals six and the number of dots on the face of the blue die does not equal one
A_6 ... the number of dots on the face of the red die does not equal six
A_7 ... the sum of dots on the faces of the red and blue die is less than or equal to twelve
A_8 ... the sum of dots on the faces of the red and blue die is greater than twelve
A_9 ... the numbers of dots on the faces of both dice is even
A_{10} ... the numbers of dots on the faces of both dice is odd
A_{11} ... one of the numbers of dots on the faces of dice is even and the second one is odd

The number of elements of sample space Ω equals $V_2^*(6) = 6^2 = 36$ (number of two-element variations of six objects with repetitions)

$\Omega = \{$ (1,1), (1,2), (1,3), (1,4), (1,5), (1,6),
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6) $\}$

$A_1 = \{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$
$A_2 = \{(1,1), (2,1), (3,1), (4,1), (5,1), (6,1)\}$

18 – Random trial, random event

$A_3 = A_1 \cap A_2 = \{(6,1)\}$ ... the intersection of A_1 and A_2

$A_4 = A_1 \cup A_2 = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),(1,1),(2,1),(3,1),(4,1),(5,1)\}$ ... the union of A_1 and A_2

$A_5 = A_1 - A_2 = \{(6,2),(6,3),(6,4),(6,5),(6,6)\}$ ... the difference of A_1 and A_2

$A_6 = \overline{A}_1 =$ { (1,1), (1,2), (1,3), (1,4), (1,5), (1,6),
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6) }

... the complement of A_1

$A_7 = \Omega$ ... sure event

$A_8 = \emptyset$ ... impossible event

$A_9 = \{(2,2),(2,4),(2,6),(4,2),(4,4),(4,6),(6,2),(6,4),(6,6)\}$

$A_{10} = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\}$

$A_{11} = \{(1,2),(1,4),(1,6),(2,1),(2,3),(2,5),(3,2),(3,4),(3,6),(4,1),(4,3),(4,5),(5,2),(5,4),(5,6),(6,1),(6,3),(6,5)\}$

Remark

- the events A_9, A_{10} are mutually exclusive ($A_9 \cap A_{10} = \emptyset$) but A_9 is not the complement of A_{10} ($A_9 \neq \overline{A}_{10}$)
- the events A_9, A_{10}, A_{11} are mutually exclusive ($A_9 \cap A_{10} = \emptyset$, $A_9 \cap A_{11} = \emptyset$, $A_{10} \cap A_{11} = \emptyset$) and also collectively exhaustive ($A_9 \cup A_{10} \cup A_{11} = \Omega$)

19 – Probability of random event

Definition

The **probability of random event** A, $A \subset \Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, is defined by the formula (classical approach):

$$P(A) = \frac{m}{n},$$

where n is the number of all possible outcomes of a random trial and m is the number of all outcomes of a random trial that meet the specification of A.

Theorem (The Kolmogorov axioms)

1. $P(A) \geq 0$; for any random event A
2. $P(\Omega) = 1$
3. $P(A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_n) + \cdots$; for any set of mutually exclusive random events $\{A_1, A_2, \ldots, A_n, \ldots\}$

Theorem (Probability characteristics)

1. $P(\varnothing) = 0$
2. $P(\overline{A}) = 1 - P(A)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

20 – Probability of random event

Example

A random trial consists of rolling two dice simultaneously and independently of one another. Calculate the probability of:
a) the sum being six
b) the sum being different from six
c) having even number in both rolls
d) having even number in both rolls and the sum equal to six
e) having even number in both rolls or the sum equal to six

The sample space of this random trial consists of 36 elements (as displayed previously).

a) $A = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$

$$P(A) = \frac{5}{36}$$

b) $B = \overline{A}$

$$P(B) = 1 - P(A) = 1 - \frac{5}{36} = \frac{31}{36}$$

c) $C = \{(2,2),(2,4),(2,6),(4,2),(4,4),(4,6),(6,2),(6,4),(6,6)\}$

$$P(C) = \frac{9}{36}$$

d) $D = C \cap A = \{(2,4),(4,2)\}$

$$P(D) = \frac{2}{36}$$

e) $E = C \cup A$

$$P(E) = P(C \cup A) = P(A) + P(C) - P(A \cap C) = \frac{5}{36} + \frac{9}{36} - \frac{2}{36} = \frac{12}{36}$$

21 – Probability of random event

Example

Of 20 light bulbs, 7 are defective. A group of 5 bulbs is chosen at random. Calculate the probability that:
a) none of them is defective
b) two of them are defective
c) at least one of them is defective

a) Choosing a group of 5 bulbs from a group of 20 we make combinations without repetition. Their number equals $C_5(20)$, which is the number of all possible results of our random trial.

The number of all outcomes that meet the specification of a) equals $C_5(13)$ (0 defective means 5 non-defective bulbs, their number is $20 - 7 = 13$).

$$P(A) = \frac{C_5(13)}{C_5(20)} = \frac{\binom{13}{5}}{\binom{20}{5}} \doteq 0.0830$$

b) The number of all outcomes that meet the specification of b) equals $C_2(7) \cdot C_3(13)$ (2 defective means 3 non-defective bulb).

$$P(B) = \frac{C_2(7) \cdot C_3(13)}{C_5(20)} = \frac{\binom{7}{2} \cdot \binom{13}{3}}{\binom{20}{5}} \doteq 0.3874$$

c) Random event C is the complement of random event A.

$$P(C) = 1 - P(A) \doteq 0.9170$$

22 – Probability of random event

Exercise

a) In a group of 72 students, 14 take neither English nor Chemistry, 42 take English and 38 take Chemistry. What is the probability that a student chosen at random from this group takes:
1) both English and Chemistry?
2) Chemistry but not English?

b) Ten married couples are in a room.
1) If two people are chosen at random find the probability that one is male and one is female.
2) If two people are chosen at random find the probability that they are married to each other.
3) If 4 people are chosen at random, find the probability that 2 married couples are chosen.

Hints

Probability of random event A

- $P(A) = \frac{m}{n}$

Permutations:

- $P(n) = n!$
- $P(n)^* = \frac{n!}{n_1! \cdot n_2! \cdots n_k!}$
- $P(n)^\circ = (n-1)!$

Variations:

- $V_k(n) = \frac{n!}{(n-k)!}$
- $V_k^*(n) = n^k$

Combinations:

- $C_k(n) = \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$
- $C_k^*(n) = \binom{n+k-1}{k}$

Factorial:

- $n! = n(n-1)(n-2)\ldots 1$
- $0! = 1$

Combinatorial number:

- $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$

23 – Probability of random event

Exercise

a) If three balls are drawn at random from a bag containing 6 red balls, 4 white balls and 8 blue balls, what is the probability that all three are red?

b) A shipment of 17 radios includes 5 radios that are defective. The receiver samples 6 radios at random. What is the probability that exactly 3 of the selected radios are defective?

Hints

Probability of random event A

- $P(A) = \frac{m}{n}$

Permutations:

- $P(n) = n!$
- $P(n)^* = \frac{n!}{n_1! \cdot n_2! \cdots n_k!}$
- $P(n)^\circ = (n-1)!$

Variations:

- $V_k(n) = \frac{n!}{(n-k)!}$
- $V_k^*(n) = n^k$

Combinations:

- $C_k(n) = \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$
- $C_k^*(n) = \binom{n+k-1}{k}$

Factorial:

- $n! = n(n-1)(n-2)\ldots 1$
- $0! = 1$

Combinatorial number:

- $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$

24 – Probability of random event

Exercise

a) Three married couples have purchased theater tickets and are seated in a row consisting of just six seats. If they take their seats in a completely random fashion, what is the probability that
1) Jim and Paula (husband and wife) sit in the two seats on the far left.
2) Jim and Paula end up sitting next to one another.

b) An experiment consists of rolling two dice simultaneously and independently of one another. Find the probability of the event consisting of having an odd number in the first roll or a total of 9 in both rolls.

Hints

Probability of random event A

- $P(A) = \frac{m}{n}$

Permutations:

- $P(n) = n!$
- $P(n)^* = \frac{n!}{n_1! \cdot n_2! \cdots n_k!}$
- $P(n)^\circ = (n-1)!$

Variations:

- $V_k(n) = \frac{n!}{(n-k)!}$
- $V_k^*(n) = n^k$

Combinations:

- $C_k(n) = \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$
- $C_k^*(n) = \binom{n+k-1}{k}$

Factorial:

- $n! = n(n-1)(n-2)\ldots 1$
- $0! = 1$

Combinatorial number:

- $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$

25 – Conditional probability, independent events

Definition

The **conditional probability** of A given that B occurs, or on condition that B occurs, is defined by the formula:

$$P(A|B) = \begin{cases} \frac{P(A \cap B)}{P(B)} & \text{if } P(B) \neq 0 \\ 0 & \text{if } P(B) = 0 \end{cases}$$

Definition

The events A and B are **independent** $\Leftrightarrow$

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

Remark

- Two events are statistically independent when the occurrence of one has no influence on the occurrence of the other.
- $P(A \cap B) = P(A|B) \cdot P(B)$ ... for any random events A, B
- $P(A \cap B) = P(A) \cdot P(B)$ ... for independent random events A, B
- $P(A \cap B) = P(A) \cdot P(B)$ $\Leftrightarrow$ random events A and B are independent

Definition

The events $A_1, A_2, \ldots, A_n$ are **mutually independent** $\Leftrightarrow$

$$P\left(\bigcap_{i \in M} A_i\right) = \prod_{i \in M} P(A_i), \ \forall M \subset \{1, 2, \ldots, n\}.$$

Remark

The following properties of mutually independent events A, B, C are satisfied:

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

26 – Conditional probability, independent events

Example

A random trial consists of rolling two dice simultaneously and independently of one another. Consider the following three events:
A ... die 1 shows 3
B ... die 2 shows 5
C ... the sum of the two rolls is 8
a) Determine whether the events A and B are independent.
b) Determine whether the events A and C are independent.

The sample space of this random trial consists of 36 elements (as displayed previously).

a) $A = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$

$$P(A) = \frac{6}{36} = \frac{1}{6}$$

$B = \{(1,5),(2,5),(3,5),(4,5),(5,5),(6,5)\}$

It is natural to assume that the two events are independent, since each event involves a different die, which has no connection with the other one. Let us determine the conditional probability $P(A|B)$ and compare it with $P(A)$ to find out whether they are equal or not (which will imply whether the events A and B are statistically independent or not).

Calculating $P(A|B)$ we suppose that B occurs and on that condition we answer our question. In case B occurs, only one element of the six possible elements meets the specification of A, which implies

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

$P(A|B) = P(A) \Rightarrow$ events A and B are independent

27 – Conditional probability, independent events

b) $A = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$

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

$C = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$

$$P(A|C) = \frac{1}{5}$$

$P(A|C) \neq P(A) \Rightarrow$ events A and B are not independent

28 – Conditional probability, independent events

Example

If 2 balls are drawn at random from a box with 5 white and 7 black balls, what is the probability that both two balls are white? Suppose that the first drawn ball
a) is put back
b) is not put back
before the second ball is drawn.

Consider the following three events:

$A_i \ldots i$-th ball is white, $i = 1, 2$
$A \ldots$ both balls are white ($A = A_1 \cap A_2$)

a) A_1 and A_2 are independent (as the results of the first and second draws are independent of one another), so we use the formula for independent events $P(A \cap B) = P(A) \cdot P(B)$ for computation:

$$P(A) = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = \frac{5}{12} \cdot \frac{5}{12} \doteq 0.1736$$

b) A_1 and A_2 are not independent (as the result of the second draw depends on the result of the first one), so we use the formula for dependent events $P(A \cap B) = P(A|B) \cdot P(B)$ for computation:

$$P(A) = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{5}{12} \cdot \frac{4}{11} \doteq 0.1515$$

29 – Conditional probability, independent events

Example

In 5 consecutive rolls of a die, find the probability of having number 6
a) in all the five rolls,
b) only in the 2nd and 4th roll,
c) just in two rolls.

Consider the following five (mutually independent) events:

A_i ... i-th roll shows 6, $i = 1, \ldots, 5$

a) $P(A)$:

$$P(A) = P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot P(A_4) \cdot P(A_5) = \left(\frac{1}{6}\right)^5 \doteq 0.0001$$

b) $P(B)$:

$$P(B) = P(\overline{A}_1 \cap A_2 \cap \overline{A}_3 \cap A_4 \cap \overline{A}_5) = P(\overline{A}_1) \cdot P(A_2) \cdot P(\overline{A}_3) \cdot P(A_4) \cdot P(\overline{A}_5) = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{1}{6}\right)^2 \doteq 0.0161$$

c) $P(C)$:

$$P(C) = P(\overline{A}_1 \cap A_2 \cap \overline{A}_3 \cap A_4 \cap \overline{A}_5) \cdot C_2(5) = P(\overline{A}_1) \cdot P(A_2) \cdot P(\overline{A}_3) \cdot P(A_4) \cdot P(\overline{A}_5) \cdot \binom{5}{2} = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{1}{6}\right)^2 \cdot 10 \doteq 0.1608$$

30 – Conditional probability, independent events

Exercise

a) Three shooters shoot independently to the same target. The probabilities of their hits are 0.8, 0.7 and 0.6. Each shooter shoots one shot. What is the probability that the target will be hit by
1) all of them,
2) non of them,
3) at least one of them,
4) exactly one of them?

b) A hockey team wins with a probability of 0.6 and loses with a probability of 0.3. The team plays three games over the weekend. Find the probability that the team:
1) wins all three games,
2) wins at least twice and does not lose,
3) wins one game, loses one, and ties one (in any order).

Hints

Conditional probability:

- $P(A|B) = \begin{cases} \frac{P(A \cap B)}{P(B)} & \text{if } P(B) \neq 0 \\ 0 & \text{if } P(B) = 0 \end{cases}$

Probability of the intersection of:
- any random events A, B:

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

- independent random events A, B:

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

31 – Conditional probability, independent events

Exercise

a) The probabilities of the monthly snowfall exceeding 10 cm at a particular location in the months of December, January, and February are 0.2, 0.4, and 0.6, respectively. For a particular winter, what is the probability
1) that the snowfall will be less than 10 cm in all the 3 months,
2) of receiving at least 10 cm snowfall in at least 2 of the 3 months?

b) An experiment consists of rolling two dice simultaneously and independently of one another. Find the probability of
1) the first roll being a 1, given that the sum of both rolls was 5,
2) the sum being 5, given that the first roll was a 1.

Hints

Conditional probability:

- $P(A|B) = \begin{cases} \frac{P(A \cap B)}{P(B)} & \text{if } P(B) \neq 0 \\ 0 & \text{if } P(B) = 0 \end{cases}$

Probability of the intersection of:
- any random events A, B:

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

- independent random events A, B:

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

32 – The total probability theorem and Bayes' theorem

Definition

The collection of sets $\{A_1, A_2, ..., A_n\}$ **partitions the sample space** $\Omega \Leftrightarrow$ the events $A_1, A_2, ..., A_n$ are mutually exclusive and collectively exhaustive.

Theorem (Total probability theorem)

If the collection of sets $\{B_1, B_2, ..., B_n\}$ partitions the sample space Ω, then the following formula is valid for any set A in the sample space Ω:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

Theorem (Bayes' theorem)

If the collection of sets $\{B_1, B_2, ..., B_n\}$ partitions the sample space Ω, then the following formula is valid for any set A in the sample space Ω:

$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$$

33 – The total probability theorem and Bayes' theorem

Example

It is known that of the articles produced by a factory, 20 % come from Machine 1, 30 % from Machine 2, and 50 % from Machine 3. The percentages of satisfactory articles among those produced are 95 % for Machine 1, 85 % for Machine 2 and 90 % for Machine 3. An article is chosen at random.
a) What is the probability that it is satisfactory?
b) Assuming that the article is satisfactory, what is the probability that it was produced by Machine 1?

Consider the following three events B_1, B_2, B_3 that partition the sample space Ω and the event A:

B_i ... the i-th article is produced by Machine i, $i = 1, ..., 3$
A ... the article is satisfactory

The following probabilities are known from the text:

$P(B_1) = 0.20$
$P(B_2) = 0.30$
$P(B_3) = 0.50$

$P(A|B_1) = 0.95$
$P(A|B_2) = 0.85$
$P(A|B_3) = 0.90$

a) The formula from the total probability theorem is used to solve question a):

$$P(A) = \sum_{i=1}^{3} P(A|B_i) \cdot P(B_i) = 0.20 \cdot 0.95 + 0.30 \cdot 0.85 + 0.50 \cdot 0.90 = 0.8950$$

b) The formula from Bayes' theorem is used to solve question b):

$$P(B_1|A) = \frac{P(A|B_1) \cdot P(B_1)}{\sum_{i=1}^{3} P(A|B_i) \cdot P(B_i)} = \frac{0.20 \cdot 0.95}{0.20 \cdot 0.95 + 0.30 \cdot 0.85 + 0.50 \cdot 0.90} \doteq 0.2123$$

34 – The total probability theorem and Bayes' theorem

Exercise

a) A certain disease affects about 1 out of 10000 people. There is a test to check whether the person has the disease. The test is quite accurate. In particular, we know that
- the probability that the test result is positive, given that the person does not have the disease, is 2 %,
- the probability that the test result is negative, given that the person has the disease, is 1 %.
A random person gets tested for the disease and the result comes back positive. What is the probability that the person has the disease?

b) An 80 % of people attend their primary care physician regularly; 35 % of those people have no health problems crop up during the following year. Out of the 20 % of people who do not see their doctor regularly, only 5 % have no health issues during the following year. What is the probability a random person will have no health problems in the following year?

Hints

Total probability theorem:

- $P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$

Bayes' theorem:

- $P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$

35 – The total probability theorem and Bayes' theorem

Exercise

a) In a certain county: 60 % of registered voters are Republicans, 30 % are Democrats and 10 % are Independents. When those voters were asked about increasing military spending: 40 % of Republicans opposed it, 65 % of the Democrats opposed it and 55 % of the Independents opposed it.
1) What is the probability that a randomly selected voter opposes increased military spending?
2) A registered voter from our county writes a letter to the local paper, arguing against increased military spending. What is the probability that this voter is a Democrat?

b) At a certain university, 4 % of men are over 6 feet tall and 1 % of women are over 6 feet tall. The total student population is divided in the ratio 3:2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman?

Hints

Total probability theorem:

- $P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$

Bayes' theorem:

- $P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$

36 – The total probability theorem and Bayes' theorem

Exercise

Machines A and B produce 10 % and 90 % respectively of the production of a component intended for the motor industry. From experience, it is known that the probability that machine A and B produce a defective component is 0.01 and 0.05 respectively. If a component is selected at random and is found to be defective, find the probability that it was made by:
1) machine A,
2) machine B.

Hints

Total probability theorem:

- $P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$

Bayes' theorem:

- $P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$

Worksheets for Statistics

Random variable

38 – Definition and types of random variable

Definition

A random variable X is a function $X : \Omega \to \mathbb{R}$ that assigns a number to every outcome of a random experiment.

Remark

We denote random variables by capital letters $X, Y, \ldots$ and their particular values by small letters $x, y, \ldots$
The range of a random variable X, denoted by R_X or M, is the set of possible values for X.

There can be two types of random variables depending on the possible set of values.

- **Discrete random variable** - the possible set of values is finite or countable

  Example of the discrete random variable:

  The random variable is $X =$ the sum of the scores on the two dice.

  The set of possible values X is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

- **Continuous random variable** - the possible set of values is a range (closed or open interval, not discrete values)

  Example of the continuous random variable:

  The random variable is $X =$ the person's height.

  The set of possible values X is $(0, \infty)$.

39 – Distribution function

All random variables have a cumulative distribution function. It is a function giving the probabilities that the random variable X is less than x, for every value x.

Definition

A cumulative distribution function $F(x)$ of the random variable X is defined as

$$F(x) = P(X < x),$$

for $x \in \mathbb{R}$.

Some important properties of $F(x)$:

- $0 \leq F(x) \leq 1$
- $P(x_1 \leq X < x_2) = F(x_2) - F(x_1)$
- $F(x)$ is a non-decreasing function:

$$\forall x_1, x_2 \in \mathbb{R} : x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$$

- $F(x)$ is a left-continuous function:

$$\forall a \in \mathbb{R} : \lim_{x \to a^-} F(x) = F(a)$$

- it has limits

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1$$

Remark

In the definition above, the "less than" sign, "$<$", is a convention, not a universally used one. The distribution function can be define as "less than or equal to" sign, "$\leq$". Then the distribution function is defined as right-continuous function: $F(x) = P(X \leq x)$.

40 – Probability function

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values.

Definition

Let X be a discrete random variable with range R_X. **Probability mass function** $p(x)$ is defined as

$$p(x) = P(X = x),$$

for $x \in R_X$.

Some properties of $p(x)$:

- $p(x) \geq 0$ for all values from $\mathbb{R}$
- the sum of $p(x)$ over all possible values of x is 1

$$\sum_{i=1}^{k} p(x_i) = 1, \text{where } k \text{ is the maximum possible value of } i$$

- cumulative probabilities are found by adding individual probabilities $p(x_i)$

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

Remark

The probability function $p(x)$ can be described by the table, the graph, the formula.

41 – Discrete random variable, probability and distribution function

Example

There are two coffee machines in the theater foyer. The probability of failure is 7 % for the first machine and 5 % for the second machine. The random variable represents the number of broken machines. Find the probability and the distribution function of the given random variable.

X = the number of broken machines

X can take three values $\{0, 1, 2\} \Rightarrow X$ is discrete random variable

Consider the following two events:

$$A_i \ldots \text{the } i \text{ -th machine is out of order}$$

The probabilities of the events A_i are known from the text

$$P(A_1) = 0.07$$

$$P(A_2) = 0.05$$

The probability function is:

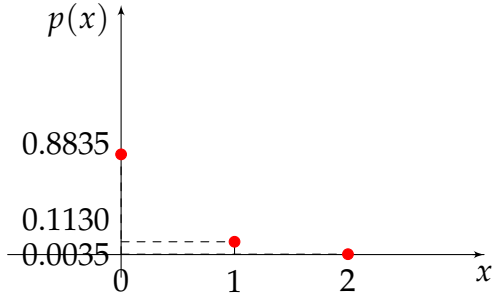
$$p(0) = P(X = 0) = P(\bar{A}_1) \cdot P(\bar{A}_2) = 0.93 \cdot 0.95 = 0.8835$$
$$p(1) = P(X = 1) = P(\bar{A}_1) \cdot P(A_2) + P(A_1) \cdot P(\bar{A}_2) = 0.93 \cdot 0.05 + 0.07 \cdot 0.95 = 0.113$$
$$p(2) = P(X = 2) = P(A_1) \cdot P(A_2) = 0.07 \cdot 0.05 = 0.0035$$

We can calculate values of the distribution function

$$\forall x \in (-\infty; 0] : F(x) = P(X < x) = 0$$
$$\forall x \in (0; 1] : F(x) = P(X < x) = P(X = 0) = 0.8835$$
$$\forall x \in (1; 2] : F(x) = P(X < x) = P(X = 0) + P(X = 1) = 0.8835 + 0.113 = 0.9965$$
$$\forall x \in (2; \infty) : F(x) = P(X < x) = P(X = 0) + P(X = 1) + P(X = 2) = 0.8835 + 0.113 + 0.0035 = 1$$

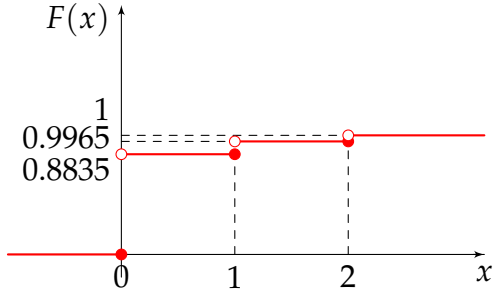
Probability function:

x	0	1	2
$p(x)$	0.8835	0.113	0.0035



Distribution function:

$x \in$	$(-\infty, 0]$	$(0, 1]$	$(1, 2]$	$(2, \infty)$
$F(x)$	0	0.8835	0.9965	1



42 – Discrete random variable

Exercise

a) The shooter has 3 bullets and shoots at the target until the first hit or until the last bullet. The probability that the shooter hits the target after one shot is 0.7. The random variable X is the number of the fired bullets. Find the probability and the distribution function of the given random variable. What is the probability that the number of the fired bullets will not be larger then 2?

b) Toss a coin 3 times. Let X be the number of heads observed. Find the probability and the distribution function of the given random variable.

Hints

The probability function

$$p(x) = P(X = x)$$

Properities of $p(x)$

- $p(x_i) \geq 0$
- $\sum_{i=1}^{n} p(x_i) = 1$

The distribution function

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

43 – Probability density function

Because for a continuous random variable $P(X = x) = 0$ for all $x \in \mathbb{R}$, the probability mass function does not work for continuous random variables. Instead, we can usually define the probability density function.

Definition

The probability density function (pdf) of the random variable X is non-negative function $f(x)$ such that

$$f(x) = \lim_{h \to 0} \frac{P(x \leq X < x + h)}{h},$$

for $x \in [a, b]$.

Some properties of $f(x)$:

- $\forall x \in \mathbb{R} : \ f(x) \geq 0$
- $f(x) = F'(x); \quad F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$
- $\lim_{x \to \infty} f(x) = 0, \quad \lim_{x \to -\infty} f(x) = 0$
- $\int_{-\infty}^{\infty} f(x) = 1$
- $P(x_1 \leq X < x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x$

Remark

The continuous random variable is represented by the area under a curve (this is known as an integral). The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

44 – Continuous random variable, probability density function

Example

The random variable X has the probability density function

$$f(x) = \begin{cases} Ce^{-2x} & \text{pro} \quad 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}.$$

Determine a constant C in order that $f(x)$ is a probability density function. Calculate the probability $P(X < 1)$ and $P(X > \frac{1}{2})$.

The probability density function has to fulfill

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \Rightarrow \int_{-\infty}^{0} 0\,\mathrm{d}x + \int_{0}^{2} Ce^{-2x}\,\mathrm{d}x + \int_{0}^{\infty} 0\,\mathrm{d}x = 1.$$

So

$$\int_{0}^{2} Ce^{-2x}\,\mathrm{d}x = 1 \;\Rightarrow\; C\int_{0}^{2} e^{-2x}\,\mathrm{d}x = C\left[\frac{e^{-2x}}{-2}\right]_0^2 = C\left(\frac{e^{-4}}{-2} + \frac{1}{2}\right) = 1 \;\Rightarrow\; C\cdot\frac{1-e^{-4}}{2} = 1,$$

we get $C = \frac{2}{1-e^{-4}} \doteq 2.04$

Using the pdf we can calculate

$$P(X<1) = \int_0^1 \frac{2}{1-e^{-4}}\cdot e^{-2x}\,\mathrm{d}x = \frac{2}{1-e^{-4}}\left[\frac{e^{-2x}}{-2}\right]_0^1 = \frac{2}{1-e^{-4}}\left(\frac{e^{-2}}{-2}+\frac{1}{2}\right) = \frac{2}{1-e^{-4}}\cdot\frac{1-e^{-2}}{2} = \frac{\frac{e^2-1}{e^2}}{\frac{e^4-1}{e^4}} = \frac{(e^2-1)e^2}{(e^2-1)(e^2+1)} = \frac{e^2}{e^2+1} \doteq 0.88$$

$$P(X>\tfrac{1}{2}) = \int_{\frac{1}{2}}^2 \frac{2}{1-e^{-4}}\cdot e^{-2x}\,\mathrm{d}x = \frac{2}{1-e^{-4}}\left[\frac{e^{-2x}}{-2}\right]_{\frac{1}{2}}^2 = -\frac{1}{1-e^{-4}}(e^{-4}-e^{-1}) = -\frac{1}{1-\frac{1}{e^4}}\left(\frac{1}{e^4}-\frac{1}{e}\right) = -\frac{1}{\frac{e^4-1}{e^4}}\cdot\frac{1-e^3}{e^4} = \frac{e^3-1}{e^4-1} \doteq 0.36$$

45 – Continuous random variable

Exercise

The random variable X has the probability density function

$$f(x) = \begin{cases} Cx^2(2-x) & x \in [0;2] \\ 0 & x \in (-\infty;0) \cup (2;\infty) \end{cases}.$$

Determine a constant C in order that $f(x)$ is a probability density function. Find a distribution function of the random variable X. Calculate the probability $P(X > 1)$.

Hints

Properities of $f(x)$

- $f(x) \geq 0$
- $f(x) = F'(x); \quad F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
- $\lim_{x\to-\infty} f(x) = 0$, $\lim_{x\to\infty} f(x) = 0$

46 – Continuous random variable

Exercise

A probability density function is given by

$$f(x) = \begin{cases} 0 & x < 1 \\ b/x^2 & x \in [1;5] \\ 0 & x > 5 \end{cases}.$$

a) What is the value of b?

b) From this obtain the probability that X is between 2 and 4.

c) What is the probability that X is exactly 2?

d) Find the cumulative distribution function of X.

Hints

Properities of $f(x)$

- $f(x) \geq 0$
- $f(x) = F'(x); \quad F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
- $\lim_{x\to-\infty} f(x) = 0$, $\lim_{x\to\infty} f(x) = 0$

47 – Numerical characteristics of random variables, the expected value

The distribution function (the probability function or the probability density function) gives us the complete information about the random variable. Sometimes it is useful to know some simpler and concentrated formulation of this information such as measures of location, dispersion and concentration.

The expected value (mean)

Definition

Let X be a random variable. The **expectation** $E(X)$ of X (the **mean** μ of X) is defined by

- $\mu \equiv E(X) = \sum_i x_i \cdot p(x_i)$, if X is a discrete random variable
- $\mu \equiv E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x$, if X is a continuous random variable

Properties:

- $E(c) = c, \quad c \in \mathbb{R}$
- $E(aX + b) = aE(X) + b \quad$ for all $a, b \in \mathbb{R}$
- $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$ for any set of random variables
- if $X_1, X_2, \ldots, X_n$ are independent: $E(X_1 \cdot X_2 \cdot \cdots \cdot X_n) = E(X_1) \cdot E(X_2) \cdot \cdots \cdot E(X_n)$

The expected value is the average value of a random variable over a large number of experiments.

The expected value (or mean) of X, where X is a discrete random variable, is a weighted average of the possible values.

The expectation is what you would expect the outcome of an experiment to be on average.Thus the mean of X is a measure of where the values of the random variable X are centered.

Worksheets for Statistics

48 – Numerical characteristics of random variables, the variance

The variance

The variance is a measure of how spread out the distribution of a random variable is.

Definition

Let X be a random variable. The **variance** of X, $\mathrm{var}(X)$, $D(X)$ or σ^2 is given by

- $\sigma^2 \equiv \mathrm{var}(X) = \sum_i (x_i - \mu)^2 \cdot p(x_i)$, if X is a discrete random variable
- $\sigma^2 \equiv \mathrm{var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\,\mathrm{d}x$, if X is a continuous random variable

The variance of a random variable X can be also defined as

$$\mathrm{var}(X) = E(X^2) - (E(X))^2.$$

Properties:

- $\mathrm{var}(c) = 0, \quad c \in \mathbb{R}$
- $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X) \quad$ for all $a, b \in \mathbb{R}$
- $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$, if X and Y are independent

The standard deviation

Definition

The **standard deviation** σ of a random variable X is defined as

$$\sigma = \sqrt{\mathrm{var}(X)}.$$

The variance and the standard deviation are non-negative. The standard deviation of X has the same unit as X, but the variance has a different unit than X - unit^2.

Worksheets for Statistics

49 – Numerical characteristics of random variables, the skewness and the kurtosis

Definition

The r**-th central moment** μ_r is defined by

- $\mu_r = \sum_i (x_i - \mu)^r \cdot p(x_i)$, if X is a discrete random variable.
- $\mu_r = \int_{-\infty}^{\infty} (x - \mu)^r \cdot f(x)\,\mathrm{d}x$, if X is a continuous random variable.

The skewness - is a measure of the lack of symmetry.

Definition

The **skewness** a_3, A is defined by

$$\mathrm{a}_3 = \frac{\mu_3}{\sigma^3}\text{, where } \mu_3 \text{ is the 3rd central moment of } X.$$

- $\mathrm{a}_3 = 0 \Rightarrow$ distribution is symmetric
- $\mathrm{a}_3 < 0 \Rightarrow$ distribution is skewed to the right
- $\mathrm{a}_3 > 0 \Rightarrow$ distribution is skewed to the left

The kurtosis - is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Definition

The **kurtosis** a_4, e is defined by

$$\mathrm{a}_4 = \frac{\mu_4}{\sigma^4}\text{, where } \mu_4 \text{ is the 4th central moment of } X.$$

The standard normal distribution has a kurtosis equal to 3. Data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. Some sources use the excess kurtosis $\bar{e} = e - 3$.

50 – Numerical characteristics of random variables, the median and the mode

The quantiles are the values which divide the distribution such that there is a given proportion of observations below the quantile.

Definition

The p**-quantile** x_p of a random variable X is such a value of X that

$$F(x_p) = p,$$

where $0 < p < 1$.

The **median** $x_{0.5}$ is the central value of the distribution, such that half the values are less than or equal to it and half are greater than or equal to it

$$P(X \leq x_{0.5}) = P(X \geq x_{0.5}) = 0.5.$$

The **quartiles** divide the distribution into four equal parts, called fourths. The **first quartile** $x_{0.25}$ is 0.25-quantile, the second quartile is the median (0.5-quantile) and the **third quartile** $x_{0.75}$ is 0.75-quantile.

The mode

- a discrete random variable X - the **mode** is the value with the greatest probability (the value x at which $P(X = x)$ reaches a maximum), it is the value of X that is most likely to occur. Distributions with only one maximum are called unimodal, those with two maxima bimodal.
- a continuous random variable - the **mode** is the point at which the probability density function reaches a local maximum, or a peak. It is not the value of X most likely to occur.

Worksheets for Statistics

51 – Discrete random variable, the numerical characteristics

Example

A discrete random variable X has the following probability distribution table

x	0	2	3	5
$P(X=x)$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{2}{7}$

Find $F(x)$. Calculate the numerical characteristics of this random variable.

The distribution function

X	$(-\infty, 0]$	$(0, 2]$	$(2, 3]$	$(3, 5]$	$(5, \infty)$
$F(x)$	0	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{5}{7}$	1

The expected value

$$E(X) = \sum_{i=1}^{4} x_i p(x_i) = 0 \cdot \tfrac{1}{7} + 2 \cdot \tfrac{1}{7} + 3 \cdot \tfrac{3}{7} + 5 \cdot \tfrac{2}{7} = \tfrac{2+9+10}{7} = \tfrac{21}{7} = 3$$

The variance + the standard deviation

$$D(X) = E(X^2) - [E(X)]^2 = \sum_{i=1}^{4} x_i^2 p(x_i) - [E(X)]^2 = 0^2 \cdot \tfrac{1}{7} + 2^2 \cdot \tfrac{1}{7} + 3^2 \cdot \tfrac{3}{7} + 5^2 \cdot \tfrac{2}{7} - 3^2 = \tfrac{81}{7} - 9 \doteq 2.57$$

$$\sigma = \sqrt{D(X)} = \sqrt{2.57} \doteq 1.6$$

The skewness + the kurtosis

$$a_3 = \frac{\sum_{i=1}^{4} (x_i - \mu)^3 p(x_i)}{\sigma^3} \doteq \frac{(0-3)^3 \cdot \frac{1}{7} + (2-3)^3 \cdot \frac{1}{7} + (3-3)^3 \cdot \frac{3}{7} + (5-3)^3 \cdot \frac{2}{7}}{1.6^3} \doteq -0.42$$

$$a_4 = \frac{\sum_{i=1}^{4} (x_i - \mu)^4 p(x_i)}{\sigma^4} \doteq \frac{(0-3)^4 \cdot \frac{1}{7} + (2-3)^4 \cdot \frac{1}{7} + (3-3)^4 \cdot \frac{3}{7} + (5-3)^4 \cdot \frac{2}{7}}{1.6^4} \doteq 2.46$$

The mode

Looking at the probability table we can see that the greatest probability is $P(X=3) = \frac{3}{7}$. So the mode is 3.

52 – Discrete random variable, the numerical characteristics

Exercise

A discrete random variable X has the following probability distribution table

x	1	2	3
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

a) Show this probability function as a graph.

b) Sketch a graph of the corresponding cumulative distribution function.

c) Find the expected value.

d) Find the standard deviation and the skewness.

e) Find the mode.

Hints

The distribution function

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

The expected value

$$E(X) = \sum_i x_i p(x_i)$$

The variance

$$D(X) = \sum_i (x_i - E(X))^2 p(x_i)$$

The standard deviation

$$\sigma = \sqrt{D(X)}$$

The skewness

$$a_3 = \frac{\sum_i (x_i - E(X))^3 p(x_i)}{\sigma^3}$$

53 – Discrete random variable, the numerical characteristics

Exercise

Three items are selected at random without replacement from a box containing ten items, of which four are defective.

a) Calculate the probability distribution for the number of defectives in the sample.

b) Sketch a graph of the corresponding cumulative distribution function.

c) What is the expected number of defectives in the sample?

Hints

The distribution function

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

The expected value

$$E(X) = \sum_i x_i p(x_i)$$

The variance

$$D(X) = \sum_i (x_i - E(X))^2 p(x_i)$$

The standard deviation

$$\sigma = \sqrt{D(X)}$$

The skewness

$$a_3 = \frac{\sum_i (x_i - E(X))^3 p(x_i)}{\sigma^3}$$

54 – Continuous random variable, the numerical characteristics

Example

The random variable X has the probability density function

$$f(x) = \begin{cases} \frac{3}{7x^4} & \text{pro} \quad \frac{1}{2} \le x \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

Calculate the expected value and the standard deviation. Find the median.

The expected value

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x = \int_{\frac{1}{2}}^{1} x \cdot \frac{3}{7x^4}\,\mathrm{d}x = \frac{3}{7}\int_{\frac{1}{2}}^{1} \frac{1}{x^3}\,\mathrm{d}x = \frac{3}{7}\left[\frac{x^{-2}}{-2}\right]_{\frac{1}{2}}^{1} = \frac{3}{7}\cdot\left(-\frac{1}{2}\right)\left[\frac{1}{x^2}\right]_{\frac{1}{2}}^{1} = -\frac{3}{14}\cdot(1-4) = \frac{9}{14} \doteq 0.64$$

The variance + the standard deviation

$$D(X) = E(X^2) - [E(X)]^2 \doteq \int_{\frac{1}{2}}^{1} x^2 \cdot \frac{3}{7x^4}\,\mathrm{d}x - 0.64^2 = \frac{3}{7}\int_{\frac{1}{2}}^{1} \frac{1}{x^2}\,\mathrm{d}x - 0.64^2 = \frac{3}{7}\left[\frac{x^{-1}}{-1}\right]_{\frac{1}{2}}^{1} - 0.64^2 = -\frac{3}{7}\left[\frac{1}{x}\right]_{\frac{1}{2}}^{1} - 0.64^2 = -\frac{3}{7}\cdot(1-2) - 0.64^2 \doteq 0.015$$

$$\sigma = \sqrt{D(X)} \doteq \sqrt{0.015} \doteq 0.12$$

The median + the first quartile - the quantile is defined by the formula $F(x_p) = p$.

To find the cumulative distribution function of X, we use $F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$, so for $x < \frac{1}{2}$, we obtain $F(x) = 0$.

For $\frac{1}{2} \le x \le 1$, we have $F(x) = \int_{\frac{1}{2}}^{x} \frac{3}{7x^4}\,\mathrm{d}x = \frac{8}{7} - \frac{1}{7x^3}$.

$$F(x) = \begin{cases} 0 & \text{for} \quad x < \frac{1}{2} \\ \frac{8}{7} - \frac{1}{7x^3} & \text{for} \quad \frac{1}{2} \le x \le 1 \\ 1 & x > 1 \end{cases}.$$

The median is 0.5-quantile $x_{0.5}$, so $F(x_{0.5}) = 0.5$

$$\frac{8}{7} - \frac{1}{7\cdot(x_{0.5})^3} = 0.5$$

$$x_{0.5} = \sqrt[3]{\frac{1}{8 - 7\cdot 0.5}} \doteq 0.606$$

55 – Continuous random variable, the numerical characteristics

Exercise

A continuous random variable X has the cumulative distribution function defined as

$$F(x) = \begin{cases} 0 & x < 0 \\ 2x/\pi & x \in [0; \pi/2) \\ 1 & x \geq \frac{\pi}{2} \end{cases}$$

a) Find the probability density function $f(x)$.

b) Calculate the expected value and the first quartile $x_{0.25}$.

c) Find the variance.

d) Calculate $P\left(\frac{1}{2} \leq X < 1\right)$.

Hints

The probability density function

$$f(x) = F'(x)$$

The expected value

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x$$

The variance

$$D(X) = E(X^2) - [E(X)]^2$$

p-quantil

$$F(x_p) = p$$

56 – Continuous random variable, the numerical characteristics

Exercise

An electrical voltage is determined by the probability density function

$$f(x) = \begin{cases} \frac{1}{2\pi} & x \in [0; 2\pi] \\ 0 & \text{for all other values of } x \end{cases}$$

a) Find its cumulative distribution function for all values of x.

b) Find the mean of this probability distribution.

c) Find its standard deviation.

d) What is the probability that the voltage is within two standard deviations of its mean?

Hints

The cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$$

The expected value

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x$$

The variance

$$D(X) = E(X^2) - [E(X)]^2$$

p-quantil

$$F(x_p) = p$$

Worksheets for Statistics

Discrete probability distribution

58 – Discrete uniform distribution

- all values of the random variable n have the same constant probability
- uniform means that each of the values is equally likely

Definition

A random variable X follows **the discrete uniform distribution** $U(n)$ if we have

$$P(X = x) = p(x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n$$

for a finite sample space of size n.

Properties:

- $E(X) = \mu = \dfrac{n+1}{2}$
- $D(X) = \sigma^2 = \dfrac{n^2-1}{12}$

Proof:

$$\mu = E(X) = \sum_{x=1}^{n} x \cdot p(x) = \sum_{x=1}^{n} x \cdot \frac{1}{n} = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

To get the variance we first calculate

$$E(X^2) = \sum_{x=1}^{n} x^2 \cdot p(x) = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

and finally

$$\sigma^2 = E(X^2) - (E(X))^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \cdots = \frac{n^2-1}{12}$$

59 – Discrete uniform distribution, typical application: Rolling one die

The possible outcomes are $\{1,2,3,4,5,6\}$, each with probability $\frac{1}{6}$. The number of possible outcomes is $n = 6$.

$\Rightarrow$ The outcomes of the roll of a fair die form an uniform distribution $U(6)$.

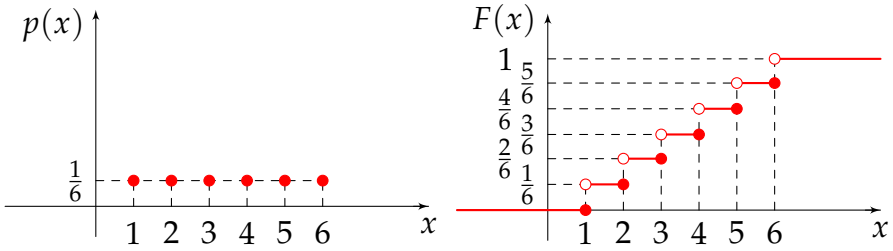
The expected value (mean): $\mu = \frac{6+1}{2} = 3.5$

The variance: $\sigma^2 = \frac{6^2-1}{12} = \frac{35}{12} \doteq 2.9$ and the standard deviation of the outcomes is $\sigma = \sqrt{\frac{35}{12}} \doteq 1.7$

Probability table:

x_i	1	2	3	4	5	6
$p(x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Graphical representation:



The probability that an odd number on the top of the die is

$$P(X = \text{odd number}) = P(X = 1) + P(X = 3) + P(X = 5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5$$

The probability that the number on the top of the die is greater than 4 is

$$P(X > 4) = P(X = 5) + P(X = 6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \doteq 0.33$$

60 – Discrete uniform distribution

Exercise

A telephone number is selected at random from a directory. Let X denote the last digit of randomly selected telephone number.

a) Find the probability that the last digit of the selected number is less than or equal to 2.

b) Find the probability that the last digit of the selected number is 3.

c) Find the probability that the last digit of the selected number is greater than or equal to 7.

d) Compute the mean and the variance of X.

Hints

The probability function is

$$P(X = x) = p(x) = \frac{1}{n},$$

where n is a number of possible outcomes.

The expected value is

$$E(X) = \frac{n+1}{2}$$

The variance is

$$D(X) = \frac{n^2 - 1}{12}$$

Worksheets for Statistics

61 – Bernoulli distribution

- an experiment with two possible outcomes: either success or failure
- success happens with probability p and failure happens with probability $1 - p$
- Bernoulli random variable - a random variable with two possible values 0 and 1

Definition

A random variable X has a **Bernoulli distribution** $Ber(p)$ if its probability distribution function is

$$p(x) = \begin{cases} p^x(1-p)^{1-x} & x \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}.$$

Properties:

- $E(X) = p$
- $D(X) = p(1-p)$

Proof:

$$\mu = \sum_{x=0}^{1} x \cdot p(x) = 0 \cdot (1-p) + 1 \cdot p = p$$

To get the variance we first calculate

$$E(X^2) = \sum_{x=0}^{1} x^2 \cdot p(x) = 0^2 \cdot (1-p) + 1^2 \cdot p = p$$

and finally

$$\sigma^2 = E(X^2) - (E(X))^2 = p - p^2 = p(1-p)$$

62 – Bernoulli distribution

Example

The random variable represents the result of the theoretical part of the exam which consists of one question. The student knows the answer to 70 % of questions. What is the probability that the student will pass the exam?

X = the result of the exam

The possible outcomes are the success (1) and the failure (0): $M = \{0, 1\}$.

The probability of the success is

$$P(X = 1) = p(1) = 0.7$$

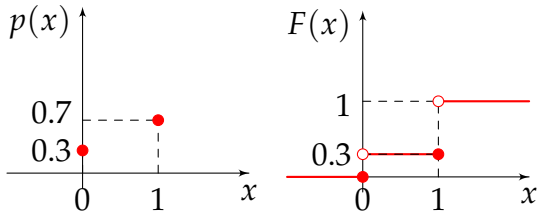
and the probability of the failure is

$$P(X = 0) = p(0) = 1 - p(1) = 1 - 0.7 = 0.3.$$

Probability table:

x_i	0	1
$p(x_i)$	0.3	0.7

Graphical representation of the probability function and the distribution function:



63 – Bernoulli distribution

Exercise

A battery has 95% chances being non-defective. We select a battery at random from a lot of batteries.

a) What is the probability that the selected battery is defective?

b) Let $X =$ selecting battery is not defective. Compute the mean and variance of the random variable X.

Hints

The probability function is

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

The expected value is

$$E(X) = p$$

The variance is

$$D(X) = p(1-p)$$

64 – Binomial distribution

- a binomial random variable represents the total number of successes in a Bernoulli experiment with n independent trials, each with a probability of success p (X denote the number of successes in n independent trials)
- Binomial distribution is one of the most important discrete distribution

Definition

A random variable X has **Binomial distribution** $Bi(n, p)$ if its probability function is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}\,, \qquad x = 0, 1, \ldots, n\,,$$

where n is the number of independent trials and p is the probability of success in each trial.

Properties:

- $E(X) = np$
- $D(X) = np(1-p)$

Remark

Excel:
$p(x) = P(X = x) = \text{BINOM.DIST}(x; n; p; 0)$
$F(x) = P(X \leq x) = \text{BINOM.DIST}(x; n; p; 1)$

Worksheets for Statistics

65 – Binomial distribution

Example

Suppose you independently throw a dart 10 times. Each time you throw a dart, the probability of hitting the target is $\frac{1}{3}$. Compute the expected value and the variance of this random variable. What is the probability of hitting the target

a) three times?

b) more than six times?

Let X denote the number of times you hit. The possible values of X are $\{0, 1, \ldots, 10\} \Rightarrow X \sim Bi\left(10; \frac{1}{3}\right)$.

a) The probability of hitting the target 3 times can be computed from the probability function of X:

$$P(X = 3) = p(3) = \binom{10}{3} \cdot \left(\frac{1}{3}\right)^3 \cdot \left(\frac{2}{3}\right)^7$$

Excel:

$$P(X = 3) = \text{BINOM.DIST}\left(3; 10; \frac{1}{3}; 0\right) \doteq 0.26$$

b) The probability of hitting the target more than 6 times can be computed from the distribution function of X:

$$P(X > 6) = 1 - P(X \leq 6) = 1 - \sum_{x=0}^{6} p(x) = 1 - \sum_{x=0}^{6} \binom{10}{x} \cdot \left(\frac{1}{3}\right)^x \cdot \left(\frac{2}{3}\right)^{10-x}$$

Excel:

$$P(X > 6) = 1 - P(X \leq 6) = 1 - \text{BINOM.DIST}\left(6; 10; \frac{1}{3}; 1\right) \doteq 0.02$$

The expected value and the variance:

$$E(X) = np = 10 \cdot \frac{1}{3} \doteq 3.3$$

$$D(X) = np(1-p) = 10 \cdot \frac{1}{3} \cdot \frac{2}{3} \doteq 2.2$$

66 – Binomial distribution

Exercise

a) Five percent of products are defective. If a lot of 60 items is ordered what is the probability that there is no defective items? What is the probability that there are at least two defective items?

b) If a basketball player takes 12 independent free throws, with a probability of 0.6 of getting a basket on each shot, what is the probability that she gets exactly 7 baskets? What is the expected number of baskets that she gets?

c) The company produces light bulbs, which are packaged in boxes of 40 for shipment. Tests have shown that 2 % of their light bulbs are defective. What is the probability that a box, ready for shipment, contains exactly 2 defective light bulbs?

d) A student is taking a quiz with 5 multiple choice questions. Each question has four options for the answer. Student, who has not studied for the quiz, randomly guesses at each answer. What is the probability that student gets two or fewer correct answers?

Hints

The probability function is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

The expected value is

$$E(X) = np$$

The variance is

$$D(X) = np(1-p)$$

Excel:
$p(x) = P(X = x) = \text{BINOM.DIST}(x; n; p; 0)$
$F(x) = P(X \leq x) = \text{BINOM.DIST}(x; n; p; 1)$

67 – Hypergeometric distribution

- trials are not independent, the probability of success changes from trial to trial
- X denotes the number of success in the random sample of size n drawn from a population

Definition

A random variable X has **Hypergeometric distribution** $H(n, M, N)$ if its probability function is

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad x = \max\{0, M-N+n\} \ldots \min\{M, n\},$$

where n is the number of trials, N is the number of units in the population and M is the number in the population classified as success.

Properties:

- $E(X) = \frac{Mn}{N}$
- $D(X) = \frac{Mn(N-M)(N-n)}{N^2(N-1)}$

Remark

Excel:
$p(x) = P(X = x) = \text{HYPGEOM.DIST}(x; n; M; N; 0)$
$F(x) = P(X \leq x) = \text{HYPGEOM.DIST}(x; n; M; N; 1)$

68 – Hypergeometric distribution

Example

A box contains 150 good light bulbs and 50 defective bulbs. We randomly choose 30 bulbs. What is the probability that

a) at least 25 bulbs are good?

b) we select 5 defective bulbs?

a) Let X denote the number of good bulbs of selected 30 bulbs. Then the probability distribution of X is hypergeometric with parameters $n = 30,\ N = 200,\ M = 150$.

The probability that at least 25 bulbs are good is

$$P(X \geq 25) = 1 - P(X \leq 24) = 1 - \sum_{x=0}^{24} \frac{\binom{150}{x}\binom{50}{30-x}}{\binom{200}{30}}$$

Excel:

$$P(X \geq 25) = 1 - P(X \leq 24) = 1 - \text{HYPGEOM.DIST}(24;30;150;200;1) \doteq 0.18$$

b) Let X denote the number of defective bulbs of selected 30 bulbs. Then the probability distribution of X is hypergeometric with parameters $n = 30,\ N = 200,\ M = 50$.

The probability that 5 bulbs are defective is

$$P(X = 5) = \frac{\binom{50}{5}\binom{150}{25}}{\binom{200}{30}} \doteq 0.1$$

Excel:

$$P(X = 5) = \text{HYPGEOM.DIST}(5;30;50;200;0) \doteq 0.1$$

69 – Hypergeometric distribution

Exercise

a) From a lot of 15 missiles, 6 are selected at random and fired. If the lot contains 4 defective missiles that will not fire, what is the probability that at most 3 will not fire?

b) A deck of cards contains 24 cards - 10 red cards and 14 black cards. 6 cards are drawn randomly. What is the probability that exactly 4 black cards are drawn?

c) A math course has 8 male and 12 female students. The teacher wants to randomly select 3 students. What is the probability that at least 2 students are females? Find the expectation of the number of females in samples of size 3?

Hints

The probability function is

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

The expected value is

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

The variance is

$$D(X) = \frac{Mn(N-M)(N-n)}{N^2(N-1)}$$

Excel:
$P(X = x) = \text{HYPGEOM.DIST}(x; n; M; N; 0)$
$P(X \leq x) = \text{HYPGEOM.DIST}(x; n; M; N; 1)$

70 – Poisson distribution

- X denote number of events that occur in a given interval of time or space
- the probability of occurrence of an event is same for each interval, the number of successes in two disjoint time intervals is independent

Definition

A random variable X has **Poisson distribution** $P(\lambda)$ if its probability function is

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, \dots,$$

where λ is the average number of successes in the given time interval or region of space.

Properties:

- $E(X) = \lambda$
- $D(X) = \lambda$

Remark

Excel:
$p(x) = P(X = x) = \text{POISSON.DIST}(x; \lambda; 0)$
$F(x) = P(X \leq x) = \text{POISSON.DIST}(x; \lambda; 1)$

If X counts the number of events in the interval [0,1] and λ is the average number that occur in unit time, then $X \sim P(\lambda t)$, that is,

$$p(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \qquad x = 0, 1, \dots.$$

If binomial distribution $n \to \infty$, $p \to 0$ such that $np = \lambda$ then binomial distribution tends to Poisson distribution.

71 – Poisson distribution

Example

A book contains 300 pages. The mean number of typing errors in a book is 1.5 per page. Find the probability that on a page chosen at random there are

a) 0 mistakes.

b) more than 3 mistakes.

Let X denote the number of mistakes on the page. Then the probability distribution of X is Poisson with parameter $\lambda = 1.5$.

a) The probability that on the page are no mistakes is

$$P(X = 0) = \frac{1.5^0}{0!} e^{-1.5} \doteq 0.223$$

Excel:

$$P(X = 0) = \text{POISSON.DIST}(0; 1.5; 0) \doteq 0.223$$

b) The probability that on the page are more than 3 mistakes is

$$P(X > 3) = 1 - P(X \leq 2) = 1 - \sum_{x=0}^{2} \frac{1.5^x}{x!} e^{-1.5}$$

Excel:

$$P(X > 3) = 1 - P(X \leq 2) = 1 - \text{POISSON.DIST}(2; 1.5; 1) \doteq 0.191$$

72 – Poisson distribution

Exercise

a) Bacteria are distributed independently of each other in a solution and it is known that the number of bacteria per milliliter follows a Poisson distribution with mean 2.9. Find the probability that a sample of 3 ml of solution contains at least 4 bacteria.

b) The mean number of accidents in a factory is known to be 3.1 per month. Find the probability that in exactly 8 of the months there were more than one accident.

c) A chemical firm produces bottles of soap. It is found over a long period of time that 1 in 35 bottles contains enough impurity to render the soap unusable. A random sample of 250 bottles is taken. What is the probability that more than 8 of them will be unusable?

Hints

The probability function is

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, \ldots$$

The expected value is

$$E(X) = \lambda$$

The variance is

$$D(X) = \lambda$$

Excel:
$p(x) = P(X = x) = \text{POISSON.DIST}(x; \lambda; 0)$
$F(x) = P(X \leq x) = \text{POISSON.DIST}(x; \lambda; 1)$

Worksheets for Statistics

73 – Discrete distribution

Exercise

a) A company is considering drilling four oil wells. The probability of success for each well is 0.40, independent of the results for any other well. What is the probability that one or more wells will be successful? What is the expected number of successes?

b) Customers arrive at a checkout counter at an average rate of 1.5 per minute. Find the probabilities that at least three will arrive during an interval of two minutes?

c) Twelve doughnuts sampled from a manufacturing process are weighed each day. The probability that a sample will have no doughnuts weighing less than the design weight is 6.872 %. What is the probability that a sample of twelve doughnuts contains exactly three doughnuts weighing less than the design weight?

d) The number of meteors found by a radar system in any 30-second interval under specified conditions averages 1.81. Assume the meteors appear randomly and independently. What is the probability of observing at least five but not more than eight meteors in two minutes of observation?

Worksheets for Statistics

Continuous probability distribution

75 – Uniform distribution

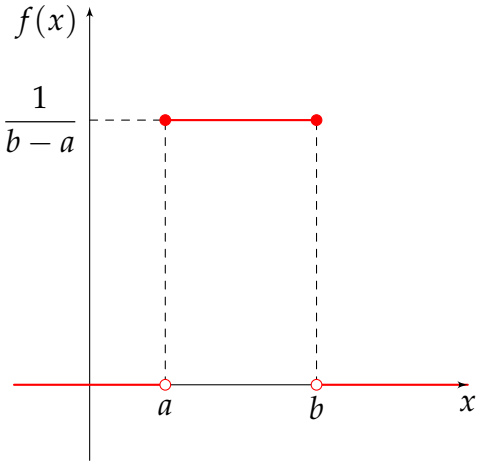
- the continuous uniform distribution is the continuous analogue of the discrete uniform distribution
- also known as rectangular distribution
- $f(x)$ is constant over the finite interval $[a, b]$ - it has equal probability for all values of X between a and b

Definition

A random variable X follows **an uniform distribution** $U(a, b)$ if its probability density function (pdf) is given by

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

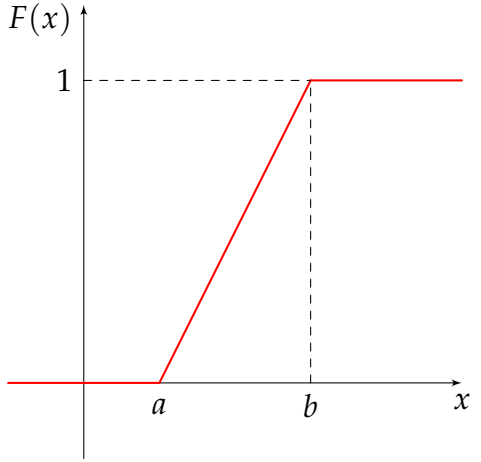
the graph of pdf



The cumulative distribution function (cdf) is:

$$F(x) = \begin{cases} 0 & x \in (-\infty, a] \\ \dfrac{x-a}{b-a} & x \in (a, b) \\ 1 & x \in [b, +\infty) \end{cases}$$

the graph of cdf



Properties:

- $E(X) = \dfrac{a+b}{2}$
- $D(X) = \dfrac{(b-a)^2}{12}$

76 – Uniform distribution

Example

The amount of time that a person must wait for a bus is uniformly distributed between zero and 20 minutes, inclusive.

a) What is the probability that a person waits fewer than 12 minutes?

b) Find the mean and the standard deviation.

c) Find the 95th percentile.

Let X denote the number of minutes (the waiting time at a bus stop). The waiting time is between 0 and 20 minutes $\Rightarrow X \sim U(0, 20)$.

a) The probability that a person waits less than 12 minutes is

$$P(X < 12) = F(12) = \frac{12 - 0}{20 - 0} = 0.6$$

b) The expected waiting time is

$$\mu = \frac{0 + 20}{2} = 10$$

and the standard deviation is

$$\sigma = \sqrt{\frac{(20 - 0)^2}{12}} \doteq 5.8$$

c) Let $x = 95\text{th percentile} \Rightarrow P(X < x) = 0.95$

$$\begin{aligned} P(X < x) &= \frac{x - 0}{20 - 0} \\ 0.95 &= \frac{x}{20} \\ x &= 0.95 \cdot 20 = 19 \end{aligned}$$

Ninety-five percent of the time, a person must wait at most 19 minutes.

77 – Uniform distribution

Exercise

The amount of time a service technician needs to repair a car is uniformly distributed between 80 and 160 minutes.

a) What is the probability that a technician needs more than two hours?

b) Compute the mean and variance.

c) Find the 75th percentile of repair time.

Hints

The density function is

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

The distribution function is

$$F(x) = \begin{cases} 0 & x \in (-\infty, a] \\ \dfrac{x-a}{b-a} & x \in (a,b) \\ 1 & x \in [b, +\infty) \end{cases}$$

The expected value is

$$E(X) = \frac{a+b}{2}$$

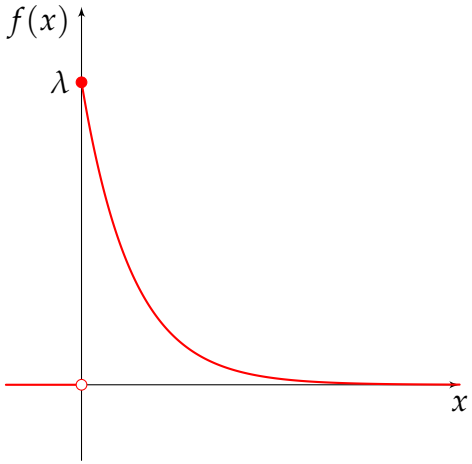
The variance is

$$D(X) = \frac{(b-a)^2}{12}$$

78 – Exponential distribution

- the single most important continuous distribution for building and understanding continuous-time Markov chains, is used for studies of reliability and of queuing theory
- a continuous memoryless distribution that describes the time between events in a Poisson process $P(X > x + x_0 | X > x_0) = P(X > x)$
- is related to the Poisson distribution, although the exponential distribution is continuous whereas the Poisson distribution is discrete

the graph of pdf



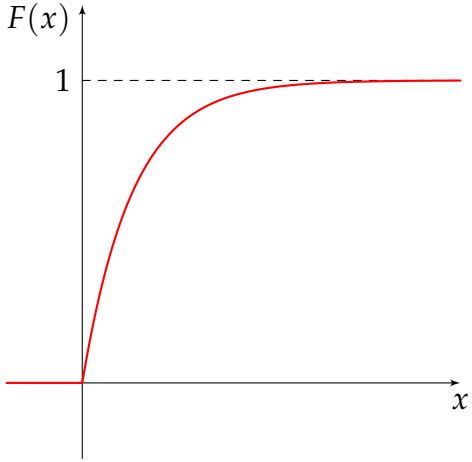
Definition

A random variable X follows **an exponential distribution** $\mathrm{Exp}(\lambda)$ if its probability density function is given by

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & x \in [0, +\infty) \\ 0 & x \in (-\infty, 0) \end{cases},$$

where λ is the intensity or the rate at which an event occurs.

the graph of cdf



For $x > 0$ the cumulative distribution function is

$$F(x) = 1 - \mathrm{e}^{-\lambda x}.$$

Properties:

- $E(X) = \frac{1}{\lambda}$
- $D(X) = \frac{1}{\lambda^2}$

Remark

Excel:
$f(x) = \mathrm{EXPON.DIST}(x; \lambda; 0)$
$F(x) = P(X \leq x) = \mathrm{EXPON.DIST}(x; \lambda; 1)$

79 – Exponential distribution

Example

The time (in hours) required to repair a machine is an exponentially distributed random variable with parameter $\lambda = 1/3$. Find

a) the probability that a repair time takes at most 5 hours.

b) the probability that a repair time takes between 1 and 3 hours.

c) the conditional probability that a repair time takes at least 7 hours, given that its duration exceeds 5 hours.

Let X denote the time required to repair a machine. The average duration of repair is 3 hours $\Rightarrow X \sim \text{Exp}(1/3)$.

a) The probability that a repair time takes at most 5 hours is

$$P(X \leq 5) = F(5) = 1 - \mathrm{e}^{-5/3} \doteq 0.81$$

Excel: $P(X \leq 5) = \text{EXPON.DIST}(5; 1/3; 1) \doteq 0.81$

b) The probability that a repair time takes between 1 and 3 hours is

$$P(1 \leq X \leq 3) = F(3) - F(1) = (1 - \mathrm{e}^{-3/3}) - (1 - \mathrm{e}^{-1/3}) = \mathrm{e}^{-1/3} - \mathrm{e}^{-1} \doteq 0.35$$

Excel: $P(1 \leq X \leq 3) = \text{EXPON.DIST}(3; 1/3; 1) - \text{EXPON.DIST}(1; 1/3; 1) \doteq 0.35$

c) The conditional probability that a repair time takes at least 7 hours, given that its duration exceeds 5 hours is

$$\begin{aligned} P(X \geq 7 | X > 5) &= \frac{P(X \geq 7)}{P(X > 5)} \\ &= \frac{1 - P(X < 7)}{1 - P(X \leq 5)} \\ &= \frac{1 - (1 - \mathrm{e}^{-7/3})}{1 - (1 - \mathrm{e}^{-5/3})} \\ &\doteq 0.51 \end{aligned}$$

The second way how to calculate problem c) is using memoryless property of exponential distribution, $P(X > x + x_0 | X > x_0) = P(X > x)$

$$\begin{aligned} P(X > 2 + 5 | X > 5) &= P(X > 2) \\ &= 1 - P(X \leq 2) \\ &= \mathrm{e}^{-2/3} \\ &\doteq 0.51 \end{aligned}$$

80 – Exponential distribution

Exercise

a) Suppose you are testing a new software, and a bug causes errors randomly at a constant rate of three times per hour. What is the probability that the first bug will occur within the first ten minutes?

b) The length of life of a certain type of electronic tube is exponentially distributed with a mean life of 300 hours. Find the probability that a tube will last more than 500 hours.

c) Suppose the mean checkout time of a supermarket cashier is two minutes. Find the probability of a customer checkout being completed by the cashier in less than three minutes.

Hints

The density function is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \in [0, +\infty) \\ 0 & x \in (-\infty, 0) \end{cases}$$

The distribution function is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \in [0, +\infty) \\ 0 & x \in (-\infty, 0) \end{cases}$$

The expected value is

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

The variance is

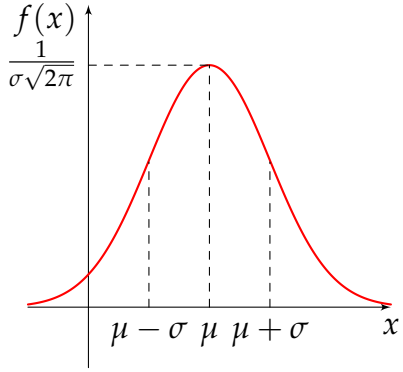
$$D(X) = \frac{1}{\lambda^2}$$

Excel:
$f(x) = \text{EXPON.DIST}(x; \lambda; 0)$
$F(x) = \text{EXPON.DIST}(x; \lambda; 1)$

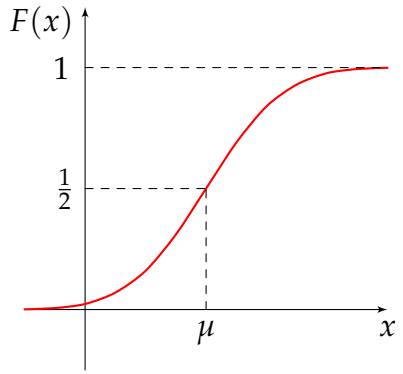
81 – Normal distribution

- it is the most important of all probability distributions. It is applied directly to many practical problems, and several very useful distributions are based on it. It is often called Gaussian distribution.
- the density is symmetrical around the mean, here are about 68 % of the observations within one standard deviation around the mean, 95 % within two standard deviations around the mean, and 99.7 % within three standard deviations of the mean

the graph of pdf



the graph of cdf



Definition

A random variable X follows **a normal distribution** $\mathrm{N}(\mu,\sigma^2)$ if its probability density function is given by

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

where μ is called the location parameter (as it changes the location of density curve) and σ^2 is called the scale parameter of normal distribution (as it changes the scale of density curve).

The cumulative distribution function is

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,\mathrm{d}t,\quad x\in(-\infty,+\infty)$$

Properties:

- $E(X)=\mu,\ D(X)=\sigma^2$
- $A=0,\ \bar{e}=0$

Remark

$f(x)=\mathrm{NORM.DIST}(x;\mu;\sigma;0)$
$F(x)=\mathrm{NORM.DIST}(x;\mu;\sigma;1)$
$x_p=\mathrm{NORM.INV}(p;\mu;\sigma)$

82 – Normal distribution

Example

A factory produces lamps, the lifetimes of which follow the normal distribution. The average lifetime of a lamp is 800 hours with a standard deviation of 40 hours.

a) What is the probability that a randomly selected lamp will last for at least 720 hours?

b) What percentage of lamps will last between 750 and 800 hours?

c) After how many burning hours would we expect 5 % of the lamps to be left?

Let X denote the lamps lifetime $\Rightarrow X \sim N(800, 40^2)$.

a) The probability that a randomly selected lamp will last for at least 720 hours is

$$P(X \geq 720) = 1 - F(720) = 1 - \text{NORM.DIST}(720; 800; 40; 1) \doteq 0.98$$

b) What percentage of lamps will last between 750 and 800 hours?

$$P(750 \leq X \leq 800) = F(800) - F(750) = \text{NORM.DIST}(800; 800; 40; 1) - \text{NORM.DIST}(750; 800; 40; 1) \doteq 0.39$$

c) After how many burning hours would we expect 5 % of the lamps to be left? This corresponds to the time at which

$$P(X > x) = 0.05$$

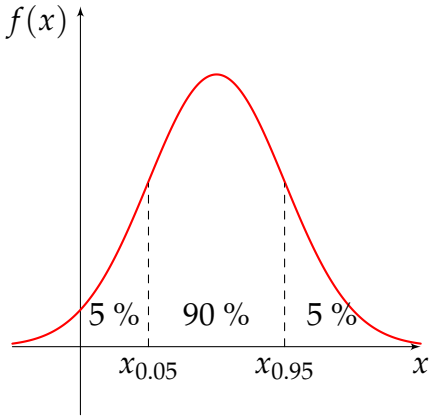
so

$$P(X < x) = 1 - 0.05 = 0.95$$

We need calculate 95% quantile of the random variable X.

$$x_{0.95} = \text{NORM.INV}(0.95; 800; 40) \doteq 866$$

Then after 866 hours of burning, we would expect 5 % of the lamps to be left.



83 – Normal distribution

Exercise

a) On a recent English test, the scores were normally distributed with a mean of 74 and a standard deviation of 7. What proportion of the class would be expected to score between 60 and 80 points?

b) It was found that the mean length of 100 parts produced by a lathe was 20.05 mm with a standard deviation of 0.02 mm. Find the probability that a part selected at random would have a length greater than 20.08 mm.

c) The average life of a certain type of motor is 12 years, with a standard deviation of 3 years. If the manufacturer is willing to replace only 4 % of the motors because of failures, how long a guarantee should he offer? Assume that the lives of the motors follow a normal distribution.

Hints

The density function is

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

The expected value is

$$E(X) = \mu$$

The variance is

$$D(X) = \sigma^2$$

Excel:
$f(x) = \text{NORM.DIST}(x; \mu; \sigma; 0)$
$F(x) = \text{NORM.DIST}(x; \mu; \sigma; 1)$
$x_p = \text{NORM.INV}(p; \mu; \sigma)$

84 – Standard normal distribution

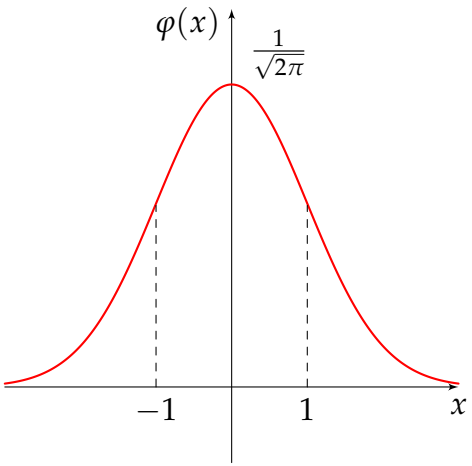
- is the special case of a normal random variable in which the mean is equal to zero and the variance is equal to one (we have the standardized situation)
- values of a cumulative distribution function are tabulated

Definition

A continuous random variable X has **a standard normal distribution** $\mathrm{N}(0,1)$ if its probability density function is given by

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}x^2}\,, \quad x \in (-\infty, +\infty)$$

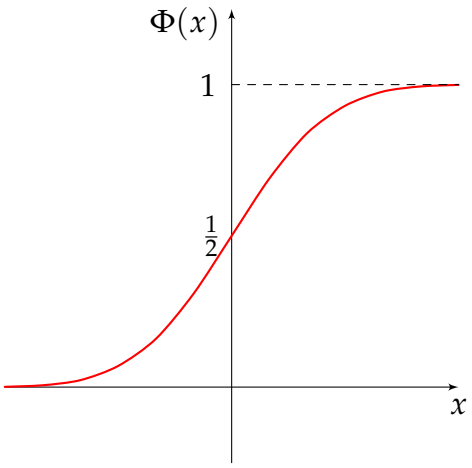
the graph of pdf



The cumulative distribution function is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}t^2}\,\mathrm{d}t\,, \quad x \in (-\infty, +\infty).$$

the graph of cdf



We can transform all the observations of any normal random variable X with mean μ and variance σ^2 to a new set of observations of another normal random variable Z with $\mu = 0$ and $\sigma^2 = 1$ using the following transformation:

$$Z = \frac{X - \mu}{\sigma}$$

$\Phi(z)$ is usually tabulated only for positive values of z due to the symmetry of Φ around zero:

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

Remark

$\varphi(x) = \mathrm{NORM.S.DIST}(x; 0)$
$\Phi(x) = \mathrm{NORM.S.DIST}(x; 1)$
$x_p = \mathrm{NORM.S.INV}(p)$

85 – Standard normal distribution

Example

Let X be the score on the IQ test. The score follow the normal distribution with a mean of 100 and a standard deviation of 15.

a) What is $P(95 \leq X \leq 115)$?

b) What is the lowest possible IQ score that a person can have and still be in the top 1 % of all IQ scores?

a) We express the distribution function of X in terms of the distribution function of a standard normal random variable Z.

$$\begin{aligned} P(95 \leq X \leq 115) &= P(X \leq 115) - P(X \leq 95) \\ &= P\left(Z \leq \frac{115-100}{15}\right) - P\left(Z \leq \frac{95-100}{15}\right) \\ &= P(Z \leq 1) - P(Z \leq -0.33) \\ &= \Phi(1) - (1 - \Phi(0.33)) \\ &= 0.8413 - (1 - 0.6293) \doteq 0.47 \end{aligned}$$

Excel:

$$P(95 \leq X \leq 115) = P(-0.33 \leq Z \leq 1) = \text{NORM.S.DIST}(1;1) - \text{NORM.S.DIST}(-0.33;1) \doteq 0.47$$

$$P(95 \leq X \leq 115) = \text{NORM.DIST}(115;100;15;1) - \text{NORM.DIST}(95;100;85;1) \doteq 0.47$$

b) What is the lowest possible IQ score that a person can have and still be in the top 1 % of all IQ scores?
If a person is in the top 1 %, then that means that 99 % of the people have lower IQ scores $P(X < x) = 0.99$. So

$$P(X < x) = P\left(Z < \frac{x-100}{15}\right) = \Phi\left(\frac{x-100}{15}\right) = 0.99$$

From table

$$0.99 \approx \Phi(2.33) \Rightarrow \frac{x-100}{15} = 2.33 \Rightarrow x = 135$$

Excel:

$$x_{0.99} = 15 \cdot \text{NORM.S.INV}(0.99) + 100 \doteq 135$$

$$x_{0.99} = \text{NORM.INV}(0.99;100;15) \doteq 135$$

A person with IQ score 135 or higher falls in the top 1 % of all IQ scores.

86 – Normal approximation

Binomial approximation
Let X be a binomial random variable with number of trials n and probability of success p. In situation, when

- n is large
- p is close to 0.5

we can use the normal distribution with $\mu = np$ and $\sigma^2 = np(1-p)$.

That is

$$X \sim Bi(n,p) \to X \sim \mathrm{N}(np, np(1-p))$$

The general conditions for using normal approximation to binomial distribution are $np \geq 5$ and $n(1-p) \geq 5$.

Poisson approximation
Let X be a Poisson random variable with mean λ. For large value of the λ we can use the normal distribution with $\mu = \lambda$ and $\sigma^2 = \lambda$.

That is

$$X \sim P(\lambda) \to X \sim \mathrm{N}(\lambda, \lambda)$$

The general condition for using normal approximation to Poisson distribution is $\lambda \geq 5$.

Continuity correction
The binomial and Poisson distributions are discrete random variables, whereas the normal distribution is continuous. We are approximating a discrete distribution with a continuous one, and so we need to make a continuity correction.

$$\begin{aligned}
P(X = a) &= P(a - 0.5 < X < a + 0.5) \\
P(X < a) &= P(X < a - 0.5) \\
P(X \leq a) &= P(X < a + 0.5) \\
P(a < X \leq b) &= P(a - 0.5 < X < b + 0.5) \\
P(a \leq X < b) &= P(a - 0.5 < X < b - 0.5)
\end{aligned}$$

87 – Normal approximation

Example

Consider tossing a coin 35 times. What is the probability of getting between 11 and 14 heads?

Let X denote the number of heads thrown $\Rightarrow X \sim Bi\left(35; \frac{1}{2}\right)$.

Since p equals 0.5, we use normal approximation to binomial distribution.

In our example, $np = 35 \cdot 0.5 = 17.5$ and $n(1-p) = 35 \cdot (1-0.5) = 17.5$.
$np \geq 5$, $n(1-p) \geq 5$ and $p = 0.5 \Rightarrow$ we can use the normal approximation $X \sim N(35 \cdot 0.5; 35 \cdot 0.5 \cdot (1-0.5))$, so $X \sim N(17.5; 8.75)$.

$$\begin{aligned} P(11 \leq X \leq 14) &= P(11 - 0.5 < X < 14 + 0.5) \\ &= P(10.5 < X < 14.5) \\ &= F(14.5) - F(10.5) \\ &= \text{NORM.DIST}(14.5; 17.5; \sqrt{8.75}; 1) - \text{NORM.DIST}(10.5; 17.5; \sqrt{8.75}; 1) \\ &\doteq 0.146 \end{aligned}$$

Example

The average number of collisions occurring in a year at a particular intersection is 54. Assume that the requirements of the Poisson distribution are satisfied. What is the probability of exactly 40 collisions in a year?

Let X denote the number of collisions per a year $\Rightarrow X \sim P(54)$.

Since λ is great (≥ 5), we use normal approximation $\Rightarrow X \sim N(54; 54)$.

$$\begin{aligned} P(X = 40) &= P(40 - 0.5 < X < 40 + 0.5) \\ &= P(39.5 < X < 40.5) \\ &= F(40.5) - F(39.5) \\ &= \text{NORM.DIST}(40.5; 54; \sqrt{54}; 1) - \text{NORM.DIST}(39.5; 54; \sqrt{54}; 1) \\ &\doteq 0.009 \end{aligned}$$

88 – Continuous distribution

Exercise

a) Diameters of bolts produced by a particular machine are normally distributed with mean 0.760 cm and standard deviation 0.012 cm. Specifications call for diameters from 0.720 cm to 0.780 cm. What percentage of bolts will meet these specifications? What percentage of bolts will be smaller than 0.730 cm?

b) The number of days ahead travelers purchase their airline tickets can be modeled by an exponential distribution with the average amount of time equal to 15 days. Find the probability that a traveler will purchase a ticket fewer than ten days in advance. How many days do half of all travelers wait?

c) A subway train on the Red Line arrives every eight minutes during rush hour. We are interested in the length of time a commuter must wait for a train to arrive. The time follows a uniform distribution. Find the probability that the commuter waits less than one minute. Find the probability that the commuter waits between three and four minutes.

Worksheets for Statistics

Descriptive statistics: summary numbers

90 – Measures of location: means

The purpose of descriptive statistics is to present a mass of data in a more understandable form. We may summarize the data in numbers as:

a) some form of average or proportion,
b) some measure of variability or spread.

Consider the data set of N measurements $x_1, x_2, ..., x_N$, which represents the complete population or the sample taken from the population.

Measures of location - describe the central tendency of the data. They include means, mode, median, and quantiles.

1. Means (averages)

Arithmetic mean
- the most popular and well known measure of central tendency
- usually the best single average to use, especially if the distribution is approximately symmetrical and contains no outliers
- disadvantage: it is much affected by outliers

Population mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Sample mean:

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

Geometric mean

$$G(x_1, x_2, ..., x_N) = \sqrt[N]{x_1 \cdot x_2 \cdot ... \cdot x_N}$$

Harmonic mean

$$\overline{x_h} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}}$$

91 – Mode, median, quantiles, sample range

2. Mode (Mo, $\hat{x}$)

- the value that appears most frequently or the midpoint of the class with the largest frequency (in case of grouped frequency approach)

3. Median (Me, $\tilde{x}$)

- the middle item in a sample of values ordered from the smallest to the largest (in case the number of items is odd) or the arithmetic mean of the two middle items (in case the number of items is even)
- advantage of median: it is not much affected by outliers

4. Quantiles (quartiles, deciles, percentiles) ($Q(p)$,x_p)

- quantiles are the cutpoints dividing a set of observations (ordered from the smallest to the largest) into equal sized groups (for example quartiles divide the range of values into four parts, each containing one quarter of the values)
- quantile $Q(p)$ is the i-th object in the ordered sample where $i = N \cdot p + 0.5$ (in case i is integer) or the arithmetic mean of the two adjacent items (in case i is not integer)

- lower quartile $x_{0.25}$ means that about 25 % of the numbers in the data set lie below $x_{0.25}$ and about 75 % lie above $x_{0.25}$
- upper quartile $x_{0.75}$ means that about 75 % of the numbers in the data set lie below $x_{0.75}$ and about 25 % lie above $x_{0.75}$

Remark

Excel:
$\overline{x}$ =PRŮMĚR(data)
$G(x_1, x_2, ..., x_N)$ = GEOMEAN(data)
$\overline{x_h}$ = HARMEAN(data)
x_p = PERCENTIL(data; p)
$\tilde{x}$ = MEDIAN(data)
$\hat{x}$ = MODE(data)

92 – Measures of variability

Measures of variability

Measures of variability describe the spread of the data. They include sample range, interquartile range, mean absolute deviation from the mean, variance, standard deviation and coefficient of variation.

1. Sample range (R)

- the difference between the largest item x_{max} and the smallest item x_{min} in the data sample

$$R = x_{max} - x_{min}$$

- disadvantage: it depends on only two items in each sample, so it does not make use of all the data
- advantage: simplicity (frequently used in quality control)

2. Interquartile range (IQR)

- the difference between the upper quartile and the lower quartile

$$IQR = x_{0.75} - x_{0.25}$$

- frequently used as a measure of variability, particularly in the Box plot

3. Mean absolute deviation from the mean (MAD)

Population mean absolute deviation from the mean:

$$MAD = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$$

Sample mean absolute deviation from the mean:

$$\widehat{MAD} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$$

93 – Measures of variability

4. Variance

- one of the most important descriptions of variability for engineers
- the variance has units of the quantity squared, for example m^2 or s^2 if the original quantity was measured in meters or seconds, respectively

Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Sample variance:

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

5. Standard deviation (σ, s)

- the square root of the variance
- the standard deviation has the same units as the original data

6. Coefficient of variation

- the ratio between the standard deviation and the mean for the same set of data, expressed as a percentage

Population coefficient of variation:

$$c_v = \frac{\sigma}{\mu}$$

Sample coefficient of variation:

$$\widehat{c_v} = \frac{s}{\overline{x}}$$

94 – Descriptive statistics

Example

Consider the sample consisting of the following nine results: 2.3, 7.2, 3.7, 4.6, 5.0, 7.0, 3.7, 4.9, 4.2.
Take the given data as a sample and calculate: mean, mode, quartiles, sample range, interquartile range, mean absolute deviation from the mean, variance, standard deviation, coefficient of variation.

Sample mean:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{9}(2.3 + 7.2 + 3.7 + 4.6 + 5.0 + 7.0 + 3.7 + 4.9 + 4.2) \doteq 4.733$$

Mode:

$$Mo = 3.7$$

Quartiles:

- the first step to find quartiles is to sort the data in order of increasing magnitude, giving the following sequence:

$$2.3, 3.7, 3.7, 4.2, 4.6, 4.9, 5.0, 7.0, 7.2$$

- the second step is to use the formula $i = N \cdot p + 0.5$ and find the i-th object in the ordered sample, which equals $Q(p)$

1. The first quartile (the lower quartile) $Q(0.25)$:

$$i = N \cdot p + 0.5 = 9 \cdot 0.25 + 0.5 = 2.75,$$

so we need to find the objects with the order numbers 2 and 3 and make their arithmetic mean

$$Q(0.25) = \frac{3.7 + 3.7}{2} = 3.7$$

2. The second quartile (the median) $Q(0.5)$:

$$i = N \cdot p + 0.5 = 9 \cdot 0.5 + 0.5 = 5,$$

so we need to find the object with the order number 5

$$Q(0.5) = 4.6$$

3. The third quartile (the upper quartile) $Q(0.75)$:

$$i = N \cdot p + 0.5 = 9 \cdot 0.75 + 0.5 = 7.25,$$

so we need to find the objects with the order numbers 7 and 8 and make their arithmetic mean

$$Q(0.75) = \frac{5+7}{2} = 6$$

Sample range:

$$R = x_{max} - x_{min} = 7.2 - 2.3 = 4.9$$

Interquartile range:

$$IQR = Q(0.75) - Q(0.5) = 6 - 3.7 = 2.3$$

Mean absolute deviation from the mean:

$$\widehat{MAD} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \overline{x}| \doteq \frac{1}{9}\left(|2.3 - 4.733| + |3.7 - 4.733| + \ldots + |7.2 - 4.733|\right) \doteq 1.148$$

Sample coefficient of variation:

$$\widehat{c}_v = \frac{s}{\overline{x}} \doteq \frac{1.568}{4.733} \doteq 0.331$$

Sample variance:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \doteq \frac{1}{8}\left((2.3 - 4.733)^2 + (3.7 - 4.733)^2 + \ldots + (7.2 - 4.733)^2\right) \doteq 2.460$$

Sample standard deviation:

$$s = \sqrt{s^2} \doteq \sqrt{2.460} \doteq 1.568$$

97 – Descriptive statistics

Exercise

a) The same dimension was measured on each of six successive parts as they came off a production line. The results were 21.14 mm, 21.87 mm, 21.53 mm, 21.37 mm, 21.61 mm and 21.93 mm. Calculate the mean, median, variance, standard deviation and coefficient of variation.
1) Consider this set of values as a complete population.
2) Consider this set of values as a sample of all possible measurements of this dimension.

b) Four items in a sequence were measured as 50, 160, 100, and 400 mm. Find their arithmetic mean, geometric mean, harmonic mean and median.

Hints

Means

- $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $G = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$
- $\overline{x_h} = \frac{1}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_N}}$

Population characteristics

- $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $MAD = \frac{1}{N}\sum_{i=1}^{N} |x_i - \mu|$
- $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$
- $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2}$
- $c_v = \frac{\sigma}{\mu}$

Sample characteristics

- $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $\widehat{MAD} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}|$
- $s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2$
- $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2}$
- $\widehat{c_v} = \frac{s}{\overline{x}}$

Worksheets for Statistics

98 – Descriptive statistics

Exercise

The temperature in a chemical reactor was measured every half hour under the same conditions. The results were 78.1°C, 79.2°C, 78.9°C, 80.2°C, 78.3°C, 78.8°C, 79.4°C. Calculate the mean, median, lower quartile, upper quartile, variance, standard deviation and coefficient of variation, considering this set of values as a complete population.

Hints

Means

- $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $G = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$
- $\overline{x_h} = \frac{1}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_N}}$

Population characteristics

- $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $MAD = \frac{1}{N}\sum_{i=1}^{N} |x_i - \mu|$
- $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$
- $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2}$
- $c_v = \frac{\sigma}{\mu}$

Sample characteristics

- $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $\widehat{MAD} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}|$
- $s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2$
- $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2}$
- $\widehat{c_v} = \frac{s}{\overline{x}}$

Worksheets for Statistics

99 – Descriptive statistics

Exercise

The times to perform a particular step in a production process were measured repeatedly. The times were 20.3 s, 19.2 s, 21.5 s, 20.7 s, 22.1 s, 19.9 s, 21.2 s, 20.6 s. Calculate the arithmetic mean, geometric mean, median, lower quartile, upper quartile, variance, standard deviation and coefficient of variation, considering this set of values as a sample of all possible measurements of the times for this step in the process.

Hints

Means

- $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- $G = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$
- $\overline{x_h} = \frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}}$

Population characteristics

- $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
- $MAD = \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|$
- $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$
- $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$
- $c_v = \frac{\sigma}{\mu}$

Sample characteristics

- $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- $\widehat{MAD} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \overline{x}|$
- $s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$
- $s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2}$
- $\widehat{c_v} = \frac{s}{\overline{x}}$

Worksheets for Statistics

Grouped frequencies and graphical descriptions

Worksheets for Statistics

101 – Stem-and-leaf display

1. Stem-and-leaf display

- suitable for exploratory analysis of fairly small sets of data

Example

Data have been obtained on the lives of batteries of a particular type in an industrial application. The following numbers represents the lives of 36 batteries recorded to the nearest tenth of a year:
4.1, 5.2, 2.8, 4.9, 5.6, 4.0, 4.1, 4.3, 5.4, 4.5, 6.1, 3.7, 2.3, 4.5, 4.9, 5.6, 4.3, 3.9, 3.2, 5.0, 4.8, 3.7, 4.6, 5.5, 1.8, 5.1, 4.2, 6.3, 3.3, 5.8, 4.4, 4.8, 3.0, 4.3, 4.7, 5.1.
Make a stem-and-leaf display for these data. Show the leaves sorted in order of increasing magnitude on each stem.

For these data we choose the digits before the decimal point (1, 2, 3, 4, 5, 6) as the stems and we put the digits after the decimal point as the leaves on its corresponding stem.

- the decimal point is not usually shown
- the leaves are often sorted in order of increasing magnitude on each stem
- the number of stems on each leaf can be counted and shown under the heading of frequency

Stem-and-leaf display

Stem	Leaf	Frequency
1	8	1
2	3 8	2
3	0 2 3 7 7 9	6
4	0 1 1 2 3 3 3 4 5 5 6 7 8 8 9 9	16
5	0 1 1 2 4 5 6 6 8	9
6	1 3	2

102 – Box plot (box-and-whisker plot)

2. Box plot (box-and-whisker plot)

- narrow box extends from the lower quartile to the upper quartile
- the median is marked by a line extending across the box
- the smallest and the largest value are marked, and each is joined to the box by a straight line, the whisker

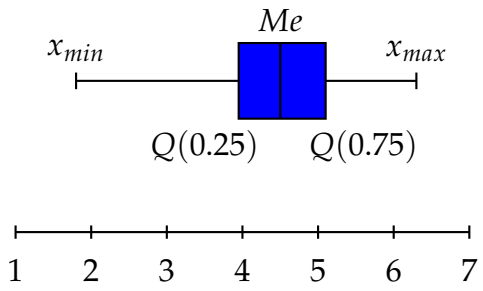
Example

Make a box plot for the data from previous example.

The first step to make a box plot is to sort the data in order of increasing magnitude and find the minimum, maximum, lower quartile, upper quartile and median. The sorted data: 1.8, 2.3, 2.8, 3.0, 3.2, 3.3, 3.7, 3.7, 3.9, 4.0, 4.1, 4.1, 4.2, 4.3, 4.3, 4.3, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.8, 4.9, 4.9, 5.0, 5.1, 5.1, 5.2, 5.4, 5.5, 5.6, 5.6, 5.8, 6.1, 6.3.

- Minimum: $x_{min} = 1.8$
- Maximum: $x_{max} = 6.3$
- Lower quartile: $Q(0.25) = 3.95$
- Median: $Me = Q(0.5) = 4.5$
- Upper quartile: $Q(0.75) = 5.1$

Box plot



103 – Bar chart (bar graph)

3. Bar chart (bar graph)

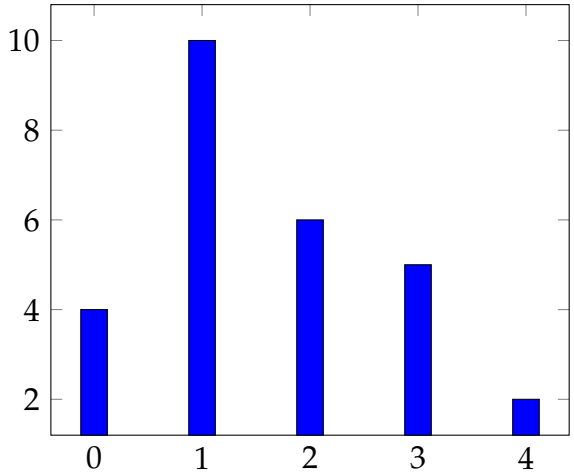
- presents discrete data with rectangular bars with heights proportional to the values that they represent
- one axis of the chart shows the specific categories being compared, and the other axis represents a measured value

Example

The numbers of defective items in successive samples of six items were detected and summarized in the frequency table below. Make a bar graph for the data.

Number of defectives x_i	**Frequency** f_i
0	4
1	10
2	6
3	5
4	2

Bar chart



104 – Graphs of continuous data

4. Graphs of continuous data

- the continuous data are divided into intervals (classes) and the frequency of occurrence for each class is counted to make the data easier to comprehend (the grouped frequency approach)
- the appropriate number of classes is given by Sturges' Rule: number of class intervals $\approx 1 + 3.3\log N$ (where N is the total number of observations in the sample or population)
- rules for the class boundaries: they must be clear with no gaps and no overlaps (e.g. if the values are stated to two decimal places, the class boundaries should end in five in the third decimal)

Example

Over a period of 60 days the percentage relative humidity in a vegetable storage building was measured. Mean daily values were recorded as shown below. Make a graph of the data.
60, 63, 64, 71, 67, 73, 79, 80, 83, 81, 86, 90, 96, 98, 98, 99, 89, 80, 77, 78, 71, 79, 74, 84, 85, 82, 90, 78, 79, 79, 78, 80, 82, 83, 86, 81, 80, 76, 66, 74, 81, 86, 84, 72, 79, 72, 84, 79, 76, 79, 74, 66, 84, 78, 91, 81, 64, 76, 78, 82

Frequency table

The total number of observations:

$$N = 60$$

The appropriate number of classes:

$$\approx 1 + 3.3\log N \doteq 6.868$$

The class width:

$$\approx (x_{max} - x_{min})/6.686 = (99 - 60)/6.686 \doteq 5.68$$

Thus we can make a frequency table with 7 classes of the width equal to 6.

105 – Frequency table

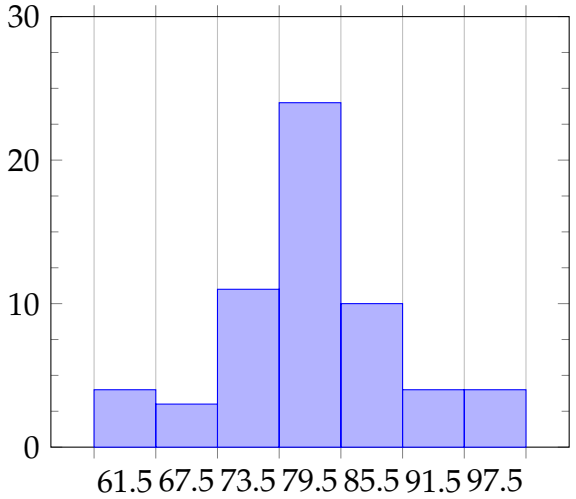
Frequency table

Lower class boundary	Upper class boundary	Class midpoint	Class frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency
58.5	64.5	61.5	4	4	0.067	0.067
64.5	70.5	67.5	3	7	0.05	0.117
70.5	76.5	73.5	11	18	0.183	0.3
76.5	82.5	79.5	24	42	0.4	0.7
82.5	88.5	85.5	10	52	0.167	0.867
88.5	94.5	91.5	4	56	0.067	0.933
94.5	100.5	97.5	4	60	0.067	1

- class midpoint: the point halfway between the corresponding class boundaries
- class frequency: the number of items in the class
- cumulative frequency: the total of all class frequencies smaller than a class boundary
- relative frequency: the class frequency divided by the total number of observations
- relative cumulative frequency: the total of all relative frequencies smaller than a class boundary

106 – Histogram

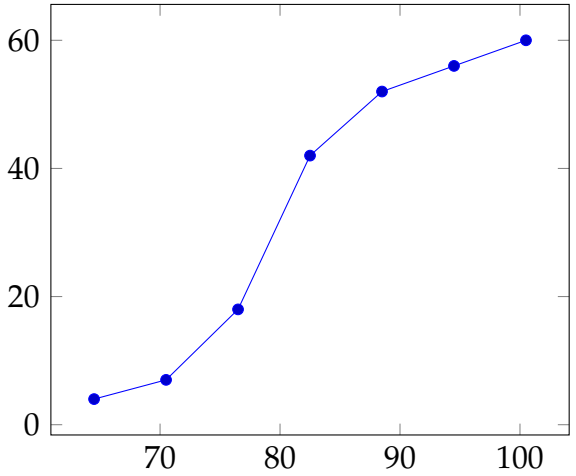
Histogram



- histogram is the bar graph in which the class frequency or relative class frequency is plotted against values of the quantity being studied (so the height of the bar indicates the class frequency or relative class frequency)
- class midpoints are plotted along the horizontal axis
- histogram for continuous data should have the bars touching one another

107 – Cumulative frequency diagram

Cumulative frequency diagram



- cumulative frequency diagram is a plot of cumulative frequency vs. the upper class boundary with successive points joined by straight lines
- cumulative frequency diagram could be changed into a relative cumulative frequency diagram by a change of scale for the ordinate

108 – Grouped frequencies and graphical descriptions

Exercise

The daily emissions of sulfur dioxide from an industrial plant in tonnes/day were as follows:

4.2 6.7 5.4 5.7 4.9 4.6 5.8 5.2 4.1 6.2 5.5 4.9 5.1 5.6 5.9 6.8 5.8 4.8 5.3 5.7

1) Make a stem-and-leaf display for these data.
2) Make a box plot for these data.

109 – Grouped frequencies and graphical descriptions

Exercise

A random sample was taken of the thickness of insulation in transformer windings, and the following thicknesses (in millimeters) were recorded:

18	21	22	29	25	31	37	38	41	39	44	48	54	56	56	57	47	38	35	36
29	37	32	42	43	40	48	36	37	37	36	38	40	41	44	39	38	34	24	32
39	44	42	30	37	30	42	37	34	37	32	24	42	36	49	39	23	34	36	40

1) Make a stem-and-leaf display for these data. Sort the data in order of increasing magnitude.
2) Estimate the percentage of all the windings that received more than 30 mm of insulation.
3) Find the median, lower quartile, and ninth decile of these data.

Hints

Sturges' Rule:

- number of class intervals $\approx 1 + 3.3 \log N$

110 – Grouped frequencies and graphical descriptions

Exercise

4) Make a frequency table for the data. Use Sturges' rule.
5) Draw a frequency histogram.
6) Draw a cumulative frequency graph.
7) Find the mode.
8) Show a box plot of these data.

Hints

Sturges' Rule:

- number of class intervals $\approx 1 + 3.3 \log N$

Worksheets for Statistics

Sampling and combination of variables

112 – Linear combination of independent variables

We often need to combine two or more distributions giving a new variable that may be a sum or difference or mean of the original variables. If we know the variance and mean of the original distributions, can we calculate the variance and mean of the new distribution?

Linear combination of independent variables:

1. Properties of mean (expected value)

- $\mu(c) = c$
- $\mu(cX) = c \cdot \mu(X)$
- $\mu(X + Y) = \mu(X) + \mu(Y)$
- $\mu(X \cdot Y) = \mu(X) \cdot \mu(Y)$

2. Properties of variance

- $\sigma^2(c) = 0$
- $\sigma^2(cX) = c^2 \cdot \sigma^2(X)$
- $\sigma^2(aX + b) = a^2 \cdot \sigma^2(X)$
- $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$

Worksheets for Statistics

113 – Linear combination of independent variables

Example

Cans of beef stew have a mean content of 300 g each with a standard deviation of 6 g. There are 24 cans in a case. What is the mean and the standard deviation of the content of a case?

Using notation:

X_i ... content of i-th can, $i = 1, \ldots, 24$
Y ... content of one case (24 cans)

We have:

$\mu(X_i) = 300,\ i = 1, \ldots, 24$

$\sigma(X_i) = 6,\ i = 1, \ldots, 24$

$\sigma^2(X_i) = 6^2 = 36,\ i = 1, \ldots, 24$

$Y = X_1 + \cdots + X_{24}$

Variables $X_1, \ldots, X_{24}$ are independent, so:

$$\mu(Y) = \mu(X_1) + \cdots + \mu(X_{24}) = 300 + \cdots + 300 = 24 \cdot 300 = 7200$$

$$\sigma^2(Y) = \sigma^2(X_1) + \cdots + \sigma^2(X_{24}) = 36 + \cdots + 36 = 24 \cdot 36 = 864$$

$$\sigma(Y) = \sqrt{\sigma^2(Y)} = \sqrt{864} \doteq 29.4$$

The mean content of a case is 7200 g, the standard deviation of the content of a case is 29.4 g.

Worksheets for Statistics

114 – Linear combination of independent variables

Example

The circumference of a board with rectangular cross-section is twice the sum of the width and thickness of the board. Knowing the variance of the width and the variance of the thickness, what is the variance of the circumference?

Using notation:

X . . . width of a board
Y . . . thickness of a board
Z . . . circumference of a board

We have:

$$Z = 2(X + Y)$$
$$\sigma^2(Z) = 2^2(\sigma^2(X) + \sigma^2(Y))$$

115 – Linear combination of independent variables

Example

An assembly plant has a bin full of steel rods, for which the diameters follow a normal distribution with a mean of 7.00 mm and a variance of 0.100 mm^2, and a bin full of sleeve bearings, for which the diameters follow a normal distribution with a mean of 7.50 mm and a variance of 0.100 mm^2. What percentage of randomly selected rods and bearings will not fit together? (If for any selection of one rod and one bearing, the difference between the bearing diameter and the rod diameter is negative, they will not fit together.)

We are interested in the difference between the bearing diameter and the rod diameter. Because both the diameters of bearings and the diameters of rods follow normal distributions, the difference will also follow a normal distribution. The mean difference will be 7.50 mm - 7.00 mm = 0.50 mm. The variance of the differences will be the sum of the variances of bearings and rods, thus 0.100 mm^2 + 0.100 mm^2 = 0.200 mm^2.

Using notation:

X ... diameter of steel rod
Y ... diameter of sleeve bearing
Z ... difference between bearing diameter and rod diameter

We have:

$X \sim N(7; 0.1)$
$Y \sim N(7.5; 0.1)$
$Z = Y - X \sim N(7.5 - 7; 0.1 + 0.1) = N(0.5; 0.2)$

and the probability that randomly selected rod and bearing will not fit together is:

$$P(Z < 0) = \text{NORM.DIST}(0; 0.5; \sqrt{0.2}; 1) \doteq 0.1318$$

Therefore, 13.18 % of randomly selected sleeves and rods will not fit together.

116 – Sampling, sample mean

Sampling

A **population** might be thought of as the entire group of objects or possible measurements in which we are interested. A **sample** is a group of objects or readings taken from a population for counting or measurement. From the observations of the sample, we infer properties of the population. For example, the sample mean, $\overline{x}$, is an unbiased estimate of the population mean, μ, and that the sample variance, s^2, is an unbiased estimate of the corresponding population variance, σ^2.

However, if these inferences are to be useful, the sample must truly represent the population from which it came. The sample must be **random**, meaning that all possible samples of a particular size must have equal probabilities of being chosen from the population. This will prevent bias in the sampling process. If the effects of one or more factors are being investigated but other factors (not of direct interest) may interfere, sampling must be done carefully to avoid bias from the interfering factors.

Sample mean

Sample mean is a random variable:

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

$X_1, \ldots, X_n$ ... random variables with the same mean μ and the same variance σ^2

n ... sample size

Characteristics of sample mean

- the mean of a sample mean: $\mu(\overline{X})$
- the variance of a sample mean: $\sigma^2(\overline{X})$
  ... is an indication of the reliability of the sample mean as an estimate of the population mean
- the standard deviation of a sample mean: $\sigma(\overline{X})$
  ... is called **the standard error of the mean**

117 – Characteristics of sample mean

1. Sampling with replacement

- $\mu(\overline{X}) = \mu$
- $\sigma^2(\overline{X}) = \frac{\sigma^2}{n}$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

$X_1, \ldots, X_n$ ... independent random variables

2. Sampling without replacement

- $\mu(\overline{X}) = \mu$
- $\sigma^2(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \cdot \frac{\sqrt{N-n}}{\sqrt{N-1}}$

$X_1, \ldots, X_n$ ... dependent random variables
N ... population size

118 – Characteristics of sample mean

Example

A population of size 20 is sampled without replacement. The standard deviation of the population is 0.35. We require the standard error of the mean to be no more than 0.15. What is the minimum sample size?

In case of sampling without replacement we use the formula:

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \cdot \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

In this case σ is 0.35 and N is 20, and substituting the limiting value of 0.15 for the standard error of the mean we obtain:

$$\frac{0.35}{\sqrt{n}} \cdot \frac{\sqrt{20-n}}{\sqrt{20-1}} \leq 0.15$$

$$\sqrt{\frac{20-n}{n}} \leq \frac{0.15\sqrt{19}}{0.35} \doteq 1.868$$

$$20 - n \leq 3.490n$$

$$n \geq \frac{20}{4.490} \doteq 4.45$$

$$n = 5$$

As the sample size n (the number of observations in the sample) must be an integer, the minimum sample size is 5. A sample size of 4 would not satisfy the requirement.

119 – Characteristics of sample mean

Example

The standard deviation of measurements of a linear dimension of a mechanical part is 0.14 mm. What sample size is required if the standard error of the mean must be no more than 0.04 mm?

Since the dimension can be measured as many times as desired, the population size N is effectively infinite and we use the formula for sampling with replacement:

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

In this case σ is 0.14 mm and substituting the limiting value of 0.04 mm for the standard error of the mean we obtain:

$$\frac{0.14}{\sqrt{n}} \leq 0.04$$

$$\sqrt{n} \geq \frac{0.14}{0.04} = 3.5$$

$$n \geq 12.25$$

$$n = 13$$

As the sample size n must be an integer, the minimum sample size is 13.

120 – Central limit theorem

Theorem (The central limit theorem)

If random and independent samples are taken from any practical population of mean μ and variance σ^2, as the sample size n increases the distribution of sample means approaches a normal distribution:

$$\overline{X} \sim N(\mu; \sigma^2/n) \text{ for } n \to \infty.$$

Remark

- if the original population is normally distributed, means of samples of any size are normally distributed (and sums and differences of normally distributed variables are also normally distributed)
- if the original distribution is not normal, means of larger samples are closer to a normal distribution
- means of samples taken from almost all distributions encountered in practice will be normally distributed with negligible error if the sample size is at least 30 (almost the only exceptions will be samples taken from populations containing distant outliers)

Example

A plant manufactures electric light bulbs with a burning life that is approximately normally distributed with a mean of 1200 hours and a standard deviation of 54 hours. Find the probability that a random sample of 36 bulbs will have a sample mean less than 1180 burning hours.

The bulb lives are normally distributed, so the mean of sample of size 36 is also normally distributed with the following characteristics:

$$\mu(\overline{X}) = \mu = 1200 \text{ (hours)}$$

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{54}{\sqrt{36}} = 9 \text{ (hours)}.$$

Then the probability that a random sample of 16 bulbs will have a sample mean less than 1180 hours is:

$$P(\overline{X} < 1180) = \text{NORM.DIST}(1180; 1200; 9; 1) \doteq 0.0131$$

121 – Sampling and combination of variables

Exercise

a) The mean content of a box of cat food is 2.50 kg, and the standard deviation of the content of a box is 0.030 kg. There are 24 boxes in a case, and there are 400 cases in a car load as it leaves the factory. What is the standard deviation of the amount of cat food contained in
1) a case,
2) a car load?

b) A coffee dispensing machine is supposed to dispense a mean of 7.00 fluid ounces of coffee per cup with standard deviation of 0.25 fluid ounces. The distribution approximates a normal distribution. What is the probability that, when 12 cups are dispensed, their mean volume is more than 7.15 fluid ounces?

Hints

Properties of mean and variance (X, Y independent):

- $\mu(c) = c$
- $\mu(cX) = c \cdot \mu(X)$
- $\mu(X + Y) = \mu(X) + \mu(Y)$
- $\mu(X \cdot Y) = \mu(X) \cdot \mu(Y)$

- $\sigma^2(c) = 0$
- $\sigma^2(cX) = c^2 \cdot \sigma^2(X)$
- $\sigma^2(aX + b) = a^2 \cdot \sigma^2(X)$
- $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$

Characteristics of sample mean:
1. Sampling with replacement

- $\mu(\overline{X}) = \mu$
- $\sigma^2(\overline{X}) = \frac{\sigma^2}{n}$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

2. Sampling without replacement

- $\mu(\overline{X}) = \mu$
- $\sigma^2(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \cdot \frac{\sqrt{N-n}}{\sqrt{N-1}}$

122 – Sampling and combination of variables

Exercise

a) Bags of sugar from a production line have a mean weight of 5.020 kg with a standard deviation of 0.078 kg. The bags of sugar are packed in cartons of 20 bags each, and the cartons are piled in lots of 12 onto pallets for shipping.
1) What percentage of cartons would be expected to contain less than 100 kg of sugar?
2) Find the upper quartile of sugar content of a carton.
3) What mean weight of an individual bag of sugar will result in 95 % of the pallets weighing more than 1200 kg?

Hints

Properties of mean and variance (X, Y independent):

- $\mu(c) = c$
- $\mu(cX) = c \cdot \mu(X)$
- $\mu(X + Y) = \mu(X) + \mu(Y)$
- $\mu(X \cdot Y) = \mu(X) \cdot \mu(Y)$

- $\sigma^2(c) = 0$
- $\sigma^2(cX) = c^2 \cdot \sigma^2(X)$
- $\sigma^2(aX + b) = a^2 \cdot \sigma^2(X)$
- $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$

Characteristics of sample mean:
1. Sampling with replacement

- $\mu(\overline{X}) = \mu$
- $\sigma^2(\overline{X}) = \frac{\sigma^2}{n}$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

2. Sampling without replacement

- $\mu(\overline{X}) = \mu$
- $\sigma^2(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
- $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \cdot \frac{\sqrt{N-n}}{\sqrt{N-1}}$

Worksheets for Statistics

Statistical inferences for the mean

124 – Hypothesis testing

We often use samples to infer some information about the population from which the samples came. One type of information is: has the population mean μ the stated value, or has it changed? Another question is to find an interval estimate for the population mean. The first type of question is answered by hypothesis testing.

Hypothesis testing - tests of significance:

1. State the null hypothesis H_0 and the alternative hypothesis H_A in terms of a population parameter.

2. State the test statistic TS (with known distribution) and the level of significance of the test p.

3. Show calculations assuming that H_0 is true, i.e. compute the observed value of the test statistic (or compute the observed level of significance p_{obs}).

4. Compute the critical value of the corresponding distribution and state the critical (rejection) region CR.

5. Find out whether the observed value of the test statistic belongs to the critical region CR or not (or compare the observed level of significance p_{obs} with the stated level of significance p) and state a conclusion: "the null hypothesis is rejected", or "the null hypothesis is not rejected".

125 – Hypothesis testing

- level of significance p is usually 0.05 (sometimes 0.01 or 0.1)
- the quantity $1 - p$ is called level of confidence (it is usually 0.95)
- we cannot prove that the null hypothesis is correct, there are two types of possible errors we can make

Errors

1. **Type I error** is to reject the null hypothesis when it is true. In the case of a mean, this occurs when the null hypothesis is correct, but an observation or sample mean is so far from the expected mean by chance that the null hypothesis is rejected. The probability of a Type I error is equal to the level of significance.

2. **Type II error** is to accept the null hypothesis when it is false. If the population mean has changed, the null hypothesis is false. But the sample mean might still by chance come close enough to the original sample mean so that we would accept the null hypothesis, giving a Type II error.

126 – Inferences for the mean when variance is known

Statistical inferences for the mean are divided into two main categories. In the first category, there is already known the value of the variance σ^2 or the standard deviation σ of the population (usually from previous measurements), and the normal distribution can be used for calculations. In the second category, we have to first estimate the variance or the standard deviation of the population, and the Student's or t-distribution is used for calculations.

Inferences for the mean when variance is known

Tests of significance

a) two-sided (two-tailed) test

- we use this type in case we are concerned with possible changes of population mean in both directions, positive and negative

$H_0 : \mu = \mu_0$ ... null hypothesis
$H_A : \mu \neq \mu_0$ ... alternative hypothesis

Test statistics:

$$TS_1 : z = \frac{x - \mu}{\sigma} \quad \sim N(0;1)$$

$$TS_2 : z = \frac{\bar{x} - \mu}{\sigma}\sqrt{n} \quad \sim N(0;1)$$

- the formula TS_1 is used in case we have only single determination x and the formula TS_2 is used if we have a sample with n observations and the mean value of $\bar{x}$

z_{obs} ... observed value of the test statistic
z_{crit} ... critical value of the standard normal distribution $N(0;1)$
$CR = (-\infty, -z_{crit}) \cup (z_{crit}, +\infty)$ ... critical (rejection) region

Conclusion:

$z_{obs} \in CR \Rightarrow H_0$ is rejected
$z_{obs} \notin CR \Rightarrow H_0$ is not rejected

127 – Inferences for the mean when variance is known

Remark

- the value of z_{crit} depends on the value of the stated level of significance p
- we can find the value of z_{crit} in statistical tables or compute it using Excel function NORM.INV:
  $z_{crit} = \text{NORM.INV}(1 - p/2; 0; 1)$

b) one-sided (one-tailed) tests

- we use this type of test in case we are concerned with possible changes of population mean in one direction:

positive:

$H_0 : \mu = \mu_0$ ... null hypothesis
$H_A : \mu > \mu_0$ ... alternative hypothesis
$CR = (z_{crit}, +\infty)$ ... critical (rejection) region
$z_{crit} = \text{NORM.INV}(1 - p; 0; 1)$

or negative:

$H_0 : \mu = \mu_0$ ... null hypothesis
$H_A : \mu < \mu_0$ ... alternative hypothesis
$CR = (-\infty, -z_{crit})$ ... critical (rejection) region
$z_{crit} = \text{NORM.INV}(1 - p; 0; 1)$

Test statistics:

$$TS_1 : z = \frac{x - \mu}{\sigma} \quad \sim N(0; 1)$$

$$TS_2 : z = \frac{\bar{x} - \mu}{\sigma}\sqrt{n} \quad \sim N(0; 1)$$

Conclusion:

$z_{obs} \in CR \Rightarrow H_0$ is rejected
$z_{obs} \notin CR \Rightarrow H_0$ is not rejected

128 – Inferences for the mean when variance is known

Example

It is very important that a certain solution in a chemical process have a pH of 8.30. The method used gives measurements which are approximately normally distributed about the actual pH of the solution with a known standard deviation of 0.020. Is there evidence at the 5% level of significance that the mean pH has changed, in case a single determination shows pH of 8.32?

The population standard deviation is known, so we can use the formulas for the test statistics from the previous page for our computations. We are concerned with possible changes of population mean in both directions, so we use a two-sided (two-tailed) test.

We have only single determination x in this case, so we will use the formula for TS_1 here.

Tests of significance:

1. Stating the null and the alternative hypothesis:

$$H_0 : \mu = 8.30$$

$$H_A : \mu \neq 8.30$$

2. Stating the test statistic TS:

$$TS_1 : z = \frac{x - \mu}{\sigma}$$

and the level of significance of the test p:

$$p = 0.05$$

3. Showing calculations assuming that H_0 is true, i.e. computing the observed value of the test statistic z_{obs}:

$$z_{obs} = \frac{8.32 - 8.30}{0.020} = 1$$

4. Computing the critical value of the corresponding distribution z_{crit}:

$$z_{crit} = NORM.INV(1 - p/2; 0; 1) = NORM.INV(0.975; 0; 1) \doteq 1.96$$

and stating the critical (rejection) region CR:

$$CR = (-\infty, -z_{crit}) \cup (z_{crit}, +\infty) = (-\infty, -1.96) \cup (1.96, +\infty)$$

5. Finding out whether the observed value of the test statistic z_{obs} belongs to the critical region CR and stating conclusion:

$$z_{obs} \notin CR \Rightarrow H_0 \text{ is not rejected}$$

Since $z_{obs} \notin CR$, we do not reject the null hypothesis. We do not have enough evidence from this calculation to say that the pH is not equal to 8.30. We could say that the difference from a pH of 8.30 is not statistically significant at the 5% level of significance.

130 – Inferences for the mean when variance is known

Example

It is very important that a certain solution in a chemical process have a pH of 8.30. The method used gives measurements which are approximately normally distributed about the actual pH of the solution with a known standard deviation of 0.020. Is there evidence at the 5% level of significance that the mean pH has changed, in case a sample of 4 determinations shows pH of 8.31, 8.34, 8.32, 8.31?

The population standard deviation is known, so we can use the formulas for the test statistics from the previous page for our computations. We are concerned with possible changes of population mean in both directions, so we use a two-sided (two-tailed) test.

We have a sample of 4 determinations $x_1, \ldots, x_4$ in this case, so we will use the formula for TS_2 here.

Tests of significance:

1. Stating the null and the alternative hypothesis:

$$H_0 : \mu = 8.30$$

$$H_A : \mu \neq 8.30$$

2. Stating the test statistic TS:

$$TS_2 : z = \frac{\bar{x} - \mu}{\sigma}\sqrt{n}$$

and the level of significance of the test p:

$$p = 0.05$$

3. Showing calculations assuming that H_0 is true, i.e. computing the observed value of the test statistic z_{obs}:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{4}(8.31 + 8.34 + 8.32 + 8.31) = 8.32$$

$$z_{obs} = \frac{8.32 - 8.30}{0.020}\sqrt{4} = 2$$

4. Computing the critical value of the corresponding distribution z_{crit}:

$$z_{crit} = NORM.INV(1 - p/2; 0; 1) = NORM.INV(0.975; 0; 1) \doteq 1.96$$

and stating the critical (rejection) region CR:

$$CR = (-\infty, -z_{crit}) \cup (z_{crit}, +\infty) = (-\infty, -1.96) \cup (1.96, +\infty)$$

5. Finding out whether the observed value of the test statistic z_{obs} belongs to the critical region CR and stating conclusion:

$$z_{obs} \in CR \Rightarrow H_0 \text{ is rejected}$$

Since $z_{obs} \in CR$, we reject the null hypothesis and accept the alternative hypothesis $\mu \neq 8.30$. At the 5% level of significance we conclude that the true mean pH is no longer 8.30 (the evidence against the null hypothesis is stronger now, because we have a sample of four determinations with the mean of 8.32 instead of only one determination with the value of 8.32).

132 – Inferences for the mean when variance is estimated from a sample

Inferences for the mean when variance is estimated from a sample

Tests of significance

Test statistic:

$$TS : t = \frac{\bar{x} - \mu}{s}\sqrt{n} \quad \sim t(n-1)$$

t_{obs} ... observed value of the test statistic
t_{crit} ... critical value of t-distribution with $n-1$ degrees of freedom $t(n-1)$

a) two-sided (two-tailed) test

$H_0 : \mu = \mu_0$ ... null hypothesis
$H_A : \mu \neq \mu_0$ ... alternative hypothesis
$CR = (-\infty, -t_{crit}) \cup (t_{crit}, +\infty)$ ... critical (rejection) region
$t_{crit} = \text{T.INV}(1 - p/2; n-1)$

b) one-sided (one-tailed) tests

$H_0 : \mu = \mu_0$ ... null hypothesis
$H_A : \mu > \mu_0$ ... alternative hypothesis
$CR = (t_{crit}, +\infty)$ ... critical (rejection) region
$t_{crit} = \text{T.INV}(1 - p; n-1)$

$H_0 : \mu = \mu_0$ ... null hypothesis
$H_A : \mu < \mu_0$ ... alternative hypothesis
$CR = (-\infty, -t_{crit})$ ... critical (rejection) region
$t_{crit} =$T.INV$(1 - p; n-1)$

Conclusion:

$t_{obs} \in CR \Rightarrow H_0$ is rejected
$t_{obs} \notin CR \Rightarrow H_0$ is not rejected

Worksheets for Statistics

133 – Inferences for the mean when variance is estimated from a sample

Example

The electrical resistances of components are measured as they are produced. A sample of six items gives a sample mean of 2.62 ohms and a sample standard deviation of 0.121 ohms. Is there evidence at the 1% level of significance that the population mean is significantly less than 2.80 ohms?

Tests of significance:
1. Stating the null and the alternative hypothesis:

$$H_0 : \mu = 2.80$$

$$H_A : \mu < 2.80$$

2. Stating the test statistic TS:

$$TS : t = \frac{\bar{x} - \mu}{s}\sqrt{n} \quad \sim t(n-1)$$

and the level of significance of the test p:

$$p = 0.01$$

3. Showing calculations assuming that H_0 is true, i.e. computing the observed value of the test statistic t_{obs}:

$$t_{obs} = \frac{2.62 - 2.80}{0.121}\sqrt{6} \doteq -3.64$$

4. Computing the critical value of the corresponding distribution t_{crit}:

$$t_{crit} = T.INV(1-p; n-1) = T.INV(0.99; 5) \doteq 3.36$$

and stating the critical (rejection) region CR:

$$CR = (-\infty, -t_{crit}) = (-\infty, -3.36)$$

5. Finding out whether the observed value of the test statistic t_{obs} belongs to the critical region CR and stating conclusion:

$$t_{obs} \in CR \Rightarrow H_0 \text{ is rejected}$$

Since $t_{obs} \in CR$, we reject the null hypothesis and accept the alternative hypothesis $\mu < 2.80$. At the 1% level of significance we conclude that the true mean electrical resistance of components is less than 2.80 ohms.

134 – Confidence interval for the mean

A sample mean is known from measurements whereas a population mean is an uncertain value for which we need an estimate. The sample mean gives a point estimate for the population mean but we often need an interval estimate. That interval corresponds to a stated level of confidence that the interval contains the true population mean.

1. Confidence interval for the mean when the population variance σ is known:

$$\mu \in \left(\bar{x} - z_{crit} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{crit} \frac{\sigma}{\sqrt{n}} \right)$$

Remark

- z_{crit} is a critical value of the standardized normal distribution $N(0;1)$
- it depends on the value of the stated level of confidence $1 - p$
- it can be found in statistical tables or computed by Excel function NORM.INV:
  $z_{crit} = \text{NORM.INV}(1 - p/2; 0; 1)$

2. Confidence interval for the mean when the population variance σ is not known:

$$\mu \in \left(\bar{x} - t_{crit} \frac{s}{\sqrt{n}}; \bar{x} + t_{crit} \frac{s}{\sqrt{n}} \right)$$

Remark

- t_{crit} is a critical value of t-distribution with $n - 1$ degrees of freedom $t(n - 1)$
- it depends on the value of the stated level of confidence $1 - p$
- it can be found in statistical tables or computed by Excel function T.INV:
  $t_{crit} = \text{T.INV}(1 - p/2; n - 1)$

Worksheets for Statistics

135 – Confidence interval for the mean

Example

A certain dimension is measured on four successive items coming off a production line. This sample gives a sample mean $\bar{x} = 2.384$ and a sample standard deviation $s = 0.048$.
a) On the basis of this sample, what is the 95% confidence interval for the population mean?
b) If instead of estimating the standard deviation from a sample, we knew the true standard deviation was 0.048, what then would be the 95% confidence interval for the population mean?

a) we use the formula for confidence interval for the mean when the population variance σ is not known:

$$\mu \in \left(\bar{x} - t_{crit} \frac{s}{\sqrt{n}}; \bar{x} + t_{crit} \frac{s}{\sqrt{n}} \right)$$

where $n = 4$, $\bar{x} = 2.384$, $s = 0.048$, $1 - p = 0.95$, $p = 0.05$, t_{crit} =T.INV$(1 - p/2; n - 1)$ =T.INV$(0.975; 3) \doteq 3.182$

$$\mu \in \left(2.384 - 3.182 \frac{0.048}{\sqrt{4}}; 2.384 + 3.182 \frac{0.048}{\sqrt{4}} \right)$$

$$\mu \in (2.31; 2.46)$$

b) we use the formula for confidence interval for the mean when the population variance σ is known:

$$\mu \in \left(\bar{x} - z_{crit} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{crit} \frac{\sigma}{\sqrt{n}} \right)$$

where
$n = 4$, $\bar{x} = 2.384$, $\sigma = 0.048$, $1 - p = 0.95$, $p = 0.05$,
z_{crit} = NORM.INV$(1 - p/2; 0; 1)$ =NORM.INV$(0.975; 0; 1) \doteq 1.960$

$$\mu \in \left(2.384 - 1.960 \frac{0.048}{\sqrt{4}}; 2.384 + 1.960 \frac{0.048}{\sqrt{4}} \right)$$

$$\mu \in (2.34; 2.43)$$

We can see that the confidence interval is appreciably narrower in part b) than in part a).

136 – Comparison of sample means

In many cases we want to compare means of two populations or samples. The two major ways of comparing means from the data that is assumed to be normally distributed are independent samples t-test and paired samples t-test.

Independent samples t-test

Requirements:
- the two random samples have been chosen separately and independently of one another
- the two estimates of variance are compatible with one another

Tests of significance

Test statistic:

$$TS : t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1-1)+(n_2-1)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad \sim t(n_1 - 1 + n_2 - 1)$$

t_{obs} ... observed value of the test statistic
t_{crit} ... critical value of t-distribution with $n_1 + n_2 - 2$ degrees of freedom $t(n_1 + n_2 - 2)$

a) two-sided (two-tailed) test

$H_0 : \mu_1 = \mu_2$ ... null hypothesis
$H_A : \mu_1 \neq \mu_2$ ... alternative hypothesis
$CR = (-\infty, -t_{crit}) \cup (t_{crit}, +\infty)$ ... critical (rejection) region
$t_{crit} = \text{T.INV}(1 - p/2; n_1 + n_2 - 2)$

b) one-sided (one-tailed) tests

$H_0 : \mu_1 = \mu_2$ ... null hypothesis
$H_A : \mu_1 > \mu_2$ ... alternative hypothesis
$CR = (t_{crit}, +\infty)$ ... critical (rejection) region
$t_{crit} = \text{T.INV}(1 - p; n_1 + n_2 - 2)$

Worksheets for Statistics

137 – Comparison of sample means

$H_0 : \mu_1 = \mu_2$ ...null hypothesis
$H_A : \mu_1 < \mu_2$ ...alternative hypothesis
$CR = (-\infty, -t_{crit})$ ...critical (rejection) region
t_{crit} =T.INV$(1 - p; n_1 + n_2 - 2)$

Conclusion:

$$t_{obs} \in CR \Rightarrow H_0 \text{ is rejected}$$
$$t_{obs} \notin CR \Rightarrow H_0 \text{ is not rejected}$$

Example

Two methods of determining the nickel content of steel are compared using four determinations by each method. The results are:
For method 1: $\bar{x}_1 = 3.285, s_1 = 0.00774$
For method 2: $\bar{x}_2 = 3.258, s_2 = 0.00960$
Assuming that the two estimates of variance are compatible, is the difference in means statistically significant at the 5% level of significance?

Tests of significance:
1. Stating the null and the alternative hypothesis:

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

2. Stating the test statistic TS:

$$TS : t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1-1)+(n_2-1)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

and the level of significance of the test p:

$$p = 0.05$$

138 – Comparison of sample means

3. Showing calculations assuming that H_0 is true, i.e. computing the observed value of the test statistic t_{obs}:

$$t_{obs} = \frac{3.285 - 3.258}{\sqrt{\frac{0.00774^2(4-1)+0.00960^2(4-1)}{(4-1)+(4-1)}\left(\frac{1}{4}+\frac{1}{4}\right)}} \doteq 4.38$$

4. Computing the critical value of the corresponding distribution t_{crit}:

$$t_{crit} = T.INV(1 - p/2; n_1 + n_2 - 2) = T.INV(0.975; 6) \doteq 2.45$$

and stating the critical (rejection) region CR:

$$CR = (-\infty, -t_{crit}) \cup (t_{crit}, +\infty) = (-\infty, -2.45) \cup (2.45, +\infty)$$

5. Finding out whether the observed value of the test statistic t_{obs} belongs to the critical region CR and stating conclusion:

$$t_{obs} \in CR \Rightarrow H_0 \text{ is rejected}$$

Since $t_{obs} \in CR$, we reject the null hypothesis and accept the alternative hypothesis $\mu_1 \neq \mu_2$. The difference in means is statistically significant at the 5% level of significance.

139 – Comparison of sample means

Paired samples t-test

Requirements:
- the two random samples are dependent and have the same number of observations ($n_1 = n_2 = n$)
- each observation from the first sample (x_i) forms a pair with one observation from the second sample (y_i)

Tests of significance
- for all pairs we compute pair differences d $(d_i = x_i - y_i)$

Test statistic:

$$TS : t = \frac{\bar{d} - 0}{s_d}\sqrt{n} \quad \sim t(n-1)$$

t_{obs}	...observed value of the test statistic
t_{crit}	...critical value of t-distribution with $n-1$ degrees of freedom $t(n-1)$

a) two-sided (two-tailed) test

$H_0 : \mu_1 = \mu_2$ $(\mu_d = 0)$	...null hypothesis
$H_A : \mu_1 \neq \mu_2$ $(\mu_d \neq 0)$	...alternative hypothesis
$CR = (-\infty, -t_{crit}) \cup (t_{crit}, +\infty)$	...critical (rejection) region
$t_{crit} = \text{T.INV}(1 - p/2; n-1)$	

b) one-sided (one-tailed) tests

$H_0 : \mu_1 = \mu_2$ $(\mu_d = 0)$	...null hypothesis
$H_A : \mu_1 > \mu_2$ $(\mu_d > 0)$	...alternative hypothesis
$CR = (t_{crit}, +\infty)$	...critical (rejection) region
$t_{crit} = \text{T.INV}(1 - p; n-1)$	

$H_0 : \mu_1 = \mu_2$ $(\mu_d = 0)$	...null hypothesis
$H_A : \mu_1 < \mu_2$ $(\mu_d < 0)$	...alternative hypothesis
$CR = (-\infty, -t_{crit})$	...critical (rejection) region
$t_{crit} = \text{T.INV}(1 - p; n-1)$	

Conclusion:

$$t_{obs} \in CR \Rightarrow H_0 \text{ is rejected}$$
$$t_{obs} \notin CR \Rightarrow H_0 \text{ is not rejected}$$

Worksheets for Statistics

140 – Comparison of sample means

Example

We decide to run a test using an experimental evaporation pan and a standard evaporation pan over ten successive days. The two types are set up side-by-side so that atmospheric conditions should be the same. A coin is tossed to decide which evaporation pan is on the lefthand side and which on the righthand side on any particular day. The measured daily evaporations are as follows.
Evaporation (mm):
Pan A: 9.1, 4.6, 14.0, 16.9, 11.4, 10.7, 27.4, 22.8, 42.8, 29.4
Pan B: 6.7, 3.1, 13.8, 16.6, 12.3, 6.5, 24.2, 20.1, 41.9, 27.7
Does the experimental Pan A give significantly higher evaporation than the standard Pan B at the 1% level of significance?

First we have to compute the differences d_i (Pan A - Pan B):
d_i: 2.4, 1.5, 0.2, 0.3, -0.9, 4.2, 3.2, 2.7, 0.9, 1.7

Tests of significance:
1. Stating the null and the alternative hypothesis:

$$H_0 : \mu_d = 0$$

$$H_A : \mu_d > 0$$

2. Stating the test statistic TS:

$$TS : t = \frac{\bar{d} - 0}{s_d}\sqrt{n}$$

and the level of significance of the test p:

$$p = 0.01$$

141 – Comparison of sample means

3. Showing calculations assuming that H_0 is true, i.e. computing the observed value of the test statistic t_{obs}:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{10}(2.4 + 1.5 + 0.2 + 0.3 - 0.9 + 4.2 + 3.2 + 2.7 + 0.9 + 1.7) = 1.62$$

$$s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2} \doteq 1.55$$

$$t_{obs} = \frac{\bar{d} - 0}{s_d}\sqrt{n} \doteq 3.31$$

4. Computing the critical value of the corresponding distribution t_{crit}:

$$t_{crit} = T.INV(1 - p; n - 1) = T.INV(0.99; 9) \doteq 2.82$$

and stating the critical (rejection) region CR:

$$CR = (t_{crit}, +\infty) = (2.82, +\infty)$$

5. Finding out whether the observed value of the test statistic t_{obs} belongs to the critical region CR and stating conclusion:

$$t_{obs} \in CR \Rightarrow H_0 \text{ is rejected}$$

Since $t_{obs} \in CR$, we reject the null hypothesis and accept the alternative hypothesis $\mu_d > 0$. The experimental Pan A give significantly higher evaporation than the standard Pan B at the 1% level of significance.

142 – Statistical inferences for the mean

Exercise

a) When a manufacturing process is operating properly, the mean length of a certain part is known to be 6.175 inches, and lengths are normally distributed. The standard deviation of this length is 0.008 inches. If a sample consisting of 6 items taken from current production has a mean length of 6.168 inches, is there evidence at the 5% level of significance that some adjustment of the process is required?

b) A cocoa packaging machine fills bags so that the bag contents have a standard deviation of 3.5 g. Weights of contents of bags are normally distributed. If a random sample of 20 bags gives a mean of 102.0 g, what is the 99% confidence interval for the mean weight of the population (i.e., all bags)?

143 – Statistical inferences for the mean

Exercise

a) Benzene in the air workers breathe can cause cancer. It is very important for the benzene content of air in a particular plant to be not more than 1.00 ppm. Samples are taken to check the benzene content of the air. 25 specimens of air from one location in the plant gave a mean content of 0.760 ppm, and the standard deviation of benzene content was estimated on the basis of the sample to be 0.45 ppm. Benzene contents in this case are found to be normally distributed.
1) Is there evidence at the 1% level of significance that the true mean benzene content is less than or equal to 1.00 ppm?
2) Find the 95% confidence interval for the true mean benzene content.

b) Two chemical processes for manufacturing the same product are being compared under the same conditions. Yield from Process A gives an average value of 96.2 from six runs, and the estimated standard deviation of yield is 2.75. Yield from Process B gives an average value of 93.3 from seven runs, and the estimated standard deviation is 3.35. Yields follow a normal distribution. Is the difference between the mean yields statistically significant? Use the 5% level of significance.

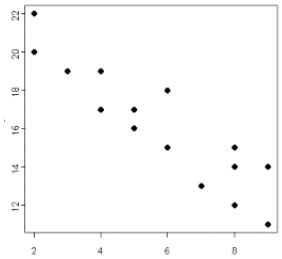
Worksheets for Statistics

Regression and correlation analysis

145 – Linear regression analysis

The goal of regression analysis is to describe the relationship between two variables based on observed data and to predict the value of the dependent variable (y) based on the value of the independent variable (x).

To better visualize the association between two data sets $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ we can employ a chart called a scatter diagram (also called a scatter plot).



The relationship can be represented by a simple equation called the regression equation. The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram.

Definition

Let (x_i, y_i), $i = 1, \ldots, n$ be n pairs of observations. The simple regression linear model of y on x is

$$y = \beta_0 + \beta_1 x_i + \epsilon,$$

where β_0 and β_1 are constant parameters (called **regression coefficients**) that we want to estimate and ϵ is the random error term.

146 – Method of least squares

In practice we will build the linear regression model from the sample data using **the least squares method**. Thus we seek coefficients a and b such that

$$\hat{y}_i = a + bx_i,$$

where a, b are some estimates for the model coefficients β_0, β_1 and $\hat{y}_i$ is the estimated (predicted) y value for i-th observation x_i.

The difference between the observed value y_i and the predicted value $\hat{y}_i$ is called as a **residual**. The i-th residual is defined as

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i).$$

The best fit line (regression line) is the line for which the sum of the distances between each of the n data points and the line is as small as possible. A mathematically useful approach is therefore to find the line with the property that the sum of the following squares is minimal.

$$\sum_{i=1}^{n} e_i^2$$

So, a, b are obtained by finding the values that minimize the sum of the squared differences between y and $\hat{y}$. This sum of the squares of the errors (residuals) for all n points is abbreviated as SSE.

$$\min \sum_i e_i^2 = \min \sum_i (y_i - \hat{y}_i)^2 = \min \sum_i (y_i - (a + bx_i))^2$$

147 – Method of least squares

To minimize a quantity we take the derivative with respect to the independent variable and set it equal to zero. In this case there are two independent variables, a and b, so we take partial derivatives with respect to each of them and set the derivatives equal to zero. We have

$$\frac{\partial(SSE)}{\partial a} = \frac{\partial}{\partial a}\sum_i (y_i - (a + bx_i))^2 = -2\left(\sum_i y_i - na - b\sum_i x_i\right) = 0$$

and

$$\frac{\partial(SSE)}{\partial b} = \frac{\partial}{\partial b}\sum_i (y_i - (a + bx_i))^2 = -2\left(\sum_i x_i y_i - a\sum_i x_i - b\sum_i x_i^2\right) = 0.$$

We get **the least squares normal equations**

$$na + b\sum_i x_i = \sum_i y_i$$

and

$$a\sum_i x_i + b\sum_i x_i^2 = \sum_i x_i y_i.$$

The solution to the normal equations results in the least squares estimators

$$a = \frac{\sum_i y_i - b\sum_i x_i}{n} = \overline{y} - b\overline{x}$$

and

$$b = \frac{\sum_i x_i y_i - \frac{1}{n}\left(\sum_i x_i \sum_i y_i\right)}{\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}.$$

148 – Linear regression analysis

Example

The table shows the results of two different tests for 8 students

1st test	80	50	36	58	72	60	56	68
2nd test	65	60	35	39	48	44	48	61

Find the best-fit regression equation of y (2nd test) on x (1st test).

We have 8 points $(n = 8)$. We calculate the following sums:

$$\sum_{i=1}^{8} x_i = 480,\ \sum_{i=1}^{8} y_i = 400,\ \sum_{i=1}^{8} x_i^2 = 30104,\ \sum_{i=1}^{8} x_i y_i = 24654.$$

The centroidal point is given by $\overline{x} = 60,\ \overline{y} = 50$.

Then

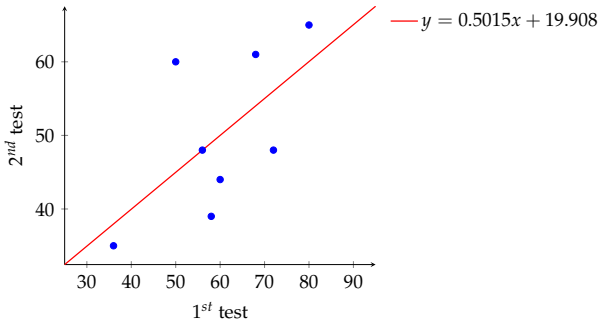
$$b = \frac{\sum_i x_i y_i - \frac{1}{n}\left(\sum_i x_i \sum_i y_i\right)}{\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2} = \frac{24654 - \frac{1}{8}(480 \cdot 400)}{30104 - \frac{1}{8}(480)^2} = 0.5015$$

and

$$a = \overline{y} - b\overline{x} = 50 - 0.5015 \cdot 60 = 19.908$$

The regression equation is

$$y = 19.908 + 0.5015\ x$$



Hints

The regression line

$$y = a + bx$$

Excel:
Insert → Charts|Scatter → Layout
and then "Linear Trend"

Excel plugin Data Analysis ToolPak:
1. Select Data → Data Analysis → Regression
2. Input Y Range - the dependent variable data;
Input X range - the independent variable data

149 – Linear regression analysis

Exercise

The shear resistance of soil (y) is determined by measurements as a function of the normal stress (x). The data are shown in the table:

x (kN m^{-2})	10	12	13	16	17	18	20	21
y (kN m^{-2})	14.08	16.94	17.68	20.68	21.72	22.80	24.79	25.67

Find the regression line of y on x. Plot the data, the regression line and the centroidal point.

Hints

The regression line

$$y = a + bx$$

The least squares estimators

$$a = \overline{y} - b\overline{x}$$

$$b = \frac{\sum_i x_i y_i - \frac{1}{n}\left(\sum_i x_i \sum_i y_i\right)}{\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2}$$

Excel:
Insert → Charts|Scatter → Layout
and then "Linear Trend"

150 – Other forms linear in the coefficients

With an extra step of calculation an important group of equations can be fitted to data by the method of least squares. For instance, equations of the form, $\log y = a + bx$, where a and b are coefficients to be determined by least squares, can be handled easily. Remember that x and y are known quantities, numbers. Then we can calculate without difficulty the value of $\log y$ for each data point. Then $\log y$ can be used in place of y, and so the regression coefficients can be calculated as before.

This modified method works for a considerable number of cases. The requirement is that the equation to which we fit data must be of the form

$$f_1(y) = a + bf_2(x),$$

where x is the only input quantity.

The two functions, $f_1(y)$ and $f_2(x)$, can be of any form and do not have to be linear, but both a and b must be coefficients to be determined by the method of least squares. Thus the fitting equation must be linear in the coefficients so that it is easy to solve for a and b.

The modified method is sometimes still considered to be simple linear regression, but then the word "simple" means that there is only one input, and the word "linear" means that the equation is linear in the coefficients.

Fitting equations amenable to the modified method include the following types:

$$\begin{aligned} y &= a + bx^3 \\ y &= a + b\sqrt{x} \\ y &= a + \frac{b}{x} \\ \frac{1}{y} &= a + b\ln x \\ \log y &= a + b\log x \end{aligned}$$

and many others.

151 – Other forms transformable to give equations linear in the coefficients

Various common forms of equations involving one input can be transformed easily to give forms of equations which are linear in the coefficients.

1. The exponential function, $y = ab^x$, can be modified suitably by taking logarithms of both sides. This gives $\log y = \log a + x \log b$. Notice that this is the form that gives straight lines on semi-log graph paper.

2. The power function, $y = ax^b$, can also be treated by taking logarithms of both sides. The result is $\log y = \log a + b \log x$. Notice that this form would give straight lines on log-log graph paper.

3. The function, $y = \frac{x}{a+bx}$, can be inverted to give $\frac{1}{y} = \frac{a}{x} + b$. An alternative is to multiply the inverted form by x to give $\frac{x}{y} = a + bx$.

It is important to note that the squares of the deviations are minimized in the transformed response variable ($\log y$ or $\frac{1}{y}$ or $\frac{x}{y}$ in the cases above) rather than y, and the graphical tests need to be applied to the transformed response variable. It is possible in some cases to apply a simple weighting function to make the variance approximately constant.

Nonlinear forms

Equations that cannot be transformed into forms linear in the coefficients can still be treated by least squares. However, now instead of applying the relations discussed to this point, iterative numerical methods must be used to minimize the sum of squares of the deviations from the fitted line. The Excel feature called Solver can be used for that calculation.

152 – Regression model validation

We describe several metrics for assessing the overall performance of a regression model.

The most important metrics are the R-squared, the adjusted R-squared and the residual standard error. These metrics are also used as the basis of model comparison and optimal model selection.

Coefficient of Determination (R Squared) R^2 is used to analyze how differences in one variable can be explained by a difference in a second variable. More specifically, R-squared gives you the percentage variation in y explained by variation in x-variable and how well the regressed values estimate the real/actual values.

The range is 0 to 1 (i.e. 0 % to 100 % of the variation in y can be explained by the x-variable).

The higher value R-squared means the better model. Clearly $0 \leq R^2 \leq 1$, so a value of R^2 closer to one indicates the better fit and value of R^2 closer to zero indicates the poor fit.

We compute the R-squared as a fraction of the variances:

$$R^2 = \frac{S_R^2}{S_T^2} = 1 - \frac{SSE}{S_T^2},$$

where

$$S_T^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{total sum of squares,}$$

$$S_R^2 = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2 \quad \text{regression sum of squares,}$$

$$SSE = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \quad \text{residual sum of squares.}$$

Total variation is made up of two parts:

$$S_T^2 = S_R^2 + SSE$$

153 – Coefficient of determination

Example

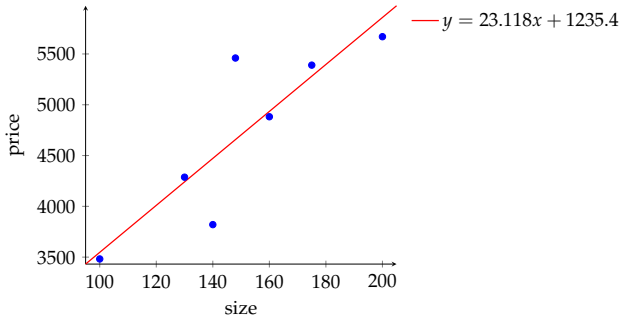
We examine the relationship between the selling price of a house and its size.

price (CZK 1000s)	4287	5460	4882	5390	3482	3820	5670
size m^2	130	148	160	175	100	140	200

Find the linear regression equation of the price on size of the house. Calculate a coefficient of determination for the data.

We can chart a regression in Excel.

- Insert → Charts| Scatter
- To add a regression line, choose "Layout" from the "Chart Tools" menu
- In the dialog box, select "Trendline" and then "Linear Trendline"
- To add the R^2 value, select "Display R-squared value on chart"



The regression equation is

$$y = 1235.4 + 23.118\ x$$

and the coefficient of determination is

$$R^2 = 0.7471,$$

so, 74.71 % of the variation in house prices is explained by variation in its size.

Hints

The regression line

$$y = a + bx$$

Excel:
Insert → Charts|Scatter → Layout
and then "Linear Trend"
and select "Display R-squared value on chart"

Excel plugin Data Analysis ToolPak:
1. Select Data → Data Analysis → Regression
2. Input Y Range - the dependent variable data;
Input X range - the independent variable data

154 – Coefficient of determination

Exercise

Consider the data obtained from a chemical process where the yield of the process is thought to be related to the reaction temperature.

temperature	50	54	56	62	67	72	75	79
yield	122	128	125	144	149	167	162	175

Find the regression line of the yield of the process on the reaction temperature. Plot the data, the regression line and calculate a coefficient of determination.

Hints

The regression line

$$y = a + bx$$

Excel:
Insert $\rightarrow$ Charts|Scatter $\rightarrow$ Layout
and then "Linear Trend"
and select "Display R-squared value on chart"

155 – Regression model validation

Adjusted Coefficient of Determination R^2_{adj} is an adjustment for the R-squared that takes into account the number of variables in a data set.

The formula is:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

where n is the number of points in the data sample and p is the number of independent regressors, i.e. the number of variables in the model, excluding the constant.

Residual Standard Deviation R_{res} is the standard deviation of the residual values, or the difference between a set of observed and predicted values. The standard deviation of the residuals calculates how much the data points spread around the regression line. The result is used to measure the error of the regression line's predictability.

The formula is:

$$R_{res} = \sqrt{\frac{1}{n-2}SSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2},$$

where n is the number of points in the data sample.

156 – Inferences for coefficients

The standard error of an estimator reflects how it varies under repeated sampling. We have:

$$s_b = \frac{R_{res}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$s_a^2 = R_{res}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

- the standard errors can be used to compute confidence intervals, these standard errors are then multiplied by appropriate values of t-distribution with $n-2$ degrees of freedom, t_α, to find confidence intervals.

There is approximately $(1-\alpha)$ % chance that the intervals

$$[a - t_\alpha\, s_a, a + t_\alpha\, s_a]$$
$$[b - t_\alpha\, s_b, b + t_\alpha\, s_b]$$

contain the true values of β_0, β_1.

- the standard errors can also be used to perform hypothesis test on the coefficients.

t-test for a population slope

$H_0 : \beta_1 = 0$ there is no linear relationship between x and y
$H_1 : \beta_1 \neq 0$ there is some linear relationship between x and y

To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{b - \beta_1}{s_b}$$

This will have a t-distribution with $(n-2)$ degrees of freedom.

157 – Inferences for coefficients

Example

We examine the relationship between the selling price of a house and its size.

price (CZK 1000s)	4287	5460	4882	5390	3482	3820	5670
size m^2	130	148	160	175	100	140	200

Compute the 95 % confidence interval for regression coefficient b. Is there a linear relationship between the selling price of a house and its size? Use t-test for a population slope.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,864341111
R Square	0,747085556
Adjusted R Square	0,696502667
Standard Error	474,7088473
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3328291,551	3328292	14,76953	0,012085361
Residual	5	1126742,449	225348,5		
Total	6	4455034			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1235,423659	922,501617	1,33921	0,238154	-1135,942241	3606,789559
X Variable 1	23,11779144	6,015378871	3,843115	0,012085	7,65476778	38,5808151

We see, that

$$b = 23.118$$
$$s_b = 6.01$$
$$t = 3.84$$
$$p = 0.01$$

Hints

The regression line

$$y = a + bx$$

Excel plugin Data Analysis ToolPak:
1. Select Data → Data Analysis → Regression
2. Input Y Range - the dependent variable data; Input X range - the independent variable data
and select "Confidence level: $1 - \alpha$ %"

t_α is computed by the formula $= T.INV(1-\alpha; n-2)$,

the 95 % confidence interval for b is $[23.118 - 2.571\ 6.01, 23.118 + 2.571\ 6.01] = [7.65, 38.56]$

Now, we find, if there is a linear relationship.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

We have two options for decision.

- $t = 3.84 > 2.57 = t_\alpha \Rightarrow H_0$ is rejected
- $p = 0.01 \Rightarrow$ p-value is less than the chosen significance level then we can reject the null hypothesis

There is sufficient evidence that the size of house affects house price.

Hints

The regression line

$$y = a + bx$$

Excel plugin Data Analysis ToolPak:
1. Select Data → Data Analysis → Regression
2. Input Y Range - the dependent variable data; Input X range - the independent variable data
and select "Confidence level: $1 - \alpha$ %"

159 – Correlation

Correlation is a measure of the association between two random variables, say X and Y. We do assume for this analysis that X and Y are related linearly, so the usual correlation coefficient gives a measure of the linear association between X and Y. In practice we work with **the sample correlation coefficient** r. It is also called the Pearson correlation coefficient.

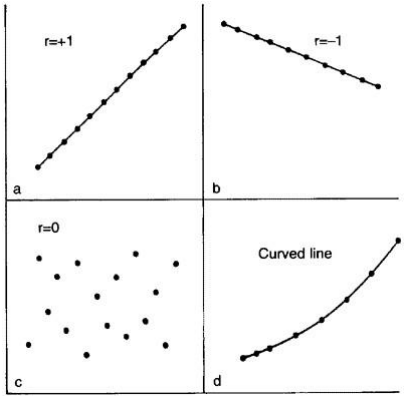
This is calculated as

$$r = \frac{\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_i (x_i - \bar{x}) \cdot \sum_i (y_i - \bar{y})},$$

which can be shown to be equal to:

$$r = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}.$$

The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of linear correlation is represented by 0.



The square of the correlation coefficient, $r^2 = R^2$, is called the coefficient of determination.

If the correlation coefficient or the coefficient of determination becomes larger for the same algebraic forms, that indicates that the relationship between the variables has become stronger.

The symbol for the population correlation coefficient is ρ.

Interpretation

$r = 0$	no linear correlation
$r > 0$	a positive correlation
$r = 1$	a perfect positive correlation
$0 < r < 1$	a partially positive correlation
$r < 0$	a negative correlation
$r = -1$	a perfect negative correlation
$-1 < r < 0$	a partially negative correlation

The strength of the association is regarded as:

$0 < \lvert r \rvert < 0.19$	very weak
$0.2 < \lvert r \rvert < 0.39$	weak
$0.4 < \lvert r \rvert < 0.59$	moderate
$0.6 < \lvert r \rvert < 0.79$	strong
$0.8 < \lvert r \rvert < 1$	very strong

160 – Correlation coefficient

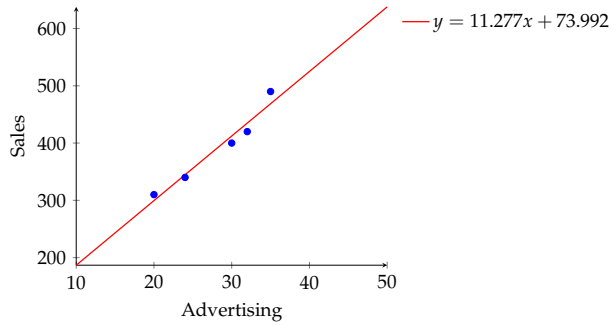
Example

A study was conducted to analyze the relationship between advertising expenditure and sales. The following data were recorded

Advertising	20	24	30	32	35
Sales	310	340	400	420	490

Find the best-fit regression equation and compute the correlation coefficient between advertising expenditure and sales.

We can use Excel software:



The regression equation is

$$y = 73.992 + 11.277\ x$$

The coefficient of determination is

$$R^2 = 0.9518$$

Finally the coefficient of correlation is

$$r = \sqrt{R^2} \doteq \sqrt{0.9518} = 0.9756$$

The correlation coefficient is 0.9756, so it indicates a strong positive correlation between advertising expenditure and sales.

Hints

The regression line

$$y = a + bx$$

The correlation coefficient

$$r = \sqrt{R^2}$$

Excel:
Insert → Charts|Scatter → Layout and then "Linear Trend"
and select "Display R-squared value on chart"

161 – Correlation coefficient

Exercise

A study of the amount of rainfall and the quantity of air pollution removed produced the following data:

Daily rainfall (0.01 cm)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1
Particulate removed ($\mu g/m^3$)	126	121	116	118	114	118	132	141

Calculate correlation coefficient between daily rainfall and particulate removed.

Hints

The regression line

$$y = a + bx$$

The correlation coefficient

$$r = \sqrt{R^2}$$

Excel:
Insert $\rightarrow$ Charts|Scatter $\rightarrow$ Layout
and then "Linear Trend"
and select "Display R-squared value on chart"

162 – Testing correlation coefficient

Let r be the observed correlation coefficient between x and y.

A test of significance for a linear relationship between the variables x and y can be performed using the sample correlation coefficient.

We wish to test the hypothesis $H_0 : \rho = 0$ (there is no significant linear relationship between x and y) against $H_1 : \rho \neq 0$ (there is a significant linear relationship between x and y).

The test statistics is

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

and it follows Student's distribution with $n - 2$ degrees of freedom.

163 – Testing correlation coefficient

Example

For a sample of eight bears, researchers measured the distances around the bears' chests and weighed the bears. The correlation coefficient between the chest size and weight of bears is $r = 0.744$ for 8 bears. Using $\alpha = 0.05$, determine if there is a positive linear correlation between chest size and weight.

The hypothesis testing problem is $H_0 : \rho = 0$ against $H_1 : \rho > 0$.

The test statistics is

$$t = 0.744\frac{\sqrt{8-2}}{\sqrt{1-0.744^2}} = 2.727.$$

Decision - 2 ways:
The critical value of t is 1.943 $(= T.INV(0.95; 6))$.

- $t = 2.727 > 1.943 = t_\alpha \Rightarrow H_0$ is rejected

- $p = 1 - T.DIST(2.727; 6; 1) = 0.0172 \Rightarrow$ p-value is less than the chosen significance level then we can reject the null hypothesis

There is sufficient evidence to conclude that there is a significant positive linear relationship between chest size and weight of bears.

Hints

The test statistics

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The critical value
Excel

$$= T.INV.2T(\alpha; n-2)$$

164 – Testing correlation coefficient

Exercise

Using $\alpha = 0.01$, determine if there is a positive linear correlation between the alcohol content and the number of calories in 12-ounce beer. How strong is this relationship?

Alcohol content (%)	4.7	6.7	8.1	4.2	5.1	5.0	5.0	4.7
Calories	163	215	222	104	162	158	155	158

Hints

Excel:
Insert $\rightarrow$ Charts|Scatter $\rightarrow$ Layout
and then "Linear Trend"
and select "Display R-squared value on chart"

The correlation coefficient

$$r = \sqrt{R^2}$$

The test statistics

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The critical value
Excel

$$= T.INV.2T(\alpha; n-2)$$

165 – Testing correlation coefficient

Exercise

Following is the data about the demand and price of a commodity for 8 periods. It was expected to estimate a linear regression for demand and price of a commodity. Test whether there is a significant negative relationship between price and demand of a product.

Demand	16	20	18	21	13	15	17	22
Price	10	8	12	6	13	9	11	7

Hints

Excel:
Insert → Charts|Scatter → Layout
and then "Linear Trend"
and select "Display R-squared value on chart"

The correlation coefficient

$$r = \sqrt{R^2}$$

The test statistics

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The critical value
Excel

$$= T.INV(\alpha; n-2)$$

Name: Worksheets for Statistics

Department, Institute: Faculty of Mechanical Engineering, Department of Mathematics and Descriptive Geometry

Authors: Petra Schreiberová, Marcela Rabasová

Place, year of publishing: Ostrava, 2021, 1st Edition

Number of Pages: 166

Published: VSB – Technical University of Ostrava

ISBN 978-80-248-4489-3

DOI 10.31490/9788024844893